IJEMD-CSAI, 2 (1) (2023), 1 – 19 https://doi.org/10.54938/ijemdcsai.2023.02.1.134

International Journal of Emerging Multidisciplinaries:
Computer Science and Artificial Intelligence

*Research Paper*
*Journal Homepage: www.ojs.ijemd.com*
*ISSN (print): 2791-0164 ISSN (online): 2957-5036*

# Big Data Management and Analytics as a Cloud Service

**Shahad Alghamdi[1*], Sarah Alghamdi[1], Yasmeen Almansour[1], Alfia Badiwalla[1]**

*[1]Prince Mohammad Bin Fahd University, Al-Khobar, Saudi Arabia*
**Corresponding author*

## Abstract

Traditional database management systems are inadequate for handling uncertain data due to design limitations. Various methods address database uncertainty, often oversimplifying and restricting representable uncertainties. Rapid data growth from technologies like IoT, multimedia, and social media has led to massive diverse data, known as big data, categorized into structured, semi-structured, and unstructured types. Coping with big data's challenges is encapsulated in the "Vs Model" with volume, velocity, and variety dimensions. It aims to gather variable information for causal understanding and informed decisions. Big data offers competitive advantages: smart decisions (69%), operational control (54%), customer insights (52%), and cost cuts (47%). This research studies big data management and cloud computing, where data and apps are accessed via the internet, saving resources globally. Cloud's simplicity and power facilitate tasks and storage. Risks include bankruptcy or breaches, yet cloud remains valuable for preserving big data. Vital factors in big data's value creation involve converting IT costs to assets and connecting big data outcomes to performance. This abstract summarizes exploration of traditional limitations, big data's essence, cloud's potential, and value creation nuances.

***Keywords***: Big data; Data Analytics; Data Processing; IoT

## 1. Introduction

Conventional database management systems are ill-suited for handling uncertain data storage, manipulation, and retrieval due to their design limitations. Consequently, various methods have been introduced to address database uncertainty, although many of these approaches tend to oversimplify and impose restrictions on the types of representable uncertainties. The rapid surge in data volume, owing to the rise of technologies like the Internet of Things, multimedia, and social media, has led to an overwhelming influx of data in diverse formats, collectively known as big data. Big data can be categorized into three types: structured, semi-structured, and unstructured data. Coping with the challenges of big data is often encapsulated in the "Vs Model" with three dimensions: volume, velocity, and variety. Volume pertains to the quantity of data generated and stored, velocity denotes the speed of data generation and processing, while variety encapsulates the array of structured, unstructured, and semi-structured data

***Email addresses***: 201600522@pmu.edu.sa (Shahad Alghamdi)

derived from multiple sources.

The idea is not to collect a huge volume of data but to collect variable information to know the relationship between cause and effect and to find answers to our questions. The information, graphs, and charts extracted from big data can show where you have been and where you will go in an orderly manner (Vighio et al., 2022). Big data provides a competitive advantage for enterprises, according to a study by BARC, big data assists in making smart decisions 69%, improved control of operational processes 54%, better understanding of customers habits and preferences 52%, and cost reductions 47% ("Big data analysis shown…", 2016).

The purpose of this research is to conduct a careful study of big data management and analytics as cloud computing. It means cloud computing in a simple way instead of storing personal files and using programs on your own device's hard drive. These things are found elsewhere and can be used over the internet, and their wide range of applications can be used to access vast areas and powerful processors that ordinary devices cannot handle. Cloud computing also provides an economical saving of resources not only for the personal user but for the world at large, saving the costs of purchasing large data storage units and devices with powerful processors. All that a computer user needs is simple, which is a high-speed internet connection, a screen, and input tools such as a keyboard and mouse, and thus the user becomes a way to take advantage of the electronic cloud. Also, by means of cloud computing, a poor-performing computer can manage complex processors inside the computer, and a mobile device can store massive amounts of pictures and videos. On the other hand, cloud computing risks such as the bankruptcy of the cloud operator or the collapse of the cloud due to a technical problem or hack, and this thing means the loss of huge data or the leakage of sensitive information, such as hospital information, banking information, or state agencies. Despite these negatives, the idea of cloud computing remains a valuable and important idea that preserves millions of big data.

## Big Data Applications

Big data has become a focal point of various projects done in societies. The condition surfaces because big data enables the creation, collection, sharing, preparing, and examining information, which augments business operations (Memon et al., 2017). For this reason, the technologies are utilized in different fields, given the significant impact they have on them.

Data continues to grow every time, which may compromise companies' capacity to store it. Cloud computing is a huge innovation that enables storage and retrieval of massive data. Cloud service is a model that enables network access such as an on-demand network to share a pool of many resources that contain application, servers, services and storage. it could be easily managed and provisioned with minimum effort and without complexity. In essence, cloud services have essential attributes that augment the storage and retrieval of data.

## Third Eye-Data

Big data accrue numerous benefits to organizations, which is a fundamental discovery in today's business generation. It assists in forecasting customers' spending patterns, noticing fraud, and abuse. Analysts can comprehend data that suit specific business requirements regardless of data volume and complexity . Data is then visualized and presented comprehensively, and internetbased firms like Google, Facebook, Twitter, and eBay use the approach (Memon et al., 2017). Data visualization organizes relevant information coherently, which promotes business processes.

### Banking

Financial institutions receive massive customer data, which necessitates the surveillance and protection of the information. Serious privacy issues may arise if inconsistencies of customer data are detected. Big data analytics assists in unearthing confidential information (Memon et al., 2017). In essence, the technology helps to organize client information in banks.

### Agriculture

Big data analytics plays a pivotal role in the control of vital agricultural elements. It helps to determine how plants would respond to certain conditions. Data on aspects, such as temperature, the quantity of water, soil, outputs, and plant sequence is analyzed (Memon et al., 2017). It offers opportunities for success in agriculture outcomes.

### Economics

Employing big data analytics in economics is crucial since it synthesizes sophisticated systems in businesses. For instance, Hadoop helps organizations to operate under low budgets (Memon et al., 2017). Economic schemes endeavor to minimize costs while maximizing revenues. Big data facilitates achieving the condition.

