INTERNATIONAL RESEARCH JOURNAL OF HUMANITIES AND INTERDISCIPLINARY STUDIES
( Peer-reviewed, Refereed, Indexed & Open Access Journal )
DOI : 03.2021-11278686 ISSN : 2582-8568 IMPACT FACTOR : 5.828 (SJIF 2022)

# Big Data Management

**Miss. A. A. Changede**
Assistant Professor
Dept. of Computer Science,
Rajarshi Shahu Mahavidyalaya, Deolali Pravara,
Tal. Rahuri, Dist. Ahmednagar (India)
E-mail: ashwinichangede2012@gmail.com

**Mrs. S. M. Hapase**
Assistant Professor &
I/C Principal
Dept. of Computer Science,
Rajarshi Shahu Mahavidyalaya, Deolali Pravara,
Tal. Rahuri, Dist. Ahmednagar (India)
E-mail: hapasesm@gmail.com

**DOI No.** 03.2021-11278686 **DOI Link ::** https://doi-ds.org/doilink/03.2022-78845899/IRJHISIC2203020

***Abstract:***

*Big Data is a collection of data that is large in volume, mostly from the various data sources. These data sets are too large to process by using traditional data processing software. But this large quantity of data gives the solution to business problems that you may not have been able to handle before. The applications of big data are cyber security, pharmaceutical drug evaluation, scientific research, weather forecasting, tax compliance, etc.*

*Big data management is the collection, administration and governance of very large volumes of both structured and unstructured data. The importance of data comes from the point of view that can be discovered by manipulating it to get the prospect that hides within. The aim of big data management is to guarantee a high level of data quality and accessibility for businessintelligence and big data analytics software. Big data management involves various processes such monitoring and ensuring the availability of all big data resources through a centralized interface, performing database maintenance, implementing and monitoring big data analytics, data quality, focusing the security of big data repositories and control access, ensuring that data are collected and stored from various resources. By using Big Data Management Tools predictions and analysis of business are becoming more accurate and interesting. This paper brings up the focus on Big Data Management. First introduction of big data, BigData Storage, Big Data Technologies, Best Policy to Big Data Management, Big Data Management (Distributed Processing Framework Hadoop and Spark, Cloud Object Storage Services, Stream Processing Engines, Cluster Management Software, NoSQL Databases, Data Lake) along with some popular tools and the benefits of using Big Data Management.*

***Keywords:*** *Big Data, Storage, Big Data Management Tools, Policies, Benefits*

## Introduction:

Due to new trends and technologies, devices and communication means like social networking sites, the huge stream of data produced. The amount of data produced at the beginning of time 2003 was 5 billion gigabytes. The same amount of data was created in couple of days in 2011 and every ten minutes in 2013.This rate is still growing drastically. Though all this information produced is meaningful and can be useful when it is processed. Black Box data, Social Media data, Stock Exchange data, Power Grid data, Transport data, Search Engine data as well as structured, Semi Structured and Unstructured data comes under the umbrella of Big Data.

