SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Investigation into the Challenges of Implementing Scalable Big Data Technologies and Techniques

Heinrich Gladwen Dankie Geiseb[1*] and Nashandi Ndinelago[2]

[1]Department of Informatics, Namibia University of Science and Technology (NUST), Brahms St, Windhoek, Namibia. E-mail: heinrichgeiseb@gmail.com

[2]Department of Informatics, Namibia University of Science and Technology (NUST), Brahms St, Windhoek, Namibia. E-mail: nnashandi@nust.na

Article Info

Volume 3, Issue 1, May 2023
Received : 11 January 2023
Accepted : 17 April 2023
Published : 05 May 2023
doi: 10.51483/IJDSBDA.3.1.2023.1-24

Abstract

The age of Big data is here, and the effective managing of this data can determine whether an organization succeeds or falls behind its competitors. In more extreme cases it can determine how effectively a country can manage a pandemic, as we have seen with the Covid-19 pandemic. Dealing with Big data present many challenges but equally there are a lot of solutions to these challenges via the use of available technologies. Furthermore, Technology in regard to managing Big data is widely available and has evolved in the past few years and there are very few excuses to why industry practitioners should not be employing scalable Big data technologies. In this study the researcher has identified the different challenges that arise when dealing with Big data and how the different technologies available on the market have been designed to deal with different challenges that Big data present. The introduction of new hardware architectures, as well as the continual accumulation of data, provide new data management difficulties. Reasoning with respect to a predefined set of resources is no longer relevant (i.e., computing, storage and main memory). Instead, using scalable technologies and techniques, data processing algorithms and processes must be designed with unlimited resources in mind. The research enquires into the different technologies currently being used by industry practitioners in Namibia and the different challenges they encounter in regard to their respective approaches. Additionally, a short online survey was also conducted to investigate the knowledge of relevant stakeholders in terms of Big data technologies. Finally, this study urges industry practitioners to desert the use of outdated technologies and to encourage one another to employ scalable technologies by providing incentives.

Keywords: *Industry practitioners, Hadoop, Scalable, Big data, Data analytics*

* *Corresponding author: Heinrich Gladwen Dankie Geiseb, Department of Informatics, Namibia University of Science and Technology (NUST), Brahms St, Windhoek, Namibia. E-mail: heinrichgeiseb@gmail.com*

1. Introduction

Rouse (2019) defines Big data as a combination of structured, semi-structured, and unstructured data that can be mined for information and applied in machine learning projects, predictive modelling, and other advanced analytics applications. The three V's are frequently used to describe Big data are:

- The tremendous volume of data in a range of environments.
- The large variety of separate data formats that are commonly stored in Big data systems
- The velocity at which information is created, gathered, and analyzed.

Velocity is further explained by Abu-Salih *et al.* (2021) as the collection of data at high speeds, near real-time, and in real-time from diverse data sources. Moreover, Abu-Salih *et al.* (2021) states that the velocity of data necessitates advanced solutions capable of storing, processing, managing, and analysing streams of heterogeneous data and inferring value on motion. Figure 1 shows how much data was produced in a minute from numerous sources such as Google, You Tube, Instagram and Facebook in 2019. Due to processing capacity restrictions, developing a Big data architecture is a recurring difficulty for developers (Rouse, 2019). According to Rouse (2019), Big data solutions must be personalized to an organization's specific demands, a do-it-yourself project that involves IT and data management teams to piece together a specific collection of technologies and tools. Data privacy, data security and ethical issues, which include information management systems, the maintenance of personal or fatal information, and data misuse, are key ongoing challenges in the Big data innovation ecosystems (Koo *et al.*, 2020). Furthermore, making data accessible to data scientists and analysts is also a challenge in managing Big data systems, particularly in distant contexts with several platforms and data repositories (Rouse, 2019). Big data is retrieved from data sources such as enterprise apps, social media feeds, email systems, and papers produced by employees. According to Rouse (2019), in order to assist analysts in finding relevant data, data management and analytics teams are increasingly creating data catalogues that include metadata management and data lineage capabilities.

Figure 1: How Much Data is Generated in Each Minute in 2019

Big data is a data-driven technology that analyzes massive volumes of data to extract usable information from the data and is able to make predictions based on that data (Koo *et al.*, 2020). Furthermore, Big data is defined by Ratra and Gulia (2019) as a set of strategies that elicit a form of integration that is used to uncover enormous unknown quantities. Additionally, Koo *et al.* (2020) claimed that Big data is being hailed as a new source of energy for corporate and technical advancements, as well as economic growth. Koo *et al.* (2020) also states that numerous economic and political interests drive Big data, particularly data integration, analysis, and mining processes. The use of Big data in a range of industries has resulted in a rapid expansion of a wide range of data resources and numerous data analysis technologies (Koo *et al.*, 2020). According to Koo *et al.* (2020) the continuous growth of the Big data market is being aided by data analysis technologies such as standardized data mining and statistical analysis technologies. Several efficient and reliable techniques are now ready to aid organizations in interpreting this vast volume of data from a variety of heterogeneous data sources, according to Saleh *et al.* (2018). Saleh *et al.* (2018) report that these variety of technologies such as Hadoop and Apache Spark allow industry practitioners the ability to process, analyse, and present the data in a way that is easy to grasp, visually appealing, and appropriate for the business language and stakeholders' goals, According to Rouse (2019), firms use Big data in their systems to enhance operations, give better customer service, build targeted marketing campaigns, and take other activities that, in turn, can raise revenue and profitability. Organizations have relied on various data sources to characterize, understand, forecast, and provision economic and commercial activity, as well as to determine the next course of action (Saleh *et al.*, 2018). Furthermore, according to Rouse (2019), medical researchers use Big data to detect disease indicators and risk factors, and doctors use it to help diagnose illnesses and medical problems in patients. Therefore, the effective mining, storing and processing of Big data by implementing scalable Big data technologies is of paramount importance to a lot if not all industries. Today, complicated analysis of vast data sources involves the employment of high-performance computing systems such as massively parallel machines or clouds (Talia, 2019). According to Talia (2019) in order to reach this goal, new design and programming concerns must be addressed and resolved. Andreea (2021) stated that a common issue is the lack of scalability, when your project starts using an increased number of resources. The importance of utilizing scalable Big data technologies is very important, and this is emphasized by the points made by Andreea (2021). Andreea (2021) reports that if your infrastructure cannot keep up with the growth of your data, bottlenecks in your Big data and analytics workloads will occur. Additionally, Andreea (2021) claims that the infrastructure will ultimately hit its resource limit in a non-scalable system. Hence this will result in migrating to a different systems that are able to process the data and according to Andreea (2021) this is a time-consuming and an intricate process that will result in significant downtime and costs.

1.1. Background

Data which is huge, difficult to store, manage and analyse through traditional databases is termed as "Big data" (Muniswamaiah *et al.*, 2019). Saleh *et al.* (2018) states that in recent years, businesses have generated, received, processed, and stored a massive amount of data from a variety of sources, including databases and the Internet. As a result, Saleh *et al.* (2018) describes Big data management as the procedure of storing, managing and handling of large amounts of data. Furthermore, Saleh *et al.* (2018) explain that, Big data has several characteristics and has a tendency to change in format; It is sometimes clear or complicated, organized or unorganized, secure or vulnerable. This makes Big data management and storage more difficult and therefore would undoubtedly necessitate assistance by advance data management techniques and technology. As a result, sophisticated, unique, and flexible analytics are required to deal with the challenges of collecting and analyzing a diverse set of Big data islands, which are rapidly growing as a result of the massive amounts of data generated by tracking sensors, social media, transaction records (Abu-Salih *et al.*, 2021). According to ELE Times Research Desk (2018) it is critical to understand your systems' transactional and analytical data processing requirements and to make appropriate selections. Additionally, it is essential to select the appropriate tools to process Big data because, data is essential for gaining important insights into target demographics and client preferences (Abu-Salih *et al.*, 2021). Every connection with technology, whether active or passive, generates new data that might be used to describe us. Ku (2021) states that with data being captured through products, video cameras, credit cards, cell phones, and other touchpoints, our data profile is growing exponentially. If analyzed correctly, these data points can explain a lot about our behavior, personalities, and life events. Furthermore, Ku (2021) explains that Companies can use these insights to

improve their products, business strategy, and marketing campaigns to better serve their target customers. But in order to reap the benefits of processing Big data companies should implement technology that can grow and adjust based on the demands imposed by Big data, these technologies include data storages, computing power, data analytic software to name a few.

A number of research papers that were reviewed followed a similar trend with regards to the purpose of the respective studies. These studies either investigated the challenges faced when dealing with Big data or the studies reviewed the different technologies available for implementation. Whereas this thesis combines the two approaches and reports the findings of a thorough study to establish the challenges that developers face when dealing with Big data and how available technologies are designed to ease these problems. The review of different technologies such as the Hadoop and Apache frameworks and how they have been developed in order to deal with the challenges Big data processing presents is of importance to this study. Also central to the research study is the investigation into the technologies industry practitioners are using and the challenges they face. Additionally, a short survey was conducted in order to explore the relative knowledge of individuals in the Information Technology industry with regards to Big data skills and technologies. This research will be conducted in Windhoek, Namibia.

1.2. Problem Statement

According to Doty (2021) data is the new gold and with the continuously growing amount of data available because of the internet, it easy to see why. It is essential to store, process and analyse this data in order to make sound decisions in very important industries like finance, medical and science. However, many companies fail to implement scalable Big data processing technologies. This is because of the challenges that arise when managing Big data. The challenges that arise when trying to implement scalable Big data applications are vast, as stated by Espinosa *et al.* (2019) this includes structured and unstructured data, data growth versus expansion, data distribution, access schemes and governance, as well as updating, searching and retrieving relevant data efficiently for analysis in the Petabyte and Exabyte range, in a variety of formats. All these challenges if left unaddressed, could contribute to more specific problems such as implementation of less scalable systems, privacy breaches, incomplete data, and ineffective and inefficient storage, hence the current study aims were to address these challenges by reviewing different technology available for Big data processing.