Firms that embrace big data analytics are better-positioned in numerous aspects. For instance, they make sound decisions and unleash improved financial performance . Numerous organizations create instruments of cloud big data analytics. It propels their operations to be topnotch by experiencing enhanced access to big data, hence, “able to deploy, administer, manage and secure” (Ara & Ara, 2016, p. 5). Cloud service facilitates networking and intelligence schemes to determine potential risks before they surface (Ara & Ara, 2016). Therefore, applying big data alongside cloud services is vital for companies.

### Key Indicators for Evaluating Big Data

Big data technologies have particular metrics regarded as indicators, for assessing their advancement. Most importantly, big data analytics are transforming persistently and fast, which compels them to utilize cloud since it has phased out the legacy systems. The indicators uncover the effect of big data on growth and development in information-based communities. They are pertinent in information data-driven societies since they reveal the progress made.

The features of cloud service contribute to big data analytics. Firstly, cloud computing allows flexibility since it is easily adjustable to meet the changing demands. Secondly, it involves limited service whereby charges apply to it. Thirdly, it is diversified since it serves numerous distinguished clients. Finally, cloud service is accessible as long as a network connection is available, and devices, such as mobile phones and laptops Therefore, big data technologies have uninterrupted access to data and storage, which facilitates analytical processes.

Various indicators provide insights into the use of big data among firms. Firstly, the number of big data and data-driven firms is considered. The number of firms that deploy big data in their operations reflects the prevalence of big data analytics. Correspondingly, the emergence of big data companies creates opportunities for the spread of big data technologies. also, the volume of investments in data-driven organizations is a primary indicator of big data. The expense of institutions on big data investments shows the continuous expansion in the use of big data technologies. Moreover, the number of data-driven research stations indicate the demand for big data technologies. When the number of research centers is high, the demand is high and vice versa. Fourthly, the number of big data-related patents insinuate the persistence of investment in big data analytics. Patents protect innovation from preserving originality. When big data-related patents are many, it shows that people are incentivized to invest in the field, which indicates growth. Finally, the number of data-related STEM (science, technology, engineering, and mathematics) imply expansion (Hajirahimova & Aliyeva, 2016). The US government and other administrations worldwide are committed to spearheading investments in STEM since it has the potential to transform economies, especially in the elite world. Particularly, investments in data-related STEM signal advancement in big data. The indicators discussed above are relevant in assessing big data.

## Comparative Evaluation of Big Data Frameworks

| **Hadoop** | **Storm** | **Flink** |
|---|---|---|
| a storage and programming platform that enables users to parse and process complex data sets within seconds | freely available data processing and computation program | The improvement of Hadoop |
| utilizes numerous software | Has the ability to utilize multiple programming languages, real-time streaming analytics, and ensuring the continuity of tasks | It has the highest processing speed |
|  | Suitable for low budget users | Costly |
|  | Recommended for beginners | suitable for large scale use such as companies |
|  |  | more effective and reliable |

Technological advancements are rapidly evolving the world of information, with the current society generating trillions of data daily. The extensive informal data set makes it difficult to utilize the data to derive meaningful information successfully. This section conducts a comparative evaluation of the existing big data frameworks, including Hadoop from Microsoft, Storm, and Flink.

The Hadoop Distributed File System (HDFS) is a storage and programming platform that enables users to parse and process complex data sets within seconds. According to Ullah, the Hadoop big data framework utilizes numerous software, including MapReduce and NameNode, which breakdown giant sets of information into meaningful blocks and categorizes them into complete files. Apache Flink is an improvement of the HDFS and offers several programming primitives. As per Ullah, it offers more features than HDFS, making it suitable for large scale use. Contrarily, the Storm data management framework is a freely available data processing and computation program. One of the outstanding advantages of this framework includes its ability to utilize multiple programming languages, real-time streaming analytics, and spout and bolt, ensuring the continuity of tasks. Of the three, Flink has the highest processing speed, as is the most costly. While Hadoop also offers significant benefits to the user as Storm, the former is freely available. For a beginner, Storm is recommendable as it is not only free but offers a range of benefits. However, corporates in need of more effective and reliable data frameworks may consider Flink for its premium services.

Cloud computing is currently on the rise, with technology enabling the availability of big data. Several big data platforms exist in the market, including Microsoft's Hadoop, Flink, and Storm. Flink is slightly costly but offers more benefits to users, while Storm is the most effective for low budget users.

## Big Data Infrastructure

Contemporary innovations focus on developing systems that are controlled to perform tasks done by human beings. In this accord, machine learning is a phenomenal advancement in big data analytics where machines are manipulated to imitate human intelligence. Big data technologies have gained popularity because they provide data collected from numerous diversified sources. particularly, the rate of data generation is high, therefore it is a primary requirement to utilize technologies that can handle the massive information capacity to obtain relevant information. Besides, the data is sophisticated, which necessitates deploying infrastructure that improves "computation time and the necessary resources to extract valuable knowledge from the data" (Salvador, Ruiz, & Garcia-Rodriguez, 2017, p. 249). While the conventional methods focused on synthesizing data in clusters, the conceptualization of the graphic processing unit (GPU) enabled multiple processing of information. GPU morphed to generalpurpose systems referred to as General Purpose Graphic Processing Units (GPGPU) . Over time, pertinent big data analytics infrastructure emerges, and the existing ones advance. Therefore, they establish the machine learning systems.

## Machine Learning

Machine learning helps to draw meaningful information from a data set. It facilitates the development of schemes focused on obtaining knowledge from data and applies it to forecast future circumstances and trends. Machine learning is grouped into several categories.

## Classification Algorithms

The scheme draws information from sample data, enabling the creation of classification rules referred to as a model. It implies that data is utilized to develop categories.

## Clustering Algorithms

Facilitates the formation of groups whose members have common features without previous information provided. Clusters are formed, and the constituents are related.

### Recommendation Algorithms

It involves forecasting desirable patterns and recommending them to the users. The trends are observed, and conclusions are made about suitable models.

### Dimensionality Reduction Algorithms

It entails reducing variables present in data without interfering with the accuracy of the concerned model. Specific variables are eliminated to an extent where the functionality of the model remains unaltered.