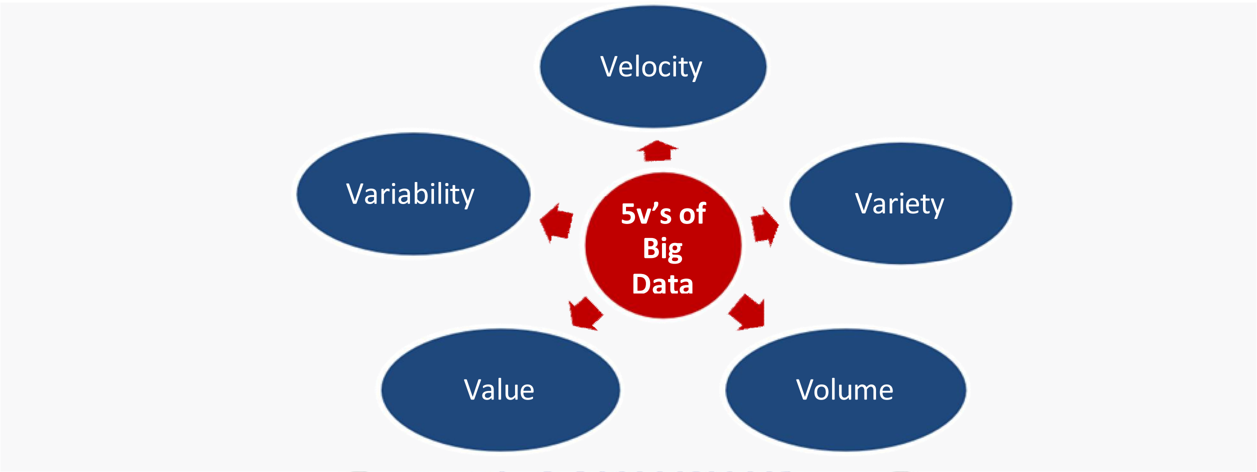
Most big data environments is step ahead of relational databases and traditional data warehouse platforms to incorporate technologies that are suited to processing and storing non-transactional forms of data. By giving more focus on gathering and analyzing big data is introducing new data architectures that often integrate data warehouses with big data systems.

Big Data Management is performing process on structured as well as unstructured data. In Big Data Management process, most of the companies must decide what data must be kept for adherence reasons, what data can be discard of and what data should be process in order to improve current business processes or provide a incredible advantage. Big data is usually composed of huge interconnected data - in addition to its volume and variety, it includes frequently flowing data and other types of data created and updated at high velocity. As a result, analyzing big data can be a daunting task. For data management teams, the biggest challenges facings are dealing with the massive data, integrating the data, improve the data quality, data preparations for data analytics applications, governing big data environment, ambiguity in big data management tool selection, and the most important is data privacy.

**What is Big Data?**

Big data refers to tools, processes, and procedures allowing an organization to create, manipulate, and manage huge data sets and storage facilities. Big data is data that cross the processing capacity of traditional database systems. The data is too large, much dynamic, or doesnot handle by existing database architectures. To gain benefits from these data, there must be an alternative way to process it. Big Data used initially introduced by the "3Vs", but now referred toas the "5Vs" as Velocity, Variety, Volume, Value, Variability. The size of data have increases rapidly as data is collected by devices such as mobile devices, cheap and numerous information- sensing Internet of things devices, remote sensing, software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. Due to such huge dataavailability data management is challenging task.

Currently, companies use Big Data to make business more enlightening and taking business decisions by enabling analytical modelers, data scientists and other professionals to analyze large volume of transactional data.



Big data is the demanding and powerful fuel that drives large IT industries in the era of 21st century. Big data is a spreading technology used in each business sector, Travel and Tourism, Financial and Banking sector, Healthcare, Telecommunication and media, Government and Military, E-commerce, Social Media, etc.

**Big Data Storage:**

Big data storage has architecture with large computational and storage capacity that collects and direct large data and supports real-time data analytics. Big data analytics operates large semi structured or unstructured data and gives the results quickly.

Big data storage demand will reach 163 Zettabytes by 2025, according to report of IDCand Seagate in 2017. The report attributes the growth to increase use of embedded systems, mobile devices, machine learning, cognitive computing, and security.

Big data collects data from various sources, which is unable to process with a relational database. The most suitable analytics engine for big data is The Apache Hadoop Distributed File System (HDFS), and is typically integrates with some key points of a NoSQL databases. HDFS spreads the analytics over hundreds or even thousands of server nodes without a performance hit. With the help of Map Reduce components distribute process against catastrophic failure. When the query arrives Map Reduce runs processing directly on data storage. Once the processing completed all the results are collected and reduces them to present single atomic result.

**Big Data Technologies:**

Big data technologies plays an important role in providing more accurate analysis, which gives more precise decision making with high ratio in operational efficiency, cost minimization, lesser risk for business. To use the power of big data, we require an infrastructure that can handle structured as well as unstructured data in real time and can preserve data integrity.

For managing big data, there are many more technologies in market from various vendors Like Amazon, Microsoft, IBM, etc. The Big data technologies are differentiate into two technologies.