An article in the Namibian published in 2019 titled "Unlocking data potential in Namibia" outlined the view of Ruu Nombanza, the founder of We Are Capable Namibia (WAC). Nombanza emphasized the importance of data and how it could be used to help improve different aspects, namely, Namibia's economic woes, by studying the past performances of policies and interventions. This article was published in 2019 and the fact that very few organizations actually implement effective scalable technologies to process their data is a problem that a lot of companies face in Namibia.

1.3. Research Objective(s)

The current study's main objective is to provide an extensive review of different literatures and practices in the industry of Big data regarding the challenges industry practitioners face when dealing with Big data and the different technologies available for the implementation of scalable applications. Furthermore, the research will investigate and give recommendations on best practices regarding the implementation of scalable Big data technologies and techniques. Additionally, the study will review the different technologies available in the industry for organizations and individuals alike to consider and utilize. Lastly, the study also addresses the following sub objectives:

- Identify the different challenges when dealing with Big data.
- To understand and investigate the traditional Big data processing, storing and analyzing applications and techniques.
- Determine how different technologies are design to overcome the different challenges.
- Investigate what technologies industry practitioners are currently using in regards to managing and analyzing large amounts of data and the challenges they face.

- Investigate the knowledge of IT experts in terms of Big data skills and tools required to perform Big data analyses.

1.4. Research Deliniation

The research study was limited to 5 industry practitioners, more specially data scientists and Big data experts that use different types of technologies to process Big data. The inclusion of data scientists and data analysts is, because of the lack of Big data experts in the region where the study was conducted.

1.5. Significance of the Study

This study will benefit industry practitioners and education institutions in effectively implementing scalable Big data technologies. It will also help industry practitioners to select the appropriate technologies and techniques to overcome the challenges they face when dealing with Big data, during the mining, storing and analyzing processes. This study will furthermore contribute to new knowledge in the field of Data management, especially the effective implementation of scalable Big data platforms in order to get the most out of the ever-growing data.

2. Literature Review

2.1. Issues and Challenges of Big Data in Information Management

2.1.1. Scalability

Scalability is one of the most prominent issues when it comes to Big data management, the challenge of scalability was examined in a study by Al-Sai *et al.* (2019). According to Al-Sai *et al.* (2019) scalability refers to the ability to supply resources in an appropriate manner to meet business needs. Additionally, Saleh *et al.* (2018) emphasized the scalability issue by stating that to address data volumes that are growing faster than computer resources, businesses must reassess their design, build, and data processing components. Moreover, Saleh *et al.* (2018) asserted that as technology moves toward cloud computing, allocating resources on expensive and large clusters requires the emergence of new data processing methods to achieve workload targets in a cost-effective manner all while dealing with system failures that may have a negative effect on data processing and recovery methods.

Furthermore, Jayashree and Abirami (2018) state that scalability is a serious issue that has persisted for many years because Big data involves significant amounts of data. The reason for this is because parallel data processing technologies that were once employed to preserve and analyze data are no longer useful when data sizes continue to grow at an exponential rate. Jayashree and Abirami (2018) emphasized that in order to keep up with the ever-increasing size and flexibility of data, new ways must be used.

2.1.2. Storage

Another prominent challenge when dealing with Big data is storage, Junghanns *et al.* (2017) (as cited in Alabdullah *et al.*, 2018) claimed that traditional relational databases are out of date and incapable of storing and processing data generated by modern business applications. Furthermore, the study by Alabdullah *et al.* (2018) reported that everyday problems like data recording, data storage costs, and synchronization issues drive data scientists to use NoSQL alternatives.

The problem of storage is also emphasized by a paper written by Jayashree and Abirami (2018) where they reported that, Some cloud models are in the early stages of development, hence the storage problem has persisted for so long. Another problem connected to storage according to Jayashree and Abirami (2018) is data replication, More data is collected from a variety of sources and devices, and it's critical to avoid data duplication, which wreaks havoc on the analysis stage. Additionally, since large amounts of data is stored in differed storage facilities, for medium and small businesses, the expense of keeping massive amounts of data remains a challenge (Jayashree and Abirami, 2018).

Furthermore, according to a study by Saleh *et al.* (2018), the location where Big data will be kept is also a huge concern for organizations. As a result, emerging storage models such as flexible web services featuring petabyte-scale data stores have difficulties in terms of ensuring minimal input/output delay from Big data repositories (Saleh *et al.*, 2018). According to the study by Saleh *et al.* (2018), when dealing with high-

volume data acquisitions and supporting a range of mixed data structures, it's tough to avoid increased delay.

2.1.3. Privacy

A literature by Saleh *et al.* (2018) named privacy as one of the major concerns in Big data. Furthermore, Saleh *et al.* (2018) stated that privacy is one of the most sensitive issues why organizations are hesitant to adopt Big data. The disclosure of personal and sensitive information to people who should not have access to it, whether done intentionally or unintentionally, is considered a violation of privacy (Saleh *et al.*, 2018). According to Saleh *et al.* (2018), privacy breaches happen when security procedures are inadequate. Furthermore, Saleh *et al.* (2018) state that the lack of privacy can jeopardize safety and diversity. Finally, Saleh *et al.* (2018) argue that because most of the information in the datasets is vital, and therefore security breaches should be a top priority for firms, as they can cause major privacy problems and risk the safety of employees and customers.

Furthermore, Koo *et al.* (2020) stated in their literature that, data must be gathered in advance with the consent of the person creating the data while securing data through active data collection. Furthermore, the collection of data used internally for a system that captures a large number of log records is subject to the consent of the creator of the data in line with an internal data ownership policy (Koo *et al.*, 2020). In addition, Koo *et al.* (2020) explain that when collecting data passively, information is typically gathered using an automation program. Hence, Koo *et al.* (2020) reported that if the information to be acquired is sensitive such as personal information. There may be legal implications, and thus the subject collecting the data should be careful and collect in accordance with the nature of the collected data.

In addition, a study by Jayashree and Abirami (2018) emphasizes the importance of privacy in Big data management by reporting that every agency and department should have its own data center, and all information mut be kept private. Additionally, they report that because certain data cannot be shared owing to privacy concerns, data sharing becomes an issue in smart cities, as data must be shared among various devices. In addition to it, Jayashree and Abirami (2018) state that on real-time applications, significant amounts of data are updated and saved in various formats every second. Hence, creating a standard data format and extracting information directly from a real-time application is tricky.

2.1.4. Polyglot Persistance

A literature by Vargas-Solar *et al.* (2017) stated that another problem that is fast becoming very common in managing Big data is the use of multiple storage spaces. This is because modern applications take a polyglot approach to persistence, where traditional databases, non-relational data stores, and scalable systems associated with the developing NewSQL movement are all used at the same time (Vargas-Solar *et al.*, 2017). As a result, application developers face issues integrating these varied databases, agile schemata, and non-standard APIs (Vargas-Solar *et al.*, 2017). Furthermore, Vargas-Solar *et al.* (2017) claim that due to the nature of schema-less data storage, developers must also ensure that the implicit schemata that these applications rely on are properly maintained. Finally, including the data schemata in the program code could cause issues with maintenance and efficiency (Vargas-Solar *et al.*, 2017). As a result, developers must manually analyze the entire source code to fully comprehend the data structures employed by these programs.

2.1.5. Lack of Skill Set

New skills are being demanded as new technologies enter the market. According to a study conducted by Saleh *et al.* (2018), accurate and actionable data mining and analysis, especially in real-time, requires a high level of technical expertise. Furthermore, Saleh *et al.* (2018) proposed that organizations form a common data analyst team, either to equip existing staff with the right skillset by providing training and obtaining necessary certifications, or to seek out new employees who are specialized in Big data and can understand data from a scientific perspective.

According to Al-Sai *et al.* (2019), the most significant problem that most organizations will face when attempting to use Big data, is preparing for Big data implementation and recruiting engineers with Big data experience. Furthermore, Al-Sai *et al.* (2019) contends that a significant constraint to extracting value from Big data would be a scarcity of skills and Big data professionals. Furthermore, Michael and Miller (2013) (as cited in Al-Sai *et al.*, 2019) reported that a lack of skills is a direct consequence of the slow adoption of Big data,

leaving a gap related to the need for adequate preparation and expertise people who support the adoption process. Finally, the study by Gao *et al.* (2015) (as cited in Al-Sai *et al.*, 2019), noted that, due to the novelty of the Big data area, it is very difficult and expensive to find and to hire Big data experts.

2.1.6. Data Availability and Accessibility

The difficulty of data availability and accessibility in Big data relates to data being available and accessible in much bigger volumes and at much quicker speeds in real time and across multiple industries (Saleh *et al.*, 2018). This massive volume of data must be processed and analyzed, and the activities may be time-consuming because analysis takes longer (Saleh *et al.*, 2018). In today's fast-paced world, analysis results are expected almost instantly. Furthermore, Saleh *et al.* (2018) note that organizations require information from a multitude of sources and, as a result, may not have enough data to perform analytics. As a result, they will probably seek or buy data from third party companies who may or may not want to share it. Finally, Saleh *et al.* (2018) claim that it is unrealistic to claim that Big data analytics always yields reliable findings, because erroneous data can lead to erroneous conclusions, leading to erroneous decision-making.

2.1.7. Inaccurate and Incomplete Data

When it comes to Big data, the data being analyzed must be complete. According to Saleh *et al.* (2018), Big data is useless if it is not used to improve decision-making. The term uncertainty can be used to explain data inaccuracy and incompleteness, and uncertainty is defined as a scenario with unknown or poor information (Hariri *et al.*, 2019). Moreover, Hariri *et al.* (2019) provided an example in which the risk of the learning algorithm producing incorrect output is significant since it was fed faulty or incomplete data for training.

Furthermore, Saleh *et al.* (2018) published a paper on the problems of Big data. According to Saleh *et al.* (2018), Big data is pointless unless it is used to better decision-making. Hence, organizations must undertake data management activities such as data gathering, extraction, and recording, data cleansing, data integration and aggregation, data representation and analytics, including modelling, analysis, and interpretation (Saleh *et al.*, 2018). Data will be used to do analysis will come from diverse sources and of different formats (Saleh *et al.*, 2018). Hence, according to Saleh *et al.* (2018) it could be comprised of incorrect information, duplication, and contradictions. Furthermore, Saleh *et al.* (2018) claims that data of extremely poor quality is unlikely to provide any useful insights or promising opportunities to an organization's precision-demanding business tasks. Additionally, Saleh *et al.* (2018) reports that data must be carefully structured in order to be analyzed in an efficient and accurate manner. Finally, Saleh *et al.* (2018) states that incomplete data can result in incorrect data analysis, which can lead to poor results, judgment, and decisions.

