Självständigt arbete på grundnivå

*Independent degree project - first cycle*

Datavetenskap
Computer Science

*Integrating Heterogeneous Data*

Gabriel Nieva
MITTUNIVERSITETET
Avdelningen för Arkiv- och Datavetenskap

Examinator: Leif Olsson, leif.olsson@miun.se
Handledare: Håkan Sundberg, hakan.sundberg@miun.se
Författare: Gabriel Nieva, gabriel.nieva@gmail.com
Huvudområde: Datavetenskap
Termin: VT 2016
Abstract

Technological advances, particularly in the areas of processing and storage have made it possible to gather an unprecedented vast and heterogeneous amount of data. The evolution of the internet, particularly Social media, the internet of things, and mobile technology together with new business trends has precipitated us in the age of Big Data and add complexity to the integration task. The objective of this study has been to explore the question of data heterogeneity through the deployment of a systematic literature review methodology. The study surveys the drivers of this data heterogeneity, the inner workings of it, and it explores the interrelated fields and technologies that deal with the capture, organization and mining of this data and their limitations. Developments such as Hadoop and its suit components together with new computing paradigms such as cloud computing and virtualization help palliate the unprecedented amount of rapidly changing, heterogeneous data which we see today. Despite these dramatic developments, the study shows that there are gaps which need to be filled in order to tackle the challenges of Web 3.0.

**Keywords:** Data Heterogeneity, Data Integration, Warehousing, ETL, Hadoop, Cloud Computing, Semantic Web, BI&A, Big Data.
Acknowledgements / Foreword

Special thanks to my family and to my parents for their support I have received from them, the children for the many endless days without playing together, and to my wife for her help including proof reading my paper without understanding much of the subject. Also thanks to Professor Håkan Sundberg at the department of Computer Science at Mid Sweden University for the guidance received in the writing of this paper.
Index

Abstract iii
Acknowledgements / Foreword iv
Index v
Terminology viii

1 Introduction 1
1.1 Background and problem motivation 2
1.1.1 A need for further research 3
1.2 Scope 4
1.3 Detailed problem statement 4
1.4 Outline 5

2 Background theory and Related work 6
2.1 Data heterogeneity – overview 6
2.2 Business Intelligence and Analytics 7
2.3 Big Data 8
2.4 Data Warehousing 9
2.5 Data Integration 11
2.6 Data Mining and knowledge discovery 12
2.7 Cloud Computing and Virtualization 14

3 Research Method 15
3.1 Research Methodology 15
3.2 Data collection and processing 16
3.3 Search strategy 18
3.4 Quality assessment 19
3.4.1 Validity 20
3.5 Societal and moral aspects 20

4 Results 22
4.1 Data Heterogeneity 22
4.2 Reasons to integrate data 24
4.2.1 Business Intelligence 24
4.2.1.1 Business Intelligence Inputs 24
4.2.1.2 On-Line Analytical Processing (OLAP) and On-Line transaction processing (OLTP) 25
4.2.1.3 Evolution of Business Intelligence 26
4.2.1.4 Emerging trends 28
4.3 Drivers and consequences of increase information disparity 28
4.3.1 Technological Advances in the it sector 28
4.3.2 Evolution of the Internet 29
4.3.2.1 Social Media 29
4.3.2.2 The Internet of Things 30
4.3.2.3 New business trends 31
4.3.3 Consequences of increase information disparity 32
4.3.3.1 Big Data

4.4 The Nature of Data - Structured vs Unstructured Data

4.5 Challenges and approaches of Integration of heterogeneous data

4.5.1 Integration approaches and frameworks

4.5.2 Service Oriented Architecture and XML

4.5.2.1 Earlier attempts to handle semi structured data

4.5.2.2 XML

4.5.2.3 SOA

4.5.3 Query Optimization

4.5.4 The Semantic Web

4.5.5 NoSQL Databases

4.5.6 Data Warehousing

4.5.6.1 The ETL Process

A. Extraction Process

B. The transformation process

B1. Data cleaning

B2. Data Staging

C. The Loading Process

4.5.6.2 Real Time ETL

4.5.6.3 ETL with Hadoop

4.5.6.4 Data warehousing trends and research directions

4.6 Cloud computing and virtualization

5 Analysis

5.1 Data heterogeneity

5.2 Reasons for data acquisition and integration – The value creation process

5.2.1 Business intelligence and data disparity – Evolution and emerging problems

5.3 Drivers for information disparity and its consequences

5.3.1 The problem of size, disparity and velocity - Big Data

5.4 Structured and unstructured data problems

5.5 Challenges and approaches of integration of heterogeneous data

5.5.1 SOA and XML contributions and limitations

5.5.2 The semantic web

5.5.3 Query languages

5.5.4 NoSQL Databases

5.5.5 Data Warehousing – In the age of Big Data

5.5.5.1 ETL process – Challenges in the age of big data

5.5.5.2 Contribution of NoSQL databases

5.5.5.3 Real Time ETL - Lowering Data latency

5.5.5.4 ETL with Hadoop

5.6 Contribution of Cloud Computing and related technology

5.7 Open Challenges and Future work

6 Conclusion

6.1 The need to integrate data and the consumers of the data

6.2 Drivers of increase information disparity and its consequences

6.3 The nature of heterogeneous data
6.4 The approaches to integration of heterogeneous data and their limitations 80
6.5 The computing paradigm most adequate to deal with vast, rapidly changing, heterogeneous data 81
6.6 Concluding remarks 81
References 82
## Terminology

### Acronyms/Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI&amp;A</td>
<td>Business Intelligence and Analytics</td>
</tr>
<tr>
<td>BPM</td>
<td>Business Process Management</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>DI</td>
<td>Data Integration</td>
</tr>
<tr>
<td>DW</td>
<td>Data Warehousing</td>
</tr>
<tr>
<td>ERP</td>
<td>Enterprise Resource Planning</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract Load Transform</td>
</tr>
<tr>
<td>HDFS</td>
<td>Hadoop Distributed File System</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>KD</td>
<td>Knowledge Discovery</td>
</tr>
<tr>
<td>OLAP</td>
<td>Online Analytical Processing</td>
</tr>
<tr>
<td>OLTP</td>
<td>Online Transaction Processing</td>
</tr>
<tr>
<td>SOA</td>
<td>Service Oriented Architecture</td>
</tr>
</tbody>
</table>
1 Introduction

The proliferation of mobile devices, the continuous increase in the volume and detail of data captured by organizations such as the rise of social media, the internet of things and multimedia is precipitating an unparalleled amount of data of despair nature which must be reconciled. This massive increase in data has yielded a new concept, that of Big Data. Due to the size and heterogeneity of this data, there is an increasing need to organize the data with some kind of structure. S. Negash (2004), also Chen, H., Chiang, R. H., & Storey, V. C. (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015).

With the development of traditional Relational Database Management System (RDMS) it was possible to capture data from business transactions. The structure of this data was relatively straightforward. Information was coming from systems which generated highly structured data such as Enterprise Resource Planning (ERP) which fit well into these RDMS. S. Negash (2004), also Chen, H., et al.(2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015).

Then Customer Relationship Management (CRM) and front end systems started to generate semi-structured data and it became more complex to deal with the information because they could not fit in rows and columns just as they were captured. New technologies were developed such as the Extensible Markup Language (XML) in order to deal with information which structure. S. Negash (2004), also Chen, H., et al. (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015).

There are many attempts to integrate heterogeneous data but amongst them data warehousing is perhaps one of the most researched areas.

The warehousing system is a good alternative to keep structured data, and it is widely used by decision makers but the disparity and the vastness of the data which we are constantly generating poses new challenges. Data is becoming available in greater and greater quantity and it can now be mined and utilized to the benefit of business organizations. As enterprise content proliferates and generates new formats, having the right information at the right time is becoming increasingly critical. Widom, J. (1995, December), Ruggles, S., Hacker, J. D., & Sobek, M. (1995), Fan, H. (2005), Ziegler, P., & Dittrich, K. R. (2004), Telang, A., & Chakravarthy, S. (2007), Kakish, K., & Kraft, T. A. (2012).

There is also research pointing out that the growth of enterprise-wide systems is blurring the distinction between front and back-end systems. These systems, according to some, work in a borderless and seamless way, but they generate different types of data, thus heterogeneity. B. Bond, et al. (2000), Rashid, M. A., Hossain, L., & Patrick, J. D. (2002) and Cap Gemini (2015).

Now, social media data and the internet of things have added to the complexity, prompting organizations to wonder how much they should invest in monitoring and analyzing this data. As a consequence companies are turning to analytics technologies not only derive
additional business value from this data, but to reduce the risks of retaining large amounts of heterogeneous information. S. Negash (2004), also Chen, H.et al. (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015).

1.1 Background and problem motivation

As suggested by Tech Target (2015), heterogeneity, in this context, can be described as information different in kind, which comes from different sources. This information must be dealt with in different ways to extract data that is meaningful. The architecture of such systems at the middleware and hardware levels is also heterogeneous so there are multiple sources of heterogeneity. See also Bello-Ortiz, G., Jung, J. J., & Camacho, D. (2016).

An important source of information heterogeneity is the evolution of the web, proliferation of social media, mobile devices, the Internet of Things, business transactions, and content distribution which have been precipitated by technological developments. Social media in particular generates an immense amount of data which is highly un-structured and different. Negash Solomon (2004). See also Chen, H., et al. (2012) and Ranjan, R., Benatallah, B., Dustdar, S., & Papazoglou, M. P. (2015).

However, the ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimize the use of resources. Seth Grimes reported that “80 % of business-relevant information originates in unstructured form, primarily text.” Negash, Solomon (2004) and Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). See also Chen, H., et al. (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015).

One of the problems that come along with this unstructured and heterogeneous data is that relational database systems were not designed to deal with unstructured data. This database approach was satisfying the needs of businesses dealing with static, query intensive data sets since these data sets were relatively small in nature. These traditional relational database systems do not answer the requirements of the increased type, volume, velocity and dynamic structure of the new data sets. Ranjan, R., Benatallah, B. et al. (2015) see also Lenzerini, M. (2002) and Negash, Solomon (2004).

The heterogeneity of the data affects in different ways and it has many ramifications. The different levels of heterogeneity at different layers require a large investment in infrastructure and also in experts who can deal with this information. Different data requires different management and it yields different problems such as the ones described above, furthermore, the disparity of in the data sources can require expertise in many different areas. That is in addition to mathematical and programmatically expertise. Chen, H., Chiang, R. H., et al. (2012), Mohamed H (2015).
1.1.1 A need for further research

The integration of heterogeneous data is a complex process but it is not new. However under the current circumstances there remain research gaps which need to be covered. As pointed out by Hashem, I. A. T., et al. (2015) numerous studies address different issues pertaining storage and processing of large amounts of data in the cloud. This data continues to increase exponentially but the improvement in processing mechanisms is slow, and only a few tools are available to deal with the issues of Big Data.

Today the technological advances and the developments which have occurred as a consequence of these have, precipitated us in the age of Big Data, where huge amounts of data are generated very rapidly. It is heterogeneous, vast and it changes constantly. Contemporary organizations require that this data be processed on the fly because some this data loses relevance as time passes and it becomes less valuable. As we shall see, mobile, cloud based applications, sensor and social media generate a constant stream of data that lack appropriate structure. Transforming and cleaning the bits and bytes generated by these sources before loading them into a warehouse for analysis remains a huge challenge. While efforts have been made towards simplifying the transformation process with some of the technologies mentioned earlier, there is still a need to understand the context of unstructured data formats, particularly when the objective is to extract accurate information. Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013), Jagadish, H. V., Gehrke, J., et al. (2014), Hashem, I. A. T., et al. (2015) and Assunção, M. D., et al. (2015) see also Azvine, B., et al. (2006, June) and Azvine, B., Cui, Z., Nauck, D. D., & Majeed, B. (2006, June)

According to Hashem, I. A. T., et al. (2015) the different state of the art technologies available such as MapReduce, a number of NoSQL databases such as MongoDB, and Query languages such as SPARQL cannot solve the current problem of storing and querying the large masses of data which we are seeing in the age of Big Data. Data quality has become an additional problem because the data is collected from many different sources and have inconsistent formats. In addition, the Apache Hadoop suite lacks solid query processing mechanisms and it lags behind in the data staging area field, which will be covered during the ETL process in the Data Warehousing section and is central to data transformation. See also Jagadish, H. V., Gehrke, J., et al. (2014), Assunção, M. D., et al. (2015), Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U. & Qu, W. (2012) as well as Demirkan, H., & Delen, D. (2013).

In order to tackle these challenges scalable interoperable and easy to use large scale distributed data mining applications, systems and other technologies need to be available. As a consequence there is the need develop new technologies and techniques which expand throughout several fields to cooperate in a joint effort to deal with this matter.
1.2 **Scope**

In this paper the focus is in heterogeneity in the context of data utilized by business organizations, and more particularly in terms of Business Intelligence and Analytics and its related branch of Big Data.

1.3 **Detailed problem statement**

The purpose of this study is to survey the problem of heterogeneity in order to explore the current approaches and their limitations as well as to propose possible solutions to a more efficient integration of data. In order to do this task I find it important to answer the following questions:

- Why do we need to integrate data and who are the primary consumers of the data?
- What are the drivers of the increase information disparity and the consequences?
- What is the nature of heterogeneous data?
- Which approaches to integration of heterogeneous data are there and what are their limitations?
- Which computing paradigm is most adequate, in dealing with vast, heterogeneous and rapidly changing data environment?
1.4 Outline

Chapter 1: Introduction – Gives some perspective and feedback about data heterogeneity in general terms.

Chapter 2: Background theory and related work – This chapter starts by giving some background in the problem of heterogeneity and it shows interrelated fields which deal with data heterogeneity and give an overview of other attempts to integrate heterogeneous data.

Chapter 3: Methodology – During this chapter there is a discussion of the method of choice to put this study together as well as an overview of ethical and societal aspects.

Chapter 4: Results – This chapter explores all the relevant fields, technologies deployed and interrelations, collecting the results for latter analysis.

Chapter 5: Analysis – During this chapter all the gatherings are analyzed and put into perspective and open challenges and research issues exposed.

Chapter 6: Conclusion – Here final conclusion is drafted and it includes some concluding remarks about the problem of data heterogeneity at the present time and suggest future directions.
2 Background theory and Related work

The search queries mentioned in the methodology section yielded a lot of literature pointing towards data warehousing and data mining, particularly during the mid-90’s. Also contemporary to this time, some researches surrounding querying unstructured and semi-structured heterogeneous data starts to emerge.

As interest in refinement and cleaning of data becomes evident, and the purpose of this cleaning seems to be no other than to make the data reliable in order to be mined for valuable patterns. It becomes critical to analyze the data warehousing process, more concretely its ETL process, which is at the core of the data warehousing approach to data integration, and it was a first attempt to integrate data originated on different sources. Kakish, K., & Kraft, T. A. (2012).

During the time interval which spans from the late 90’s and the beginning of the 2000’s research surrounding the semantic web and integration by web services emerges and it becomes evident that interest expands also to semantic data integration. The concept of business intelligence starts emerging, which uses data warehousing as a means to integrate data, and which makes extensive use of web services. This is why background on SOA and related technologies becomes necessary Ziegler, P., & Dittrich, K. R. (2004), Parapatics, P. (2007, June) and Telang, A., & Chakravarthy, S. (2007).

The quantity of data generated during the semantic web 2.0 starts becoming problematic because conventional system used previously to integrate heterogeneous data can no longer process such despair amount of data. It becomes apparent that new ways to deal with this rapidly changing vast and heterogeneous data are necessary; this marks the beginning of the age of Big Data. As time progresses further other related fields start emerging particularly that of Big Data Analytics which is related to Business Intelligence and Analytics and it deals with the questions of vast, rapidly changing heterogeneous data. In order to understand the efforts made towards integrating massive amounts of data it becomes critical to understand the field of Big Data and the computing paradigms deployed in such projects such as Cloud computing, virtualization, Hadoop and related technologies W. Kaisler, S., et al. (2013, January) and Chen, J., et al. (2013), latter Chen, H., Chiang, R. H., et al (2012), Mohammed, J.(2015) and Hashem, I. A. T., et al. (2015) as well as Assunção, M. D., et al. (2015) and Che, D., et al. (2013), Khan, et al. (2014).

These are the reasons behind the chosen theory and related work. The following fields are not surprisingly very neatly tight together and as we shall see later, the integration of and processing of the increasing amounts of large volumes of heterogeneous and rapidly changing data cannot be achieved without a close cooperation amongst these fields.

2.1 Data heterogeneity – overview

One of the most important challenges for integrating different autonomous data sources is the heterogeneity which can appear at different levels. The hardware on which two information sources are developed, the network protocols, the software, the data and the query languages may be different. Kermanshahani, S. (2009),
Within the data integration community as well as in the data mining, problems that might occur as a consequence of heterogeneous data sources are classified along two categories: structural heterogeneity and semantic heterogeneity. Structural heterogeneity implies that different sources store the data into different structures and Semantic heterogeneity refers to the content of the data items and their intended meanings. It is perceived that semantic heterogeneity is the most complicated Liu, H., & Dou, D. (2008) and Kermanshahani, S. (2009).

Richard Hull (1997) cited by Kermanshahani, S. (2009) broadly categorizes heterogeneity into two perspectives: Platform (Systems) and Logical (Semantic). The platform perspective contains hardware, data model, DBMS and APIs that are supported and in which network communication protocols, middleware, and standards such as ODBC JDBC and CORBA can help manage this type of heterogeneity. On the other side we have Semantic heterogeneity which deals with the different ways in which similar world entities are modelled. Semantic heterogeneity is focused in the representation of data at the logical level. See also de Ferreira Rezende, F., Hergula, K., & Daimler-Benz, A. G. (1998, August) and Ziegler, P., & Dittrich, K. R. (2004).


Set aside the intrinsic semantic and structural categorizations of data heterogeneity, the problem of heterogeneous data is that disparate data as it is generated might be of little value as is. As a consequence data has to be processed and integrated, in order to be able to mine and extract meaning from it. The result of the integration and mining efforts are trends, patterns and other insights which are valuable to an organization. We shall see that to these ends, organizations recur to the field of BI&A which is a field that amalgamates a set of technologies and methods that are used to capture, processes, integrate and extract meaning from data which is later used in order to take organizational decisions.


### 2.2 Business Intelligence and Analytics

Business Intelligence and Analytics (BI&A) was born within the industrial world in the early 90’s. Chen, H., Chiang, R. H., et al. (2012) points out that BI&A is a data centric approach and it has its roots in the data base management field. Data management and warehousing

As suggested by Watson, H. J., & Wixom, B. H. (2007), after moving data from a set of source systems into an integrated data warehouse, i.e. what is commonly referred to as Data Warehousing, it gains value to an organization and is defined as Business Intelligence at the point where data is extracted and analyzed. See also Shollo, A., & Kautz, K. (2010).

As suggested by a Gardner survey, pointed out by S. Negash (2004), the strategic use of Business Intelligence can be classified as follows:

- Corporate Performance Management
- Optimization of customer relation
- Monitoring business activities
- Decision support


As pointed out by Mohammed, J. (2015), Chen, H., Chiang, R. H., et al. (2012) and also by Watson, H. J., & Wixom, B. H. (2007), Business Intelligence became a popular topic within the business and IT communities during the 1990s. During the late 2000 Business Analytics was introduced. More recently Big Data and Big Data Analytics have been used to describe super large data sets and analytical techniques that are extremely large and complex and are outside the scope of what traditional systems can handle.

S. Negash (2004), Chen, H., Chiang, R. H., et al. (2012) and Lim, E. P., et al. (2013) and later Mohammed, J. (2015) point out that data warehousing is a first step in any attempt to mine data or be used in any BI&A system, see also Watson, H. J., & Wixom, B. H. (2007).

2.3 **Big Data**

Advances in technology, particularly in data storage capacity and processing power and networking, together with the evolution of the web and more particularly social media and the internet of things have contributed to an ever increasing volume of data that can now be analyzed by those companies who have the knowhow and the computing power to do it. Other development in corporate transactional systems such as ERP and CRM has contributed to the ever increasing volume of data. The evolution of the web has also contributed to this trend with the development of social media and the sensor based applications. W. Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January) and Chen, J., et al. (2013), latter Chen, H., Chiang, R. H., et al. (2012), Mohammed, J.(2015) and Hashem, I. A. T., et al. (2015) as well as Assunção, M. D., et al. (2015) and Che, D., Safran, M., & Peng, Z. (2013, April).
Big Data is characterized by the so-called 3V’s: Volume, Velocity, and Variety. Big Data Analytics is a related area of B&IA and similarly it requires that the information be extracted from multiple sources, transformed to fit the analytical needs and loaded into a warehouse. Intel (2013), Dong, X. L., & Srivastava, D. (2013, April), Chen, J., et al. (2013) and Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014).

The management of Big Data, and the intelligent use of large, heterogeneous datasets, is becoming increasingly important for competition and are affecting all sectors. Srinivas B., & Togiti B. (2015). See also Hashem, I. A. T., et al. (2015).

The concept of Big Data is not a new concept; it has in fact been around from the earliest days of computing. It originally meant the volume of data which could not be efficiently processed with traditional databases, methods and tools Kaisler, S., et al. (2013). See also Gudivada, V. N., Baeza-Yates, R. A., & Raghavan, V. V. (2015).