Machine learning encompasses several libraries. They help implement specific algorithms in the previous categories. The systems used in machine learning libraries are discussed below. Firstly, Apache mahout implements numerous machine learning algorithms, including clustering and classification, by utilizing Apache Hadoop. Secondly, MLlib is a component of Apache Spark that performs machine learning algorithms, such as, regression, classification, dimensionality reduction and clustering, (Salvador, Ruiz, & Garcia-Rodriguez, 2017). Finally, FlinkML offers scalable algorithms, facilitating designing systems. It implies that the FlinkML library has a tool for developing suitable schemes.

Machine learning incorporates several components to make data meaningful. Hence, it synthesizes information, enabling the specific systems to adapt to heightening demands . Machine learning allows flexibility in handling the data to suit the prevailing condition.

Some platforms allow big data processing. big data entails sophisticated data gathering approaches such that it is difficult for the ordinary database schemes to synthesize it . Terabytes and petabytes are the typical volumes of the data used. Platforms such as MapReduce, Apache Hadoop, Apache Spark, Apache Storm, Apache Flink function with big data analytic systems to process the complex data (Salvador, Ruiz, & Garcia-Rodriguez, 2017). To conclude with, the infrastructure mentioned above are integrated with big data to improve the operationalization of the technologies.

### Big Data Computing and Clouds Trends and Future Direction

In today’s world, organizations generate large volumes of data from monitoring user activity, instrumented business processes, and sensors. Business organizations can gain a competitive advantage by managing these vast volumes of data. Notably, companies can use data mined from social network websites to understand consumer products and preferences. In this case, businesses use information from product evaluations, tweets, and data from social networking sites to predict customer wants, understand the needs of customers, and optimize the use of resources (Abou et al. 34). However, managing data in companies is particularly challenging because it requires expensive analytics solutions that mine both structured and unstructured data to help organizations gain insights from the data they have gathered from various sources. Big data is a new paradigm closely linked to cloud computing, where organizations only pay for the resources they need. Big data describes an enormous amount of digital data resulting from `numerous sources such as digital cameras, social networking sites, numerical modelling as well as sensors.

Clouds offer organizations an opportunity to use computing resources in a pay-as-you-g o fashion. In computing, clouds improve elasticity and availability, as well as cost reduction. Moreover, clouds ensure organizations do not spend money maintaining high-cost IT infrastructure, which they do not use most of the time. Therefore, clouds are an ideal platform for implementing scalable analytics. In this article, the aim is to explore big data computing and clouds. Moreover, the report will examine trends and direction in big data computing and clouds.

The relationship between cloud computing and big data results from the fact that people need storage technology, which meets the need for rapid data growth on the low-cost and highly reliable clouds. More specifically, the cloud of a high capacity storehouse while big data is the product that must be stored in it. As a result, it is impossible to create storehouses without a product stored in them. Typically, the traditional relational databases are incapable of processing multiple data sources, and thus, big data is essential in today’s world. Therefore, the leading research works currently ongoing in the fields of big data and cloud computing are how proper processing power and high capabilities for analysis can be obtained. Without analyzing data quickly, it is impossible to realize all the benefits of big data and cloud computing.

In the current world, exploring data to divide customers into various segments, understand their behavior, gain insights from multiple sources is key to achieving a competitive advantage. Even though decision-makers love to use data to base their actions and decisions on non-obvious patterns, predicting the future through mining data is not as simple as it appears. For this reason, organizations use data mining and knowledge discovery in data tools to extract non-obvious patterns and information from a large data set. Particularly, data mining seeks to discover previously unknown patterns among various datasets by applying methods from numerous fields such as database management systems, machine learning and artificial intelligence, and statistics.

Cloud computing offers several services using different models, for example, SaaS, PaaS, and IaaS. Infrastructure as a Service, otherwise known as IaaS, is online services that abstract a user from knowing infrastructure details. In this case, a user utilizes physical computing resources, security, backup, scaling, and data partitioning without necessarily knowing how they are provisioned (Assunção et al 7). Accordingly, through this service, such services as applications and operating environments are offered to users. On the other hand, Software as a Service (SaaS) is where users no longer have to own Software, but when they need to use it, they acquire it from clouding computing service providers. Notably, the Software is a property of the service provider, and a user will thus only pay period subscriptions or every time they need to use. Platform as a Service (PaaS) is where service providers develop standards and toolkits for distribution channels, payment, and development. In this case, through PaaS models, cloud service providers build computing platforms such as programming-language execution environments, operating systems, web servers, and databases. Some of the operating platforms provided through PaaS include Windows/NET. and J2EE.

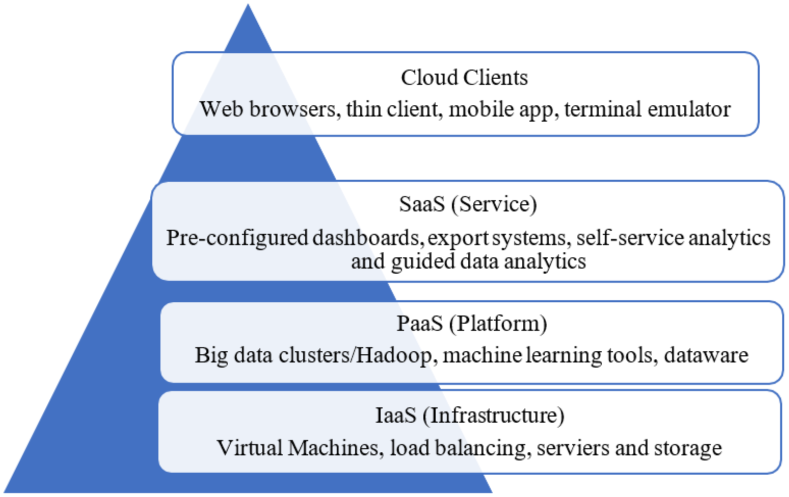

Cloud Clients
Web browsers, thin client, mobile app, terminal emulator
SaaS (Service)
Pre-configured dashboards, export systems, self-service analytics and guided data analytics
PaaS (Platform)
Big data clusters/Hadoop, machine learning tools, dataware
IaaS (Infrastructure)
Virtual Machines, load balancing, serviers and storage