Lastly, a paper by Jayashree and Abirami (2018) name incomplete data as one of the persisting problems when dealing with Big data. Jayashree and Abirami (2018) reported that data uncertainty is caused by missing information, which is a big concern that must be handled. Furthermore, uncertainty arises from incomplete data during data analysis, which must be addressed (Jayashree and Abirami, 2018).

2.2. Scalable Technologies for Big Data Proccesing

2.2.1. Big Data Processing Ecosystems

2.2.1.1. Hadoop Ecosystem

Hadoop is an Apache-led open-source project. It was created with the goal of processing large amounts of data in a timely, efficient, and cost-effective manner (Ajah and Nweke, 2019). Hadoop can deal with both unstructured and structured data. Additionally, Ajah and Nweke (2019) report that Hadoop is user friendly and extremely scalable. Moreover, Ajah and Nweke (2019) state that Hadoop YARN broadens Hadoop's ability to support a wide range of applications, minimize Hadoop's limitation to only running MapReduce applications. As a result, real-time analytics, dynamic querying and data streaming applications are all possible with Apache Hadoop due to YARN (Ajah and Nweke, 2019). Ajah and Nweke (2019) also mentioned Common, which is an important component of the Hadoop ecosystem. Ajah and Nweke (2019) reports that the Common package includes features and tools for a variety of activities, including error detection, codec compression, proxy user authorization and input/output utilities. One of the main disadvantages of Hadoop has to do with security. Furthermore, according to DataFlair (2019), because Hadoop is developed in java and java is the most commonly used programming language is vulnerable to cyber-attacks. Additionally, the

problem that Hadoop has with dealing with small files is made clear by (DataFlair, 2019), because, a small file is nothing but a file which is significantly smaller than Hadoop's block size which can be either 128 MB or 256 MB by default. Hence, these large number of small files overload the Namenode.

Figure 2 shows that MapReduce is part of the technologies use in a Hadoop ecosystem. MapReduce is a distributed programming model for batch processing that employs key-value pairs (Ajah and Nweke, 2019). Additionally, Ajah and Nweke (2019) state the responsibilities of MapReduce as resource scheduling and job management. Moreover, Ajah and Nweke (2019) report that MapReduce is made up of two main components: the mapper and the reducer. According to Ajah and Nweke (2019) the mapper filters and transforms data. Additionally, stated by Ajah and Nweke (2019) is that MapReduce receives data blocks from big HDFS files. Part of the Big data management category in the Hadoop is Hive. According to Ajah and Nweke (2019), Hive is a SQL-like Hadoop interface that was developed at Facebook. Hive also allows SQL users to generate MapReduce jobs using conventional SQL commands and relational table structures, even if they aren't familiar with MapReduce (Ajah and Nweke, 2019). Furthermore, Ajah and Nweke (2019) assert that Hive handles all data as though it belonged in tables and allows users to define tables on top of the data files. Another component of the Hadoop ecosystem is Pig. Pig is a data flow scripting language that was developed at Yahoo (Ajah and Nweke, 2019). Moreover, according to Ajah and Nweke (2019), Pig is a program that turns scripts into MapReduce jobs. The piggy bank is a type of storage used by Pig scripts, for loading and unloading massive volumes of data into and out of Hadoop, the tool employs an agent. Flume also belongs to the Big data processing category and through the usage of agents, is highly suited for obtaining weblogs from multiple sources (Ajah and Nweke, 2019). According to Ajah and Nweke (2019), Flume is equipped with a number of connectors that make it simple to create robust and dependable agents. Flume is also extremely scalable across a significant number of machines (Ajah and Nweke, 2019). Sqoop is another component of the Hadoop ecosystem. Scoop is a tool for transferring data into and out of relational databases, according to Ajah and Nweke (2019). Sqoop is a combination of the words SQL and Hadoop (Ajah and Nweke, 2019). Furthermore, Ajah and Nweke (2019) state that Scoop is an excellent tool for importing or exporting data among any RDBMS and the Hadoop Distributed File System (HDFS).

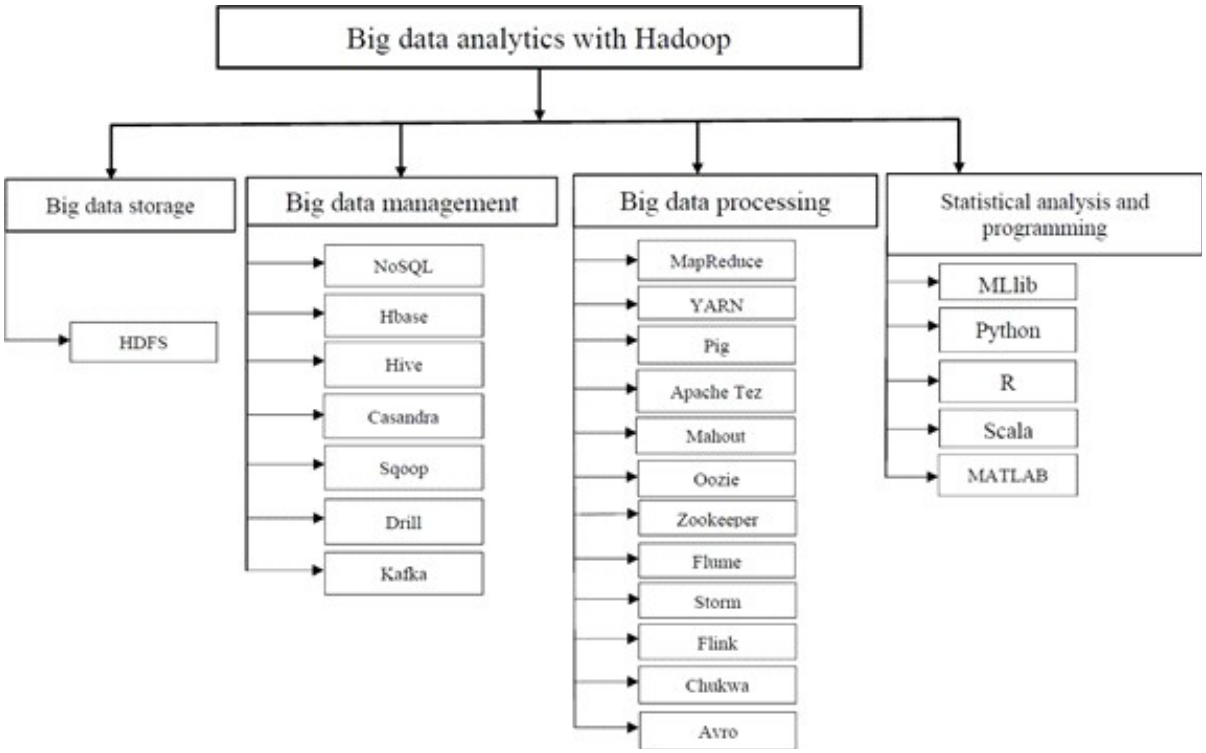

Figure 2: Hadoop Ecosystem

Source: https://doi.org/10.3390/bdcc3020032

Furthermore, Oozie is a component of the Hadoop ecosystem. Oozie is a Hadoop cluster workflow and coordination tool (Ajah and Nweke, 2019). Oozie, according to Ajah and Nweke (2019), runs on a supercomputer platform. It enables jobs to run concurrently while awaiting input from other jobs. According to Ajah and Nweke (2019), one of the benefits is that Oozie includes a sophisticated scheduling tool. As a result, this enables the supercomputing platform to coordinate jobs that are awaiting other requirements (Ajah and Nweke, 2019). Finally, HBase is a popular columnar NoSQL database implemented on top of Hadoop (Ajah and Nweke, 2019). Ajah and Nweke (2019) report HBase is an Apache project that uses Google's Big Table data storage paradigm. Furthermore, according to Ajah and Nweke (2019) HBase presents data in a columnar format and is schema-less.

2.2.1.2. Apache Spark Ecosystem

According to Shaikh *et al.* (2019), Apache Spark is a sophisticated Big data processing tool that uses a hybrid framework. Furthermore, according to Shaikh *et al.* (2019), Apache Spark is a hybrid framework that supports stream and batch processing capabilities. More importantly, Shaikh *et al.* (2019) claim that despite the fact that Apache Spark follows many of the same principles as Hadoop's MapReduce engine, it beats it in terms of performance.

Furthermore, Apache Spark has a few drawbacks. According to DataFlair (2019), Apache Spark lacks file management systems and thus must rely on other databases, and when it does use Hado's HDFS, it inherits all of the HDFS's drawbacks, such as the inability to deal with small files. Another disadvantage is the cost; according to DataFlair (2019), in-memory capability can become a bottleneck when we want cost-efficient Big data processing. Spark requires a lot of RAM to run in-memory, so its price is quite high.