However, the original definition focused on structured data, but this conception has shifted in recent years and most experts have come to realize that most of the information worldwide comes from unstructured data. This data is what yields most value to an organization. Kaisler, S., et al. (2013), see also Gandomi, A., & Haider, M. (2015), see also Hendler, J. (2014).

As a consequence, the hegemony of structured data in analytics has been altered by the arrival of semi-structured data such as XML, RSS, Feeds, and unstructured data such as the human language, text and other data difficult to be classified such as audiovisual data, data coming from other devices and multidimensional data such as the one that can be extracted from a data warehouse. This implies that we are before a so vast data mix as never seen before. Russom, P. (2011). See also Kaisler, S., et al. (2013).

The most visible application of Big Data Analytics has been in business enterprises. It is estimated that a retailer fully utilizing the power of analytics can increase its operating margin by 60%. Utilizing new opportunities (for e.g., location-aware and location-based services) leads to significant potential for new revenues. Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). See also Assunção, M. D., et al. (2015).

Due to the wide and diverse data sources and the speed and quantities of which it is acquired, heterogeneity is a characteristic as well as a key factor of Big Data. Wu, X., Zhu, X., et al. (2014) and Labrinidis, A., & Jagadish, H. V. (2012).

Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Kaisler, S., et al. (2013).

2.4 Data Warehousing

There has been a struggle from the 1980 to recognize and store information in the most efficient way. The early 80's marks the beginning of the attempt to design systems that allow interoperability of databases. John Miles Smith, et al. (1982).
The University of Minnesota made the first successful attempt to build a data integration system for the Integrated Public Use Microdata Series (IPUMS). This system was driven by structured Meta data and IPUMS used a data warehousing approach which extracts, transforms and loads data from heterogeneous sources into a single view or global schema which yielded what was to be known as ETL (Extract, Transform and Load). Widom, J. (1995, December), Ruggles, S., et al. (1995), Fan, H. (2005).

A formal and definition comes from W. H. Inmon (2002), cited by Hao Fan (2005). This definition states that:

A data warehouse is a subject-oriented, integrated, nonvolatile and time-variant collection of data in support of management’s decisions.

The implications of the definition are summarized in table 1 below:

<table>
<thead>
<tr>
<th>Characteristics of data warehousing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject oriented</td>
</tr>
<tr>
<td>Information that is relevant to the organization</td>
</tr>
<tr>
<td>Integrated</td>
</tr>
<tr>
<td>data comes from multiple sources.</td>
</tr>
<tr>
<td>Nonvolatile</td>
</tr>
<tr>
<td>data is not updated in real time.</td>
</tr>
<tr>
<td>Time Variant</td>
</tr>
<tr>
<td>the information collected is from a past point</td>
</tr>
</tbody>
</table>

*Table 1* Defining characteristics of data warehousing. Source: Hao Fan (2005).


*Figure 1* Simplified illustration of a traditional data warehouse architecture. Source: DW Mantra (2007)
Jennifer Widom (1995) points out that the ETL process consumes most of the data warehouse resources; this approach entails a large investment in infrastructure as well and it is an, often complex, process that requires highly skilled staff.

There are many applications that benefit from integrated information. For example, in the area of Business Intelligence (BI), where integrated information can be used for querying and reporting on business activities, for statistical analysis, online analytical processing (OLAP) and data mining, forecasting, aiding in decision making and enterprise planning, all in order to gain competitive advantage. Ziegler, P., & Dittrich, K. R. (2004) and S. Negash (2004), later Watson, H. J., & Wixom, B. H. (2007), Chen, H., Chiang, R. H., et al. (2012), Mohammed, J. (2015), Parapatics, P. (2007, June), Kakish, K., & Kraft, T. A. (2012), Oracle (2005).

Organizations use data warehousing to support strategic and mission-critical applications. Data deposited into the data warehouse must be transformed into information and knowledge and appropriately disseminated to decision-makers within the organization and to critical partners in various capacities within the organizational value chain. March, S. T., & Hevner, A. R. (2007), Parapatics, P. (2007, June).

### 2.5 Data Integration

Data integration attempts to combine heterogeneous data, structured and unstructured, often simultaneously from different sources. The goal is building a homogeneous interface which can provide end users with querying capabilities over these autonomous sources. This is done with a set of data transformation and integration tools, it helps in analyzing and putting together meaningful data which resolves heterogeneity in data structure and semantics. It involves combining data residing in different sources and providing users with a unified view of these data. This raises many challenges among which the heterogeneity of data sources, the fragmentation of data, the processing and optimization of queries appear to be the most important. Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., & Widom, J. (1995, March), Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., & Widom, J. (1995, December), Abiteboul, S. (1997, January), Kermanshahani, S. (2009), Lezerini M. (2002), Ziegler, P., & Dittrich, K. R. (2004), Cali, A., Calvanese, D., De Giacomo, (2013) and Telang, A., & Chakravarthy, S. (2007) see also Jagadish, H. V., Gehrke, J., et al. (2014).

This process occurs in many situations, which include both commercial, for example when companies need to merge their databases or there is a need to reconcile information which comes from different sources such as front and back end systems, and scientific domains, for example when combining research results from different bioinformatics repositories. Ziegler, P., & Dittrich, K. R. (2004), Chen, H., Chiang, R. H., et al. (2012).

The unified view of this data is structured according to a global schema, which represents a level of the integrated and reconciled data and which provides the elements for expressing queries over a data integration system. This abstraction frees the user from the knowledge on where the data resides and how the data is structured and also on how the data is to be merged and reconciled to fit the global schema. Cali, A., Calvanese, D., et al. (2013).
Interest in these systems has been growing during the last years and many companies still face the problem of integrating data residing in different heterogeneous sources. Companies who build a Data Warehouse, engage in Data Mining. Cali, A., et al. (2013).

As pointed out by Telang, A., & Chakravarthy, S. (2007), Data Integration is not a new topic and it has actually been around for the last two decades and many techniques have evolved since then. Some of these developments can be attributed to the development of different technologies in the field of Warehousing, Database Systems, Artificial Intelligence, Information Retrieval and Cloud Computing and Virtualization, among other, which are necessary to accomplish the integration effort.

Other developments in the area of information retrieval, for example, the arrival of XML have contributed to the data integration effort. XML is a markup language that serves as a uniform file format to transfer information. This markup language defines a set of rules for encoding documents in a format that is human readable. Parapatics, P. (2007, June). XML has addressed many issues of data integration it focuses on documents and the language is widely used for the representation of arbitrary data structures such as those used in web services. Besides it has contributed in other ways such as being able to deal with semi-structured data. However, other issues such as the issue of semantic heterogeneity are more complicated to solve. See D., Halevy, A. Y., & Weld, D. S. (2001), Cruz, I. F., & Rajendran, A. (2003), Halevy, A., Rajaraman, A., & Ordille, J. (2006, September) and latter Parapatics, P. (2007, June) and Madria, S., Passi, K., & Bhowmick, S. (2008).

One of the most complicated problems in dealing with data integration is the aspect of semantic heterogeneity. As described by Elmagarmid and Sheth, (1999) and cited by Kermanshahani, S. (2009), semantic heterogeneity refers to the difference in meaning and the use of data. See also Telang, A., & Chakravarthy, S. (2007) and Ziegler, P., & Dittrich, K. R. (2004).


2.6 Data Mining and knowledge discovery

The process of Data Mining is a process that comes after the Data Warehousing effort. With increasing processing power Data Warehousing became the basis for Data Mining. So Data Mining and Data Warehousing often go hand in hand. Parapatics, P. (2007, June) and Bill Palace (1996).

Data Mining is a step in the knowledge discovery process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data. Fayyad, U., et al. (1996).
Because there are different kinds of data and databases used in different applications, one of the big challenges of a knowledge discovery system is that it should be able to extract valuable insights from different kinds of data. Databases contain structured data and complex data objects such as hypertext, multimedia data, spatial and temporal data, transaction data and other types of data which these systems should be able to handle.

Chen, M. S., Han, J., & Yu, P. S. (1996). As defined by Hand, D. J., Mannila, H., & Smyth, P. (2001), Data Mining is the process of analyzing observational datasets to find new relationships and to summarize the often large scale data, in ways that are both understandable and useful to the data owner.

These relationships and summaries that are discovered and prepared through the Data Mining Process are referred to as models and patterns. Examples of which include linear equations, rules, clusters, graphs, trees and recurrent patterns in time series. Hand, D. J., et al. (2001).

Data Mining is performed generally by companies with a strong consumer focus, i.e. retail, financial, communication, and marketing organizations. It enables these types of organizations to determine relationships among “internal factors” like price, product positioning, staff skills and so on and “external factors” such as economic indicators, competition and customer demographics. Bill Palace (1996). See also S. Negash (2004), also Chen, H., et al (2012), Lim, E. P., et al (2013), and Mohammed, J., (2015).

Advances in the field of computer science technology and the developments of the Web have contributed to the evolution of Data Warehousing and Data Mining. Some of the most important factors are processing and data storage capacity but in particular the evolution of the Internet. Parapatics, P. (2007, June). See also S. Negash (2004), also Chen, H., et al (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015).

Drifting towards Web 2.0 social applications became popular. Due to increased connectivity, of computer systems and the possibility of machine to machine communication an increased number of distributed, heterogeneous data sources has become available. Parapatics, P. (2007, June), S. Negash (2004), also Chen, H., et al (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015). As a consequence, the internet has become the largest data source today, and web mining has become an important field. The evolution of the Internet, aided by technological advances and stimulated by the social media has given place to different additional trends such as that of multimedia mining and most importantly an increase in real time knowledge extraction which includes real-time warehousing as well as real-time data mining. Fayyad, U., et al. (1996).
an increase in real time knowledge extraction which includes real-time warehousing as well as real-time data mining. Fayyad, U., et al. (1996).

2.7 Cloud Computing and Virtualization

Cloud Computing refers to both the application and the hardware and system software delivered as a service over the internet.

Virtualization is a process of resource sharing and isolation of the underlying hardware to increase computer resource utilization, efficiency and scalability. Hashem I. A. T., et al. (2015).

Cloud Computing is a significant shift in modern computing. Service for the enterprise applications has become a powerful architecture to perform large-scale and complex computing. This new computing paradigm provides virtualized resources, parallel processing, security, and data service integration with scalable data storage. It can significantly minimize the cost and restriction for automation and computerization by individuals and enterprises and provide reduced infrastructure, maintenance cost and a more efficient management and user access. Hashem, I. A. T., et al. (2015).

Cloud Computing is a fast growing field and has emerged as one of the major players. This technology offers an affordable and effective architecture for large scale complex enterprise applications Sharma, S., Tim, U. S, et al. (2015) and Hashem, I. A. T., et al. (2015).
3 Research Method

Within the context of research it is important to discern between research methods and research methodology. Research methods can be defined as the various ways of conducting a research. Research methodology on the other side is concerned as how a research is to be done scientifically and it is rooted in the existence of methods. Kothari, C. R. (2004).

According to Kothari, C. R. (2004), there are different basic options in regard to the research methods which can be deployed in a research. Some are listed in table 2 below:

<table>
<thead>
<tr>
<th>Research types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive vs Analytical Research</td>
<td>Includes surveys, and fact finding techniques of different kinds. The major purpose of descriptive research is description of the state of affairs as it exists at present</td>
</tr>
<tr>
<td>Applied vs Fundamental Research</td>
<td>Applied research aims at finding a concrete solution of a problem while fundamental research is mainly concerned with generalizations and with the formulation of a theory.</td>
</tr>
<tr>
<td>Quantitative Vs Qualitative Research</td>
<td>Quantitative research deal with issues that can be expressed in terms of quantity while qualitative research deals with phenomena related or involving quality or kind.</td>
</tr>
<tr>
<td>Conceptual vs. Empirical</td>
<td>Conceptual research is related with abstract ideas or theories. Generally used in philosophy while empirical researches relies on experience or observation alone.</td>
</tr>
<tr>
<td>Other types</td>
<td>Other types are variations of one or more of the above.</td>
</tr>
</tbody>
</table>

Table 2 shows different types of research methods Source: Kothari, C. R. (2004).

However, most of the above can be classified into two basic approaches to research: the quantitative and the qualitative approaches. Kothari, C. R. (2004).

For the purpose of this study a qualitative research was performed because it better allows capturing the knowledge needed to understand the subject in question and permits filling gaps from knowledge captured from similar studies.

3.1 Research Methodology

This research is based on a literature review methodology. As defined by Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. (2012), a literature review represents a logically argued case founded on comprehensive understanding and knowledge of a topic. This type of research methodology approach is normally deployed as part of an academic assessment. See also Cronin, P., Ryan, F., & Coughlan, M. (2008) and University of Bedfordshire (n.d.).
Two types of literature reviews were considered. The first option considered was a systematic literature review, and the second one a traditional narrative literature review.

The systematic literature review is a method recommended for aggregating evidence. This was the primary objective of this research i.e. aggregating previously existing evidence which came from previous researches. In addition it works well within the computer science field. B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009) and Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007).

The University of Berdforshire (n.d.) points out that a key aspect of a systematic literature review is the systematic part of it. This implies that there must be an order and structure on how the research in question has been undertaken. See also Cronin, P., Ryan, F., & Coughlan, M. (2008) and Onwuegbuzie, A. J., et al. (2012).

On the other side, the traditional or narrative literature critiques and summarizes a body of literature and draws conclusions about the topic in question. Furthermore, its purpose can be attributable to providing the reader with comprehensive background knowledge and to underline the directions and significance of new research Cronin, P., Ryan, F., & Coughlan, M. (2008) which better coincided with the purpose of this study.

The chosen methodology is a systematic literature review because the intention behind this research is to focus on summarizing and synthesizing the arguments and ideas that exist from previous studies and analyzing critically. This is done in order to give background knowledge into the topic of data heterogeneity without adding a substantial new contribution which according to Ramdhani, A., et al. (2014) differentiates it from an academic research paper. Consequently the literature review is made up of relevant studies and knowledge that addresses the subject area of data heterogeneity Cronin, P., et al.(2008).

Performing a narrative literature review was necessary mainly because the purpose is to provide the reader with knowledge and inspire research ideas by identifying gaps, which is the purpose of the narrative or traditional literature review, as described by Cronin, P., et al. (2008). However, I found elements of the systematic literature review option also useful, in particular in terms of data collection, and chronology.

As pointed out by Parhoo (2006) cited by Cronin, P., Ryan, F., & Coughlan, M. (2008) a systematic review should detail the time frame within which the literature was selected as well as the methods used to evaluate and synthesize the findings in order for the reader to assess the reliability and the validity of the review. There were advantages in analyzing the articles chronologically. This was done in order to gain perspective into what is interesting during a specific time and to keep track of what challenges have been met. See also Ramdhani, A., et al. (2014).

### 3.2 Data collection and processing

This research starts with the identification and collection of literature. A collection of 149 articles were gathered of which 58 were dropped for the research purposes. The literature was collected from the period stretching from 1995 all the way to 2016 and it originates Mostly from Google Scholar.
This literature consists mainly on journals, white papers, articles, previous literature reviews and reports on overlapping fields which revolve about the problem of integrating heterogeneous data and the related fields that deal with it. This is in line with the data collection procedures which are mentioned by Cronin, P., Ryan, F., & Coughlan, M. (2008) and Ramdhani, A., et al. (2014) for the systematic literature review.

Once the main subjects were identified a further more detailed round of queries was performed containing the queries in table 3 below.

<table>
<thead>
<tr>
<th>Query search on Google Scholar</th>
<th>Business Intelligence</th>
<th>Data Warehousing</th>
<th>Data Integration</th>
<th>Data Mining</th>
<th>Big Data</th>
<th>Cloud computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Intelligence and Analytics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business intelligence and data integration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real time business intelligence problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research gaps in data integration + business intelligence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business Intelligence and Big Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Warehousing and Heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The ETL process and data heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Staging and data heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Cleaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data warehousing and semantic heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NoSQL databases and data heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>data staging and data heterogeneity + big data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The problem of Data Heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data integration and data heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data integration and Semantic Heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data integration challenges and solutions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data integration evolution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data integration approaches</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data mining and data heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Mining and Big Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data mining and business intelligence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Mining and Data Warehousing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Issues and Problems with Data mining and heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Mining and Big data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Integration and Big Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Warehousing and Big Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big Data integration problems and issues</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big Data and Cloud Computing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big Data Processing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research gaps in Big Data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloud Computing and Data Heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cloud Computing and Data Integration
ETL and Cloud Computing
Cloud Computing: Issues and Challenges
Cloud Computing research gaps

Table 3 Shows the search queries deployed on this research and the fields identified using such queries.

The data was collected longitudinally, i.e. over time, as well as cross sectional, i.e. at the same point in time. Of the six data collection methods i.e. literature search, file review, natural observation, surveying, expert opinion and case studies the first was the one deployed.

The data was triangulated by means of gathering data from several sources i.e. Academic journals, white papers, articles, books and other literature reviews.

As a way to facilitate the classification of the review literature for its analysis different folders were created and named with the year of the research papers, and the results pertaining to those years were put in that folder.

Inclusion and exclusion criteria

In order to assess whether it contained relevant material, the abstracts for of each respective paper were examined as well as the conclusion and the keywords, when found. The document was also searched for the following keywords:

- Data heterogeneity,
- Heterogeneous data integration,
- Data heterogeneity,
- Data integration,
- Disparate data,
- Data differences,
- Data reconciliation.

If the data did not contain explicit or implicit information, i.e. concepts relevant to the subject of integration of heterogeneous data, in the abstract and it did not contain any of these keywords either after the abstract or throughout the body, or variations of it, such as in the endings, for example heterogen or dispar, the research was discarded.

3.3 Search strategy

Two iterations of search were performed

1. The search was conducted in such a way that the first iteration of search consisted on documents, general research reports, published, papers and books in the area. So as to explore theories and concepts related to the area in question and examine generalizations derived from the issues considered.
2. The second iteration consisted on specific studies in the area of interest in a chronological perspective. In order to gather inputs to answer the questions of this research and see the progression of the area in question.

The search was conducted on resources that ranged the following:

- Technical reports
- Academic journals
- Literature reviews
- Books

As suggested by Cronin, P., Ryan, F., & Coughlan, M. (2008) and Ramdhani, A., et al. (2014) in conducting a record of the framework it is important to keep records of the keywords and methods used in searching the literature. The search criteria consisted on queries to google scholar in which a variety of statements where used which are summarized in table 3 at the end of this chapter. Where possible, i.e. when an article was perceived relevant for the purpose of this study, its related articles were followed and more abstracts were read. The queries to the database used where performed during different time intervals ranging from 1995 to 2016.

Finding the source of the problem was a first step towards understanding the problem of data heterogeneity. As a way to achieve this objective, the causes of data heterogeneity were researched first. In order to do this a variety of search criteria was deployed to gain an overview of the reasons for such problematic.

Amongst the first journals to come up upon searching with these queries, there are already some abstracts pointing out to data cleaning and the ETL process which was clearly identifiable to the data-warehousing activity.

The review of these articles yielded knowledge that helped determine the causes of this heterogeneity and pointed out to the direction of data integration efforts. From these readings, it became evident that the reason to undertake such effort was to extract value from it, which in turn pointed out to the direction of Data Mining, Business Intelligence and Analytics and other related technologies such as cloud computing and virtualization which are also related.

3.4 Quality assessment

As pointed out by Onwuegbuzie, A. J., et al. (2012) using literature from multiple sources, allows for the reviewer to combine information from multiple sources and types which in turn are beneficial because more data can be extracted and more meaning can be generated enhancing in this way the quality of the synthesis.

To this end, more than one source for the same topic was selected. After this, the information was contrasted and even put into chronological perspective, which implied contrasting the results with other results from previous years on the same topics, for example, data warehousing in 1995, data warehousing in 2000 and data warehousing on 2010. Even though sometimes it yielded very similar results, it was indeed possible to extract additional details and cross check the validity of the arguments used. In this way it
was also possible to discern what improvements were made if any and what challenges remained unsolved.

As suggested by Cronin, P., Ryan, F., & Coughlan, M. (2008) journals are more up to date I have favored the use of journals because the data in them better represents the period of time in which they were published.

### 3.4.1 Validity

The validity of this study can further be assessed because the study is conducted with a known methodology i.e. Systematic Literature review and Traditional Literature review, this implies that it has an appropriate time scale, and it is also conducted following a methodology which is known to yield positive results in aggregating previous knowledge, in particular in the computer science field i.e. the Systematic Literature review as proposed by Brereton, P., Kitchenham, B. A., et al (2007) and later B., Brereton, et al (2009).

**Triangulation**

Patton, M. Q. (1999) point out that the purpose of triangulation in a qualitative research is to increase the credibility i.e. validity of the results. This is so because by combining multiple observers, theories, methods, and data sources, researchers can try to overcome the skepticism that singular methods, lone analysts, and single-perspective theories or models bring along.

There are different types of triangulation possible:

- Analyst triangulation- Utilizing multiple analysts to review findings.
- Methods triangulation – checking out the consistency of findings
- Triangulation sources.- Examining the consistency of different data sources within the same method .
- Theory / perspective triangulation to interpret the data.