*Fig 1*: *Cloud Computing models*

Integrating big data in cloud computing is crucial because it has several advantages that would improve its performance. However, there are several issues to be addressed before big data cloud computing becomes mainstream. In this case, even as big data and clouds become prevalent in today's world, business organizations and providers must be aware of challenges and risks when deploying solutions linked to the two trends. One of the most fundamental issues in deploying big data on a cloud environment is security. Typically, there exist some security vulnerabilities which result from creating new and unfamiliar platforms. For example, platform heterogeneity is a well known cloud security vulnerability that requires that organizations wishing to use clouds for significant data deployment to learn existing security practices and tools that work on each platform (Zanoon, Abdullah, and Sufian 6971). Security tools include encryption, access control, Cloud Clients Web browsers, thin client, mobile app, terminal emulator SaaS (Service) Pre-configured dashboards, export systems, self-service analytics and guided data analytics PaaS (Platform) Big data clusters/Hadoop, machine learning tools, dataware housing, dashboarding, operating systems IaaS (Infrastructure) Virtual Machines, load balancing, serviers and storage intrusion detection, event logging, monitoring, and authentication. Additionally, business organizations should also be aware of various security policies and consolidation plans that each platform has.

## Results and Discussion

In conclusion, cloud computing and big data have a complementary relationship. Notably, cloud computing and big data are part of a distributed network technology. However, even as companies move towards big data cloud computing, such issues as integration, performance, and security must be addressed. Despite that, cloud computing offers flexibility and reliability in the processing and managing data; customers should be assured that their data is safe in the clouds. Furthermore, the rapid progress towards technological development must be aligned to customer requirements.

- **Challenges and Solutions to Big Data in The Cloud**

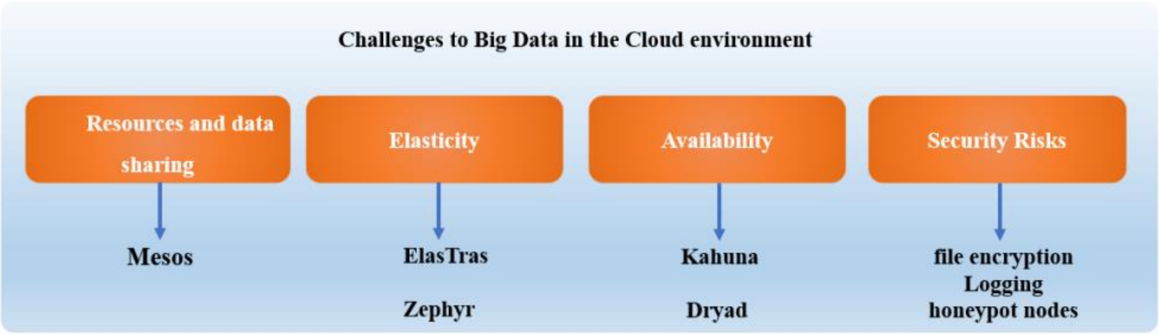


Figure 1. challenges and solutions to big data in the cloud environment

Big data is extremely large and complex data sets created by various devices, systems, equipment, business applications, and many more. These sets of data are difficult to be stored, processed, and analyzed using traditional database management tools because of their large volume and characteristics. This copious data should be stored using a cloud computing service to be processed and managed efficiently. Merging big data with cloud computing presents great opportunities, but there are some hurdles that need to be overcome.

- **Security Risks**

According to the annual state of the cloud survey conducted by RightScale (2018), security risks were the top concern of technical professionals in 2018 as 77% of respondents stated in the survey. Big data usually contain confidential information such as medical and identifiable financial information, and the breaches of these critical data will definitely lead to serious legal consequences and huge reputational damage to the organization. (Inukollu, Arsi, & Ravuri, 2014) suggested various security measures and techniques to improve the security of cloud computing services such as file encryption, logging, honeypot nodes.

Because the data is present in the machine’s cluster, a hacker can steal and spy on sensitive data. Thus, the data must be encrypted and stored behind a robust firewall. This way, even if the data got stolen, the hacker will not be able to extract meaningful information from it. In addition, All the map reduce jobs that modify the data, and the information of users who are responsible for these jobs should be logged. These logs should be inspected regularly to find if there is any data manipulation by unauthorized users. Another mechanism is honeypot, a honeypot is a trap node that appears in a cluster as a regular node. This node traps the attacker and necessary actions would be taken to eliminate the attack.

- **Availability**

The strict requirement of availability is tied with the system's ability to tolerate faults. The system should be able to sustain both transient failures such as CPU availability and persistent failures such as network outages. A reliable and highly available cloud service is essential for maintaining customers' satisfaction and confidence. Even though distributed environment such as MapReduce has high fault tolerance and availability, fault detection is still a problem because of the large size of the cluster. Tan et al.(2010) discussed that time-based detection of faults for MapReduce systems is difficult due to several factors such as input and cluster size.

Ford et al. (2010) presented a fault and availability analysis of Google cloud storage system. The system made up of three layers, including BigTable, GFS, and Linux file system. The fault tolerance at all these layers is significant enough to provide high availability to the cloud. The authors found that a node can become unavailable for many reasons such as overloading of the node or network failure. However, they found that only 10 percent of the failures lasted more than 15 minutes. They mentioned that transient failures do not have a huge effect on cloud availability because of the high replication strategies. The authors found that most of the nontransient failures occur in bursts due to rack failures. The analysis has assisted in developing analytical models for improving availability, replication strategies, and choices for data placement.

Kahuna is a fault detection tool to detect performance problems. In normal situations, the nodes in MapReduce perform symmetrically, and a node that behaves differently will cause an error. The similarity between MapReduce and Kahuna is detecting performance errors by observing unusual behaviors and characteristics. Moreover, Dryad is a scheduler and fault tolerance model. The graph is automatically mapped to physical resources by the Dryad execution framework. Deterministic vertices indicate an error. However, non-deterministic vertices indicate that executions are errors-free (Tan et al., 2010).