1. **Operational Big Data:**

It focuses on operational competence for real-time, interactive workloads where data is basically captured and stored like MongoDB. NoSQL Big Data systems are designed to take benefits of new cloud computing architecture which works on huge computation with efficiently and affordable price. Some NoSQL system can works on patterns and trends based on real-time data with less amount of coding and make independent from data scientists as well as well-planned infrastructure.

**Examples:**

Online ticket booking system e.g. trains, buses, movies and flights, etc.

Online shopping from e-commerce websites like Amazon, Flipkart, Walmart, etc.

2. **Analytical Big Data:**

In that, Massively Parallel Processing (MPP) database systems and Map Reduce that provide analytical capabilities for backdated and complex analysis with includes all of the data comes from various data sources. Map Reduce provides upgraded method of analyzing data over the capabilities of SQL. Map Reduce system is designed for single server to thousand of high and low power machines.

**Examples:**

Weather forecasting data.

Medical health records.

**Difference between Operation and Analytical Big Data:**

| **Key Points** | **Operational Big Data** | **Analytical Big Data** |
|---|---|---|
| **Latency** | 1 ms - 100 ms | 1 min – 100 min |
| **Concurrency** | 1000 – 100000 | 1 – 10 |
| **Technology** | NoSQL | MapReduce, MPP Database |
| **Access Pattern** | Writes and Reads | Reads |
| **Data Scope** | Operational | Retrospective |
| **End User** | Customer | Data Scientists |
| **Queries** | Selective | Unselective |

**Best Policy to Big Data Management:**

1. **Develop the master plan and roadmap:**

Business should start by creating a master plan for big data that includes business goals, data requirements and applications and system deployments, review of data management skills and procedures.

2. **Develop and implement best architecture:**

   The big data architecture includes various types of systems and tools that support data management phases, from ingestion, processing and storage to data quality, integration and preparation work.

3. **Determine Business goals and requirements:**

   Data management teams continuously collaborate with data scientists, data analysts and business users to make guarantee that big data get business requirements to enable moreaccurate decisions about data.

4. **Be adaptable on managing data:**

   Data scientists basically need to know how they manipulate data for predictive analytics, machine learning, and other types of big data applications.

5. **To Access and Governance control:**

   Big data focus on users access control, data privacy and protection**.** Somehow, it helps organizations to obey data privacy laws regulating the collection and use of private data. but well-controlled data can also lead to best-quality and higher level of accurateanalytics.

**Big Data Management:**

There are great mixture of platforms and tools for managing big data, with both open source and commercial versions. The list of big data technologies that can be install, frequently in combination with one another, includes distributed processing frameworks Hadoop and Spark, Cloud Object Storage Services, Stream Processing Engines, Cluster Management Software, NoSQL databases, Data Lake.

1. **Distributed Processing Framework Hadoop and Spark:**

Hadoop and Spark is the most famous data processing framework for big data architecture. Hadoop and Spark have open source platforms for processing, managing and analyzing the big data. Most controversy on using Hadoop verses Spark go around optimizing big data environment for real time processing or batch processing. But these two frameworks simplify the differences as Apache Hadoop and Apache Spark.Hadoop initially used with only batch applications and now use with iterative querying and real-time analytics. Spark also initially developed for batch job processing. Most of the organizations work with both platforms for various big data use cases. Spark application are build on the top of Hadoop’s YARN resource management technology as well as Hadoop Distributed File System (HDFS).

### 2. Cloud Object Storage Services:

Cloud object storage is a platform for storing unstructured data in cloud. Object storage is decided to be best because of its flexibility and easily integrate into multiple petabytes to support huge data. The architecture stores and manages data as an object. The object storage software design includes globally unique identifiers for every objects including metadata. An object identifier is an address of an object which is helpful for finding objects on distributed system. Object storage including metadata can be accessed by using APIs, HTTP and HTTPS.