While falling short on deploying several analysts to perform the study, as mentioned earlier, more than one source for the same topic was selected. This source triangulation and the different theories utilized enhance the validity of findings. This is also done by deploying two methodologies and examining the results from several perspectives.

### 3.5 Societal and moral aspects

The London School of Economics and Political Science (2016) points out that although any particular empirical investigation may be modest in scope, if it entails human participants, it is nonetheless essential that staff, students and supervisors should consider and address any ethical implications that may pertain to the project.

This research is conducted following a qualitative methodology in which it was not necessary to collect data from human subjects by means of interviewing or experimenting. The subject matter of this research might very well have social implications such as privacy, security, data integrity and other aspects inherent from the subject of Integration of Heterogeneous Data.
The contribution of this research subject to society is significant because it shows combinations of technologies and disciplines that allow businesses to better streamline the delivery of products and services towards specific needs and specific characteristics in a much more effective way which in turn contributes to make businesses more prolific.

However, the technologies discussed within this research can contribute to increase vulnerabilities in terms of privacy and security for the individual. Great care should be placed when manipulating data, which shall be done in any case in good faith and with good ethics.

As suggested by Logan University (2016), and also by Bell, E., & Bryman, A. (2007), this project is conducted by reporting the findings with honesty. It is not the intention to misinform and / or mislead by intentionally misinterpreting the sources of information. On the other side appropriate credit has been given when utilizing work done by other persons.
4 Results

After iterating through the literature gathered following the methods described in the methodology section it was possible to gather background knowledge based on a sufficiently ample population size of relevant related works and theoretical background which serves the purpose so as to extract results from the literature subject of investigation and further discussion and analysis.

4.1 Data Heterogeneity

Heterogeneity restricts the efficiency of data conversion. Thus if such format conversion is made more efficient the application some fields which require data integration such as Business Intelligence and Analytics, and Big Data can benefit from it. Chen, J., et al. (2013) and Chen, M., et al. (2014).

Several kinds of heterogeneity can be identified including disparity in hardware and operating systems, data management software, data models, schemas, and semantics, syntax, middleware, user interfaces, and business rules and integrity constraints Ziegler, P., & Dittrich, K. R. (2004).

Liu, H., & Dou, D. (2008) point out that Data heterogeneity falls along two broad lines:

- Structural Heterogeneity
- Semantic Heterogeneity

Härder, T., Sauter, G., & Thomas, J. (1999) explains that Structural heterogeneity relates to where the data is stored i.e. different sources store the data into different structures. These systems could be flat files, legacy systems, other databases and other systems that are more frequently than not distributed geographically.

Generally, application programs require access to data which frequently require un-normalized schemas to gain speed. These are tailored to specific access profiles. As a consequence of this, the structures of the schemas differ with the applications and their access characteristics. Härder, T., Sauter, G., & Thomas, J. (1999)

What happens at the business level is that organizations tend to have data distributed along departmental and functional lines and it leads to fragmented data resources and services and the emergence of the so called islands of information. In other words, data is managed by disparate management systems from different vendors and different operating systems that might use different network protocols. Ferreira Rezende, et al. (1998, August).

Ferreira Rezende, et al. (1998, August) states that corporate data resources are made up of multi-vendor database (DB) servers, legacy and current data, and relational and non-relational data sources and such autonomous data sources have no ability to relate data from these heterogeneous data sources within the organization.
As explained by Kermanshahani, S. (2009), a way to deal with this kind of heterogeneity is through Network communication protocols and the standards such as ODBC, JDBC and CORBA can help to mitigate the effects of this disparity through what we call Middleware which is a generic term referring to a system layer of software that tries to overcome this type of heterogeneity.

One type of heterogeneity which is deemed, by some, easy to solve is the syntactical heterogeneity. This type of heterogeneity can be overcome with technologies such as XML which provide a uniform representation of information which is encapsulated in a standard language. Parapatics (2007) and Ziegler, P., & Dittrich, K. R. (2004).

The heterogeneity of semantics, on the other side, which is considered the most complex, is the other aspect that prevents systems to be reconciled in a simple way and it has to do with the meaning of things. Härder, T., Sauter, G., & Thomas, J. (1999) point out that the DB design is influenced by the needs of a particular application to optimize performance at run time. Integrity constraints are frequently embedded, distributed and replicated, and as a result it prevents a uniform system control and enforcement of data semantics, so at the schema level, only a partial description of the application semantics can be conceived. This causes that it is not possible to fully automatize capturing all aspects of semantics. See also Liu, H., & Dou, D. (2008).

In addition data can be structured or structured. And it must be dealt with simultaneously. Data can be in the form of videos, files of all kinds, memos, spreadsheets, phone conversations coming from all kinds of sources. Semi-structured data is more complicated to deal with because of indexing issues i.e. it is not easily searched using existing tools for conventional databases. S. Negash (2004).


Kondylakis, H., et al (2009, November), Kermanshahani, S. (2009) and Shvaiko, P., & Euzenat, J. (2013) suggest that ontology matching is the main technology to overcome the problem of semantic heterogeneity and data integration approaches that are based on ontologies or meta data are also known as semantic data integration approaches and schema mapping is the solution adopted to solve heterogeneity in a data integration system.

Hashem, I. A. T., et al (2015), Roche, M. (2015), Hendler J. (2014) suggest that heterogeneity is the result of the growth of almost unlimited different sources of data and problem gets bigger as the data grows, in fact, as we shall see later on, data heterogeneity is one of the defining factors of Big Data. See also Jagadish, H. V., Gehrke, J., et al. (2014).

Hashem, I. A. T., et al. (2015) further suggests that data from multiple sources are generally of different types and representation forms and significantly interconnected and the more breadth in diversity and volume the more complicated to deal with.
In summary, heterogeneity arises from the deployment of different technologies and methods to store data. Simultaneously technological advances contribute to the proliferation of this data. There are different types of heterogeneity which can be considered amongst which the semantic heterogeneity is the most complicated to deal with and for which ontologies are the best alternative. The structural heterogeneity can be dealt with networking protocols and standards, particularly through the use of middleware. Finally the evolution of information systems is causing that this task is more complex, particularly as the data grows and becomes more disparate.

4.2 Reasons to integrate data

Data acquisition in itself does not have any value, and does not add value to an organization if it is not processed. Since the data comes from disparate sources and formats i.e. it is heterogeneous, in order to have any value at all it has to first be integrated, which is frequently done in a data warehouse. Then it has to be Mined and Analyzed and finally it becomes useful for BI&A and decision takers which are the primary consumers of this data. S.Negash (2004), Watson, H. J., & Wixom, B. H. (2007), Kumar, A., Tyagi, A. K., & Tyagi, S. K. (2014), Chen, H., Chiang, R. H., et al. (2012) and Mohammed, J. (2015).

4.2.1 Business Intelligence

Business Intelligence and Analytics (BI&A) and the related field of Big Data Analytics have become increasingly important in both the academic and the business communities over the past two decades and probably the biggest beneficiaries of data integration. Organizations within the business realm, resource to Business Intelligence and Analytics (BI&A) and mine data because it can provide knowledge and insights which can help increase performance. It is decision takers through Business intelligence and analytics which consume this data. This data is frequently found dispersed throughout multiple sources and Business Intelligence and Analytics and its related branch of Big Data Analytics require the integration of numerous sources of disparate data in order to gain value from the discovery of valuable business patterns and trends. This is true now more than ever. S. Negash (2004), Watson, H. J., & Wixom, B. H. (2007), Lim, E. P., et al. (2013), Kumar, A., et al. (2014), Chen, H., Chiang, R. H., et al. (2012) and Mohammed, J. (2015).

4.2.1.1 Business Intelligence Inputs

S. Negash (2004) points out that Business Intelligence is an evolution of previous systems designed to support decision making. See also Herschel, R. T., & Jones, N. E. (2005).


<table>
<thead>
<tr>
<th>Inputs of Business Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLAP</td>
</tr>
<tr>
<td>Real time data warehousing</td>
</tr>
<tr>
<td>Data Mining</td>
</tr>
<tr>
<td>DSS</td>
</tr>
<tr>
<td>ERP</td>
</tr>
<tr>
<td>Visualization</td>
</tr>
</tbody>
</table>
Some of the data generated by these systems is structured and other unstructured and thus heterogeneous and BI&A systems have to deal with it simultaneously S. Negash(2004).

Seufert, A., & Schiefer, J. (2005) points out that that the real time characteristic is only possible and mainly influenced by the refresh cycle of the data warehouse system in which the data is stored. This will be investigated during the Data Warehousing Section.

Azvine, B., Cui, Z., Nauck, D. D., & Majeed, B. (2006, June) as well as S. Negash (2004) coincide in that Business Intelligence includes an amalgam of different features which clearly indicate that it has many different characteristics and heterogeneous as well. Some characteristics are oriented to data reporting and visualization, other are oriented to business performance management and yet other are oriented to data extraction transformation and integration. These can be classified in three main categories:

1. Data warehouses
2. Analytical tools
3. Reporting tools.

This amalgam of different interoperating technologies and challenges required can be grouped in layers as summarized in table 5 below:

<table>
<thead>
<tr>
<th>Three layers of BI stack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational Layer</td>
</tr>
<tr>
<td>Data Integration Layer</td>
</tr>
<tr>
<td>Analytics</td>
</tr>
</tbody>
</table>

Table 5 illustrates the different layers of the BI stack. Source: Azvine, B., Cui, Z., Nauck, D. D., & Majeed, B. (2006, June)

4.2.1.2 On-Line Analytical Processing (OLAP) and On-Line transaction processing (OLTP)

Chaudhuri, S., & Dayal, U. (1997) point out that in order to facilitate analysis and visualization, at the integration layer, the data in a warehouse is modeled multidimensional i.e. warehouse, product, location, sales person and so on, and cubes are frequently extracted from the data warehouse and made available to managers for specific decision making situations. This data is usually segregated from operational data See also March, S. T., & Hevner, A. R. (2007) and Kakish, K., & Kraft, T. A. (2012) and earlier Chen, M. S., et al (1996).

Online Analytical Processing (OLAP) operations at the analytic layer include ROLLUP which entails increasing the level of aggregation and DRILL DOWN which entails decreasing the

OLTP applications at the operational layer, automate operating data processing such as order entry, transactions, purchases, inventory status and so on which are structured and recurrent atomic transactions of the everyday operations, so the databases gathering this data are normally very large. These two OLAP and OLTP are different in nature i.e. OLTP is normalized and the other is de-normalized i.e. also heterogeneous Chaudhuri, S., & Dayal, U. (1997) and Herschel, R. T., & Jones, N. E. (2005) and Herschel, R. T., & Jones, N. E. (2005).

The data feed to the analytics layer comes from the Data Integration Layer which is where the Data Warehousing effort resides. This feed to the analytics layer requires data from different operational systems and sources i.e. are heterogeneous and they are fused together. At the integration layer it is important to take in consideration that the data layer needs to provide quality and accurate data to the analytics layer. Azvine, B., Cui, Z., Nauck, D. D., & Majeed, B. (2006, June). In regards to data quality March, S. T., & Hevner, A. R. (2007) add that it includes consistency and timeliness.

4.2.1.3 Evolution of Business Intelligence

S. Negash (2004) points out that this is due to the technical advances in hardware and software, the emergence of data warehousing, with which is neatly connected to BI&A, as well as other advances in data cleaning and the information generated by other different systems. Furthermore to understand the source of heterogeneity it is important to put into perspective the evolution of the web and BI&A applications. To this end, Chen, H., Chiang, R. H., et al. (2012), Lim, E. P., Chen, H., & Chen, G. (2013) and later Mohammed, J. (2015) identify three stages of evolution of BI&A. They point out that BI&A 1.0 can be traced to data management and warehousing efforts, design of data marts and ETL tools. During the early 2000s Web 2.0 (the semantic web) brought about an explosion of web content, web analytics, and user generated content and social media. The semantic web saw the birth of BI&A 2.0. Finally, the Internet of Things and the proliferation of mobile devices during 2008 onwards have led towards Web 3.0 and although there is not yet a BI&A 3.0 application in the market, we seem to be headed to BI&A 3.0. See table 6 below for the development stages of BI&A.

<table>
<thead>
<tr>
<th>BI&amp;A Version</th>
<th>Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI 1.0</td>
<td>RDBM Data management and warehousing efforts, ETL</td>
</tr>
<tr>
<td>BI&amp;A 2.0</td>
<td>Semantic Web 2.0 – Social Media</td>
</tr>
<tr>
<td>BI&amp;A 3.0</td>
<td>Semantic Web 3.0 – Internet of things</td>
</tr>
</tbody>
</table>

Within this context, the current research framework can be classified in five critical areas which are enumerated as follows:

1. Big Data & Analytics
2. Text Analytics
3. Web Analytics
4. Network Analytics
5. Mobile Analytics

All of which can contribute to BI&A 2.0 and 3.0. Based on these findings Chen, H., et al. (2012), Lim, E. P., et al. (2013) propose a research framework which is summarized in table 7 below.

<table>
<thead>
<tr>
<th>Foundational technologies</th>
<th>[Big]Data Analytics</th>
<th>Text Analytics</th>
<th>Web Analytics</th>
<th>Network Analytics</th>
<th>Mobile Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDBMS</td>
<td>Information retrieval document representation</td>
<td>Information retrieval document representation</td>
<td>Cloud services</td>
<td>Link mining</td>
<td></td>
</tr>
<tr>
<td>Data Warehousing</td>
<td>Query processing</td>
<td>Computational representation</td>
<td>Cloud computing</td>
<td>Community detection</td>
<td></td>
</tr>
<tr>
<td>ETL</td>
<td>Relevance feedback</td>
<td>Search engines</td>
<td>Social search and mining</td>
<td>Dynamic Network modelling</td>
<td></td>
</tr>
<tr>
<td>OLAP</td>
<td>User models</td>
<td>Web crawling</td>
<td>Reputation systems</td>
<td>Agent-based modeling</td>
<td></td>
</tr>
<tr>
<td>BPM</td>
<td>Search engines</td>
<td>Website ranking</td>
<td>Social media analytics</td>
<td>Social -based modeling</td>
<td></td>
</tr>
<tr>
<td>Data Mining</td>
<td>Enterprise search Systems</td>
<td>Search log analysis</td>
<td>Web visualisation</td>
<td>Social influence and information diffusion models</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
<td>Recommender systems</td>
<td>Web based auctions</td>
<td>ERGMs</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td></td>
<td>Web services</td>
<td>Internet monetization</td>
<td>Virtual Communities</td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td></td>
<td>Mobile services</td>
<td>Social marketing</td>
<td>Criminal / Dark networks</td>
<td></td>
</tr>
<tr>
<td>Association analysis</td>
<td></td>
<td>Mashups</td>
<td>Web privacy security</td>
<td>Social / Political analysis</td>
<td></td>
</tr>
<tr>
<td>Anomaly detection</td>
<td></td>
<td></td>
<td></td>
<td>Trust and reputation</td>
<td></td>
</tr>
<tr>
<td>Neural networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generic algorithms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>multi-variate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>statistical analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristic search</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Emerging research         | Statistical machine learning | Information Extraction | Statistical NLP | Link mining |
|                          | Sequential temporal mining  | Question-answering systems | Text NLP | Mobile web services |
|                          | Spatial mining              | Option mining          | Sentiment / affect analysis | Mobile pervasive apps |
|                          | Mining high speed data streams and sensor data | Multilingual analysis | Web stystematic analysis | Mobile sensing apps |
|                          | Process mining              | Web visualisation       | Multimedia analysis | Mobile innovation |
|                          | ricacy preserving data mining | Web based auctions     | Text visualisation | Mobile social networking |
|                          | Network web mining          | Internet monetization   | Multimedia IR | Mobile visualisation |
|                          | In-memory dbms              | Social marketing        | Mobile IR | HCl |
|                          | Parallel dbsms              | Web privacy security    | Hadoop | Mobile social networks |
|                          | Cloud computing             |                         | Map reduce | mobile visualization |
|                          | Hadoop                      |                         |               | / HCI |
|                          | Map reduce                  |                         |               | Personalization and |
|                          |                             |                         |               | behavioral modeling |
|                          |                             |                         |               | Gamification |
|                          |                             |                         |               | Mobile advertising |
|                          |                             |                         |               | and marketing |

Table 7 shows research framework proposed by Chen, H., Chiang, R. H., et al. (2012).

Wu, L., et al. (2007, June) point out that business intelligence is evolving in order to be able to provide advanced analytics, which can include data mining, near-real time measures, scalability to support petabytes of data and thousands of concurrent users through an unprecedented massively parallel processing, clouds and virtualization.
4.2.1.4 Emerging trends

There are new trends that can be observed in terms of Business intelligence and they can be classified along 3 different lines.

1. Industry trends
2. Data trends and
3. Platform technology trends

See Lim, E. P., et al. (2013).

Lim, E. P., et al. (2013) further point out that amongst the most visible industry trends it can be noted that BI&A is a high priority technology item for CIO, visibly because, through it, businesses are gaining important insights that help them increase their business volume.

The most important data trends over the past years has been that of Big Data which is spreading throughout our society prompted by the ever growing amount of web, social media and sensor generated data. New science, discovery, and insights can be obtained from the highly detailed, contextualized, and rich contents of relevance to businesses and organizations Chen, H., et al. (2012), Lim, E. P., et al. (2013), Mohammed, J. (2015).

Lastly Lim, E. P., et al. (2013) explains that at the platform technology level, that of cloud computing and mobile computing are critical. A cloud computing platform is one that is built upon a large number of low cost computers to meet the needs of storing and computing big data in BIA applications.

Summarizing this section, these results show that the data in itself as it is gathered has no value. From the perspective of a business organization, this data is utilized by decision makers through the deployment of Business Intelligence and Analytics, a field in which data integration is central. Business intelligence input are an amalgam of many different systems which draw data from different sources and these different autonomous sources makes it complicated to integrate it. The fact that this data is used by decision makers implies that the data has to be accurate. Research trends are shifting towards trying to understand how to deal with the vast amounts of data which are being generated today which is constantly changing and highly heterogeneous as a consequence of technological advancements.

4.3 Drivers and consequences of increase information disparity

The different causes of data heterogeneity are a combination of technological advances in the IT sector, particularly in the areas of data storage and processing, which have, in turn, untied other developments such as the evolution of the Internet, which has evolved as a consequence of such advances and has given place to a number of important phenomenon in particular Social Media and the Internet of things. Different Business trends seem also seem to contribute to data heterogeneity. Khan, et al. (2014).

4.3.1 Technological Advances in the it sector

Technological advances in the form of increase in computing power and electronics, in particular processing and storage, has led to an increase in the amount and diversity of data

To this we would have to add other developments in the software and networking realm which makes it possible to gather data and propagate it automatically such as XML and SOA. Halevy, A., Rajaraman, et al. (2006), Ziegler, P., & Dittrich, K. R. (2004) and Telang, A., & Chakravarthy, S. (2007).

### 4.3.2 Evolution of the Internet

Along with the incredible advances in technology the web has changed and evolved. See S. Negash (2004), also Chen, H., et al. (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015). See table 8 to see a summary of the evolution of the web.

<table>
<thead>
<tr>
<th>The semantic web</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web type</strong></td>
</tr>
<tr>
<td>Web 1.0</td>
</tr>
<tr>
<td>Semantic Web 2.0</td>
</tr>
<tr>
<td>Semantic Web 3.0</td>
</tr>
</tbody>
</table>


### 4.3.2.1 Social Media

Bello-Orgaz, et al. (2016) points out that Social Media is feature rich and as such it generates very large heterogeneous data sets. See also Yaqoob, I., Chang, V., et al. (2016). Table 9 below shows some of the most popular social media platforms and some features and highlights.

<table>
<thead>
<tr>
<th>Social Media Types</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>Facebook</td>
</tr>
<tr>
<td>Twitter</td>
</tr>
</tbody>
</table>
Table 9 shows some characteristics of known social media. Source: Brandwatch (2016) and Programmable Web (n.d.).

Most of these platforms have one or more API with which to interact, some of them open source. Other such as YouTube has also an unlimited number of third party API’s that deal with different types of issues. ProgramableWeb (n.d.).

4.3.2.2  The Internet of Things

Stankovic, J. A. (2014) points out that advances in wireless networking technology and the greater standardization of communications protocols make it possible to collect data from sensors and wireless devices with different capabilities. Stankovic, J. A. (2014) adds that in this mosaic it is necessary to address the need for heterogeneous systems that have sensing, acting, communication, cognitive, processing and adaptability features and includes sensors, actuators, nano-electronics circuits, embedded systems, algorithms, and software embedded in things and objects. These different types of issues create data heterogeneity.

Rose, K., Eldridge, S., & Chapin, L. (2015) points out that in their report “Unlocking the Potential of the Internet of Things”, the McKinsey Global Institute 24 describes the broad range of potential applications in terms of “settings” where Internet of Things (IoT) is expected to create value for industry and users. A summary of these are summarized on table 10 below.
As pointed out by Khan, et al. (2014), the planet's population exceeds 7.2 million as of the writing of this report. And over 2 billion is connected to the internet. As a result of technological advances and the evolution trends that are taking place millions of people are generating a tremendous amount of data through the use of internet connected devices, in particular remote sensors, and social media, which produce a constant stream of heterogeneous data.

Stankovic, J. A. (2014) points out that if we continue the current trend and the numbers of smart devices to increase trillions of different things will be on the Internet generating a massive amount of rapidly changing heterogeneous data.

