**LITERATURE SURVEY**

**GLOBAL SALES DATA ANALYTICS**

**ABSTRACT :**

Big data is currently a buzzword in both academia and industry, with the term being used to describe a broad domain of concepts, ranging from extracting data from outside sources, storing and managing it, to processing such data with analytical techniques and tools. This thesis work thus aims to provide a review of current big data analytics in an attempt to highlight big data analytics’ importance to decision making. Due to the rapid increase in interest in big data and its importance to academia, industry, and society, solutions to handling data and extracting knowledge from datasets need to be developed and provided with some urgency to allow decision makers to gain valuable insights from the varied and rapidly changing data they now have access to. Many companies are using big data analytics to analyse the massive quantities of data they have, with the results influencing their decision making. Many studies have shown the benefits of using big data in various sectors, and in this thesis work, various big data analytical techniques and tools are discussed to allow analysis of the application of big data analytics in several different domains.

**Big data analytics (BDA): tools and methods :**

* Big data storage and management
* Big data analytics processing
* Big data analytics

**Big data storage and Management :**

➢ The big data environment accepts and demands all possible data sources. On the other hand, EDW approaches data sources with caution, as it is more streamlined towards supporting structured data (Elgendy and Elragal, 2014; Hartmann et al., 2019).

➢ Due to increasing number of data sources and data analyses possible, big data storage requires agile databases to give analysts the opportunity to produce and adapt to data easily and quickly (Elgendy and Elragal, 2014; Hartmann et al., 2019).

➢ A big data repository must be deep, allowing analysts to analyse the datasets deeply by using complex statistical methods (Elgendy and Elragal, 2014; Hartmann, T. et al., 2019).

**BIG DATA ANALYSING PROCESSING :**

a) Fast data loading: limited interference between disk and network, to speed up query execution.

b) Fast query processing: workloads are heavy, therefore real-time requests should be processed as quickly as possible to satisfy user requirements. The data placement structure should also have the ability to process multiple queries as query volumes increase. Sarah Al-Shiakhli 20

c) Highly efficient utilisation of storage space: as user activities grow rapidly, they need scalable storage capacity and computing power. As disk space is limited, it is necessary to manage data storage during processing and address the space issues adaptively.

d) Strong adaptivity to highly dynamic workload patterns: the underlying system should be highly adaptive, as data processes have different workload patterns and the analysing of big datasets has many different applications and users, with different purposes and methods (Elgendy, N. and Elragal, A., 2014).



Big Data Analytic Processing

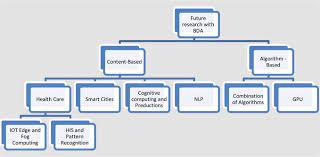
After the big data storage, comes the analytic processing. According to [10], there are four critical requirements for big data processing. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and

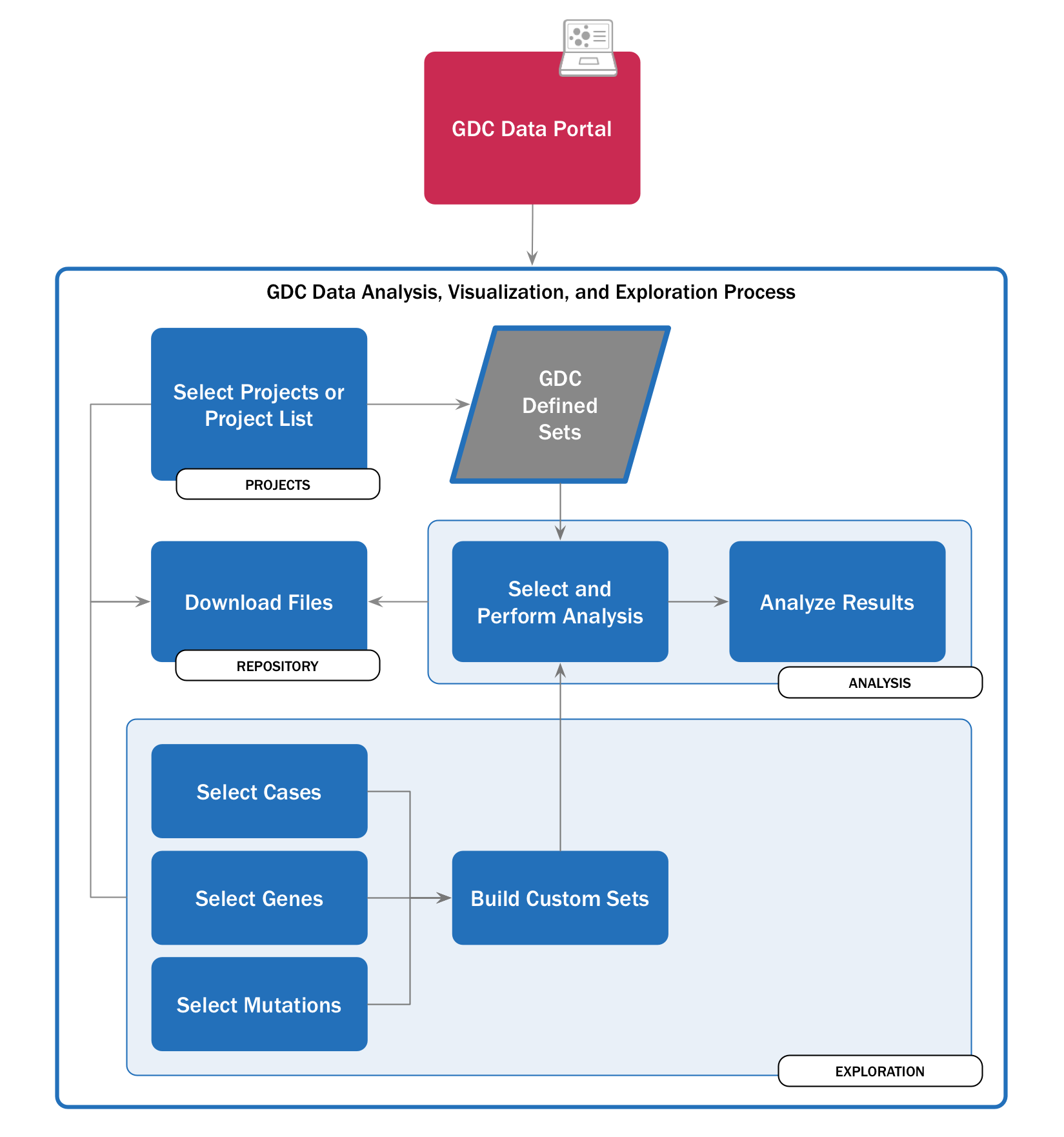
real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase. Additionally, the third requirement for big data processing is the highly efficient utilisation of storage space. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space