As illustrated in Figure 3 the spark ecosystem consists of Spark SQL. According, Shaikh *et al.* (2019) Spark SQL is formerly known as Shark. Spark SQL is a distributed framework that works with structured and semi-structured data (Shaikh *et al.*, 2019). Additionally, it facilitates analytical and interactive application for both streaming and historical data which can be accessed from various sources such as JSON, Parquet and Hive table (Shaikh *et al.*, 2019). Another component of the Apache Spark ecosystem is Spark Streaming. Shaikh *et al.* (2019) reports that, Spark Streaming enables users to process streaming of data in real time. In order to perform streaming analysis, Spark streaming enhances the fast-scheduling capability of Apache Spark by inserting data into mini batches. As seen in Figure 3 MLlib is part of the

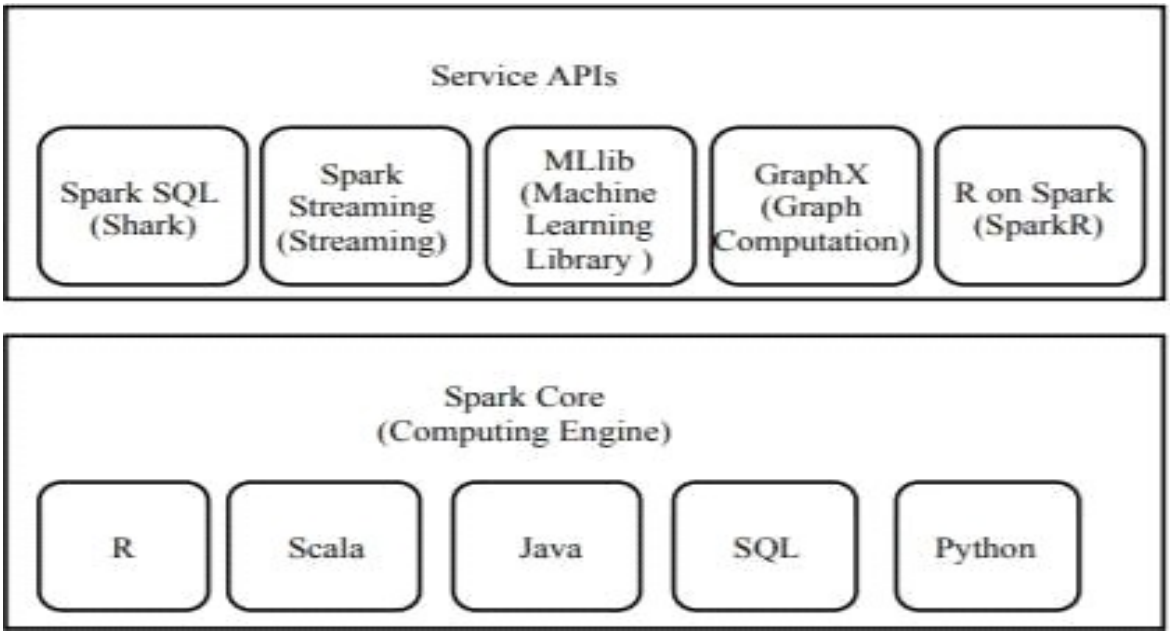

Figure 3: Apache Spark Ecosystem

Source: https://www.researchgate.net/publication/339176824_Apache_Spark_A_Big_Data_ Processing_Engine

Apache Spark ecosystem. MLlib provides excellent machine learning algorithms at a fast rate, making machine learning simple to use and scale (Shaikh *et al.*, 2019). These machine learning algorithms include regression models, classification models, clustering and linear algebra. Finally, Shaikh *et al.* (2019), also report that MLlib is a library that can be used in Python, Java and Scala. Figure 3 shows that GraphX is also part of the Apache ecosystem. GraphX, according to Shaikh *et al.* (2019), is a graph computation engine that allows for the large-scale compilation, manipulation, transformation, and execution of graph-structured data. Additionally, GraphX is a collection of Spark RDD APIs that make it easier to create directed graphs (Shaikh *et al.*, 2019). Furthermore, one of the most important components in the Apache Spark ecosystem is the Spark Core. Hence, many Apache Spark features are developed on top of the Spark core (Shaikh *et al.*, 2019). Additionally, Shaikh *et al.* (2019) reports that Spark core provides a vast range of APIs as well as applications for programming languages such as Scala, Java, and Python APIs to facilitate the ease of development. Furthermore, Shaikh *et al.* (2019) claim that in-memory processing is incorporated in Spark core to improve speed and overcome the MapReduce problem. Finally, Shaikh *et al.* (2019) reports on SparkR. Shaikh *et al.* (2019) states that SparkR is a R package that allows data scientists to use the Spark engine directly from the R shell. Furthermore, SparkR's Data Frame is the fundamental unit of SparkR, same as the Data Frame is the basic data structure for data processing in R. According to Shaikh *et al.* (2019) SparkR can do things like selection, filtering, and aggregation with huge datasets.

2.3. Data Mining Technologies

2.3.1. RapidMiner

RapidMiner allows you to access, load, and analyze any sort of data (Ratra and Gulia, 2019). Thus, according to Ratra and Gulia (2019), data, such as text, static images, and media, can either be structured and unstructured. Furthermore, Ratra and Gulia (2019) state that RapidMiner extracts valuable information from structured data and also convert unstructured data into structured formats. Furthermore, Ratra and Gulia (2019) state that RapidMiner supports over 35 different file formats, including URL's, ARFF and SAS. It has wizards for connecting to databases such as Microsoft Excel and Access, as well as CSV files (Ratra and Gulia, 2019). It can connect to MongoDB and Cassandra NoSQL databases, as well as Dropbox, web pages, PDF files, and a variety of other applications (Ratra and Gulia, 2019).

2.3.2. Mozenda

According to Ratra and Gulia (2019), Mozenda offers technology, which can be given as code (SaaS and on-premises choices) or as a managed service, that allows customers to take unstructured internet data and turn it into structured formats for data analysis. Mozenda is as a tool that does internet scraping. Internet scraping is a technique for extracting data from the internet using a computer (Ratra and Gulia, 2019). Furthermore, Ratra and Gulia (2019) explain that Mozenda converts website data into structured data using a point-and-click code tool.

2.3.3. Octoparse

Ratra and Gulia (2019) state that, Octoparse is the ultimate tool for information extraction its cable of web locomotion, information movement, and information mining. Hence, the Octoparse internet scraping program, users will have the ability to convert the entire "internet" into structured information (Ratra and Gulia, 2019). Additionally, Ratra and Gulia (2019) report that the Octoparse team has never slowed down in making information more available and prepared for everyone in order to achieve autonomous internet scraping in the real world.

2.4. Data Storage Technologies

2.4.1. Hadoop Distributed File System (HDFS)

HDFS is a filesystem designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware (White, 2015). According to Abu-Salih *et al.* (2019), HDFS is best suited for write-once, read-many-times processing, whereas Kudu can be used to address fast-data (insert/update) issues. Furthermore, Abu-Salih *et al.* (2021) claims that expensive, highly reliable hardware isn't required; instead, readily available hardware on the market is a viable option when constructing Hadoop Ecosystems. There is no Single Point of Failure (SPOF) issue in HDFS, because one of HDFS's design goals is to handle

SPOF, so when a workstation fails in a Hadoop cluster, there will be no visible obstruction to users (Abu-Salih *et al.*, 2021). One issue to be aware of is that HDFS may not be the best choice when an application demands low-latency data access (Abu-Salih *et al.*, 2021). According to Abu-Salih *et al.* (2021), if the data set comprises a large number of small files, it is not a good idea to utilize HDFS to store the data since HDFS stores files to a block that is often set to 128 MB (by default) or 256 MB, resulting in storage waste.

2.4.2. Apache Spark

Apache Spark is an open source technology developed by the AMP science lab at UC Berkley (Ratra and Gulia, 2019). Ratra and Gulia (2019) define Apache Spark as a framework with stream processing abilities. Furthermore, Apache Spark according to Ratra and Gulia (2019) is built on planned exploitation of number of Hadoop's MapReduce engine's principles. In addition, according to Ratra and Gulia (2019), Spark focuses entirely on accelerating process and instruction execution workloads by providing full in-memory computation and processing enhancement. Hence, the speed of spark is 100 times faster than Hadoop's. According to Ratra and Gulia (2019), Resilient Distributed Datasets (RDDs) is a model used by Apache Spark.

2.4.3. MongoDB

MongoDB is a JSON-based database that was first released in 2009. It is developed in C++ and is based on JSON documents (Ratra and Gulia, 2019). According to Ratra and Gulia (2019), the MongoDB database is used to store data that does not have a set schema. Ratra and Gulia (2019) state that MongoDB does not have a predefined format like tables in relational databases, but rather data is stored in BSON form documents. BSON objects are JSON-like objects that have been binary encoded (Ratra and Gulia, 2019). MongoDB is a NoSQL database that is written in C++. According to Ratra and Gulia (2019), MongoDB is designed specifically for information storage and retrieval.

2.4.4. HBase

HBase is a column-oriented distributed database implemented on top of the Hadoop file system (Ibtisum, 2020). Furthermore (Ibtisum, 2020) reports that Hbase is an open-source project and is horizontally scalable. Additionally (Ibtisum, 2020) states that HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data. It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop File System (Ibtisum, 2020). Lastly, (Ibtisum, 2020) suggest that users can store data in HDFS either directly or indirectly through HBase, using HBase, the data consumer accesses the data in HDFS at random.

Table 1: Comparison Between Hadoop, Spark and MongoDB

Criteria	Hadoop	Spark	MongoDB
Processing Model	Batches	Mini batches	Single document
Latency	High latency	Very low latency	Low
Fault tolerance	Uses replication	Uses RDD	Replica set
Supported languages	Java	Scala, Java, Python	C++
Batch framework	HDFS	Core Spark API	BSON
Security	High secure	Less secure	Secure
Advantages	Storage and processing speed, low cost, secure	Scalable, high speed, powerful caching	Expandable
Weakness	Name Node' is the single point of failure	Expensive, less no. of algorithm, small files problems	Fault tolerance issue
Hardware cost	Cost is more	Cost is more	Cost effective because it is a single product

Source: https://doi.org/10.35940/ijeat.b2360.129219

2.5. Data Analysis Technologies

2.5.1. Python

Python, according to Ginde *et al.* (2017), is a general-purpose programming language designed for readability. Furthermore, according to Ginde *et al.* (2017), Python's greatest strength is its diverse set of libraries for doing various tasks such as data extraction, text processing, and machine learning methods. Furthermore, Python has a number of built-in tools and frameworks for data analysis, including Pandas, Scipy, Scikit-learn, and Numpy. This makes data pre-processing and analysis simple, but scalability is a big concern (Ginde *et al.*, 2017). In addition to Python's scalability limitations, Ginde *et al.* (2017) claims that Python's performance in big batch processing does not make it a desirable tool. According to Ginde *et al.* (2017), this is because most analysis operations and calculations take occur in-memory, resulting in much worse performance when compared to distributed systems as data is massively scaled.

2.5.2. R Language

R is mainly designed for statical analyses. Its primary goal is to improve mathematical modeling and assist in the development of analytical prototypes (Ginde *et al.*, 2017). Furthermore, Ginde *et al.* (2017) claims that R offers a less complicated way to access various machine learning and mathematics functions. Furthermore, Ginde *et al.* (2017) claims that one downside of R is its slowness in analysis and data processing. According to Ginde *et al.* (2017), R is slower in performance than its rival Python.