**4.3.2.3 New business trends**

Data is becoming available in greater and greater quantity and can now be use to streamline the organization’s front and back end operations to increase performance. There is research pointing out that the growth of enterprise-wide systems is blurring the distinction between front and back-end systems i.e. Customer Relationship Management Systems (CRM) and Enterprise Resource Planning (ERP). These systems, according to some, work in a borderless and seamless way, but they generate different types of data; structured on the one side and unstructured on the other. Furthermore, the evolution of such systems has led to ERP II in which CRM is integrated and it also provides for social media. B. Bond, et al. (2000), Rashid, M. A., et al. (2002), Cap Gemini (2015).

New marketing strategies are being developed that promote a multichannel integration approach to marketing. This is also possible because of the reach of technological advances.
Some already talk about Omni-channel integration. An Omni-channel integration approach is that in which all available communication channels to reach the customers are exploited. This type of marketing trend is important within business because of their potential connection with the semantic web. Hillerbrand, E. (2016) see also TechTarget (2014) and Simon Knox, Stan Maklan, Adrian Payne, Joe Peppard, Lynette Ryals (2003).

There are other business trends which contribute to this data disparity. Chen, H., et al. (2012) points out that business performance management (BPM) using scorecards and dashboards help analyze and visualize a variety of performance metrics but in addition to these well-established business reporting functions, statistical analysis and data mining techniques need to be adopted for association analysis, data segmentation and clustering, classification and regression analysis, anomaly detection, and predictive modeling in various business applications. See also Golfarelli, M., et al. (2004, November).

4.3.3 Consequences of increase information disparity


4.3.3.1 Big Data

The constant increase in computational power promoted by technological advances has generated an overwhelming flow of data. The amount of data generated by some known sources such as Facebook, Google Earth, Filckr or Youtube far exceeds what was possible to store 20 years ago. If we take Facebook as an example in 2012 it served serves 570 billion page views, stores 3 billion new photos, and manages 25 billion pieces of content monthly Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U. & Qu, W. (2012).

Originally Big Data can be characterized by the so called 3V’s Velocity, Variety, and Volume. However, just in recent years there are some that have added additional two V’s, to the already existing three, i.e. Veracity and Value, as defining characteristics of big data. Dong, X. L., & Srivastava, D. (2013, April) points out that this is due to the fact that the data is useless and does not add value unless it is accurate and Hashem et al. (2015) go as far as writing that value is the most important aspect of big data. See also Che, D., et al. (2013) and Gandomi, A., & Haider, M. (2015).

As suggested by Che, D., et al. (2013) heterogeneity in big data means that it is an obligation (rather than an option) to accept and deal with structured, semi-structured, and even entirely unstructured data simultaneously on the fly.

Srinivas, B., & Togiti, B. (2015) point out that for enterprise level data mining the data is normally large and usually distributed and parallel computing is required. However in the case of Big Data, the mining effort is so large that the capacity of such systems is insufficient, which requires a framework what relays on cluster computers with a high performance computing platform.
Processing technologies identifiable to big data include Hadoop, Map Reduce, NoSQL Databases, cloud dataflow, Apache Spark, Machine Learning. Table 11 below shows list of these Hashem, I. A. T.et al. (2015) and Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U. & Qu, W. (2012).

<table>
<thead>
<tr>
<th>Name</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop</td>
<td>Open source distributed open platform</td>
</tr>
<tr>
<td>Map Reduce</td>
<td>Simplify tasks (Divide and conquer)</td>
</tr>
<tr>
<td>Relational Databases</td>
<td>Used for rapid querying of structured data used by ETL</td>
</tr>
<tr>
<td>NoSQL Databases</td>
<td>Conveniently deal with heterogeneity in a scalable way.</td>
</tr>
<tr>
<td>XML</td>
<td>Deals with unstructured data and widely used by web services and in cloud computing</td>
</tr>
<tr>
<td>Cloud Computing</td>
<td>Deals with the problem of scalability and infrastructure.</td>
</tr>
<tr>
<td>Distributed computing</td>
<td>To deal with data which is distributed in distinct geographical locations</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Artificial intelligence</td>
</tr>
<tr>
<td>Streaming Processing</td>
<td>Approach for real time business intelligence</td>
</tr>
<tr>
<td>Clustering</td>
<td>Storage and database</td>
</tr>
</tbody>
</table>


Due to the difference between the services of the application layer, the characteristics of multi-source heterogeneous data also are different with different services, so our data storage and management must use corresponding methods for differing services, for example Big Table, the key value of NoSQL and so on. Liu, Y., Wang, Q., & Chen, H. Q. (2015).


Big data evaluation is driven by fast-growing cloud-based applications developed using virtualized technologies. However, further from facilitating the computation and processing of Big Data, cloud computing serves also as a model. Hashem, I. A. T., et al. (2015)

In summary, there is a trend towards more data disparity. Technological advances have allowed us to store more data, which is growing larger and larger, and it changes very rapidly and the integration of all this disparate data is getting increasingly difficult. Technological advances, particularly in data processing, networking and storage have impacted the World Wide Web, which has evolved towards what we know today as the
semantic web, where things such as Social Media, The internet of things and a sharp explosion of mobile technology has precipitated us into what we know today as the age of Big Data.

4.4 The Nature of Data - Structured vs Unstructured Data

BI&A requires experts to deal with heterogeneous data, in particular structured and semi-structured data simultaneously. This term, i.e. semi-structured data, is used for all data which does not fit in a relational structure. There is data which is highly unstructured but it is referred to semi-structured because most of the data has some inherent structure. S. Negash (2004).

Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., & Widom, J. (1995, December) refer to Semi Structured data to data which has no absolute schema fixed in advance in which the structure can be irregular or incomplete and point out two common examples in which it can arise:

1. When data is stored in sources that do not have a rigid structure such as the World Wide Web.

2. When data is combined from several heterogeneous data sources, which could be structured, and when new sources are frequently added.


Semi-structured data is more complicated to deal with, particularly because it is not easily searched using existing tools for conventional databases. S. Negash (2004).

However, suggested by a Meryl Lynch report cited by S. Negash (2004), 85 % of all business information exists in the form of semi-structured. See also Chen, H., et al (2012), Lim, E. P., et al. (2013), and Mohammed, J., (2015). See table 12 below for a summary of sources of semi-structured data.

<table>
<thead>
<tr>
<th>Sources of Semi Structured Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business processes</td>
</tr>
<tr>
<td>Chats</td>
</tr>
<tr>
<td>E-mails</td>
</tr>
<tr>
<td>Graphics</td>
</tr>
<tr>
<td>Image files</td>
</tr>
</tbody>
</table>

Table 12 shows some examples of semi-structured data. Source: S. Negash (2004).

Chen, H., Chiang, R. H., & Storey, V. C. (2012) point out that an important amount of the unstructured data received by organizations is in text format which, for example, can include mail, communication, web pages and social media content among other.
As pointed out by S. Negash (2004), structured and semi-structured data types can be further segmented by looking at the front and back end sources within the organization. See also B. Bond, et al. (2000), Rashid, M. A., et al. (2002), Cap Gemini (2015). The data source – data type dimension are illustrated in table 13 below.

<table>
<thead>
<tr>
<th>Source / Type</th>
<th>Back-end</th>
<th>Front-end</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>ERP</td>
<td>CRM</td>
</tr>
<tr>
<td>Semi-structured</td>
<td>Business Processes</td>
<td>News Items</td>
</tr>
</tbody>
</table>

Table 13 illustrates data type dimension in Business Intelligence. Source: S. Negash (2004).

Semi-structured data tends to be business oriented while structured data tends to be mostly technical. The traditional warehousing architecture has been used by Business Intelligence for its structured data. Table 14 below shows a meta-data example of structured and semi-structured data.

<table>
<thead>
<tr>
<th>Focus</th>
<th>Derivation</th>
<th>Administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business (mostly semi-</td>
<td>➢ What does it mean?</td>
<td>➢ What training is available?</td>
</tr>
<tr>
<td>structured)</td>
<td>➢ Is it relevant?</td>
<td>➢ How fresh is the data?</td>
</tr>
<tr>
<td></td>
<td>➢ What decisions can I take?</td>
<td>➢ Can I integrate it?</td>
</tr>
<tr>
<td>Business (mostly</td>
<td>➢ How was it calculated?</td>
<td>➢ What business rules were applied?</td>
</tr>
<tr>
<td>structured)</td>
<td>➢ Are the sources reliable?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➢ What business rules were applied?</td>
<td></td>
</tr>
<tr>
<td>Technical (Mostly</td>
<td>➢ Filters</td>
<td></td>
</tr>
<tr>
<td>structured)</td>
<td>➢ Aggregates</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➢ Calculations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➢ Expressions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➢ Capacity planning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➢ Space allocation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➢ Indexing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➢ Disk utilization</td>
<td></td>
</tr>
</tbody>
</table>

Table 14 shows metadata for structured vs unstructured data. Source: S. Negash (2004).

S. Negash (2004) point out, however, that not all the data in a BI system is structured. There is also the possibility that data is semi-structured or not structured at all. This is corroborated also by Chen, H., Chiang, R. H., et al. (2012), Lim, E. P., et al. (2013) and later Mohammed, J. (2015). See figure 2 below for an illustration of the architecture of semi-structured data.

Figure 2 illustrates typical architecture of semi-structured data. Source: S. Negash (2004).
It is important to note that web data sets are usually less structured and processing such data at scale poses a big challenge Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., ... & Qu, W. (2012).

Summarizing, these findings show that there is a wide spectrum of possibilities in regards of data. On the one side there is structured data which fits well relational database systems and models, and to the other side there is highly unstructured data which does not fit RDBMS. Semi structured data tends to be business oriented, and structured data more technical. Semi-structured and structured data are more complicated to deal with and is more abundant on the World Wide Web where data tends to be quite unstructured.

4.5 Challenges and approaches of Integration of heterogeneous data

4.5.1 Integration approaches and frameworks

As we saw earlier, there are different types of heterogeneity that can arise when attempting to integrate heterogeneous data sources these can be caused by hardware and operating systems, data management software, data models, schemas and semantics, middleware user interfaces, business rules and so on. Of these semantics and data models are important hot topics today. S. Negash (2004) and Hillerbrand, E. (2016).

As pointed out by Telang, A., & Chakravarthy, S. (2007) integration of heterogeneous data is not a new topic, it has actually been around for the last two decades and many techniques have evolved since them. Telang, A., & Chakravarthy, S. (2007) further suggest that the problem now is more complex because of the rise in data sources and data types and the difference in semantics. Ziegler, P., & Dittrich, K. R. (2004) add that there is a need to develop semantics for integrable data so that we can achieve a meaningful output because we are seemingly falling behind. Table 15 shows the most prominent approaches for data integration that have been developed so far.

<table>
<thead>
<tr>
<th>Approaches for integrating heterogeneous data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>Mediated query systems</td>
</tr>
<tr>
<td>Portals</td>
</tr>
<tr>
<td>Data warehouses</td>
</tr>
</tbody>
</table>

* Halevy, A., Franklin, M., & Maier, D. (2006, June) point out that data integration systems require semantic integration before any services can be provided before any services can be provided. Data Space systems are not per se a data integration approach; rather, they are more of a data co-existence approach.

Furthermore, of the above approaches the only one to address the problem of semantic heterogeneity is the ontological approach, which remains a hot topic today. Telang, A., & Chakravarthy, S. (2007) and Ziegler, P., & Dittrich, K. R. (2004). See also Hearst, M. A., Levy, e (1998) and Hillerbrand, E. (2016).

As pointed out by Saggion, H., Funk, A., Maynard, D., & Bontcheva, K. (2007) it is possible to use Ontology based information extraction techniques for Business intelligence. Which they proof with their MUSING project, which according to them, is develops a new generation of BI tools and modules based on semantic-based knowledge and natural language processing (NLP) technology to mitigate the efforts involved in gathering, merging, and analyzing information.

In all there have been many frameworks developed which attempt to address the problem of integration of heterogeneous data-sources some of them are shown in table 16 below. It is worth noting that most of the models and frameworks have used wrappers and mediators.

<table>
<thead>
<tr>
<th>Framework</th>
<th>Challenge Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Havasu</td>
<td>Imprecise query specification, query optimization and source-statistics collection.</td>
</tr>
<tr>
<td>Metaquerier</td>
<td>Exploration and integration of deep-web sources</td>
</tr>
<tr>
<td>Mainfold Project</td>
<td>Query efficiency</td>
</tr>
<tr>
<td>Ariadne</td>
<td>Includes structured and semi-structured data sources</td>
</tr>
<tr>
<td>Whirl</td>
<td>Use of information retrieval and Artificial intelligence In incorporating single domain queries and semi structured static sources.</td>
</tr>
<tr>
<td>Ontologic (RDF)</td>
<td>Extraction of implicit and hidden knowledge – Semantic differences</td>
</tr>
<tr>
<td>TSIMMIS</td>
<td>Data heterogeneity from structured and unstructured sources</td>
</tr>
<tr>
<td>InfoMaster</td>
<td>Provided integrated access to structured information sources</td>
</tr>
<tr>
<td>Tukwila</td>
<td>Interleaved query planning and execution, adaptive operators, and a data source catalog.</td>
</tr>
<tr>
<td>Whirl</td>
<td>Query language deploys information retrieval with SQL-like queries about the similarity of names.</td>
</tr>
<tr>
<td>GIS</td>
<td>For a geographical or spatial integration</td>
</tr>
<tr>
<td>InfoMosaic</td>
<td>Query specification</td>
</tr>
</tbody>
</table>


4.5.2 Service Oriented Architecture and XML

4.5.2.1 Earlier attempts to handle semi structured data

There have been earlier attempts such as TSMMIS project to handle unstructured and semi structured data. The differences with other attempts of integrating information sources is that the TSIMMIS project focused on providing integrated access to very diverse information which would be structured and unstructured, it required human participation and it was not fully automatic Hammer, J., et al.(1995) and Quass, D., et al. (1995, December).


4.5.2.2 XML

Martinez, J. M. P., Berlanga, R., Aramburu, M. J., & Pedersen, T. B. (2008) underline the importance of XML-based technology in the development of the web as the largest source of information. XML provides a means of information interchange between applications, as well as a semi structured data model for integrating information and aid in knowledge retrieval.

Ziegler, P., & Dittrich, K. R. (2004) point out that XML has become the industry standard for data exchange and it has made it possible through Web Services to provide interoperability between different software applications running on different platforms, furthermore, it has enabled the automation of such by making it possible to be read by machines which in itself has contributed to data interchange and data propagation. See also Birant, D. (2011).

Halevy, A., Rajaraman, et al. (2006) point out that XML fueled the desire for data integration, because it offered a common syntactic format for sharing data among data sources. Furthermore, D., Halevy, A. Y., & Weld, D. S. (2001) point out that it is less technical which makes it more suitable for business. XML is a language that deals with semi-structured data, however they also point out that XML did nothing to address the semantic integration issues. Halevy, A., Rajaraman, et al. (2006) also explain that the Tsimmis Project was the first to illustrate the benefits of semi-structured data in data integration. Users utilizing XML can represent the same thing in different ways and this creates heterogeneity. Cruz, I. F., & Rajendran, A. (2003). See also Firat, A., Madnick, S., & Grosof, B. (2002).

XML is widely used by BI&A because human readable and business oriented. As pointed out by Cao, L., Zhang, C., & Liu, J. (2006), perhaps one of the setbacks that is most visible in the field of BI&A is the lack of semantic integration mechanisms, which according to Saggion, H., et al. (2007) seems to be palliated with the deployment of Ontology based extraction.
Integrating Heterogeneous Data

Gabriel Nieva

2016-10-22


<table>
<thead>
<tr>
<th>Data Exchange Languages</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML</td>
<td>YES</td>
</tr>
<tr>
<td>Extensible Markup Language which provides encoding and document rules</td>
<td></td>
</tr>
<tr>
<td>Resource Description Framework (RDF)</td>
<td>NO</td>
</tr>
<tr>
<td>Ontology Language (OWL)</td>
<td>NO</td>
</tr>
<tr>
<td>ATOM</td>
<td>NO</td>
</tr>
<tr>
<td>SKOS</td>
<td>NO</td>
</tr>
<tr>
<td>JSON</td>
<td>NO</td>
</tr>
<tr>
<td>REBOL</td>
<td>NO</td>
</tr>
<tr>
<td>GELISH</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 17 shows the different data exchange languages. Source: Wikipedia (2016).

Of these languages above only XML and Gelish have what is called semantic verification. Semantic verification is necessary to check the absence of errors related to the meaning of artifacts which have been constructed with a modeling language IGI Global (n.d).
4.5.2.3 SOA


Parapatics, P. (2007, June) point out that Service Oriented Architecture (SOA) works as an implementation of context independent data which enforces data independence on a very high level, so services can be applications as well as data resources. Loose coupling is achieved by encapsulating a specific service behind a well-defined interface. See also Ziegler, P., & Dittrich, K. R. (2004), see also Birant, D. (2011).

Informatica (2004) suggests that the SOA architecture is beneficial for data integration purposes because data integration technology takes advantage of the layer of abstraction that enables its components and services to be wrapped and reused without extensive hand coding. However, SOA by itself cannot address the issue of heterogeneity. See also Demirkan, H., & Delen, D. (2013).


Parapatics, P. (2007, June) further points out that most of the information on the Semantic Web (Web 2.0) is human understandable only. With the evolution of SOA more and more data is available in the Internet which is understood by machines and can therefore be processed automatically. This automation is an important premise for real time analytics. See Seufert, A., & Schiefer, J. (2005) and Azvine, B., et al. (2006, June).

Devi, C. P., Venkatesan, V. P., Diwahar, S., & Shanmugasundaram, G. (2014) point out that the information integration systems built on SOA or web services have centered on the following applications illustrated in table 18 below.

<table>
<thead>
<tr>
<th>Focus of Information Integration Systems built on SOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Frameworks</td>
</tr>
<tr>
<td>Workflow Based models</td>
</tr>
<tr>
<td>Application Specific models and Frameworks</td>
</tr>
<tr>
<td>Information as a service</td>
</tr>
</tbody>
</table>

Table 18 shows the frameworks that SOA based information integration systems focus on. Source: Devi, C. P., et al. (2014).

4.5.3 Query Optimization

As suggested by Ziegler, P., & Dittrich, K. R. (2004), see also also by Telang, A., & Chakravarthy, S. (2007), Kermanshahani, S. (2009) and later Hashem, I. A. T., et al. (2015), one of the most important challenges in heterogeneous data integration is that of query processing. This is because of the heterogeneity between the data - legacy database systems, websites, web-services, etc.
There has been further research in terms of Query optimization, which remains an active research field in this matter. Telang, A., & Chakravarthy, S. (2007) point out that there has been a number of querying languages designed over the decade extending from 1997 to 2007. These querying languages extend the basics of relational algebra and allow access to structured data. Some of these are included in table 19 below.

<table>
<thead>
<tr>
<th>Query Languages available from 1997 to 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>SQL</td>
</tr>
<tr>
<td>OOQL</td>
</tr>
<tr>
<td>Whirl</td>
</tr>
<tr>
<td>SEMQL</td>
</tr>
<tr>
<td>STRUDEL</td>
</tr>
<tr>
<td>LOREL</td>
</tr>
<tr>
<td>UNQL</td>
</tr>
<tr>
<td>Vague</td>
</tr>
</tbody>
</table>


As suggested by Telang, A., & Chakravarthy, S. (2007), during the past decade there many query languages have been developed. However, according to Telang, A., & Chakravarthy, S. (2007), these languages have, had limited success, in incorporating imprecise user queries posed on a single-domain (or fixed set of multiple domains).

Some other newer query languages have been developed which address different heterogeneity problems of structured, semi-structured and unstructured data such as XPath and Xquery and also semantic differences such as SPARQL and RDF Query Language. Intel (2013), Madria, S., et al. (2008) and Makris, K., Bikakis, N., Giodasis, N., Tsinaraki, C., & Christodoulakis, S. (2009, September) see also Workman, M. (2016). Table 20 below shows just a few of them.

<table>
<thead>
<tr>
<th>Newer query languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Language</td>
</tr>
<tr>
<td>XQuery</td>
</tr>
<tr>
<td>XPath 3.0</td>
</tr>
<tr>
<td>SPARQL and RDF Query Language</td>
</tr>
<tr>
<td>HiveQL</td>
</tr>
<tr>
<td>Language Integrated Query (LINQ)</td>
</tr>
</tbody>
</table>

Chen, H., Chiang, R. H., et al. (2012) and later Mohammed, J. (2015) point out that ever since the advent of Web 2.0 the internet has undergone a dramatic explosion of content, and its biggest contributor is the web itself, which generates a vast amount of poly-structured and heterogeneous data. The current line of research points out to an extremely difficult problem to solve that remains at large which is that of semantic differences. See also Srivastava, K., et al. (2012).

With the success of 2.0 most IT companies tend to store and analyze an ever larger amount of data such as search logs, crawled web content and click streams from a variety of web services. However, these data sets are usually non-relational or less structured. And processing such semi-structured data at scale is a big challenge Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U. & Qu, W. (2012).