- **Elasticity**

Big data cloud should be able to scale according to the state of the system. That is, the cloud should be capable to scale to meet the needs during the moment of high-demand and shrink during low-usage periods. These adjustments are supported through virtualization, where virtual machines are moved from one physical machine to another to enable flexible load balancing and resource provisioning. While these adjustments are established at the infrastructure layer, issues arise at the application layer because of service interruptions that may occur due to live migration. Scaling out means partitioning of the database, and affecting query processing during the migration process.

The solution is ElasTras, which is a transactional distributed database for the cloud. The data store provides on-demand elasticity through transactional managers, which are able to allocate and de-allocate resources (Das et al., 2009). Zephyr adds to ElasTras capabilities by integrating live migration. It eliminates cloud outage by simultaneously allowing transactions at the source and destination. The migration process entails the transfer of metadata to the destination, so when the metadata transition is complete, a new transaction is initiated at the destination and existing transactions are being completed at the source (Elmore et al., 2011).

- **Resource and Data Sharing**

There are many platforms for big data computing. However, there is no platform that is efficient and optimal for all the big data applications. Hindman et al. (2011) Reported a scenario where Yahoo and Facebook users would like to build multiple clusters for their usage. The solution was to set up separate clusters for each application and transfer data between them. However, this approach causes data duplication.

Hindman et al. (2011) Suggested Mesos, which enable users to share several frameworks in the same cluster. It works as an intermediary between the cluster and framework, where it provides resources to each framework. In addition, it is responsible to schedule the resources' tasks. However, this approach might cause failure under high scalability requirements.

- **Big Data Management & Analytics Benefits**

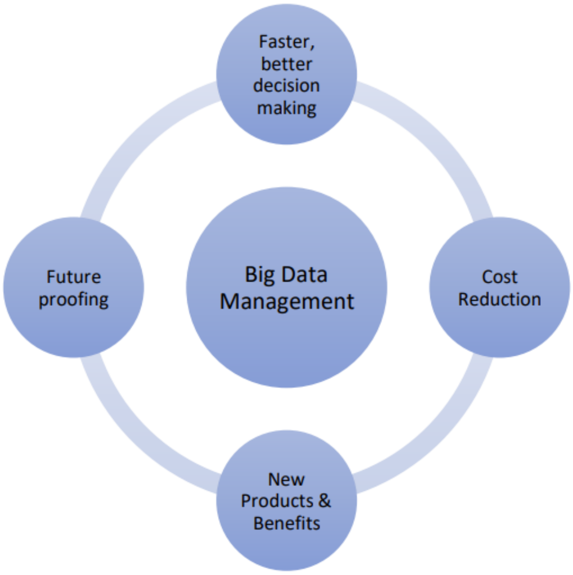


*Fig 2: Big Data Management*

Cloud computing and big analytics is the backbone of today's world. When combined with bandwidth and connectivity, cloud computing is expected to bring the internet of things into reality. With billions of sensors, machines, as well as other components networked, enterprises must keep adjusting their expertise and skills to compete and provide value to their customers However, as data increases exponentially, there are several challenges posed in its handling and analysis. As a result, big data management and analytics solutions must be developed to ensure that its management demands are met. Today, there is no doubt that cloud computing is the future. In the future, enterprises will rise and fall depending on their ability to generate store, and analyze Big Data Management Faster, better decision making Cost Reduction New Products & Benefits Future proofing data in the clouds. Notably, organizations that will succeed in the future will be capable of adopting cloud computing, which allows the usage of virtualized resources and dynamic scalability through the internet. Through cloud computing, organizations can acquire platforms as a service, Software as a service, and infrastructure as a service where processing power and memory are allocated to a business depending on its computing resources they need.

From a management point of view, big data creates enormous challenges, including the ever-increasing complexity, irrelevance, insecurity, and risks. The limitations and benefits of accessing big data in the clouds result from the inability or complexity of analyzing medical records, financial data, social media interactions, genetic sequences, and government records. As a result, if an organization wants to benefit from big data and cloud computing, it must develop efficient and effective analytic applications, services, frameworks, and programming tools. Therefore, big data management and analytics as a cloud service are expected to help organizations control costs, offer greater flexibility, and efficiently handle rapid storage growth requirements. In a century where business organizations run data-intensive computing workloads, developing and deploying data analytics and management in the cloud is the best option because they can meet scalability needs and decrease the number of computing devices needed to store big data.

Cloud computing service providers allow modern business organizations to store relational data on cloud servers owned and managed by providers in their data centers. As a result, a company does not have to invest in disks as well as DBMS software because they rely on cloud computing service providers. With the amount of data which needs to be analyzed growing exponentially, experts have been looking towards growing data analytics technology, including machine learning and data mining. However, these technologies have been in the past been shown not to offer the desired scalability or ubiquitous access for big data analysis. However, by implementing big data management and analytics as a cloud service, these shortcomings can be overcome. Particularly, cloud-based infrastructure would offer organizations with scalable hardware, platform, and software resources to execute big data analytic workflows.

There exist numerous analytic workflow management systems today that many of them have evolved from being domain-specific. The systems also vary in how they support visualization and collaboration. While the analytical workflow management systems were capable of offering all capabilities to traditional relational databases, they are incapable of executing significant data analytics demands in today's world. As a result, deploying big data management and analytics as a cloud service provides organizations with ubiquitous access, on-demand computing, storage, memory resources, online collaboration, and fault tolerance. Therefore, cloud platforms are excellent platforms for big data analytics and management because they offer big data query and management tools. Additionally, even though data security is a significant risk in cloud platforms, recent developments in such techniques as intrusion detection, access control, encryption, and data anonymization have led to more reliable solutions in the areas (Zulkernine et al. 2). The growth and benefits of cloud computing is a testament that it has begun to influence the big data field.

In a digital world, large amounts of data are being generated, stored as well as shared. For example, webpages, data warehouses, and blogs generate massive audio, text, and video streams which require efficient management. The massive generation of complex and pervasive data implies creating, storing, sharing efficiently, and analyzing it to get useful information. For maximum benefits, organizations should use big data management and analytics as a cloud service because it has many benefits such as scalability, familiar development models, and reliability. Additionally, big data analytics and management as a cloud service allow modern organizations to optimize resources optimization, and thus they only pay for what they require. As a result, a company does not have to invest in hardware, software and platforms which they do not need more often because they only but them when need arises.