necessitates that data storage be well managed during processing, and issues on howto store the data so that space utilisation is maximised be addressed. Finally, the fourth requirement is the strong adaptivity to highly dynamic workload patterns. As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the

underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns.

Map Reduce is a parallel programming model, inspired by the “Map” and “Reduce” of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions [6]. According to EMC, the MapReduce paradigm is based on adding more computers or resources, rather than increasing the power or storage capacity of a single computer; in other words, scaling out rather than scaling up [9]. The fundamental idea of MapReduce is breaking a task down into stages and executing the stages in order to reduce the



Regression Analysis is a mathematical tool used to discover correlations between several variables based on experimental or observed data. Where analysis defines the relationships between variables as non-random, such analysis may make the correlations between variables appear simpler and more regular

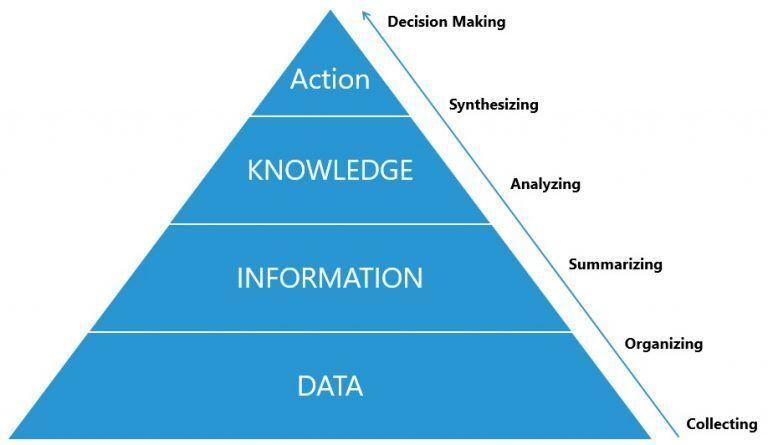
Quality Management and Improvement :

Especially for the manufacturing, energy and utilities, and telecommunications industries, big data can be used for quality management, in order to increase profitability and reduce costs by improving the quality of goods and services provided. For example, in the manufacturing process, predictive analytics on big data can be used to minimise the performance variability, as well as prevent quality issues by providing early warning alerts. This can reduce scrap rates, and decrease the time to market, since identifying any disruptions to the production process before they occur can save 224 N. Elgendy and A. Elragal significant expenditures [4]. Additionally, big data analytics can result in manufacturing lead improvements [17]. Furthermore, real-time data analyses and monitoring of machine logs can enable managers to make swifter decisions for quality management.

Also, big data analytics can allow for the real-time monitoring of network demand, in addition to the forecasting of bandwidth in response to customer behaviour. Moreover, healthcare IT systems can improve the efficiency and quality of care, by communicating and integrating patient data across different departments and institutions, while retaining privacy controls [4]. Analysing electronic health records can

improve the continuity of care for individuals, as well as creating a massive dataset through which treatments and outcomes can be predicted and compared. Therefore, with the increasing use of electronic health records, along with the advancements in analytics tools, there arises an opportunity to mine the available de-identified patient information for assessing the quality.

Conclusion



The purpose of this study was to offer a literature review on the topic of big data analytics. This began with the presentation of a general background to the topic, including big data definitions and characteristics, followed by a review of big data analytics tools and methods. This thesis presented data analysis techniques characterised in four sections: supervised, unsupervised, semi-supervised, and reinforcement learning. Some analytics techniques were also presented, such as clustering, correlation, regression, and factor analytics, and some big data tools and platforms such as Hadoop, Apache Mahout, and R were explained in relation to these. Big data storage, management, and analytics processing were also discussed, and some emergent advanced data analytics techniques further examined.