4.5.4 The Semantic Web

As suggested by Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., ... & Qu, W. (2012) and also by Chen, H., et al.(2012). The web is today one of the largest if not the largest source of data and it poses additional challenges because one the one side the data sets found on the web are very large and on the other side because these datasets tend to not to have much structure which poses additional challenges because they don’t fit well Relational Database Management Systems (RDBMS).

Shvaiko, P., & Euzenat, J. (2013) further point out that the solution for dealing with semantic heterogeneity is ontology matching. The semantic web provides semantic description or meta-data about the resources themselves. These metadata allow intelligent agents to work with it. To this effect RDF/RDFS has been introduced by the World Wide Web Consortium (WTC). RDF is a semantic web language which supports the needs of ontology language for the web resources. OWL in turn was designed on the RDF graph model and the semantic web found description logics for it. Suwanmanee, S., et al. (2005), Jean-Mary, Y. R., Shironoshita, E. P., & Kabuka, M. R. (2009) and Husain, M. F., et al. (2009, December). See also Bergamaschi, S., et al (2001), Hakimpour, F., & Geppert, A. (2001, October), Cruz, I. F., & Rajendran, A. (2003), and later Suwanmanee, S., et al. (2005) and Workman, M. (2016). Table 21 below illustrates a summary of this technology.

<table>
<thead>
<tr>
<th>Components of the Semantic Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Descriptor Framework (RDF)</td>
</tr>
<tr>
<td>RDFS Chema (RDFS)</td>
</tr>
<tr>
<td>Simple Knowledge Organization System (SKOS)</td>
</tr>
<tr>
<td>SPARQL- an RDF query language</td>
</tr>
<tr>
<td>Notation3 (N3), designed for human-readability</td>
</tr>
<tr>
<td>N-Triples</td>
</tr>
</tbody>
</table>
Turtle (Terse RDF Triple Language) | A Turtle document allows writing down an RDF graph in a compact textual form.
---|---
Web Ontology Language (OWL) | Provides vocabulary for property and class description and relationships between classes.
Rule Interchange Format (RIF) | An XML language for expressing web rules which computers can execute.
XML (Markup language) | Provides elemental syntax w no semantics.
XML Schema | Restrict the structure and content contained in XML documents.

**Table 21** shows some components of the semantic web. *Source*: Data FCQL (2016).

In the Semantic Web context information may is processed from combinations of multiple heterogeneous data sources with different representations of a common domain. This led to the proliferation of ontologies in specific knowledge domains expressed in OWL. Suwanmanee, S., et al. (2005).

Some also argue that OWL has its limitations. Hardcoding their descriptions limits the flexibility and usability of a detection mechanism. Alnusair, A., & Zhao, T. (2009, October).

This is also something corroborated by Workman, M. (2016), who suggest that there is a trend towards moving from programmed logic towards dynamically generated and interpreted logic and new definitions for semantic technologies.

The World Wide Web Consortium (W3C) is creating new markup, for example the Resource Description Framework (RDF). This markup is developed to enrich information and enable intelligent systems in an attempt to make a better use of metadata. Workman, M. (2016).

As pointed out by Workman, M. (2016) there is a family of web ontology markup languages in the marketplace which many businesses are adopting.

These approaches combined, however, have some limitations. Kondylakis, H., et al (2009, November) explain that despite extensive work with ontology based data integration, there is a persistent problem which goes frequently ignored, which is the fact that ontologies change over time, rendering the mappings obsolete.

On the other side, current semantic web frameworks, for example, Jena have scalability problems. Because these run on single machines and cannot handle huge amounts of data. Furthermore, storing a huge number of RDF triples and querying them efficiently remains a challenge. However, it is possible to store RDF data in in filesystems like Hadoop and Data mining algorithms are being rewritten in different forms to take the advantage of the MapReduce technology, technology which will be explored later on in the warehousing section. Husain, M. F., Doshi, P., Khan, L., & Thuraisingham, B. (2009, December).

In any case, progress in this area has been furthered by combining expertise from the areas of Database Systems, Artificial Intelligence, and Information Retrieval. This can be exemplified with the ASMOV project which provides Automatic Ontology Matching with Semantic Verification with the deployment of Artificial Intelligence. Jean-Mary, Y. R., et al.

4.5.5 **NoSQL Databases**

As suggested by Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., ... & Qu, W. (2012) and also by Chen, H., et al. (2012). The web is today one of the largest if not the largest source of data and it poses additional challenges because one the one side the data sets found on the web are very large and on the other side because these datasets tend to not to have much structure which poses additional challenges because they don’t fit well Relational Database Management Systems (RDBMS).

NoSQL data bases better address the problem of heterogeneity because they allow for the collection of structured as well as unstructured data. Furthermore, these databases are suitable and frequently utilized in processing large scale and high concurrent applications because their dictionary like data structure allows them to read and write the data very efficiently. In addition they support mass storage. Aside of this they are easy to expand and because they are open source they are cheap. Furthermore, they Gudivada, V. N., Baeza-Yates, R. A., et al. (2015), Han, J., Haihong, E., Le, G., & Du, J. (2011, October), Leavitt, N. (2010). See also Seera, N. K., & Jain, V. (2015).

Just to give an example, Facebook uses NoSQL databases and a query language called Fast Query Language (FQL), which process data in (value, pairs) in order to access its data. A simple such query can look like the one shown in figure 3 below. This type of database plays an important role in the integration of vast, heterogeneous and very rapidly changing amounts of data.

```
SELECT uid, name FROM user WHERE uid = me()
```

```json
{
    "Id": "10153002785216936",
    "name": "Gabriel Nieva"
}
```

*Figure 3* illustrates query using FQL. Source: Facebook Graph API Explorer (2016) Simple Query Result Screenshot.

One very popular query language analogous to these databases is the so called SPARQL which is the W3C recommended query language for RDF. A SPARQL query defines a graph pattern P that is matched against an RDF graph G. One peculiarity of SPARQL is that it is Schema-Less so it works very fast. Huang, J., Abadi, D. J., & Ren, K. (2011), Schätzle, A., et al. (2011, June), Liu, C., et al. (2012, May).
The basis of all graph patterns is the so-called Triple Patterns. A Triple Pattern is an RDF triple where subject, predicate and object can be variables. This is one of the key technologies of the semantic web. Schätzle, A., et al. (2011, June), Liu, C., et al. (2012, May), Huang, J., Abadi, D. J., & Ren, K. (2011). Some of the distinguishing characteristics of NoSQL databases are that they are deployable in cluster environments and in parallel processing and they are open source, highly scalable and use (key, value) pairs.

There are some limitations identifiable to NoSQL databases these are shown in table 22 below.

<table>
<thead>
<tr>
<th>Limitations of NoSQL Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overhead and complexity</td>
</tr>
<tr>
<td>Reliability</td>
</tr>
<tr>
<td>Consistency</td>
</tr>
<tr>
<td>Unfamiliarity with the technology</td>
</tr>
<tr>
<td>Limited Eco structure</td>
</tr>
<tr>
<td>Limitation of key value pairs</td>
</tr>
</tbody>
</table>

Table 22 illustrates some limitations of NoSQL databases. Source: Leavitt, N. (2010).

Perhaps one of the most important is that proposed by Uppsala University professor Tore Risch (n.d) cited by Leavitt, N (2010) who point out that they do not natively support the ACID test which provides for consistency and accuracy.

4.5.6 Data Warehousing

As seen before, Data Warehousing is an essential component of BI&A where data is integrated and a first step towards the BI&A effort. Data sources in a data warehouse usually come from multiple sources and conform to multiple schemas while the data warehouse has a single schema. The heterogeneous source data has to be transformed therefore into the data warehouse schema before loading into the data warehouse. This is normally done through the Extract Transform Load (ETL) process. J. Windom (1995), Fan H. (2005), Kakish, K., & Kraft, T. A. (2012). Figure 21 below illustrates the architecture of a data warehouse.

4.5.6.1 The ETL Process

Within the warehousing effort, the data extraction process is one of the most complicated and difficult parts in building a data warehouse and it demands a lot of sources, including time. It is the process of extracting, transforming, cleansing and loading the data onto the data warehouse in order to be used by decision takers in the organization. Jennifer Widom (1995), Intel (2013). See also Kakish, K., & Kraft, T. A. (2012). Figure 4 shows the different stages of the traditional ETL process.
This ETL process has been used for the past 20 years and there are a large percentage of companies which are still using it. Intel (2013). See also Kakish, K., & Kraft, T. A. (2012).

**A. Extraction Process**

The first part of this process entails bringing out the data from the source systems. These systems can be spread out geographically so the system will most likely be distributed over a geographic area. Anand, N., & Kumar, and M. (2013) Intel (2013).

The systems will contain heterogeneous information, this disparity can be due to the data interchange formats in which the information is carried and also in the type of data which can be structured, semi-structured and unstructured. Relational Database Management Systems contain structured data, XML can handle semi-structured data, flat files contain records that have no structured relationships it can also contain formats such comma separated values which can be read very fast. There can be other formats such as IMS, ISAM, and EBCDIC. Anand, N., & Kumar, M. (2013) and Intel (2013). Table 23 shows a summary of these formats.

<table>
<thead>
<tr>
<th>Common formats to carry data</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDMS</td>
</tr>
<tr>
<td>XML</td>
</tr>
<tr>
<td>Flat Files</td>
</tr>
<tr>
<td>IMS</td>
</tr>
<tr>
<td>ISAM</td>
</tr>
<tr>
<td>Virtual Storage Access</td>
</tr>
<tr>
<td>EBCDIC</td>
</tr>
<tr>
<td>CSV</td>
</tr>
</tbody>
</table>

Table 23 illustrates different data interchange formats. Anand, N., & Kumar, M. (2013) and Intel (2013).

**B. The transformation process**

The second step in the ETL process is the transformation process. The transformation process is responsible to make data complete, correct, consistent and unambiguous. To this end it will undergo many manipulations such as moving, splitting, translating, merging, sorting and so on. This is the place where data cleaning takes place. Simitsis, A. L. K. I. S. (2004), Kakish, K., & Kraft, T. A. (2012), and Intel (2013).
In this stage the heterogeneous data has to be transformed into the data warehouse schema before being loaded into the data warehouse. Simitsis, A. L. K. I. S. (2004), Hao Fan (2005) and Oracle (2005), Kakish, K., & Kraft, T. A. (2012).

**B1. Data cleaning**

In the process of data extraction from heterogeneous sources, there can be erroneous and inconsistent information which has to be reconciled before loading it to the data warehouse. The data cleaning process is a process which deals with detecting and removing errors and inconsistencies in order to improve the data quality. The need for data cleaning increases when trying to integrate multiple data sources in data warehouses, federated database systems and or global web based information systems. Rahm, E., & Do, H. H. (2000) and Fan, H. (2005).

The process of data cleaning can be subdivided into single-source and multi-source. The single source cleaning sub process takes care of data arriving from a single data source and the multi-source cleaning process involves several data sources and might include merging data from multiple sources. Rahm, E., & Do, H. H. (2000) and Kakish, K., & Kraft, T. A. (2012).

See figure 5 below for an illustration of the cleaning process.

![Figure 5](image)

**Figure 5** illustrates the data cleansing process. Source: Super Develop (2014).

This process is iterative and data cleaning is typically performed in a separate data staging area before loading the transformed data into the warehouse Rahm, E., & Do, H. H. (2000).

**B2. Data Staging**

Data is stored in this temporary area and then deleted once the data is loaded in the data warehouse, data mart or other repositories. The staging area keeps whole copies of the data sources and brings them under the control of the data warehouse administrator. The main function of a data staging area is to provide increase efficiency in the ETL process; it is essentially an auxiliary area of volatile data, which is employed for the purpose of data transformation, reconciliation and cleaning. Rahm, E., & Do, H. H. (2000), Simitsis, A. L. K. I. S. (2004), Oracle (2005). The data in the staging area may be heterogeneous and contain “dirty” (e.g. duplicate or inconsistent) data. Hao Fan (2005).
The thought behind staging data is that it should be possible to restart some of the phases independently of each other achieving in this way certain de-coupling. For example, if the transformation process fails then it should not be necessary to restart from the extract stage, only from the transform stage. In order to make data processing faster the data is aligned at this point. What this really implies is that data is dumped to an intermediate location, i.e. staging area, where it can then be read by the next processing stage. See Javlin Data Solutions (2015). Table 24 below shows some of the functions of a Data staging area.

<table>
<thead>
<tr>
<th>Functions of a data staging area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Comparison</td>
</tr>
<tr>
<td>Change detection</td>
</tr>
<tr>
<td>Consolidation</td>
</tr>
<tr>
<td>Alignment</td>
</tr>
<tr>
<td>Transformation</td>
</tr>
<tr>
<td>Data cleaning</td>
</tr>
<tr>
<td>Aggregation</td>
</tr>
<tr>
<td>Data archiving and troubleshooting</td>
</tr>
</tbody>
</table>


The Staging and ETL phases are considered to be most crucial stage of data warehousing where maximum responsibility of data quality efforts resides. These areas are also used by the ETL process to store intermediate results and the data in this this staging area may be heterogeneous and contain dirty data. Fan, H. (2005). See also Anand, N., & Kumar, M. (2013).

C. The Loading Process

The objective of the ETL process is to convert the data stream from different sources into a single format which is then loaded to the data warehouse so it can be used by its clients.

This can be done in batch processes or row by row. During the load time the system is unavailable to customers. Third generation ETL tools are using techniques to achieve real time data without causing downtime Kakish, K., & Kraft, T. A. (2012). See also Oracle (2005), Oracle (2014) and Intel (2013).

Batches for data warehouse loads used to be scheduled weekly; today’s businesses have an ever increasing demand for real-time load processing and this process can be performed with more frequency, for example instead of weekly it can be performed daily and even twice a day, achieving in this way near-real time Kakish, K., & Kraft, T. A. (2012), see also Oracle (2005), Oracle (2014) and Intel (2013).

4.5.6.2 Real Time ETL

Data loses value as it ages and in time it can become obsolete. In today’s business world it is imperative to have real time data to be Aavant-guard and in addition to this the data has to be accurate; data of poor quality is harmful Oracle (2012) and Azvine, B., et al. (2006, June). Azvine, B., Cui, Z., Nauck, D. D., & Majeed, B. (2006, June) explains that two main reasons make real time business intelligence necessary:
1. Environments in which businesses operate are constantly changing.
2. Advances in technology make the real-time possibility more reachable.

Seufert, A., & Schiefer, J. (2005) further points out that an organization’s survival depends on the construction and integration of knowledge which fosters the adaptation to a rapidly changing environment. This knowledge comes from a data warehouse which is a repository into which the data relevant to the organization is stored. See also March, S. T., & Hevner, A. R. (2007).

Azvine, B., Cui, Z., Nauck, D. D., & Majeed, B. (2006, June) suggest that with the vast amounts of heterogeneous, data that we see today achieving real time is very difficult. The data feeds to the analytics layer require data from different operational systems and very often fused data from these systems.

Seufert, A., & Schiefer, J. (2005) further point out that the reduction of the data latency can be achieved in two ways:

1. Not performing the ETL in the data warehouse (ETL entails periodic updates)
2. Deploying a so called event driven update

This line of thought is corroborated also by Azvine, B., Cui, Z., Nauck, D. D., & Majeed, B. (2006, June) who concretizes this aspect by stating that real time can mean:

1. The requirement to obtain zero latency within a process
2. That a process has access to information whenever it is required

ETL tools have gone from being mostly batch oriented to accommodating real time. In loading the data in real time there cannot be a system downtime. This implies that the load process can coincide with the periods of highest data usage. Kakish, K., & Kraft, T. A. (2012).

As of 2005 Oracle offered two kinds of transformations possibilities in data warehousing which are illustrated in table 25 below.

<table>
<thead>
<tr>
<th>Transformations types</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi Stage transformation</td>
<td>Staging tables exists</td>
</tr>
<tr>
<td>Pipelined transformation</td>
<td>No staging tables exist</td>
</tr>
</tbody>
</table>

Table 25 illustrates two possible approaches to transformation. Source: Oracle (2005).

The multistage transformation made use of staging tables with which it is easier to monitor and restart the transformation process because it stores the incremental result of each step. This requires what is called Data Staging Area and it is used for the purpose of data transformation, reconciliation and cleaning. Zuters, J. (2011, October). However, this method is complex.

The opposite happens when using the pipelined transformation method. When deployed the pipelined transformation method will require that the source data and the target data

As of 2014 Oracle provides various solutions for different near real-time data integration use cases. Oracle (March 2014). See table 26 below for a summary.

<table>
<thead>
<tr>
<th>Description</th>
<th>Batch</th>
<th>Mini-Batch</th>
<th>Micro-Batch</th>
<th>Real-Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data is loaded in full or incrementally using an off-peak window</td>
<td>Data is loaded incrementally using intra-day loads.</td>
<td>Source changes are captured and accumulated to be loaded in intervals.</td>
<td>Source changes are captured and immediately applied to the DW</td>
<td></td>
</tr>
<tr>
<td>Latency</td>
<td>Daily or higher</td>
<td>Hourly or higher</td>
<td>12 min &amp; higher</td>
<td>Sub-second</td>
</tr>
<tr>
<td>Capture</td>
<td>Filter Query</td>
<td>Filter Quality</td>
<td>CDC</td>
<td>CDC</td>
</tr>
<tr>
<td>Initialization</td>
<td>Pull</td>
<td>Pull</td>
<td>Push, then Pull</td>
<td>Push</td>
</tr>
<tr>
<td>Target Load</td>
<td>High Impact</td>
<td>Low Impact, load. Frequency is tuneable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Load</td>
<td>High Impact</td>
<td>Queries at peak times necessary</td>
<td>Some to none depending on CDC technique</td>
<td></td>
</tr>
</tbody>
</table>

Table 26 illustrates different architectures possible to load data warehouses. Source: Oracle (March 2014).

There are other ways however to lower the data latency, Near Real Time ETL is achieved by updating the tables with more frequency. In the Direct Trickle Feed, True Real Time is achieved by continuously moving the changed data from the source systems by inserting or updating them in the fact tables. In the Trickle and Flip the data is inserted or updated into staging tables which are in the same format as target tables. With Trickle and Flip technique the data warehouse can access updated data instantly by getting a copy from the staging tables into the fact tables. The update window can be from hours to minutes. The problem with the Trickle Feed approach is that constant updates on the tables by the querying OLAP tools end up degrading the query performance of the data warehouse. Kakish, K., & Kraft, T. A. (2012).

With the external real time data cache (RTDC) the real time data is stored outside the data warehouse. This resolves the query contention and scalability problem by directing queries to the RTDC and as it is located outside on a separate cache it does not require additional loads, but it would require additional infrastructure. Kakish, K., & Kraft, T. A. (2012) and Langseth, J. (n.d.).


In general, third generation ETL tools use techniques without causing downtime. Table 27 below shows a summary of the different ways and major highlights of each technique:

<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Near Real Down Time ETL</td>
<td>Updates performed frequently</td>
<td>Weekly</td>
</tr>
<tr>
<td>Direct Trickle Feed</td>
<td>No Staging, data is inserted directly in the fact tables</td>
<td>Instantly</td>
</tr>
<tr>
<td>Trickle and Flip</td>
<td>Updated into staging tables which are in the same format as target tables</td>
<td>Hours / minutes</td>
</tr>
<tr>
<td>External Real Time Data Cache</td>
<td>Data is stored outside data warehouse in an external real time data cache (RTDC).</td>
<td>Seconds</td>
</tr>
<tr>
<td>Integrating OLAP and OLTP</td>
<td>The physical data residing in OLAP is in its de-normalized form. And the one on OLTP is in the 3rd normal form.</td>
<td>N/A</td>
</tr>
<tr>
<td>BI&amp;A technique</td>
<td>Transfer from operational data stores to the op data warehouse occurs in real time in data event stream formats.</td>
<td>Instantly</td>
</tr>
</tbody>
</table>

Table 27 shows the different types of Real time ETL techniques. Source: Kakish, K., & Kraft, T. A. (2012).

Kakish, K., & Kraft, T. A. (2012) point out that one approach which has been proposed for real time business intelligence consists on that the operational data stores deliver the data to the data warehouse in streams. This reduces the reliance on batched or offline updating, and the staging part, so more structured and unstructured data can be updated simultaneously very rapidly. See also Intel (2013). Figure 6 below shows the architecture of such proposal.

Figure 6 illustrates the event driven technique proposed for real time BI&A. Source: Kakish, K., & Kraft, T. A. (2012).

Kakish, K., & Kraft, T. A. (2012) further suggest that Event driven real-time such as the one proposed by the BI&A approach have two major challenges:
1) Queries running continuously
2) Algorithms cannot be recurrent

4.5.6.3 ETL with Hadoop

As suggested by Intel (2013) many things have changed in the database realm over the past two decades. Databases have become far more powerful and RDBMS today support very complex transformations in SQL, including in-database mining, and quality validation, cleaning profiling, statistical algorithms and more. RDBMS engines are also optimized for disk I/O and this increases throughput, in all as long as RDBMS hardware scales up so does the system performance.

As the data increases in size and disparity processing it can be a daunting task. However, as pointed out by Che, D., et al. (2013) heterogeneity in big data also means that it is an obligation (rather than an option) to accept and deal with structured, semi-structured, and even entirely unstructured data simultaneously on the fly.