Table 2: Suitable Technologies to Tackle the Challenges

Challenge	Technology	How it Solves the Challenge
Data Availability and Accessibility	Mozenda, Octoparse and RapidMiner	Allows users to access and convert unstructured web data into structured
Inaccurate and Incomplete Data	Apache Spark Ecosystem, language and Python	Machine Learning algorithms
Lack of Skill Set	Apache Spark Ecosystem	Developers can use programming languages they know
Scalability	Apache Spark Ecosystem and Hadoop Ecosystem	Can add and remove different components as desired
Storage	HDFS, MongoDB	No-SQL
Privacy	Hadoop Ecosystem	Proxy authentication
Polyglot Persistence	Hadoop Ecosystem	Programmers do not need to care for achieving distributed processing

3. Research Methodology

3.1. Exploratory Research

The researcher chose exploratory research, because this type of approach allows the researcher to be creative in order to gain the most amount of insight on a subject. Additionally, the researcher wants to use an outside audience for this research, this will give the researcher a good opportunity to know what works and what not a productive method to use is. Additionally, it makes for a clearer grasp about what the aims of a research team should be during the course of an investigation. Furthermore, anyone performing research from outside sources will benefit from having this material on hand.

Additionally, according to QuestionPro (2018), exploratory research is a method of research that is used to study a problem that is not well defined. The researcher does exploratory research in order to better understand a problem that is already existing, even though the research will not provide conclusive results. The researcher begins with a broad concept and uses this study as a tool to discover challenges, and this study can serve as the foundation for future study. Dudovskiy (2016), states that a key consideration here is that the researcher should willing to shift course in response to the discovery of fresh data or insight. Figure 4 below depicts the general flow of the exploratory research method and the three steps it consists of:

Figure 4: Exploratory Research Approach

The first process includes identify the problem and this was where the researcher identified the subject of the challenges faced when implementing scalable Big data technologies. The researcher addressed this subject by employing a variety of approaches to answering the questions. The second step was when the researcher conducted preliminary study and discovered that the problem had not been adequately handled, the researcher developed research objectives rather than a hypothesis. The final procedure entails additional research; once the data has been gathered, the researcher will continue the study by descriptive investigations. Quantitative approaches are used to investigate the subject further and determine whether or not the information is correct.

Furthermore, the exploratory research method has its advantages and disadvantages. The advantages of exploratory research are as follows: It provides the researcher with a great deal of flexibility since it allows the researcher to adjust to changes as the research advances, typically, the cost is minimal, lays the groundwork for future studies and finally, it assists the researcher in determining whether a topic is worthwhile to pursue, and it can finally assist the researchers in exploring other probable explanations for the problem that may be further examined in detail (QuestionPro, 2018).

Additionally, the disadvantages of exploratory research are a few and they are as follows: It is inconclusive, despite the fact that it can point researchers in the proper direction for determining the answers, as exploratory research typically uses small sample sizes, it cannot be representative of the overall population, when conducting this form of study, the data acquired may be old and out of date if indeed the secondary research approach is used and ultimately, the most serious criticism leveled towards exploratory research is that it relies on qualitative data, which might be biased (QuestionPro, 2018).

3.2. Population

Population is defined by Beaudry and Miller (2016) as a group of people about whom a researcher wishes to learn more. Given that definition the researcher defined a target population, these are a group of specific individuals within a larger group that are best suited to give proper primary data. The target group includes industry practitioners in the industry of Computer Science namely data scientist.

3.3. Sampling

Sampling can be explaining as a principle used to pick a specific group of individuals within a wider population (Dudovskiy, 2016). Additionally, Beaudry and Miller (2016) explain sampling as a subset of a population that is meant to represent this population. The sample is made up of 5 industry practitioners both male and female. The sample size of the questionnaire is restricted to 5 because of the lack of Big data experts and Data scientists in Namibia. The sample size for the online survey is made up of 15 to 20 individuals working in the Information Technology industry. The survey has a larger sample size, because it focuses on different Information Technology professionals that should now a bit about the different Big data technologies.

Furthermore, the researcher made use of the non-probability sampling method. The reason for selecting this sampling method is because the researcher utilized the convenience sampling approach in particular. This choice is further supported by the fact the researcher works in the Information Technology.

3.4. Methods and Tools for Data Processing

3.4.1. Methods to Collect Data

To gather the information needed for the investigation, the researcher used both the primary and secondary research methods. The primary research methods used by the researcher was a questionnaire and an online

survey. The questionnaire was distributed via email to all the participants, all participants that which were selected are experts in the field of data analyses. The advantage of this is that an expert can supply you with significant insights that a broad public source cannot. Furthermore, the online survey was distributed via a website called survey monkey, this was to give the respondents an easily accessible tool for quick responses.

Additionally, the researcher used the secondary research method to improve understanding regarding the topic. The secondary research method used was literature research, online research and case study and tutorial research. The availability of a vast amount of material in libraries, Internet sources, and even commercial databases is an advantage of this research. A large amount of data is easily available on the internet, and the researcher obtained it as needed. The researcher analyzed the case in relation to all the variables present in earlier cases against this case, which in this case is the challenges faced when implementing scalable Big data technologies.

3.4.2. Data Analysis

The researcher used Python's jupyter notebook to do data analysis. This software allowed the researcher to create different visualizations of the data collected via the primary research method which was the questionnaires. Additionally, Microsoft excel was used to do various task such as preparing and managing the data collected via the questionnaires in order to do different visualizations, as well as creating the datasets from the data collected. These findings were then presented in graphs and tables to support the research.

3.5. Delineations and Limitations

This research has a few factors limiting its effectiveness and the biggest one is the lack of previous studies addressing the challenges faced when implementing scalable Big data technologies. A literature review forms the bases of any research, because it helps to identify the scope of works that have been done so far in a specific research area. The sample size also plays an important role when looking at the effectiveness of the research being conducted and the smaller the sample size the less representative it is to the general population which is the case for this research. Experience of the researcher in implementing the data collection methods will also affect the quality of data being collected, because the researcher has little experience in primary data collection. Lastly, given all these limitations the research will be well conducted and will provide effective answers to the research objectives. Finally, the researcher would like to make one suggestion to all interested parties in regard to Big data. This suggestion is that any future research done in relation to scalability in Big data use more of a constructive research approach, because this was not possible in this research because of the lack of skills of the researcher.

3.6. Risk/Feasibility Analysis and Ethical Considerations

The researcher did strictly abide to research ethics. During the course of the study, the researcher did:

- Maintain a high level of impartiality throughout the research during conversations and analysis.
- Allow voluntary participation of respondents and furthermore, participants have the right to decline participating in the study at any point if they so desire.
- Inform respondents sufficiently about participating in the study in order to allow individuals to accept the implications if any they arise.
- Use APA referencing to acknowledge the contributions of other authors to this study.
- Respect the privacy of participants and additionally the anonymity of respondents will be always respected.

4. Results and Analysis

4.1. Introduction

The researcher used data visualization to analyze the data. The methods used to collect data were questionnaires and surveys, therefore the data collected was quantitative. The reason for visualizing the data is the fact that it is a very small data set, because of its target population Big data specialist. The technologies used to perform the data analysis are Excel workbooks and a jupyter notebook with its accompanying technologies like Plotly and Pandas.

4.2. The Datasets

Figure 5 shows the data sets that were created from the questionnaire data. These data sets were created manually for each question in the questionnaire

4.3. Technologies

The bar graph (Figure 6) shows the technologies currently being used by the 5 industry practitioners. The technologies mostly used are Python, R Language, Power Bi and Excel with at least 2 participants naming them as the technologies they use currently. The rest of the technologies Such as the very Hadoop and Spark are only use by an individual participant.

4.4. Type of Storage

Figure 7 shows that out of the 5 participants that answered the questionnaire 80% make use of scalable approach when it comes to storing data. On the other hand, none of the participants make use of only storing data "In-house". The Hybrid and Cloud approach allows practitioners the ability to only use the amount of

	Bottlenecks	Participants
1	High CPU usage	5
2	Low memory	4
3	High disk I/O	1
4	High disk usage	1

	Drawback	Participants
1	Insufficient Storage	3
2	Volume (Processing s…	2
3	Finance	1
4	Lack of skilled profes…	1
5	Other	2

Figure 5: Performance Bottlenecks and Drawbacks Faced by Practitioners

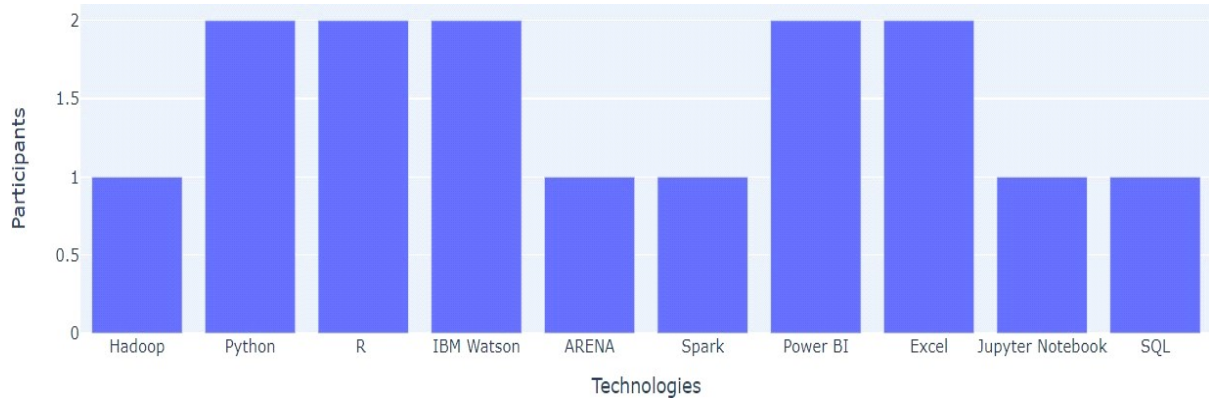

Figure 6: Technologies Used by Practitioners

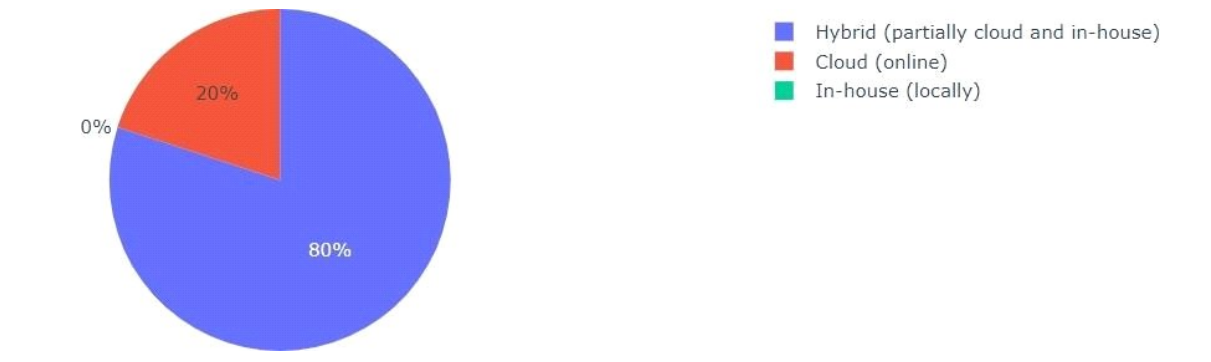

Figure 7: Type of Storage Used by Practitioners

storage they require based on the amount of data they are using for certain project and if need be, they can acquire additional storage if the need arises. Whereas the local approach gives a fix amount of storage size and if the need arises to increase the storage additional storage hardware is needed.