### 3. Stream Processing Engines:

It is a big data technology that focuses on the real time processing of continuous stream of data in motion. A stream processing simplifies the parallel hardware and software by restricting the performance of parallel computation. It is also helpful for fraud detection.

### 4. Cluster Management Software:

It is based on automated management of complete Hadoop clusters by tracking and troubleshooting of distributed Hadoop systems and immediate reporting. Apache Mesos, form Apache software foundation; Kubernetes, founded by Google Inc.; Heartbeat, FormLinux; Docker Swarm, Red Hat cluster suite, Nomad, from Hashicorp; Rancher, from Rancher Labs; Trinityx form Cluster Vision Solutions are the cluster management softwares.

### 5. NoSQL Databases:

NoSQL databases are the best solution for big data management problems. Big data userlike Amazon, Google, Facebook etc., were first faced the limitations of Relational Database Management System and because of that they give focus on development of NoSQL databases. Google File System (GFS), Map Reduce, Big Table, Sharding, AJAX (Asynchronous JavaScript and XML) are the different trend and technologies form a foundation for NoSQL database further development.

### 6. Data Lake:

A data lake is much more flexible than data warehouse. Most of the organizations use Hadoop File System because it works with big data where Data Lake is likely to be used.A data lake is a centralized repository designed to store, process and secure huge amountof structure data from relational databases, semi structured data like CSV, logs, XML, JSON, etc., and unstructured data like emails, documents, PDFs, etc. It can store data in its original format and analyze in multiple ways, ignoring the size of data.

### Big Data Management Tools:

Oracle Data Management Suite, SAP Data Management, IBM Infosphere Master Data Management Server, Microsoft Master Data Services, Dell Boomi, Talend, Tableau, Amazon Web Services-Data Lakes and Analytics, Google Cloud, Looker BI, Cassandra, Chartio, Alooma,Panaply, Blendo, Informatica Powercenter, Informatica MDM Reference 360, Collibra, Profisseetc. are the some most popular big data management tools use by organizations.

**Big Data Management Benefits:**

1. **Improved customer services:**

   Improved customer service was the most-common benefit of big data management, citedby 56 percent of respondents.

2. **Increased revenue:**

   In Experian study, 61 percent of those surveyed report that data management efforts were helping their organizations increase revenue by improving data quality.

3. **Enhanced Marketing:**

   One of the best benefits as sales and customer service, marketing also get a boost frombig data management.

4. **Increase efficiency:**

   According to the Experian survey, 57 percent of those surveyed express maintaining high quality data helped them increase efficiency.

5. **Cost Saving:**

   Big data Management gives high level of financial benefits because of proper data management.

6. **Enabling new applications:**

   Organizations have more confidence in their data it increases their innovation and inspires them to create new applications.

7. **Improved accuracy for analytics:**

   Good data management policies is that it increases the accuracy and reliability of bigData Analytics.

**Conclusion:**

Big data has scientific, qualitative and economics importance in research for getting betters the accuracy in result. Data is produced at every moment of time which is always increases in amount. According to the research of IDC's Digital Universe Study data increases 44 folds to 35ZB per year from 2009 to 2020. There are some big data management policies which helps to improve to manage high volume of data management which leads to high data quality. Big data provides organization with much more options with variety of technologies and tools like Hadoop distribution, stream processing, cluster storage management, Data Lake, NoSQL, etc.

There are some tremendous benefits of big data which is introduced in this paper. The upcomingtrends are data-driven so it is important to analyze, develop and install data management system that accomplishes business needs.

**References:**

1. https://www.researchgate.net/publication/280933768_A_Review_of_Big_Data_Management_Benefits_and_Challenges
2. https://www.researchgate.net/publication/315670193_Big_Data_Management_and_Analysis
3. https://www.techtarget.com/searchdatamanagement/definition/big-data-management
4. https://www.datamation.com/big-data/big-data-management/
5. https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/
6. https://svitla.com/blog/18-best-data-management-tools
7. https://www.javatpoint.com/big-data-technologies
8. http://www.appperfect.com/services/big-data-services/big-data-cluster-management.php
9. https://www.google.co.in/