Apache Hadoop is an open source distributed software platform for storing and processing data. It is written in java and runs on a cluster of industry-standard servers configured with direct-attached storage. Husain, M. F., Doshi, et al. (2009, December) and Intel (2013)

Using Hadoop it is possible to store petabytes of data reliably on tens of thousands of servers while scaling performance cost-effective by adding inexpensive nodes to the cluster. Intel (2013).

A framework called Map Reduce is central to the scalability of Apache Hadoop. MapReduce contributes to solve data-parallel problems for which the data set can be sub-divided into mall parts and processed independently, i.e. it follows a divide and conquer strategy which makes it easy for ordinary developers to deploy parallel programming constructs and not just to high performance computing experts Intel (2013) and Husain, M. F., Doshi, et al. (2009, December).

MapReduce splits inputs in multiple chunks, each of which is assigned a map which can process the data in parallel. Each map task reads a set of (key, value) pairs and generates a transformed set of (key, value) pair result which are then shuffled and sorted. The framework sends the intermediate (key, value) pairs to the reduce tasks which in turn group them into final results. MapReduce has monitoring mechanisms which like Job Tracker and Task Tracker to schedule tasks, monitor them and restart them as needed if they fail Intel (2013).

With Apache Squoop it is possible to automate the process of transferring data between Hadoop and a relational database in a concurrent way i.e. parallel and simultaneously check for errors. Intel (2013).

Hadoop is designed for scalability, to aid in this task it includes the Hadoop Distributed File System (HDFS) which includes fault tolerance characteristics. This system stores large files into blocks of 64 or 128 MB and replicates the blocks on three or more servers. In addition
it provides APIs to for MapReduce applications to read and write data concurrently. So the capacity and data nodes can be scaled by adding data nodes and a single Name Node mechanism manages data placement and monitor server performance and availability. See figure 7 below to see the Hadoop architecture Intel (2013) and Husain, M. F., et al. (2009, December). M. F., et al. (2009, December) further suggests that it is possible to store huge amount of semantic web data in Hadoop clusters which answer queries fast enough even when built by cheap commodity class hardware.

Figure 7 illustrates the Hadoop Architecture. Source: Intel (2013).

Other components in the Apache Hadoop system are very useful for the ETL process. Table 28 below includes a comprehensive list of Apache Hadoop Components.

<table>
<thead>
<tr>
<th>Apache Hadoop Components</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache MapReduce</td>
<td>To sub-divide data sets into small parts</td>
</tr>
<tr>
<td>Hadoop Distributed File System (HDFS)</td>
<td>Designed for scalability and fault tolerance by clustering.</td>
</tr>
<tr>
<td>Apache Fume</td>
<td>Distributed system for collecting, aggregating and moving large amounts of structured and unstructured data from multiple sources</td>
</tr>
<tr>
<td>Apache Squoop</td>
<td>To import data from MySQL or Oracle into the Hadoop Distributed File System (HDFS).</td>
</tr>
<tr>
<td>Apache Hive and HiveQL</td>
<td>SQL dialect which supports a subset of the SQL syntax</td>
</tr>
<tr>
<td>Apache Pig</td>
<td>Programming language that can be extended by the developer with his own functions written in Java Python and other.</td>
</tr>
<tr>
<td>ODBC / JDBC Connectors</td>
<td>To connect with common databases</td>
</tr>
</tbody>
</table>

Table 28 shows a few components of Hadoop and MapReduce  Source: Intel (2013).
Hadoop also contributes in terms of query processing by providing HiveQL. Although a bit slow, it facilitates a dialect of SQL which provides a subset of SQL syntax. Intel (2013). The fact that it is slow is viewed as a handicap by some who suggest that Hadoop lacks as of yet a viable query processing strategy for such a vast and heterogeneous amount of data. See also Hashem I.A.T, et al. (2015).

MapReduce still falls short in such things as time-varying graphs and dynamic networks, real-time processing requirements and scalable stream data processing. However, it is a frequently used computing paradigm when dealing with big data projects. Gudivada, V. N., et al. (2015).

Although it has been around for quite a while the ETL process is not dead and great amount of companies continue to use it. Hadoop contribution to the ETL consists of two major advances, first of all, its capacity to take massive amounts of data without specifying a schema and secondly of the possibility to offload the transformation of raw data by scalable parallel processing Intel (2013). Figure 8 bellow illustrates the ETL process with Hadoop.

Besides these there are a number of other features which make Hadoop attractive for the web. Amongst other Hadoop SPARQL provides a Web interface for users to submit queries through Web browser, Liu, C., Qu, J., et al. (2012, May). SPARQL is the W3C recommended query language for RDF, Schätzle, A., et al. (2011, June) and Huang, J., Abadi, et al. (2011). When used in conjunction with NoSQL databases the process of staging is skipped and data is loaded as is. See also Oracle (2005), Kakish, K., & Kraft, T. A. (2012), Intel (2013) and later Oracle (2014), Parapatics, P. (2007, June).

### 4.5.6.4 Data warehousing trends and research directions

Many developments in the data warehousing have been made but this field continues to evolve. Some of the most important trends drift towards processing large amounts of data in real time, other focus on the semantic web and yet other on data quality. Table 29 below sows some of the trends and research directions in the warehousing field Parapatics, P. (2007, June), Martinez, J. M. P., et al. (2008), Singh, R., & Singh, K. (2010), Kakish, K., &

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
<th>Field of Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge warehouses</td>
<td>The user is an expert in decision making</td>
<td>Semantic Web</td>
</tr>
<tr>
<td>Data warehouse federation</td>
<td>A set of data warehouses distributed in different physical locations</td>
<td>Big Data</td>
</tr>
<tr>
<td>Hadoop</td>
<td>Free computational tool of MAD kind. Good to interact with large storage systems and big data</td>
<td>Big Data</td>
</tr>
<tr>
<td>Map Reduce</td>
<td>Program that can generate and process large data sets</td>
<td>Big Data</td>
</tr>
<tr>
<td>Cheetah</td>
<td>Another Data base system designed for advertising online which works on top of MapReduce</td>
<td>Big Data</td>
</tr>
<tr>
<td>Near Real Time ETL</td>
<td>Snapshots sources, logged sources, database log scraping, log sniffing, time stamped sources and lockable sources</td>
<td>BI</td>
</tr>
<tr>
<td>DW quality</td>
<td>Divided in data model quality, presentation quality and data quality</td>
<td>Data Integration</td>
</tr>
<tr>
<td>Quality inside the data Warehouses Architecture (QuiDA).</td>
<td>To transfer data quality indicators from input databases to the spatial data warehouse, integrating quality dimensions within the warehouse. Uses syntactic ratio correction.</td>
<td>Data warehousing</td>
</tr>
</tbody>
</table>


In summary, the data integration efforts have revolved around developing technologies which address different problems by incorporating technologies which make it possible to deal with structured and unstructured data such as XML and SOA. Different frameworks and approaches have been developed to deal with different aspects. However, just one i.e. the ontological approach deals with the highly unstructured data from the web. The ontological or semantic approach with newer frameworks and query languages such as OWL, RDF and SPARQL and the development NOSQL databases have contributed to make it possible to process the increasing amount of heterogeneous and ever changing data from
the internet. Other technological improvements in the data warehousing field, which has transitioned from batch processing to highly concurrent and parallel processing, have contributed, and thanks for example to developments such as Hadoop, and its suit components which deploy methods that make it possible to process this incredible amount of poly-structured data in real-time. There seems to be a balance between speed and accuracy. In addition the lack of structure and unpredictability of the sources in the Internet pose a challenge.

4.6 Cloud computing and virtualization

Cloud computing is a model for allowing ubiquitous convenient and on demand network access to a number of configured computer systems resources such as for example networks, server, storage, application and services. Hashem, I. A. T., et al. (2015).

Cloud computing is attractive to business owners as it eliminates the requirement for users to plan ahead to provide for infrastructure and increase resources when there is an increase in the need. UJ, U., Antony, P. J., & Sachin, D. N. (2016).

In addition virtualization makes it adequate for handling huge amounts of data i.e Big Data, because it makes the may systems required for dealing with such data look as one. Cloud based analytics platforms will become common for these type of projects. Seera, N. K., & Jain, V. (2015) and Talia, D. (2013).

When it comes to the interaction of disparate systems the virtualization feature of cloud computing is very interesting. Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., & Qu, W. (2012) point out that it abstracts disparate systems. They go further to suggest that cloud computing comprises a set of technologies and trends that are used in combination in order to make applications modular, dynamic and consumable.

Hashem, I. A. T., et al. (2015), suggest that in a cloud environment users can store a wide spectrum of data, from structured to non-structured. Those formats that are highly structured fit well the relational model but semi-structured data is more difficult to deal with and require multiple systems.

Gong, C., Liu, J., Zhang, Q., Chen, H., & Gong, Z. (2010, September) point out that although it is based on previous research in HPC, virtualization, utility computing and grid computing, cloud computing has its own characteristics see Table 30 below for a comparative summary of Cloud Computing.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Cloud computing</th>
<th>Grid computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service oriented</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Loose Coupling</td>
<td>Yes</td>
<td>Half</td>
</tr>
<tr>
<td>Strong fault tolerant</td>
<td>Yes</td>
<td>Half</td>
</tr>
<tr>
<td>Business model</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ease of use</td>
<td>Yes</td>
<td>Half</td>
</tr>
</tbody>
</table>
Enterprises generate huge volumes of data and the types of data change enormously. Because of this there is the need for a shorter processing speed and data virtualization can maximize the return of data Van der Lans, R. (2012).

As pointed out by Wu, L., et al. (2007, June) in a cloud computing environment, data-mining is the process of extracting structured information from unstructured and unstructured or semi-structured web data sources. See also UJ, U., Antony, et al. (2016).

Some important cloud mining techniques such as the so called Feature Extraction produces new attributes as linear combination of existing attributes and it is applicable for text data, latent semantic analyst, data decomposition, projection and pattern recognition. Another such technique is clustering which is useful for exploring data and finding natural groupings. UJ, U., Antony, et al. (2016).

Sharma, S., Tim, U. S., Gadia, S., & Wong, J. (2015) points out that in a cloud architecture storage, as well as application and infrastructure are abstracted, i.e. they do not reside in the client entity. This does away with the need of infrastructure and it allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable simple and secure services for their users. See also Hashem, I. A. T., et al. (2015), Assunção, M. D., et al. (2015) and later UJ, U., Antony, P. J., & Sachin, D. N. (2016) and UJ, U., Antony, et al. (2016). See also Seera, N. K., & Jain, V. (2015).

Kaisler, S., Armour, F., et al. (2013, January) point out that an open research question in cloud computing is whether for any given algorithm, there is a fundamental limit to its scalability. They suggest that every algorithm has an inflection point i.e. at which performance ceases to increase linearly and start decreasing. While this is known for specific algorithms in specific machines they argue that general computational solutions are not yet known, in particular, for unstructured data.

There is even the possibility that some algorithms cannot deliver a solution in a reasonable amount of time and yet other are what we call incomplete algorithms, which entail that it does not provide a guarantee that it will eventually find a satisfying assignment or declare that the given formula is not solvable. Kautz, H. A., Sabharwal, A., & Selman, B. (2009).

Although very attractive computing paradigm, especially for Big Data projects as pointed out by Hashem, I. A. T., et al. (2015), some of the areas in the field of cloud computing remain a challenge. See figure 31 below for a list of research challenges and open research issues.

<table>
<thead>
<tr>
<th>TCP / IP based</th>
<th>Yes</th>
<th>Half</th>
</tr>
</thead>
<tbody>
<tr>
<td>High security</td>
<td>Half</td>
<td>Half</td>
</tr>
<tr>
<td>Virtualization</td>
<td>Yes</td>
<td>Half</td>
</tr>
</tbody>
</table>

Table 30 Illustrates the characteristics of cloud computing. Source: Gong, C., et al (2010, September)
Research challenges and open research issues in Cloud Computing

<table>
<thead>
<tr>
<th>Research challenges</th>
<th>Open research issues</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Availability</strong></td>
<td>Refers to the resources of the system accessible on demand by an authorized individual</td>
</tr>
<tr>
<td><strong>Data Integrity</strong></td>
<td>Means that data can be modified only by authorized parties</td>
</tr>
<tr>
<td><strong>Transformation</strong></td>
<td>Turning data to a form suitable to analysis. Which is an obstacle in the adoption of big data</td>
</tr>
<tr>
<td><strong>Data quality</strong></td>
<td>The multi-source nature of data has become a source of problems for cloud computing</td>
</tr>
<tr>
<td><strong>Heterogeneity</strong></td>
<td>Variety is the result of virtually unlimited different sources of data. The challenge is how to handle multiple data sources and types.</td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
<td>Continue to hamper users who outsource their private data into the cloud storage.</td>
</tr>
<tr>
<td><strong>Legal</strong></td>
<td>Laws and regulations must be established to preserve the personal and sensitive information of users.</td>
</tr>
<tr>
<td><strong>Governance</strong></td>
<td>Data governance embodies the exercise of control and authority over data-related rules of law, transparency, and accountabilities of individuals</td>
</tr>
<tr>
<td><strong>Data Staging</strong></td>
<td>Consequence of gathered from different sources do not have a structured format.</td>
</tr>
<tr>
<td><strong>Distributed Storage Systems</strong></td>
<td>Include the capability of current cloud technologies to provide necessary capacity and high performance to address massive amounts of data</td>
</tr>
<tr>
<td><strong>Data analysis</strong></td>
<td>Information from large amounts of data requires scalable analysis algorithms to produce timely results.</td>
</tr>
<tr>
<td><strong>Data Security</strong></td>
<td>Security threats are magnified by the volume, velocity, and variety of big data.</td>
</tr>
</tbody>
</table>


In short, these results show that cloud computing is based on a Service Oriented Architecture. SOA provides for loose coupling, and deals better with the unstructured data that comes from the World Wide Web. Cloud computing is strongly fault tolerant, easy to use and it follows a business model which is why it is preferred by business users. Because it is TCP-IP based it can be distributed geographically and its virtualization feature abstracts the complexities of the mix of different systems which are needed by an organization which in turn can consume “on demand”.

Cloud computing in combination with the technologies such as Hadoop and its suit components and other development such as RDF and its analogous databases and queries i.e. RDF and NoSQL and SPARQL contributes positively to the integration of massive amounts of data. Furthermore, it stems from this evidence that this is the a better computing paradigm which addresses the requirements which arise when dealing with vast, heterogeneous and rapidly changing data such as the what we are seeing today in the age of Big Data.
5 Analysis

Once conclusive results have been extracted from a sufficiently large sample size of literature, following the methodology proposed in the theory section, it is now possible to proceed to the analysis section, in which further contrast of the different approaches and theories is performed with the objective of answering the research questions made earlier.

In order to provide alignment the analysis is performed in the same sequence in which the results were gathered. Furthermore, during the analysis section the research questions posed in the detail problem statement will be answered in the order in which they were made.

In Section 5.1 through 5.6 the value creation process and the reasons for data acquisition and integration are reviewed, and the drivers of information disparity and the nature of data structured vs unstructured are analyzed. Furthermore, the most important approaches to integration of heterogeneous data are shown and their limitations evaluated. Finally in the last section the most important aspects on cloud computing and virtualization are explained. All in hopes to answer the research questions proposed.

This analysis based from other existing analyses made previously by other researchers in their respective works so it is a secondary analysis performed in hopes to aggregate information which is in line with the chosen method and the methodology deployed in this study.

5.1 Data heterogeneity

Many efforts have been made to tackle the two categories of data heterogeneity explained earlier Structural Heterogeneity and Semantic Heterogeneity. In regards to the former it has been extended with cloud computing which expands the possibilities reached earlier through middleware and networking protocols and standards such as ODBC, JDBC and CORBA.

As we have seen, mobile, cloud based applications, sensor and social media generate a constant stream of data that lack appropriate structure. Transforming and cleaning the bits and bytes generated by these sources before loading them into a warehouse for analysis remained a huge challenge. As suggested earlier and corroborated by Hashem, I. A. T., et al. (2015) Efforts have been made towards simplifying the transformation process with some of the technologies mentioned earlier, but there is still a need to understand the context of unstructured data formats, particularly when the objective is to extract accurate information.

Scalability remains another obstacle. The lack of cloud computing features to support RDBMS which is associated with enterprises has made RDBMs less attractive for the deployment of large scale applications in the cloud in favor of NoSQL databases which on the other side are less accurate. These issues should be further explored.
The Data Staging Area field is probably one of the most important open search issues which pertain to the heterogeneous nature of data, which was also corroborated by Hashem, I. A. T., et al. (2015). The problem arises because data gathered from different sources do not have a structured format. Staging areas are a necessary mechanism to cleanse and transform the data, relating this particular i.e. data staging areas go on to argument that the problem arises from gathering from multiple sources.

Ontology matching is the main technology to overcome the problem of semantic heterogeneity and schema mapping is the solution adopted to solve heterogeneity. These approaches combined work to an extent. However, they have some limitations. Besides the syntactic and structural heterogeneity, there is a persistent problem which is the fact that ontologies change over time. In addition, ontologies constructed for the same domain by different experts can be different, it is necessary to analyze subclass and superclass relationships for each entity as well as the lexical correspondences, so it is necessary for ontologies to have instances to improve comparison. This has to be performed automatically.

The problem is just getting bigger. The evolution of information systems is causing that data is more and more heterogeneous and the task of integrating it is more complex, particularly as the data grows.

5.2 Reasons for data acquisition and integration – The value creation process

Data is gathered to extract value from it and decision makers are, through BI&A, the primary consumers of this data, which is used to increase performance in organizations. It is evident that before any data becomes valuable it has to be integrated. Furthermore, it follows that if the data is to be used in decision making then this data has to be accurate. If the data is not accurate then the decisions may be based on the wrong premises. It follows from this that the purpose of the data gathering is not just having a massive amount of data, rather having useful data. This data integration occurs at the integration layer on the BI&A stack i.e. data warehousing and it is then fed to the analytical layer where patterns are mined.

This is corroborated by Watson, H. J., & Wixom, B. H. (2007), who point out that after acquiring data, it gains value to an organization and is defined as Business Intelligence at the point where data is extracted from source systems into an integrated data warehouse where it is analyzed and important patterns are found through mining.

However the data integration effort is complicated by the disparity in formats of the data gathered as well as their vastness. Through the warehousing effort and in particular through the ETL process, data is cleansed in order to gain accuracy and transformed in order to fit a unified schema i.e. that of the data warehouse.

As mentioned earlier, Hand, D. J., et al. (2001) suggested that Data Mining deals with data which has already been collected. This means that the objectives of the data mining exercise play no role in the data collection strategy. In order to extract meaningful information the data sample has to be sufficiently large and because of the often large scale
nature of the data there are other problems that arise in terms of analysis techniques, timeliness, storage and other housekeeping issues.

As pointed out earlier, and suggested by Wu, L., et al (2007, June), BI&A is evolving to include Data Mining, Near Real time measures, scalability to support petabytes of data and thousands of concurrent users, cloud, virtualization and massively parallel processing to support Big Data Analytics. This can be seen in table 7 where Lim, E. P., et al. (2013) proposed a research framework to cover all the aspects of Business Intelligence and Analytics. It is evident that the emerging research and the foundation technologies shown in table 7 derive a clear and tangible need for several fields to coordinate efforts in order tackle the problem of heterogeneity. Figure 9 below summarizes graphically these fields and their interrelationships.

![Figure 9](image-url)

**Figure 9** illustrates the value creation process and the interrelationships of each field which is necessary to accomplish this. Note: black borders indicate related fields.

### 5.2.1 Business intelligence and data disparity – Evolution and emerging problems

BI&A Systems are complex, and often distributed geographically so there is a variety of inputs and components that are interrelated in this process and cause additional heterogeneity and the need to deploy distributed technologies.

Table 4 shows that BI&A is tightly dependent upon different data integration approaches, data warehousing, data mining and large complex heterogeneous systems also distributed geographically. As mentioned earlier S. Negash (2004) suggested that business intelligence
is an evolution of previous systems designed to support decision making and was particularly impacted by the development of hardware and software as well as the emergence of Data Warehousing advances in Data cleaning and information generated from other systems.

This was also corroborated by Wu, L., et al. (2007, June) who point out BI&A is evolving to include Data Mining, Near Real time measures, scalability to support petabytes of data and thousands of concurrent users, cloud, virtualization and massively parallel processing. Which corroborate the fact that a number of related fields that have to cooperate to tackle the problem of heterogeneity.

It is visible that technological developments and their corresponding impact has also affected BI&A. Contrasting table 6 and table 8 it shows the evolution of BI&A has drifted along with the evolution of the web towards an ever increasing amount of data heterogeneity, leading to an ever more complex challenge. Furthermore, the data shown in these two tables show that BI&A 2.0 and BI&A 3.0 depart from BI&A 2.1 in that it transitions from the structured to the unstructured and heterogeneous of the Semantic Web, to include social media and mobile and sensor generated data. Amongst the key characteristics of BI&A 1.0 were that it was DBMS based structured content and it required ETL processing and OLAP which are highly structured. In today’s world however the unstructured and heterogeneous coexist.

While being able to mine such amount of disparate data can bring many benefits to an organization. This increase disparity is not in line with Relational Database Systems which were thought for highly structured data. As a consequence, new alternatives are necessary to benefit from it.