4.5. Drawbacks

The above line graph (Figure 8) shows that 3 out of the 5 participants link their technology's drawbacks to the lack of storage it possesses. Whereas 1 participant stated that the drawback of their systems is financial, which means the cost to implement and maintain these technologies are not feasible to the implementor. Additionally, 1 out of the 5 participants claimed that the lack of skilled professionals is a key concern, this implies that the individuals task with implementing these technologies do not have the required knowledge. Two Participants chose other with regards to the drawbacks linked to their technologies; one response given for 'other' was as follows:

4.6. Performance Bottlenecks

Figure 10 is a pie chart that shows the common performance bottlenecks experienced by the 5 participants. The participants could select more than one answer in regard to collecting this data. As the pie charts shows

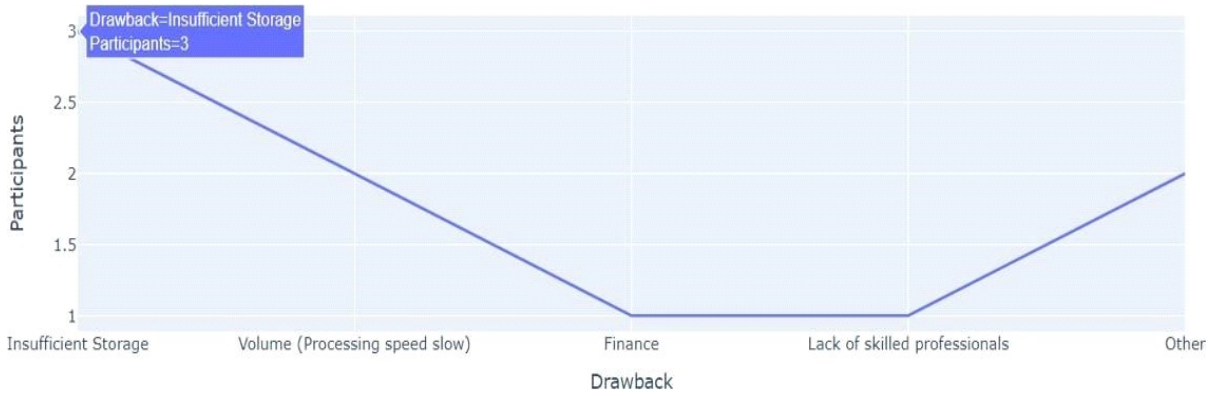

Figure 8: Drawbacks Faced by Practitioners

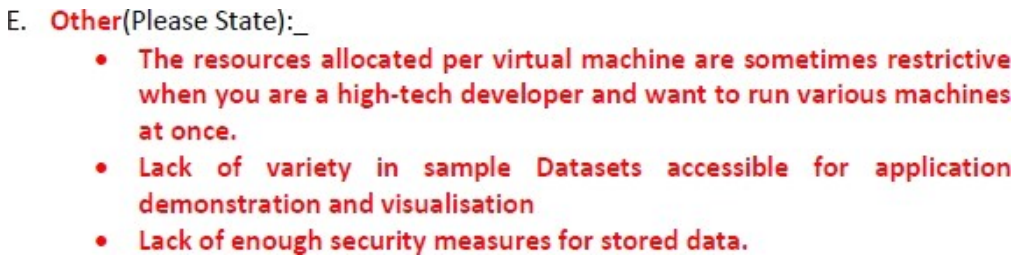

Figure 9: Other Challenges Faced by Practitioners

Figure 10: Performance Bottlenecks and Faced by Practitioners

45.5% of the participants chose High CPU usage as the most common performance bottleneck they experience, and the least experience performance bottlenecks are High disk usage and High disk I/O.

4.7. Skilled Profesionals

Did the parties that develop the platform receive training before development started?

Figure 11 shows the responses the participants gave regarding the question of whether the parties involved in implementing the technologies received any sort of training before the undertaking of the project. 60% percent of the respondents chose yes which equates to 3 out of the 5 participants. This question also led to the researcher exploring the knowledge of IT professionals around the subject of Big data tools.

4.8. Scaling

Figure 12 shows what aspect of their technology the industry practitioners would like to improve. For this question participants could select both scaling approaches. 3 out the 5 participants chose to scale vertically which means that they would like to upgrade the CPUs they are currently using. They could also be considering vertically scale the memory, storage, or network speed. Additionally, vertical scaling may also be referring to replacing a server entirely or moving a server's workload to an upgraded one if they are using one. Whereas a total 4 out 5 participants chose to scale horizontally this refers to adding additional nodes or machines to the existing infrastructure. 2 participants chose both approaches.

4.9. Knowledge Regarding Big Data Tools

This online survey was provided to different types of Information Technology specialist like software developers, web developers, system analyst and RPA developers. The visualizations below represent the stats of their responses based on each question of the survey. The participants were asked various question relating to the skills and technologies required to perform Big data analytics.

Figure 11: Was Training Required?

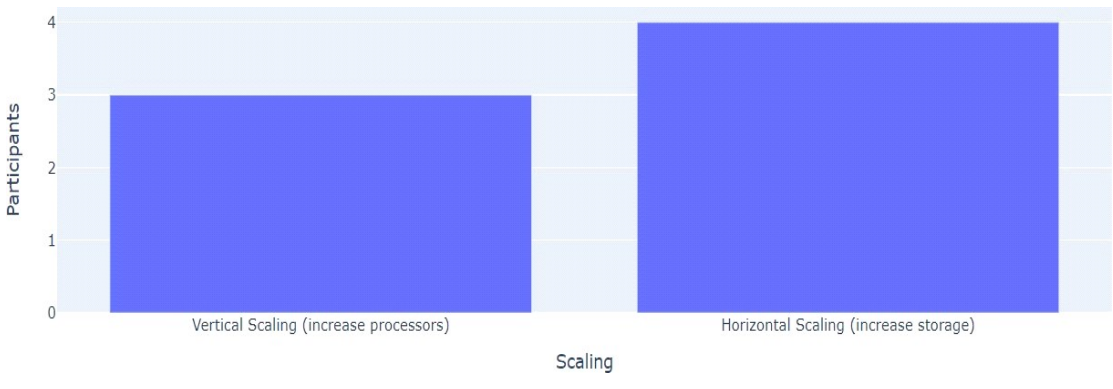

Figure 12: Desired Scaling Approach?

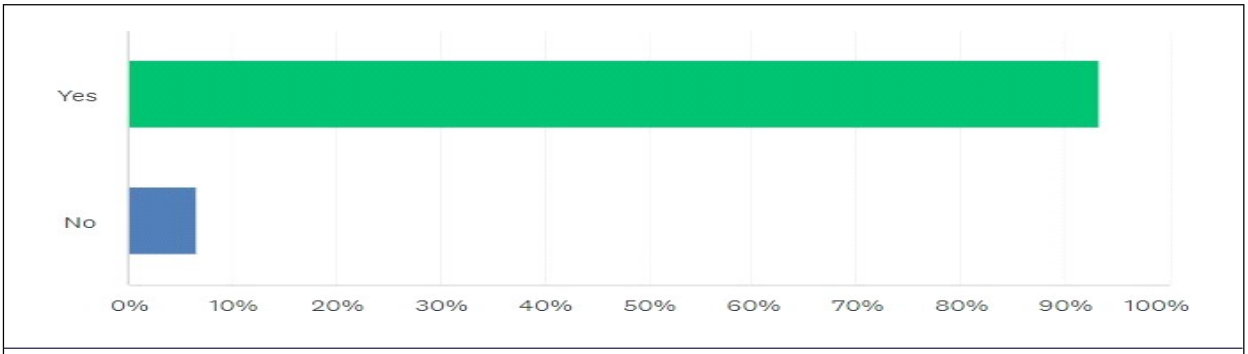

Figure 13: Are You Familiar with Big Data?

Q1. Are you familiar with Big data?

Figure 13 shows the responses of 15 IT professionals that were asked whether they are familiar with Big data. The responses show that majority of IT professionals are aware of what Big data is because 93.33% of respondents said they are familiar with Big data.

Q6. Do you have any knowledge of performing general (basic) statistical and quantitative analysis?

Figure 14 shows that 73.33% of respondents have knowledge of performing basic statical analyses. This type of analyses is the groundwork any sophisticated statical analyses and anyone trying to perform Big data analyses requires this knowledge.