- **Data, Information and Knowledge Management**

Every day huge amounts of data is entered into an information system and stored in the database and these data are of various functions of the business like financial information system will generate financial data these data can be probed later to generate a report and are capable to recover on failures. These data when processed into sets according to context, it provides information and these data and information can be used interchangeably to make business decisions. In the world of accelerating changes and the growing competition organizations have loaded responsibility on the managers to make business decisions quickly and efficiently based on the data that was analyzed. These quick and efficient decision making processes of the business can be very beneficial to the business and have a major impact on the business.

These management decisions are divided into at three broad levels within the organization:

- Operational decisions

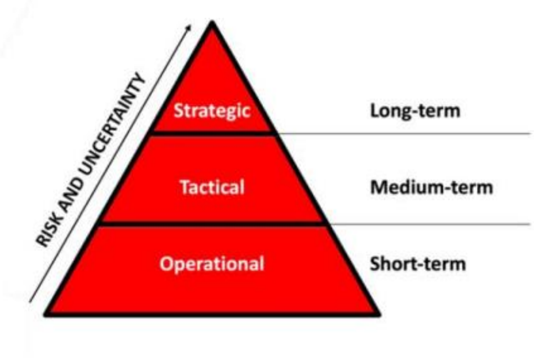
These are the decisions that affect the day to day business of the organization and are for short duration of time and are made frequently also are repetitive in nature as well as have a significant impact on the resource efficiency.

- Tactical decisions

These decisions are generally taken by the middle managers, these tactical plans cover shorter time frames and are related with less uncertainty and therefore are at lower risk as compared to strategic planning. They involve implementation of the policies within the organization and resource acquisitions and utilizations to accomplish the organizational goals.

- Strategic decisions

These decisions are taken by top management, and affects the whole organizations business, it involves deciding and developing strategic plans to accomplish strategic goals and are typically for long term as well as are infrequent. They involve a lot of risk and uncertainty.



All these decisions are made from processed data, that is organized, structured and presented in a given context, the data or the information required at different level for decision making differs from each other, each type of decision has its own uniqueness. The process of organizing, gathering, and analyzing, using, sharing, and maintaining an organization's information and knowledge is known as Knowledge management. There are numerous benefits of the knowledge management it helps in achieving business objectives for instance improving competitive advantage, sharing insights, refining performance, enhancing innovation, and continuously improving the organization.

- **Management of Big Data on Cloud**

With the increase in the organization's data the computing requirements grow into more and more complex, to eliminate this complexity and to process these data efficiently there has been a development of new computing module called MapReduce (T.suryakanthi, 2016) which is used for processing and generating large data sets on clusters of computers it can decrease the data using key value pairs. The MapReduce system comprises two essential functions, specifically Map and Reduce.

- Map: In the map function it takes a set of data and transforms it into another set of data, where individual elements are broken down into tuples (key/value pairs)

- Reduce: The Reduce function is always performed after the map job, it takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

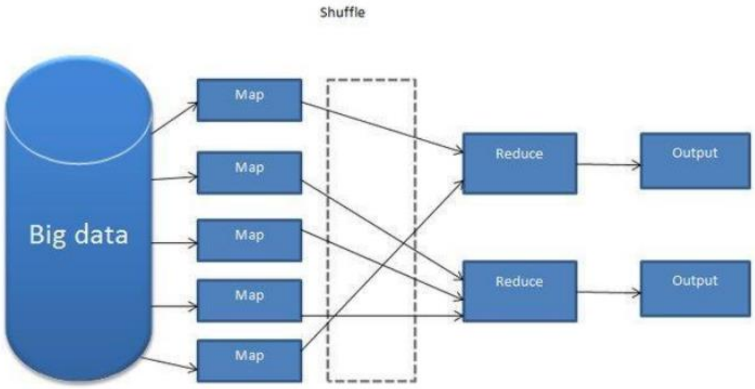
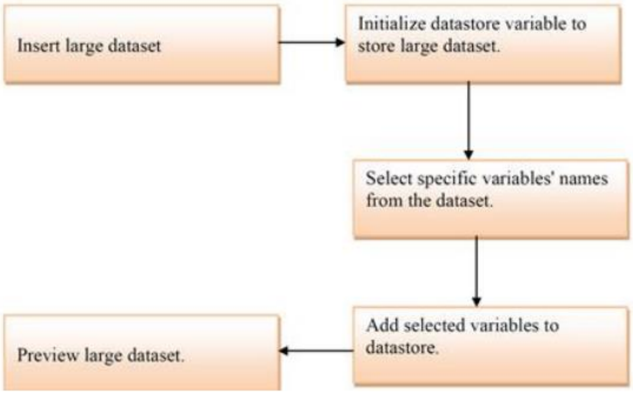


Figure: The MapReduce architecture (Mohammed Elmogy, 2016)

The MapReduce function has mainly three steps:

- Reading Large dataset

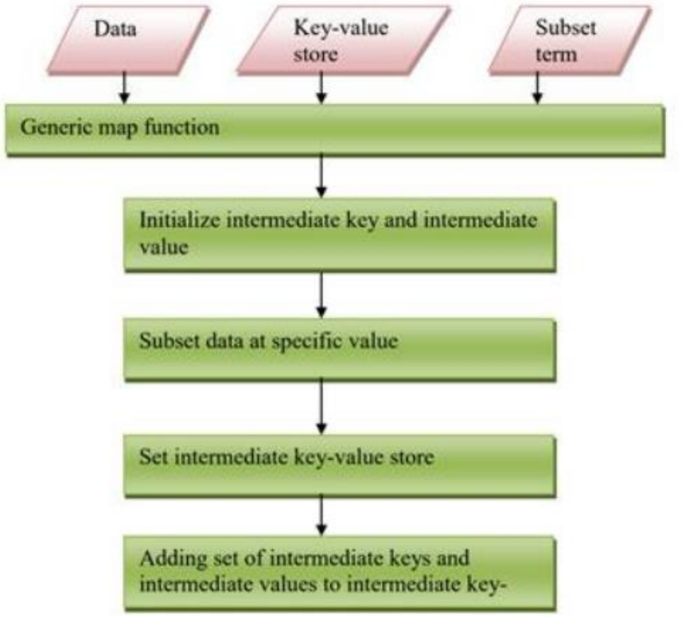
Data store is created using a dataset with CVS extension which is a comma-separated value file, that permits data to be saved in a tabular format. Then, Specific variable name is selected from the dataset which allows the user to retrieve the data using commands.