As these results have shown, however, BI&A has also evolved along the evolution of the web and in its evolution path it has adopted other techniques which make it possible for it to mine and extract valuable data from sources such as social media or the web itself and in order to do so it deploys techniques such as Web Services, XML, Metadata, Semantic modelling techniques and ontology frameworks such as RDF which together with Data Warehousing enable a broader spectrum of sources from where to gather valuable insights.

5.3 Drivers for information disparity and its consequences

As shown earlier, one of the more complex problems when integrating several autonomous data sources is the heterogeneity of these sources. Kermanshahani, S. (2009).

Today we have seen a sharp increase in data disparity due primarily to the development of new technologies, in particular processing and storage which, in turn, have led to other developments such as the development of the Internet, and the evolution of existing techniques, business systems and trends.

Looking at tables 9 and 10 we can see a list of the different types of data which is being generated by social media in table 9 and in table 10 we can see the breadth of the internet of things, which spans through very many aspects of our society i.e. human, homes, medical, offices, factories, cities, and so on. Essentially what has happened is that after the
explosion of information created by the semantic web 2.0 the data sources have become more various, diffuse, scattered, and very vast amounts of information are processed and stored. The internet of things, social media and other business trends pretty much derived from the evolution of technology are contributing to this amount of despair data. This evolution is also giving way to new business trends and practices which also contribute to this data disparity because they themselves become generators of data. As a consequence, in the present day the problem at hand is far more complex than the one we had earlier.

As commented earlier Hashem, I. A. T., et al. (2015), Roche, M. (2015), Hendler J. (2014) coincided in that this heterogeneity is the result of the growth of almost unlimited different sources of data and problem gets bigger as the data grows. What happens essentially is that data from multiple sources are mostly of different types and representation forms and they are significantly interconnected and represented inconsistently. As a consequence when access is needed to different source systems, the sources and their data need to be re-arranged and put together. This implies that the general goal in terms of data integration is to provide a homogeneous, unified view on data from different sources, something in which Ziegler, P., & Dittrich, K. R. (2004) agreed, who further pointed out that information systems are not designed for integration. In addition, effective large scale analysis requires a large collection of heterogeneous data from multiple sources but machine analysis algorithms expect homogenous data and are not good at understanding disparity. As a consequence of this limitation, data must be structured first prior to data analysis.

5.3.1 The problem of size, disparity and velocity - Big Data

The problem of integrating heterogeneous data gets even bigger with highly unstructured data such as blogs, video, audiovisual content, text and natural language, all of which present even more in the context of social media and the internet of things which are generated by the Worldwide Web. To all this, we would have to add semantic and syntactic differences. In addition we would have to consider that social media is used by large segments of the population worldwide which implies that the data is very vast and changes all the time.

This generates an ample spectrum of data types. Some social media entities, like Facebook, use graph based NoSQL databases, other use MySQL, other are SOA based and use XML other JSON, just to give an example. Because of the very large heterogeneous dataset from social media, one of the major challenges is to identify the valuable data and how analyze it to discover useful knowledge improving decision making of individual users and companies which is corroborated by Bello-Orgaz, et al. (2016). Gathering huge amounts of data in it is not useful, even if it is integrated. The data has to be relevant.

This complexity and dramatic increase in data capture and disparity is giving way to what is known today as Big Data. Big data has been also precipitated by unprecedented technological advances and the evolution of the Internet together with Internet of things and social media. However, in this case the data is so vast and disparate, and it changes so rapidly that it is no longer possible to process it with the existing techniques such as the traditional ETL process.
For one thing, as mentioned previously and corroborated by Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U. & Qu, W. (2012) RDBMS used in the traditional ETL process are not suitable for the unstructured data. In addition to this, web data sets do not match the relational model and processing these data sets at scale implies additional problems. Furthermore, traditional distributed file systems cannot satisfy providers such as Google, Amazon, Yahoo and Microsoft.

### 5.4 Structured and unstructured data problems

There is a wide spectrum between structured and unstructured data and in between we can find data with some degree of structure on one end and other with hardly any structure at all on the other end, for which new approaches are necessary.

BI&A Systems have to deal with structured and unstructured data simultaneously. The formats in which the data arrive might not be determined upfront and when the data arrives it has to be structured and integrated in order to extract value from it.

Some data can be classified as semi-structured, because there is some inherent structure to it which, to a great extent can be attributed to the advent of Extensible Markup Language (XML) which allows encoding documents in a way that is human readable and at the same time it can be computer processed which promotes automation and it contributes to the propagation of data used in conjunction with Web Services.

Table 12 above shows some examples of unstructured data which includes such things as e-mails, memorandums, marketing material, reports of all kinds, audiovisual presentation and content of all kinds, including websites, spreadsheets and so on. S.Negash points out that the importance of these types of data is such that 85% of all business information exists as semi-structured data. Furthermore, data found on the web is less structured; facts corroborated by Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U. & Qu, W. (2012). Some of it, in fact, has no structure at all, for example, natural language, making it complicated to integrate. For this data which is unstructured and does not fit in rows and columns it is necessary to device classification techniques and taxonomies. Perhaps realizing of the potential of unstructured and semi-structured data businesses are forced to deploy solutions in order to extract value from it. Another problem which such data entails is that it is difficult to automatize, often requiring some degree of manual intervention.

The traditional ETL integration process works well with the problem of integrating structured data. Structured data fits the traditional data warehousing architecture. However, the problem with the traditional integration process increases dramatically when incorporating semi-structured data. This problem however has been partially overcome with the advent of data interchange formats such as XML and SOA.

With increase information disparity and massive amounts of data that changes constantly it becomes necessary to device new systems that allow for the massive ingestion of data structured and unstructured simultaneously and gives results in real time. This poses a challenge that has to be solved.
5.5 Challenges and approaches of integration of heterogeneous data

There can be the case where data has some structure in which case it can be dealt with XML and SOA and there can be the case where the data has no structure at all, such as for example natural language, in which case it requires that it be integrated by means of metadata and taxonomies. Table 12 showed some examples of semi-structured data. Most of the data received in an organization is in text format which, for example, can include mail, communication, web pages and social media content among other.

5.5.1 SOA and XML contributions and limitations

XML and SOA are commonly used together because they address the complexity issues by providing a uniform way of carrying the data i.e. through XML and abstracting complexities but not all solutions provided for dealing with such information address the problem of semantic heterogeneity. In the case of XML, for example, users can model the same data in different ways.

There have been important advances in dealing with semi-structured data the most representative of which is XML. XML has contributed to make it possible to deal with semi-structured data and as pointed out by and Hashem, I. A. T., et al. (2015) it is now possible together with the help of Service Oriented Architectures to deal with these data formats. It turns out however that the only two data exchange languages which have semantic verification are XML and GELISH. However, these technologies help integrating information across systems with different representations of data.

Refer back to Table 17 for a list of data exchange languages. Of the data exchange languages shown in this list, most of these except for XML and Gelish provide for semantic verification. The semantic verification is important because it permits error detection related to the meaning of artifacts. This is probably the reason that XML and XML related languages are used to address the problem of semantic differences. This implies that perhaps XML has contributed to the semantic web, at least through its syntax.

Informatica (2005) suggested that while the convergence of data integration and SOA can reduce it complexity, SOA by itself cannot address the issue of heterogeneity. If a SOA based system lacks a data integration foundation it can aggravate the problems. So it is necessary to work hand in hand with a data integration foundation. With a data integration in the mix a system can cleanse the data and resolve conflicting relationships and hierarchies amongst different entries. SOA has offered a good alternative to data integration particularly in minimizing complexities and simplifying its use. On the other side it has contributed to generate more and more data on the web. This can be corroborated by Parapatics, P. (2007, June) who suggests that SOA has contributed to the increase of information on the web by making it understandable by machines and thus being processed automatically with the subsequent increase in volume.

Looking back to table 18 for information systems based on SOA, it is visible from it that most of the focus of systems based on this architecture is put on semantic frameworks, workflow based models and in delivering information as a service. SOA has helped in applications which are business oriented and also those that have been oriented towards dealing with the problem of semantics in particular ontology based, and in addition it
contributes in the realm of cloud computing which we shall see later on. All of which are Integration approaches contemporary to the age of the semantic web.

In the web there are gigantic amounts of data which tends to be unstructured which does not fit the rows and columns of relational models. According to Martinez, J. M. P., et al. (2008) the web has been enriched with semantic annotations. XML can aid with the semi-structured data but then it has trouble when it comes to semantic heterogeneity with which according to Cao, et al (2006) it was not prepared to deal with.

Referring back to table 17 for a view of the different formats and contrasting this data with the results of table 9 and 10 which showed the internet of things and social media, it is evident that current meta-search engines try to integrate information from multiple sources of multiple domains which are constantly generating data at every level of our society worldwide, so data is emanating from many different contexts. Consequently, it is far more complicated to integrate because the repositories can be in various formats for example XML, HTML, Web Databases, sensor data and all kinds of files. Its numbers are growing exponentially.

Table 15 shows a list of approaches to integrate data and Table 16 for the different frameworks developed. In these tables it can be seen that so far there have been many attempts to integrate data and many frameworks developed but none of the integration methods shown helped to resolve the problem of semantic heterogeneity with the exception of the Ontological approach. Probably because of these reasons Ziegler, P., & Dittrich, K. R. (2004), mentions that explicit and precise semantics of integrable data are required in order to achieve semantically correct and meaningful integration output. To this end, the ontological approach helps in blending heterogeneous schemas across multiple domains. The semantic web research community has focused on the problem of semantic integration and the WC3 continues to develop markup to satisfy this need.

### 5.5.2 The semantic web

The mediation played an important role in the Semantic Web context in which information may not be processed from a single data source, but instead from combinations of multiple heterogeneous data sources with different representations of a common domain. OWL became the determinant standardization effort of the international research community in this area. This led to the proliferation of ontologies in specific knowledge domains expressed in OWL.

There are a number of instances where OWL fails and this technology does not work well in gigantic datasets. Some also argue that OWL is hard-coded against specific design patterns but anything that goes beyond those patterns cannot be expressed and it limits the flexibility and usability of a detection mechanism.

Because of the randomness and unpredictability of the information which we find on the web and the disparity of their formats it is very difficult to know which design pattern is most appropriate to use at a particular time. According to the World Wide Web Consortium the structural specification of OWL is a graph structure. In order to check constraints or execute rules on those graph structures, a graph language is needed; SPARQL seems to
fulfill this task. We shall see later on that Hadoop handles this unpredictability with a combination of tools which make it possible to take in all this unpredictable data.

Because of this unpredictability a trend towards moving from programmed logic towards dynamically generated and interpreted logic and new definitions for semantic technologies. This is confirmed by Workman, M. (2016) who suggests that there is a family of web ontology markup languages in the marketplace which many businesses are adopting, which clearly shows an effort is being done in the areas of semantic differences and it will allow us to create better and smarter information systems which exploit the semantics of data.

World Wide Web Consortium (W3C) is actively creating new markup to address the issue of semantic heterogeneity, this markup is developed to enrich information and enable intelligent systems in an attempt to make a better use of metadata.

One of these efforts can be characterized by RDF. The Resource Description Framework (RDF) its schema-free model makes it attractive because of its flexibility for describing entities in a way that many different data publishers can add arbitrary information about the same entity or create links between disparate entities. As we saw earlier and corroborated by Huang, J., Abadi, et al. (2011), Hadoop is able to take large amounts of disparate data even from the web.

5.5.3 Query languages

Although some techniques developed earlier such as global schemas, schema integration, domain specific wrappers and global transactions have produced positive steps, they are very complex and have never reached the stage of maturity required for deployment and usage and as commented earlier and corroborated by Telang, A., & Chakravarthy, S. (2007) as well as Ziegler, P., & Dittrich, K. R. (2004), they don’t address the problem of multi-domain. This complexity implies that development of such technologies can have been discouraged and unlike other which were open source, lacked a large community to support it.

A popular query language which gaining momentum is SPARQL, which is analogous to NoSQL and it is the language recommended by W3C for RDF. NoSQL databases have very attractive characteristics for processing very large and different data which lack structure such as the one in the web. In addition SPARQL is better equipped to deal with semantic heterogeneity. SPARQL is schema less and because of this it is very fast, however, as we shall see in the next section its schema less feature might lead to some trouble when it comes to accuracy.

5.5.4 NoSQL Databases

The poly-structured data which is generated today in the World Wide Web does not fit well in relational databases. We are drifting towards Web 3.0 and while structured data fits relational database systems, semi-structured data does so only to an extent and unstructured data does not work well at all because it cannot fit into rows and columns and require a complex format. Hashem, I. A. T., et al. (2015).
In addition, semantic differences seem to be more predominant in the unstructured data of the web and it poses a problem which OWL and latter RDF have tried to solve. However, Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U. & Qu, W. (2012) pointed out that processing such data at scale poses also a big challenge.

As a consequence of the evolution of the web as well as BI&A applications and practices it has incorporated the use of content data which is particularly difficult to fit in rows and columns. This is so because the structure of heterogeneous data is rather unpredictable. Take, for example, the case of content from sources such as social media i.e. natural language and, video, photographs, music and so on.

To tackle this problem we have seen the emergence of NoSQL which although popular now, have existed for a long time. However, their use has been precipitated by the surge of unstructured data in the web. Besides providing SQL querying capabilities they make it possible to deal with other formats and together with languages such as SPARQL contribute to the problem of semantic heterogeneity. Furthermore, NoSQL databases allow the large scale collection of data in different formats at unprecedented speeds which make them ideal for cloud computing and in particular the Hadoop suite. Besides it makes it suitable for the web and it is already being used by large companies in the social media realm which normally has highly heterogeneous data, such as for example Facebook which uses the graph API. See figure 3 for an example query to a NoSQL database used by Facebook. From this figure it is evident that it has close resemblance to a SQL query, only that it yields a value, pair result.

It is important to note that SPARQL which is analogous to NoSQL databases is the W3C recommended query language for RDF, it addresses the issue of semantic heterogeneity. We discussed earlier that SPARQL is schema less and because of this it is very fast. However, however this poses a problem when accuracy in the data is needed such as in circumstances when the data is used in decision support systems i.e. in a business intelligence context.

Because they do not natively support the ACID test (atomicity, consistency, isolation, durability) the reliability is low. Furthermore, when deployed within the context of data warehousing together with SPARQL the process of staging is skipped and when using them, data is loaded as is, because of the absence of a schema. This can be a source of inconsistencies and errors which are not affordable to decision makers and it is perhaps the reason why relational databases are usually used by businesses and often for transactions that require great precision.

May be it is a combination of both that is required in order to address both needs, relational databases which provide accuracy for business decision support and NoSQL for ingesting vast amounts of ever changing data very rapidly.

5.5.5 Data Warehousing – In the age of Big Data
Business systems such as ERP generate highly structured data while other like CRM generates semi-structured data. The ETL process has worked well in solving the already
Integrating Heterogeneous Data

Gabriel Nieva

complex problem of integrating structured and with the help of XML it has been possible to add some sort of structure to the data which was unstructured.

5.5.5.1 ETL process – Challenges in the age of big data

There is extensive research on warehousing and integration particularly in dealing with highly structured business transactions for which the classic data warehousing strategies will work well. However, per definition data warehouses provide historical data and not real time. Widom, J. (1995, December), Fan H. (2005), Oracle (2005), Parapatics, P. (2007, June), Kakish, K., & Kraft, T. A. (2012), Intel (2013).

However, in today's world most of the data adds value only if it is processed in real time all the time. That is, that it is permanently available and that the data stream has no interruptions Oracle (2012). What would the effect of this entire heterogeneous data gathering be otherwise? Take for example a sensor measuring a patient’s blood parameters during a clinical trial for a pharmaceutical company. If it stops working the doctors or researches might be unable to see an important condition which requires immediate attention. Furthermore interruptions in such stream of data might render it useless altogether. What was today can be different tomorrow, or even after taking the medicine. Every minute and its seconds count.

Furthermore, the constant availability of the data is necessary for business critical applications and the reduced latency is necessary because data loses its value as it ages, for example take a sensor measuring the passenger traffic at a subway station. If it is not real time, the management might be unable to provide for the increase traffic in the in an immediate manner. This can include security, more trains, cleaning, and other staffing requirements.

The same can be said at an investment bank. Detecting a sharp increase in the volume of transaction in real time is of paramount importance. Without this information the bank might run out of liquid assets in a peak of withdrawals. In this way, today the conventional ETL approach for data integration becomes less useful.

The ETL process is what controls the flow of data between the many sources and other BI applications. As these evolve and become more complex and large the ETL tools have to accommodate this evolution by keeping up with the pace. See Figure 3 for an illustration of the ETL process.

Considering that the web is today one of the largest if not the largest source of data we need to address the challenges posed by the data coming from the web, these datasets tend to not to have much structure and don’t fit well Relational Database Management Systems (RDBMS).

This implies that in order for BI&A to capitalize on the value of the data found on the web, a data warehouse would have to be able to manage such large amounts of data which are normally in the order of petabytes and do not fit well relational databases. On the other side, in order to provide real time the data would have to be extracted and processed very rapidly from its sources.
As suggested by Parapatic, P. (2007, June), the more frequently the ETL process is performed the finer is the data granularity in the data warehouse and the finer data granularity the more accurate analysis and therefore better suited for decision support. However, the traditional ETL with its relational databases is not suitable for such an immense amount of data because it requires data to be structured so that it can fit a unified schema and this process creates a bottleneck situation.

So there seems to be a compromise between speed and accuracy, which explains recent research trends in the field of data quality. The quality is degraded as the data sources increases and the faster we process the data also decreases its accuracy. However, to be useful for decision makers the data has to be accurate, and it cannot have lost its value because it’s old.

5.5.5.2 Contribution of NoSQL databases

In summary, NoSQL databases have very attractive characteristics for processing very large and different data such as the one in the web. They are designed to be deployed on distributed cluster computers and offer choices for data consistency and support parallel programming using MapReduce. In addition SPARQL is popular in Big Data Cloud-based Projects.

While they support SQL like query languages, these are databases that allow storage and retrieval mechanisms different from relational databases. Many NoSQL systems are designed to be deployed in distributed-cluster computers, provide built in support for parallel processing using Apache Hadoop's Map Reduce and feature many application programming interfaces (API).

NoSQL databases have very attractive characteristics for processing very large and different data such as the one in the web. They are designed to be deployed on distributed cluster computers and offer choices for data consistency and support parallel programming using MapReduce.

5.5.5.3 Real Time ETL - Lowering Data latency

It is evident from the data gathered that there is an increasing demand for real time warehousing from BI&A applications. The problem is how to overcome the issue of heterogeneity with such a gigantic amount of data.

In table 27 we have seen some options to real time ETL. It is evident from these results that there are different ways deal with data latency, for example by means of Near Real Time ETL or by Trickle and Flip. Near real time is not an easy task but near real time is much simpler to achieve, it suffices to increase the update interval of the data warehouse, for example, instead of weekly updates, daily updates, but in today’s world a matter of minutes can cost much resources. With the Direct Trickle and Flip the lack of staging areas make can degrade the quality of data. We shall remember that it is in the staging area is employed for the purpose of data transformation, reconciliation and cleaning and that data has to be as accurate as possible to serve be fed to decision support systems.
The problem with the Trickle Feed approach is that constant updates on the tables by the querying OLAP tools end up degrading the query performance of the data warehouse, mainly because the staging tables are duplicated and the copy is swapped with the real-time tables which causes an overhead.

To overcome this it is possible to deploy a multi-staged trickle and flip. The ‘Multi-stage Trickle & Flip’ approach reduces amount of data to be copied and possible collisions between loading and querying activity but it is complex to implement, in addition the very existence of staging tables slows down the whole process, this is where the transformations and cleaning occurs.

External real time data cache provides information up to the second, but since it resides outside the data warehouse it may require additional infrastructure, in particular another database and its associated maintenance server, management and other things.

On the other side, integrating OLAP and OLTP raises important questions related to the logical database and whether physical data residing in OLAP and OLTP repositories can be integrated in data warehouses in the same database environment. For one thing the data which resides in OLAP is deormalized i.e. could have a certain degree of redundancy and inconsistencies, and Relational On Line Analytical Processing (ROLAP) is normalized with the 3rd normal form and it is best to keep them separate.

In addition Chaudhuri, S., & Dayal, U. (1997) suggests that the operational databases are tuned to support the known On Line transaction processing (OLTP) workloads and that executing complex queries against the operational databases would result in a very slow and unacceptable performance.

As commented earlier and suggested by Kakish, K., & Kraft, T. A. (2012) Event driven real-time such as the one proposed by the BI&A approach have two major challenges:

3) Queries running continuously
4) Algorithms cannot be recurrent

For starters, running queries continuously, which is required by the query processing once events arrive from input streams, might not be feasible, with huge amounts of data, for example, such as the one we are seeing in Big Data, the constant querying might cause a large overhead which would require a big processing power and slow down the real-time feature. Perhaps, however, this could be overcome by cloud computing. Distributed Systems platforms such as Hadoop allow for the processing of massive amounts of heterogeneous data. But the question remains would the data be accurate? It would have to be in order to work for decision support.

Secondly the algorithms which require iterating over the data are not possible if the data is streamed. Mining techniques frequently require such iterations.