Q3. Which of these technologies are you familiar with? (Choose one or more)

Figure 15 Shows that most respondents are familiar with SQL and this no surprise because this is the most widely used technology in regard to storing data and query for specific information. 73.33% of the respondents

Figure 14: Knowledge of Performing Basic Statistical and Quantitative Analysis

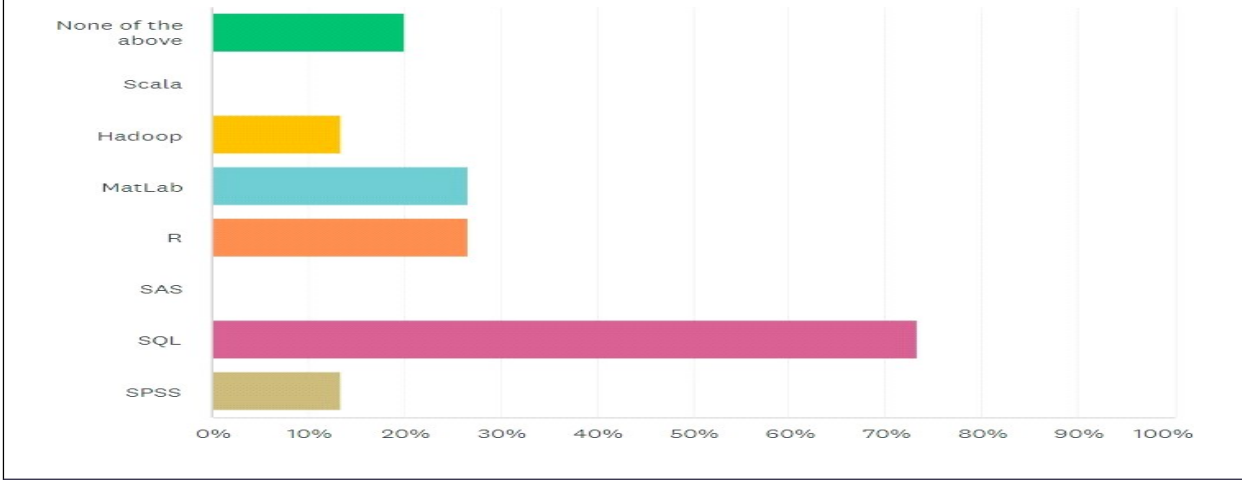

Figure 15: Which of the Following Data Mining Technologies Have You Used Before?

stated that they know SQL but more positively 26.63% of participants are knowledgeable with R and Matlab which are very important technologies in Big data analytics.

Q4. Which of the following data mining technologies have you used before?

Figure 16 shows that not many IT professionals have actually used some of the listed Big data mining technologies. With 80% of the respondents stating that they have not used any of the listed Big data technologies.

Q5. Do you have any knowledge in NoSQL databases? If yes, select

Based on the results presented by Figure 17 most the participants are familiar with the different types of NoSQL databases. These bouts well fore organization that want to take the no SQL rout when deploying a database for Big data Storage. A total of 9 out the 15 participants are familiar with at least 1 type of No SQL data base with 7 out of the 9 choosing MongoDB.

Q8. Do you have any knowledge with the Machine Learning (ML) algorithm building process?

Figure 18 shows that 68% of the participants have not used any type of ML algorithms. This shows that not many IT professionals use ML algorithms to do data analysis. Of the 3 respondents that said yes 1 indicated

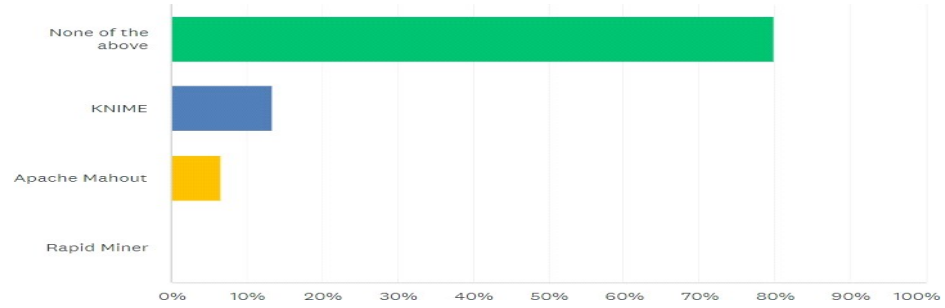

Figure 16: Knowledge of Big Data Mining Tools

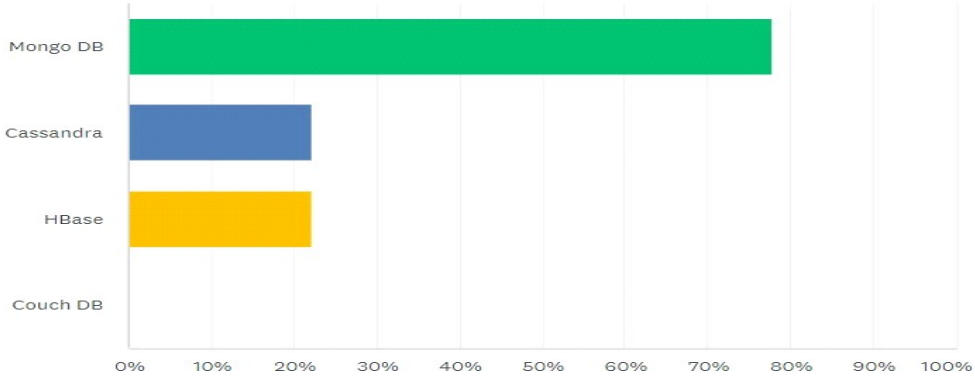

Figure 17: Knowledge of NoSQL Databases

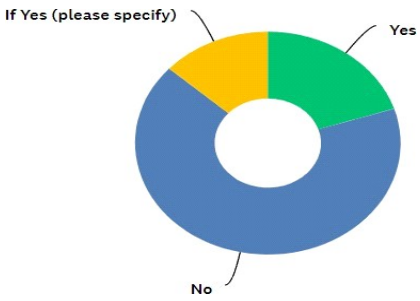

Figure 18: Knowledge of Machine Learning Algorithms

that he/she used the k nearest neighbour algorithm. This algorithm is very useful when dealing with incomplete data one of the key challenges Identified in this research.

5. Conclusion and Recommendations

5.1. Conclusions Based on the Findings

To achieve the study's objectives, an issue was identified: the challenges faced when dealing with Big data prohibits industry practitioners from using scalable Big data technologies. A detailed literature analysis was done to offer context for the issues faced by industry practitioners when dealing with Big data, as well as the solutions available to address these challenges. The research design featured the following essential components in order to achieve the study's objectives: Industry practitioners were given self-administered questionnaires, while participants in the Information Technology industry were given an online survey.

With regard to the first aim, the results of the study identified a variety of different challenges, namely, Privacy, Scalability, Incomplete data, Data availability, Polyglot persistence and a lack of required skills from industry practitioners. All these challenges arise, because of the characteristics of Big data which include volume, variety and velocity. These challenges than result in practitioners opting for less scalable and simpler technologies, resulting in many industries not really reaping the full rewards of Big data analysis.

In accordance with the second aim of study, the study identified and reviewed two Big data processing ecosystems, the Hadoop ecosystem and the Apache Spark ecosystem. The outcome of the results were the advantages and disadvantages of the two ecosystems. Additionally, the results also showed what these two systems are comprise of and how each component in the ecosystem is design to fulfil a specific purpose whether processing, storing or analyzing the data.

As to the third aim, the literature review grouped the technologies into three categories, Data acquisition technologies, Data storage technologies and Data analysis technologies. The review further explains how these technologies designed in order to deal with a multitude of challenges.

For the fourth objective, the results revealed that very few industry practitioners use any the technologies reviewed in this study. Only 3 technologies (R, Haddop and Spark) reviewed in the study are being utilized by the participants, with one participant making use of two of those technologies. Additionally, the results revealed that 45% of the participants named High CPU usage as the most common performance bottleneck they experience. With High disk usage and High disk I/O chosen by 9.09% of the participants. Furthermore, the results showed that 3 out of the 5 participants chose insufficient storage as the disadvantage of the system they are currently using.

The final research objective had to do with how much knowledge individuals in the information Technology industry have about the different Big data technologies. Most IT related jobs required the use of different technologies to do data analytics whether it be web developers tracking how much traffic a website receives or a security analyst monitoring log incident on a network this is all data that could be used to improve performance. The findings from the survey shows that 14 out of the 15 individuals are familiar with the term big which is a very positive indication. However, as the questions became more specific about the knowledge regarding the tools required to perform Big data analysis the responses became less positive. One of the aspects which the participants where not really knowledgeable ware the data acquisition tool such as RapidMiner and KNIME. Out of the 15 participants 80% were not familiar with any of the technologies listed. Additionally, having the ability to perform basic statical analysis is crucial, because it forms the foundation of performing more complex analysis such as machine learning. In regard to being able to perform basic statical analysis 73.33% of the participants indicated that they are able to do it.

The overall results of the present research provide convincing evidence of how the available technologies can deal with the different challenges that arise when dealing with Big data. In the researched involving industry practitioners, the lack of scalable Big data technologies implemented emerged undoubtedly. The best and most effective way of dealing with these challenges is making sure to choose the correct technologies for

the specific Big data project. The Hadoop ecosystem and Apache Spark ecosystem are undoubtedly important tools for effective Big data processing, because of their ability to integrate different technologies to make use of their individual strengths. Practitioners must therefore look beyond the simple quick fixes and research on technologies that will enable them to manage Big data in an effective manner.

Given these findings, it seemed reasonable to conclude by saying that there is a general need to adopt the new technologies available for Big data processing, because the age of Big data has arrived and is an essential part of any organization. If the results of the study are a true representation of the overall population's approaches with regards to Big data processing, then the need to implement scalable Big data technology is of paramount importance within the research area (Namibia). Failure to implement any form of scalable technology whether for Big data mining, storing or analysing will most certainly result in ineffective Big data processing.

I believe that this research, despite the small sample size, makes an important contribution to providing guidance on why industry practitioners should start implementing scalable Big data technologies in Namibia.

5.2. Recommendations

If industry practitioners do not put effort into reviewing and investigating old and new technologies available for Big data management to suit their requirements, a lot of industries and organization will not realise the full potential of Big data. Effective data analysis is a necessary element in any organization that strives to improve different aspects such as customer service, financial projections and resource management. The findings of the research indicated the need for individuals to receive exposure to Big data very early on, because to most it's a term they are familiar with but do not understand.