Insert large dataset
Initialize datastore variable to store large dataset.
Select specific variables' names from the dataset.
Add selected variables to datastore.
Preview large dataset.

The block diagram of MapReduce read data (Mohammed Elmogy, 2016)

- Generic Map Function

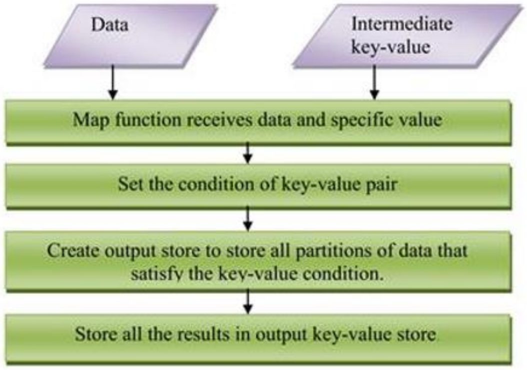
The Generic Map Function facilitates the programmer to set any pair of key-value pairs for the designated dataset. Once the intermediate key and intermediate value are set the dataset is subset at this precise value. Lastly, a set of key-value stores is obtained in the keyvalue store.


Data
Key-value store
Subset term
Generic map function
Initialize intermediate key and intermediate value
Subset data at specific value
Set intermediate key-value store
Adding set of intermediate keys and intermediate values to intermediate key-

The block diagram of generic map function. (Mohammed Elmogy, 2016)

- **Map Function**

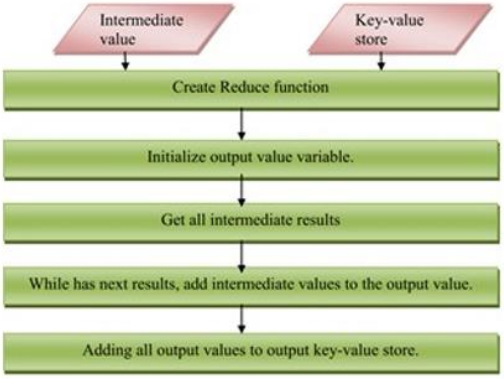
The Map function acquires a table with a selected variable name from the data which then extracts the subset of data that validates the condition value of the selected key.


Data
Intermediate key-value
Map function receives data and specific value
Set the condition of key-value pair
Create output store to store all partitions of data that satisfy the key-value condition.
Store all the results in output key-value store

The block diagram of map function. (Mohammed Elmogy, 2016)

- **Reduce Function**

This function obtains the subset data results acquired from the Map function and combines them into a single table. The Reduce returns one key and one value.

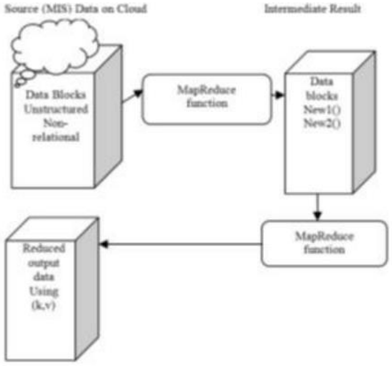

Intermediate value
Key-value store
Create Reduce function
Initialize output value variable.
Get all intermediate results
While has next results, add intermediate values to the output value.
Adding all output values to output key-value store.

The block diagram of reduce function.(Mohammed Elmogy, 2016)

- **Cloud and MapReduce**

With the help of this MapReduce data model, the data is stored in cloud architecture which will help in reduction of the data. It allows us to think about undertaking information and their operational data which is being stored on the server. This data can be structured, unstructured or semi structured data.

The data firstly divided into a block of chunks using Hadoop which is an open source, Java based framework used for storing and processing big data. The data is stored on inexpensive servers that run as clusters. These blocks of chunks are then sent to the MapReduce function to further reduce the data into blocks. These data are then sorted and shuffled to produce the intermediate data which again undergoes under an MapReduce function to develop final Reduced output data.


Source (MIS) Data on Cloud
Intermediate Result
Data Blocks Unstructured Non-relational
MapReduce function
Data blocks New1() New2()
MapReduce function
Reduced output data Using (k,v)

Data model for big data processing

- **Big Data Value Creation Model**

For companies to generate value from big data, they can adopt distinct approaches. While the hierarchical method illustrates data accumulation to create knowledge, the sequential model describes a step-by-step procedure from data collection to distribution to the proponents (Ylijoki & Porras, 2018). In essence, the two approaches can be deployed to transform information knowledge and then knowledge to value, but the choice depends on businesses that use big data.

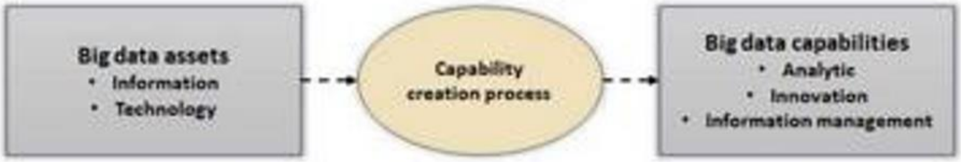


Figure 2. Converting big data capabilities into big data impacts (Ylijoki & Porras, 2018).

Big data analytics has importance in organizations. The elite world is greatly dependent on access to information. Big data ascertains that the complex data is organized meaningfully for use to create value in companies. Massive data assets and capabilities are requirements for companies to have significant big data effects . The capability development process converts information to knowledge . Capability is a component of the factors that describe a company's competitive advantage . Most importantly, the availability of big data capabilities in firms reflects increased value (Ylijoki & Porras, 2018). It is expected that organizations have big data for information to be realized. Furthermore, the acquired information turns to knowledge, and ultimately, transformation is realized.