As mentioned before, BI&A's primary users are decision makers; this was corroborated by a Gartner report cited by S. Negash (2004) which said that one of the major user groups of
BI&A is decision makers. Decision Makers expect reliable information, i.e. the data must be accurate in order for them to use them and extract value. So these doubts can contribute to decision makers being wary of this way to go and therefore further investigation would be necessary to determine the validity of this procedure.

These findings suggested that in order to keep it up it was necessary to go from a batch processing strategy to a single threaded processing that extracts and loads data in real time through continuous parallel processes. This is particularly true given the fact that the problem with the ETL process is compounded as the volume and disparity of the data increased. In addition, data changes quickly and old data becomes obsolete. This was one of the premises or 3 Vs of Big Data i.e. Velocity, which implies Rapidly Changing data. This rapid change of data demands real time.

This is when apache Hadoop comes in handy. Apache Hadoop is a system that can ingest massive amounts of disparate data very fast because of its subsystems, for one thing it uses of NoSQL databases and SPARQL which is schema less. This system and subsystems not only make it possible to ingest massive amounts of heterogeneous data. Furthermore, it can be deployed using the cloud computing paradigm, making it possible to do away with the problem of infrastructure. Cloud computing in turn uses SOA as its basis and this it conceals complexities making it attractive to business users and allowing them to pay as you go. In any case the problem of accuracy remains. Let us see why.

**5.5.5.4 ETL with Hadoop**

In regards to heterogeneity, in the age of big data it is imperative to deal with structured, semi-structured, and even entirely unstructured data simultaneously and on the fly. This, is in itself is a daunting task considering the volume of data which we are seeing today. As explained previously and suggested by Intel (2013), databases have become far more powerful and RDBMS today support very complex transformations in SQL, including in-database mining, and quality validation, cleaning profiling, statistical algorithms and more. However, despite all the progress made in the RDBMS field, these database systems were not conceived for these tasks.

During the initial overview of heterogeneity, it was explained and affirmed by Chen, J., et al. (2013) and Chen, M., et al. (2014) that Heterogeneity restricts the efficiency of data conversion. To this end, Apache Hadoop and its associated components contribute in making it possible to confront the problems that vast and heterogeneous data brings along.

Refer to figure 6 for Hadoop architecture and table 28 for an overview of Hadoop Components. Apache Hadoop and MapReduce, with its divide and conquer approach, provide an array of tools that make it easier to process vast amounts of despair data structured and unstructured on the fly. For one thing with Apache Flume it is now possible to collect, aggregate and move large amounts of data from multiple resources into HDFS or another central storage. This makes it possible to ingest vast amounts of heterogeneous data both structured and unstructured. Furthermore, with Apache Sqoop it is possible to automate the process of transferring data between Hadoop and a relational database in a concurrent way i.e. parallel and simultaneously check for errors. Intel (2013). This error
checking is essential; let us refer back to the need for decision makers to possess accurate information. With the Warehousing approach granularity was achieved with more ETL iterations and this yielded accuracy. There seems to be a compromise between speed and heterogeneity vs accuracy, that is, the less heterogeneous the data and the slower the more accurate. However, in today’s world, waiting even one day can mean loss of millions of dollars. As suggested earlier and corroborated by Parapatics, P. (2007, June), the more frequently the ETL process is performed the better suited for decision support, and not specifying schemas can wear out the reliability of the data, making it less suitable for decision support.

HiveQL which is a dialect of SQL which provides a subset of SQL syntax and it provides a simple way to query. As commented during the section on Hadoop with ETL, which was corroborated by Intel (2013) HiveQL does not seem to have a sufficiently low latency to process huge amounts of disparate data. For similar reasons Hashem I.A.T, et al. (2015) also suggest that Hadoop lacks as of yet a viable query processing strategy for such a vast and heterogeneous amount of data. This suggests that there may be a bottleneck situation which can be pinpointed to the query processing strategy. Furthermore, NoSQL databases with their (value, pair) also have their limitations i.e. if keys are the only way to retrieve and store data; it might be difficult to use them in any kind of business application especially for analytic purposes.

Furthermore, the database problem with data heterogeneity in large datasets is partially solved with NoSQL databases which don’t have a predefined schema. Hadoop is a Distributed File System and as such it can handle all this kind of data because the content of the files is irrelevant to it, i.e. it is a file system, and in this way it allows for the ingestion of large datasets of heterogeneous data. However, if the data is to be processed by a system which requires some relational operation it will not work properly.

Although HiveQL is slow, there is an active community that is permanently enhanced by a developer community in order to enable a low latency solution to query processing. To this end they there is also the development of HBase and HDFS and Pig Latin, which is a procedural programming language that provides a high-level abstraction for MapReduce. MapReduce can now be extended with user defined functions written in Python, Java and other languages. These languages make it easier for an ordinary developer to program so the complexity problem is addressed as well. This implies that it will be easier to develop further, perhaps even for open source to proliferate. These developments will perhaps address different problems and in the end solve some of the dilemmas that we have such as Accuracy vs Disparity and vastness.

In the ETL arena Hadoop contributes to the ETL process in two very important ways.

1. The possibility to upload massive amounts of data without specifying a schema,

2. The possibility to offload the transformation of raw data by scalable parallel processing Intel (2013).

See figure 7 for an illustration of offload ETL with Apache Hadoop.
In addition referring back to the ETL process using Hadoop seen in figure 10 we can see that it is also able to process information from the web, from social media and from sensor logs addressing in this way the semantic web issues, social media and huge amount of data from the internet of things. As stated earlier, it is possible to store huge amount of semantic web data in Hadoop clusters which answer queries fast enough even when built by cheap commodity class hardware. Husain, M. F., et al. (2009, December). The latter is an additional advantage because it decreases cost.

However, not all the issues are solved with MapReduce, particularly when dealing with very large and heterogeneous data sets such as those found in Big Data. As stated earlier, and confirmed by Gudiwada, V. N., et al. (2015), it falls short, for example, in dealing with time-varying graphs and dynamic networks, real-time processing requirements, and scalable stream data processing which pose additional challenges. In any case, in dealing with big data it is common to use NoSQL databases and Hadoop together with cloud computing.

Much research is underway in this field, some of which to address the issues of real time, quality Hadoop and MapReduce.

5.6 Contribution of Cloud Computing and related technology

Enterprises are increasingly turning to BI&A to find useful patterns and valuable information that they can use to maximize profits. New Challenges such as the one Big Data have emerged which pose a big challenge to this BI&A effort. In order to make some sense of the data gathered it must first of all be integrated, and then mined for useful patterns. This is a gigantic and very complex task which requires plenty of resources, concretely in infrastructure and expertise, cooperation among different fields and interaction of many disparate systems.

As mentioned during section 4.1 on data heterogeneity, data from multiple sources are generally of different types and representation forms and significantly interconnected and the more breadth in diversity the more complicated to deal with Hashem, I. A. T., et al. (2015). For example, with the internet of things and social media it is almost impossible to know in advance which data formats should be expected. This renders many methods and techniques invalid, for example, the programmatic methodology with its object orientation described earlier which OWL follows. This is so, because it requires predetermined design patterns which will not work because of this unpredictability. The type of data that will come up is simply not possible to anticipate.

The problems of integrating large datasets of heterogeneous data are many. For starters it requires a lot of computing power. Some of the problems, like the one which has to do with infrastructure, are solved by deploying cloud solutions. SOA promotes loose coupling by encapsulating a specific service behind a well-defined interface i.e XML and it allows automatic data interchange between applications and it provides a data model for semi structure data. So thanks to XML it is possible to process semi-structured data and in addition, SOA Web Services which are based on XML standards facilitate the delivery of software applications as a service automatically. Cloud computing is based on SOA and it uses these technologies. From this it follows that if cloud computing is Service Oriented i.e. uses SOA, and then it is suited to deal with deal with these complexities.
Furthermore, as mentioned earlier, it seems that the combination of several database systems need to work together, Relational Database Systems for reliability and NoSQL databases for speed and ingesting large amounts of structured and unstructured data, to this we have to add multiple other types of systems which need to coexist to make it possible to integrate the rapidly changing and vast heterogeneous data which we see today. This is where cloud computing with its virtualization and encapsulation of complex aspects comes in handy.

When it comes to the interaction of disparate systems the virtualization feature of cloud computing is very interesting. Because of the many systems that are involved in integrating heterogeneous data, virtualization is almost necessary in order to give the illusion of homogeneity. So Cloud Computing not only does away with the need of heavy investment in infrastructure, it also helps hiding the many complexities of these systems by delivering parts as a service, it helps dealing with unstructured and heterogeneous data and in addition it gives the illusion of homogeneity when in fact the computing requirements and systems to achieve the task of integrating large heterogeneous datasets are highly heterogeneous themselves.

As explained before Cloud computing is based uses SOA and as the result gathering have shown SOA provides a high level of abstraction and the ease of use makes it very attractive to environments where the experts are more business oriented and less technically oriented i.e. oriented towards a computing science realm. Furthermore, using a cloud computing architecture the storage, the application itself and the required infrastructure reside outside in a different realm to where the clients are located, which implies that it can be outsourced and thus the cost of the infrastructure required to process large amounts of data is lower because it can be paid on an “as you go” basis.

Looking back at table 11 we can see that Big Data projects frequently utilize some of the technologies deployed in cloud and the strategies used, i.e. MapReduce’s divide and conquer make it easier to deal with large datasets and heterogeneous data types. Furthermore, advances have been made with the advent of NoSQL databases, which are based on a dictionary data type which is much more efficient that the hashing of the relational databases. NoSQL databases are designed to be deployed on distributed cluster computers in cloud environments and offer choices for data consistency and support parallel programming using MapReduce.

As discussed during the result gathering on cloud computing, Kaisler, S., Armour, F., et al. (2013, January) pointed out that there seems to be an open research question in cloud computing which is that of scalability. These limits, they argued, are known for specific algorithms and corresponding running times with specific implementations on specific machines at specific scales, but there is no well-known limit particularly when it comes to unstructured data.

Perhaps this is due, at least in part, to the unpredictability of the data. Running times of some algorithms are much higher than other. There is even the possibility that some algorithms cannot deliver a solution in a reasonable amount of time and yet other are what we call incomplete algorithms, which entail that it does not provide a guarantee that it will
eventually find a satisfying assignment or declare that the given formula is not solvable. Since the data which we acquire today grows and changes very rapidly, it is unpredictable and specific algorithms, just as best suited specific design patterns, cannot be predicted.

The “on demand” characteristic of cloud computing also bears a certain degree of unpredictability, and this can imply that the scalability cannot be pinpointed with precision. An important thing to consider is that as the demand for more real time access to data continues to increase it is necessary to provision the availability of cloud services. For example in a case of a peak, there needs to be resources to cover the client’s needs.

So cloud computing seems a solid computing solution, which together with Apache Hadoop and its suit of components better address the problem of integrating heterogeneous data. However it still has some limitations some known and some not so known. Looking back at table 31 there are some open challenges and research issues in Cloud Computing which are directly linked with the heterogeneous nature of data and are related to the very efforts of data integration which have been already covered in this paper. These are:

- Availability,
- Transformation,
- Data quality,
- Heterogeneity itself,
- Data staging,
- Distributed storage systems
- Data analysis.

For one thing availability addresses the issue commented earlier. The resources are accessible on demand which makes it difficult to predict. Transformation is exactly what happens at the staging tables which are not suited to process these massive, heterogeneous and rapidly changing data, quality itself is what is gained through the process of staging i.e during the ETL process which was discussed previously. We shall remember also that some of the Hadoop components deployed through cloud platforms such as SPARQL is schema less and it skips the staging process altogether. And other things such data analysis and distributed storage are a problem consequence of the issues to integrate heterogeneous data and linked also to scalability. Information from large amounts of data requires scalable analysis algorithms which produce timely results and distributed storage which is palliated with Haddop HDFS is also linked with scalability.

5.7 Open Challenges and Future work

The survey conducted has identified several areas where further work is required in order to tackle the problem of integrating heterogeneous data which are exposed as follows:

1. **Query Languages** – Query languages have to be simpler for the user to use, yet powerful. Complex query programming for the databases can be difficult. This goes also for the query languages as well as other APIs required to integrate data. In addition, faster querying mechanisms are needed to keep up with the vastness and disparity of the data. As of right now Hadoop lacks adequate querying mechanisms.
2. **Database** – Structured data remains more reliable for decision making because of its reliability. The ETL process provides granularity and this granularity provides reliable data. Further work in the ETL would be required to make it faster but more accurate. Concretely, to make staging areas more effective and efficient at handling unstructured data from different sources.

3. **Unstructured data** – Needed computational solutions, particularly using unstructured data.

4. **Quality** – Faster implies less reliable. More ETL more granularities implies better for decision making. Further work is required on quality, one of the quality problems increases with different data sources and the problem is aggravated for the need to process data in real time.

5. **Cloud Computing** – Together with concurrent programming can provide the necessary speed for structured and unstructured data to be processed simultaneously very rapidly, but Cloud Computing remains uncertain with respect to scalability. In addition, the heterogeneity of the massive amount of data, the heterogeneity of the sources and the increased demand for real time pose other challenges such as that of availability which require further investigation.

6. **Cheaper solutions needed** – Open source solutions provide a cheap alternative to software but their continuity depends on a developing community to keep it alive, the larger, the better, since it is not driven by the going concern of commercial solutions.

7. **Improvements in Artificial Intelligence** – For example to make querying and matching automatic, recognize errors better. Also in semantic heterogeneity to provide better matching alternatives for ontology matching.

8. **Need for standardization** – The more is not necessarily the better. Standardizing certain aspects could lead to reaffirm the course of action to take. More solutions in the market can mean more disparity.

9. **RDF** – While addressing the requirements of the semantic web and successes demonstrated such as being able to process RDF with Hadoop. Storing huge number of RDCF triples and being able to query them is a challenging task. Furthermore, besides the syntactic and structural heterogeneity, there is a persistent problem which is the fact that ontologies change over time, rendering the mappings obsolete, which needs further research.
6 Conclusion

The purpose of this study was to survey the problem of heterogeneity in order to explore the current approaches and their limitations as well as to propose possible solutions to a more efficient integration of data. In order to do this task it was helpful to answer the questions formulated in the detailed problem statement section:

- Why do we need to integrate data and who are the primary consumers of the data?
- What are the drivers of the increase information disparity and the consequences?
- What is the nature of heterogeneous data?
- Which approaches to integration of heterogeneous data are there and what are their limitations?
- Which computing paradigm is most adequate, in dealing with vast, heterogeneous and rapidly changing data environment?

It has been shown that heterogeneity arises from the deployment of different technologies and methods to store data and it has been sharply accentuated by technological improvements which not only have allowed us to process and store more data, but also generate far more data than before.

6.1 The need to integrate data and the consumers of the data

The survey has shown that decision makers through Business Intelligence and Analytics are the primary consumers of data and expect accurate and real time data which permits them to take action immediately. However, raw data remains useless without being able to make some sense of it. In order to perform any data mining effort the data has to be structured first and useful patterns have to be found because if the data is not useful, it would not add any value. The study also shows that because vast amounts of data are being generated very fast, and in changing very rapidly data loses its value quickly. Because of this data has to be processed in real time. This answers the question of why we need to integrate and it also clarifies who are consumers of the data.

6.2 Drivers of increase information disparity and its consequences

It is clear from the study that technological advances in processing, storage and networking, which led to the evolution of the Internet and its transition from Web 1.0 to Semantic Web 3.0 and in particular social media and the Internet of Things, have contributed to the vast and rapidly changing heterogeneous data that we have to deal with today.

Social media is available worldwide and it is utilized by large segments of the population. Through social media there is the possibility to publish rich content of all kinds such as video, audio, text, pictures and all kinds of files which are generated constantly in a rather unpredictable way. The Internet of Things, in particular sensor data and mobile machines, generate a constant stream of data which changes very quickly. This rapid change demands that the data be processed in real time because the data loses value as it ages and it becomes obsolete.
Some business trends and systems such as CRM, ERP have also evolved and generate vast amounts of disparate data and marketing strategies such as Omni-Channel Communication, which advocates for the seamless integration of these systems contribute to the dissemination of disparate data.

These trends have precipitated us into what we know today as the age of Big Data. What happens essentially is that data is so vast and it changes so rapidly that it cannot be processed by conventional methods. It requires new technologies and methods. This answers the question of the drivers that cause data heterogeneity and it shows its consequences.

6.3 The nature of heterogeneous data
A survey of the different types of heterogeneity that can arise has been performed. The many technological advances have produced a sharp increase of data which is unstructured. The web, mobile cloud based applications and social media generate a constant stream of data that lack appropriate structure and makes it necessary to device different ways to deal with data.

On the one side the study shows that heterogeneity arises from the deployment of different technologies and methods to store data. Simultaneously technological advances contribute to the proliferation of this data. There are different types of heterogeneity which can be considered, amongst which the semantic heterogeneity is the most complicated to deal with and for which ontologies are the best alternative. The structural heterogeneity can be dealt with networking protocols and standards, particularly through the use of middleware.

There is a wide spectrum of possibilities when it comes to the type of data that we should expect from today’s information systems. This data has to be processed in different ways and structured in one way or another in order to mine value from it. To one extreme we have structured data and to the other extreme we have unstructured data. It becomes clear that while structured data fits well the relational model, unstructured data does not fit well the rows and columns of RDMS. Semi-structured data tends to be business oriented and it is critical to the organization. In this type of data there is some inherent structure and it is possible to process with the advent of XML. Ontology based data integration also has some limitations, there is a persistent problem which goes frequently ignored, which is the fact that ontologies change over time, rendering the mappings obsolete.

6.4 The approaches to integration of heterogeneous data and their limitations
A survey of the different frameworks, which deal with different aspects of data heterogeneity, has been performed.

The survey shows that only one approach, the ontological, deals with the highly unstructured data from the web. The ontological or semantic approach, with newer frameworks and query languages such as OWL, RDF and SPARQL and the development NOSQL databases, have contributed to make it possible to process the increasing amount of
heterogeneous and ever changing data from the internet were data is highly unstructured and semantically difficult to integrate.

Improvements to palliate the ETL process limitations have also been covered. The warehousing approach is a central part of Business Intelligence and Analytics, particularly when processing such heterogeneous and vast amount of data through its ETL process. NoSQL databases, Apache Hadoop and its components, SparkQL, and other developments such as the possibility to process data from the web though RDF in combination with the ETL process using Hadoop, contribute to the integration of heterogeneous data. The transitioning from batch processing to highly concurrent and parallel processing techniques with these technologies make it easier to process highly poly-structured data in real time. This answers the question of which approaches are there and what are their limitations.

6.5 The computing paradigm most adequate to deal with vast, rapidly changing, heterogeneous data

This paper shows how cloud computing conceals complexities and makes things easier, more attractive to business oriented users and automated because its architecture is based on SOA. It is evident that virtualization did away with a problem of heterogeneous infrastructure, and it is clear that since the systems that are needed for this task can reside outside the organization they can be contracted as needed, i.e. on demand.

The study also shows that cloud computing together with the Apache Hadoop and its component suit is the best computing paradigm to solve the problem of integration of heterogeneous and rapidly changing data. Hadoop has the capacity to ingest massive amounts of data, both structured and unstructured, without specifying a schema. Hadoop offers the possibility to offload the transformation of raw data by scalable parallel processing which is a departure from the batch processing strategies traditionally deployed. In addition Hadoop is simple and it is possible for any programmer to use. Despite this, the problem of query optimization remains an issue. Hadoop is tainted by some as not having adequate query strategies because apache hive is slow. These aspects answer the question of the most adequate paradigm in dealing with vast, heterogeneous and rapidly changing data.

6.6 Concluding remarks

The integration of heterogeneous data is a difficult problem to solve, it is compounded by the massive amount of data that we are generating and it requires cooperation of several interrelated disciplines which use different, yet at times overlapping technologies, more concretely, Data Integration, Data Warehousing, Data Mining, Business Intelligence, its related field of Big Data and Cloud Computing.

Integration of heterogeneous structured and unstructured data and information from distributed, heterogeneous virtual clouds need further research. Much research has been performed on the issue of heterogeneity but further progress in this area would be significantly accelerated by combining expertise from the areas of Database Systems, Artificial Intelligence, Information Retrieval and Cloud Computing, in particular the issues pertaining to scalability.
References


Facebook (n.d.) “Graph API” Available at: https://developers.facebook.com/docs/graph-api last Accessed: 08-09-2016;


Firat, A., Madnick, S., & Grosof, B. (2002). “Knowledge integration to overcome ontological heterogeneity: Challenges from financial information systems”.


Integrating Heterogeneous Data

Gabriel Nieva

2016-10-22


IGI Global “Semantic Verification” Available at: http://www.igi-global.com/dictionary/semantic-verification/34563 Last Accessed:


Oracle (2005) “Data Warehousing Guide, 10g Release 2 (10.2)” Available at: https://docs.oracle.com/cd/B19306_01/server.102/b14223.pdf Last Accessed: 2016-08-4


Parapatics, P. (2007, June). “Evolution of Data Warehousing and Data Mining”. In Seminar, MIT.


Tech Target “Omni Channel Integration” Available at: [http://searchcio.techtarget.com/definition/omnichannel](http://searchcio.techtarget.com/definition/omnichannel) Last Accessed: 2016-08-11