In light of the foregoing, the researcher desires to suggest some recommendations that, if followed, could result in some positive modifications to Namibia's existing approaches when it comes to Big data management. Industry practitioners and important stakeholders in Big data should:

- Replace current, less scalable techniques to Big data processing in favor of new approaches based on a wide range of scalable technologies that will enable businesses and industries to realize Big data's full potential. This entails launching aggressive education and awareness campaigns in order to maximize the efficiency of Big data processing.
- Take more practical approaches when it comes to addressing different aspects of Big data. This implies that all stakeholders should employ a more technical and practical approach when conducting research in this field. This will allow future researchers the groundwork to base their research on.
- Encourage each other to implement Big data processing ecosystem like Hadoop and Spark. This means giving organizations that employ Big data ecosystems incentives such as increase budgets to help extend the knowledge around Big data.

5.3. Indications for Further Research

The study has identified various researchable areas that individuals working to create scalable Big data technology should look into further. The study's findings demonstrated how existing solutions are geared to address the majority of the issues that arise when dealing with Big data. However, this research does not provide solutions to one of the problems identified, privacy. Hence, there is an urgent need to address the problem of privacy when it comes to Big data. Privacy is a problem that cannot necessarily be solved by the introduction of new a technology. This study can be used as a starting point for future in-depth investigations. More research is needed to determine the best approaches to deal with the issue of privacy.

References

Abu-Salih, B., Wongthongtham, P., Zhu, D., Chan, K. and Rudra, A. (2021). Chapter 2 Introduction to Big data Technology. https://arxiv.org/ftp/arxiv/papers/2104/2104.08062.pdf

Ajah, I. and Nweke, H. (2019). Big data and Business Analytics: Trends, Platforms, Success Factors and Applications. *Big Data and Cognitive Computing*, *3*(2), 32. https://doi.org/10.3390/bdcc3020032

Alabdullah, B., Beloff, N. and White, M. (2018). Rise of Big data-Issues and Challenges. Retrieved October 18, 2021, from https://core.ac.uk/download/pdf/159767357.pdf

Al-Sai, Z. A., Abdullah, R. and Husin, M. Heikal. (2019). Big data Impacts and Challenges: A Review. *IEEE Xplore,* April 1. https://doi.org/10.1109/JEEIT.2019.8717484

Andreea, J. (2021). *Scalability: Essential in Running Analytics and Big Data Projects*. [online] DATAVERSITY. Available at: https://www.dataversity.net/scalability-essential-in- running-analytics-and-big-data-projects/# [Accessed 26 Nov. 2021].

Beaudry, J. and Miller, L. (2016). Research Literacy: A Primer for Understanding and using Research. *Faculty and Staff Books*. https://digitalcommons.usm.maine.edu/facbooks/232/

DataFlair (2019). Top Advantages and Disadvantages of Hadoop 3. DataFlair, February 28. https://data-flair.training/blogs/advantages-and-disadvantages-of-hadoop/

Doty, D. (2021). *Data Is Gold For Publishers In A New Media World Order*. Forbes. https://www.forbes.com/sites/daviddoty/2021/07/29/data-is-gold-for-publishers-in-a-new-media-world-order/?sh=7f391a19bd2a

Dudovskiy, J. (2016). *The Ultimate Guide to Writing a Dissertation in Business Studies: A Step- by-Step Assistance.*

Shaikh, E., Mohiuddin, I., Alufaisan, A. and Nahvi, I. (2019, November). *Apache Spark: A Big data Processing Engine*. ResearchGate; unknown. https://www.researchgate.net/publication/339176824_Apache_Spark_A_Big_Data_ Processing_Engine

ELE Times Research Desk (2018). *Big Data: A Guide to Choosing the Right Big Data Analytics Tools*. ELE Times. https://www.eletimes.com/big-data-a-guide-to-choosing-the-right- big-data-analytics-tools

Espinosa, J. A., Kaisler, S., Armour, F. and Money, W. (2019). *Big data Redux: New Issues and Challenges Moving Forward,* January 8. Scholarspace.manoa.hawaii.edu. https://doi.org/10.24251/HICSS.2019.131

Gao, J., Koronios, A. and Selle, S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. Semantic Scholar. https://www.semanticscholar.org/paper/Towards-A-Process-View-on-Critical-Success-Factors-Gao-Koronios/247bfe6fa3365d74bd98c2c460785d62c3d7561d

Ginde, G., Saha, S., Aedula, R. and Mathur, A. (2017). *Big Data Analytics,* 1st Edition [Review of Big data Analytics].

Hariri, R. H., Fredericks, E.M. and Bowers, K.M. (2019). Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0206-3

Ibtisum, S. (2020). A Comparative Study on Different Big Data Tools. Retrieved October 18, 2021, from https://library.ndsu.edu/ir/bitstream/handle/10365/31657/A%20Comparative%20St udy%20on%20Different%20Big%20Data%20Tools.pdf?sequence=1&isAllowed=y

Jayashree, K. and Abirami, R., (2018). *Big Data Technologies and Management*. Https://Services.igi-Global.com/Resolvedoi/Resolve.aspx?Doi=10.4018/978-1-5225-5829-3.Ch009. https://www.igi-global.com/gateway/chapter/205406

Ratra and Gulia (2019). Big Data Tools and Techniques: A Roadmap for Predictive Analytics. *International Journal of Engineering and Advanced Technology*, 9(2), 4986-4992. https://doi.org/10.35940/ijeat.b2360.129219

Michael, K. and Miller, K.W. (2013). Big data: New Opportunities and New Challenges [Guest editors' introduction]. *Computer*, *46*(6), 22-24. https://doi.org/10.1109/mc.2013.196

Muniswamaiah, M., Agerwala, T. and Tappert, C. (2019). Big Data in Cloud Computing Review and Opportunities. *International Journal of Computer Science and Information Technology*, *11*(4), 43-57. https://doi.org/10.5121/ijcsit.2019.11404

Junghanns, M., Neumann, M. and Rahm, E. (2017). Management and Analysis of Big Graph Data: Current Systems and Open Challenges, in Handbook of Big data Technologies, pp. 457-505.

Koo, J., Kang, G. and Kim, Y.G. (2020). Security and Privacy in Big data Life Cycle: A Survey and Open Challenges. *Sustainability*, *12*(24), 10571. https://doi.org/10.3390/su122410571

Ku, L. (2021). The Impact of Big Data in Business. https://www.plugandplaytechcenter.com/resources/impact-big-data-business/

Rouse, M. (2019). *What is Big Data? - Definition from WhatIs.com*. SearchDataManagement. https://searchdatamanagement.techtarget.com/definition/big-data

Saleh, S.H., Ismail, R., Ibrahim, Z. and Hussin, N. (2018). Issues, Challenges and Solutions of Big data in Information Management: An Overview. *International Journal of Academic Research in Business and Social Sciences*, 8(12), 1382-1393.

Talia, D. (2019). A View of Programming Scalable Data Analysis: From Clouds to Exascale. *Journal of Cloud Computing*, 8(1). https://doi.org/10.1186/s13677-019-0127-x QuestionPro. (2018, December 3). *Exploratory Research: Definition, Methods, Types and Examples | QuestionPro*. QuestionPro. https://www.questionpro.com/blog/exploratory-research/

Vargas-Solar, G., Zechinelli-Martini, J.L. and Espinosa-Oviedo, J.A. (2017). Big Data Management: What to Keep from the Past to Face Future Challenges? *Data Science and Engineering*, *2*(4), 328–345. https://doi.org/10.1007/s41019-017-0043-3

White, T. (2015). Hadoop: The Definitive Guide; Storage and Analysis at Internet Scale. Beijing O'reilly Media.

Apendix A

Questionnaire

Industry Practitioners

Name:_____**(Not required)**

Specialisation: Computer Science & Cybersecurity _____

1. What type of Big data mining platform or technologies are you currently using to obtain and store data?

   Hadoop, iCloud, OneDrive, Google Drive, GitHub, and IBM Watson

   _____

2. What setup are you using to store your data? (Choose one)

   A. In-house (locally) B. Cloud (online) C. Hybrid (partially cloud and in-house)

3. What are the drawbacks of the platforms or technologies currently being used?

   The cost for some of the platforms is not competitive and the lack of other functionalities and capabilities across different platforms.

4. What is the main cause of the drawback mentioned (mentioned above)?

   A. Insufficient Storage B. Volume (Processing speed slow)

   C. Finance D. Lack of skilled professionals E. Other(Please State):

   - The resources allocated per virtual machine are sometimes restrictive when you are a high- tech developer and want to run various machines at once.
   - Lack of variety in sample Datasets accessible for application demonstration and visualisation
   - Lack of enough security measures for stored data.

5. Do you experience any performance bottlenecks with your current setup?

   A. High CPU usage B. Low memory

   C. High disk I/O D. High disk usage

6. Have you considered scaling your Big data platform? If yes, which scaling option did you opt for?

   A. Vertical Scaling (increase processors) B. Horizontal Scaling (increase storage)

7. Did the parties that develop the platform receive training before development started?

   Such information may not necessary be availed to clients, the assumption is probably they did.

   Some of the cloud providers do not have profiles of who is behind the scenes.

   Of course, for my personally developed platforms, I do offer training for the team I work with

8. How much did building your platform initially cost you(estimate)?

   That varies across the different platforms in use, on average it costed me US$150 to initially setup one.

9. What type of data is being captured or used by the Big data platform you are currently using?

   Text, Audio and Video files

10. Are you experiencing any problems other than the ones mentioned above with your Big data platform in regard to scalability? If yes mention

    Yes, not enough data analytics tools are available to provide visualisation. To the huge datasets I am building every day.

11. What technologies are you using to perform your Big data analytics?

    Mostly scripting using Python , R and ARENA simulation platform

12. How do you obtain your datasets?

    Personally created, and in most cases from Datasets repositories on GitHub and other similar platforms if its relevant to my research and application needs

13. What is the data used for? (If other state)

    A. Reports B. Decision Making

    C. General Analyses D. Other:__Teaching and Research purposes

Cite this article as: Heinrich Gladwen Dankie Geiseb and Nashandi Ndinelago (2023). Investigation into The Challenges of Implementing Scalable Big Data Technologies and Techniques. *International Journal of Data Science and Big Data Analytics*, 3(1), 1-24. doi: 10.51483/IJDSBDA.3.1.2023.1-24.