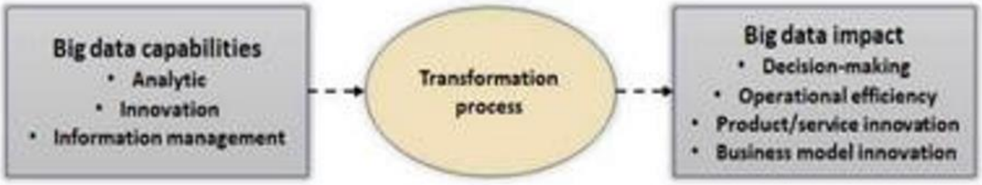


Figure 3. Converting big data capabilities into big data impacts (Ylijoki & Porras, 2018).

Some factors are vital in big data value creation. Firstly, companies have to invest to realize big data assets. The perspective focuses on converting the costs incurred on information technologies to assets, which have economic value . In essence, the accumulated information in the process of investment is a valuable asset in big data. Secondly, big data effects should be considered intermediate outcomes, which imply a need to bridge the results to actual performance measures . In essence, companies' performance relies on the competition process, which links organizations to "its industry, ecosystem, competitors and customers" (Ylijoki & Porras, 2018, p. 12). It is crucial to convert the big data results to actual performance, which constitutes the value creation model.

## 2. References

[1] Ara, A., & Ara, A. Cloud for big data analytics trends. IOSR Journal of Computer Engineering, **18(05),** 01-06. doi:10.9790/0661-1805040106 (2016).

[2] Ali, H., Ekmogy, M., & Barakat, S.(2016). A big data processing framework based on Mapreduce with applications to Internet of things. Ciência e Técnica Vitinícola, **31,** 2-25. Retrieved from https://www.researchgate.net/publication/305489358_A_BIG_DATA_PROCESSING_FRAMEWORK_BASED_ON_MAPREDUCE_WITH_APPLICATION_TO_INTERNET_OF_THINGS (2016).

[3] Abou El-Seoud, Samir, et al. "Big Data and Cloud Computing: Trends and Challenges." International Journal of Interactive Mobile Technologies (iJIM) 11(2), 34-52 (2017).

[4] Assunção, Marcos D., et al. "Big Data Computing and Clouds: Trends and Future Directions." Journal of Parallel and Distributed Computing **79,** 3-15. Big data analysis shown to increase revenues and reduce costs. (2016, October 19). Retrieved from http://barc-research.com/big-data-analysis-shown-to-increase-revenues-and-reduce-costs/ (2015).

[5] Das, S., Abbadi, A., & Agrawal, D. ElasTraS: An Elastic Transactional Data Store in the Cloud. ArXiv, abs/1008.3751 (2009).

[6] Elmore, A.J., Das, S., Agrawal, D., & Abbadi, A. Zephyr: live migration in shared nothing databases for elastic cloud platforms. SIGMOD '11, 2011.

[7] Hajirahimova, M., & Aliyeva, A. Some indicators of big data. IOSR Journal of Engineering, **6(10),** 1-6 (2016).

[8] Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A., Katz, R., Shenker, S., & Stoica, I. Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center. NSDI, 2011.

[9] Inukollu, V., Arsi, S., Ravuri, S. Security Issues Associated with Big Data in Cloud Computing. International Journal of Network Security & Its Applications. **6(3),** 45-56. 10.5121/ijnsa.2014.6304 (2014).

[10] Memon, M. A., Soomro, S., Jumani, A. K., & Kartio, M. A. Big data analytics and its applications. Retrieved from https://www.researchgate.net/publication/320345031_Big_Data_Analytics_and_Its_Applications (2017).

[11] Ouf, Shimaa, and Mona Nasr. "Cloud Computing: The Future of Big Data Management." International Journal of Cloud Applications and Computing (IJCAC) **5.2,** 53-61 (2015).

[12]RightScale. State of the Cloud Report. Retrieved from https://www.suse.com/media/report/rightscale_2018_state_of_the_cloud_report.pdf (2018).

[13] Salvador, J., Ruiz, Z., & Garcia-Rodriguez, J. Big data infrastructure: A survey. doi:10.1007/978-3-319-59773-726 (2017).

[14] Tan, J., Pan, X., Marinelli, E., Kavulya, S., Gandhi, R., & Narasimhan, P. Kahuna: Problem diagnosis for Mapreduce-based cloud computing environments. 2010 IEEE Network Operations and Management Symposium - NOMS 2010, 112-119 (2010).

[15] Ullah Saeed et al. Big data in cloud computing: a resource management perspective. Scientific Programming Journal. Retrieved from https://www.hindawi.com/journals/sp/2018/5418679/ (2018).

[16] Van Gils, Teun & Ramaekers, Katrien & Caris, An & De Koster, René. Designing Efficient Order Picking Systems by Combining Planning Problems: State-of-the-art Classification and Review. European Journal of Operational Research. **267.** 1-15. 10.1016/j.ejor.2017.09.002. (2018).

[17] Vighio, M. S. et al. A Relational Data Model for Uncertain Data. International Journal of Emerging Multidisciplinaries: Computer Science and Artificial Intelligence. **1(2),** https://doi.org/10.54938/ijemdcsai.2022.01.2.141 1-2 (2022)

[18] Ylijoki, O., & Porras, J. A recipe for big data value creation. Business Process Management Journal, **25(5),** 1-22. doi:10.1108/BPMJ-03-2018-0082 (2018).

[19] Zanoon, Nabeel, Abdullah Al-Haj, and Sufian M. Khwaldeh. "Cloud Computing and Big Data Is There A Relation Between the Two: A Study." International Journal of Applied Engineering Research **12.17,** 6970-6982. (2017).
[20] Zulkernine, Farhana, et al. "Towards Cloud-Based Analytics-As-A-Service (Claaas) For Big Data Analytics in The Cloud." 2013 IEEE International Congress on Big Data. IEEE, 2013.