Educational Administration: Theory and Practice
2022, 28(4), 346-349
ISSN: 2148-2403
https://kuey.net/

Research Article

"Challenges Of Big Data Management"

Mr. Mofikul Islam[1*],

[1*]Assistant Professor, Department. Of Business Administration, AMU Centre Murshidabad,
Email: profmofikulmcamba@gmail.com

Citation: Mr. Mofikul Islam, (2022), "Challenges Of Big Data Management", *Educational Administration: Theory and Practice, 28(4),* 346-349
Doi: 10.53555/kuey.v28i4.7989

ARTICLE INFO

ABSTRACT

In this present time and in the near future, Big Storage holds great promise in the realm of technical storage and retrieval. Hadoop and various other big data frameworks have introduced a data storage model fundamentally distinct from traditional database management systems. Consequently, numerous enterprise applications have come to rely on these innovative architectural features in their database management systems, which has led to an emerging need for further development. Drawing from years of experience dealing with Hadoop workloads, we have encountered the network challenges inherent in HDFS and similar file systems. Big-data storage systems serve as crucial platforms in these domains, providing the necessary scalability and reliability for storing and processing vast amounts of data. Recently, workflow-based data handling has been achieved by employing either application specific overlays that map the output of one task to serve as the input of another in a sequential pipeline or, more recently, by leveraging the Map Reduce programming model. However, it is evident that the Map Reduce model does not suit every scientific application. When deploying a large-scale workflow across multiple data centres, geographically distributed computation encounters bottlenecks due to data transfers, resulting in high costs and significant latencies. In the current context, the consumption of storage space by big-data systems supports the storage and processing of large data sets. It fills a crucial gap between the self-describing data commonly used by scientists for data distribution and sharing and the increasingly important big-data systems crucial for scientific analysis. Building upon this approach, we have extended two important and widely used big data systems, utilizing them to directly support the storage and analysis of scientific data stored in self-describing formats.

Key Words: Hadoop*, HDFS, Map Reduce, Self-describing,* business intelligence, scenario

1. Introduction

Big data is defined as a large amount of data that requires new technologies and arrangements to extract value from it through capturing and analysis processes. We live in an era of data deluge, given the unprecedented amount of data produced, collected, and stored in recent years. It poses a great challenge for industries to derive benefits from this influx of data. While Big Data is undoubtedly perceived as a significant blessing, it also presents significant challenges due to large-scale datasets. The sheer volume of data often makes it impossible to run analytics using a central processor and storage. Therefore, distributed processing with parallelized multiprocessors, preferably storing the data in the cloud, is suitable. Moreover, as the size of data grows exponentially, not all current algorithms are efficient or scalable enough to handle such large volumes of data. Designing more accurate and intelligent models to meet market needs presents both huge opportunities and challenges to various communities including analytics, machine learning, data mining, distributed and high-performance computing, etc. It is believed that this special issue will provide a timely collection of novel research results beneficial for researchers and practitioners working in these communities. This special issue focuses on all aspects of big data and targets a mixed audience of researchers from all these communities. The major issues in Big Data shouldn't be confused with the problems; it is crucial to recognize and manage them. These issues are related to specific characteristics, such as:

Data Volume: As data volume increases, the value of different data records decreases in proportion to their age, type, richness, and quantity, among other factors. Existing social networking sites generate terabytes of data daily, making it challenging to handle using traditional systems.
Data Velocity: Our traditional systems often lack the capability to analyse data that is constantly in motion. The rapid expansion of E-commerce has increased the speed and richness of data used in various business transactions. Managing data velocity involves more than just a bandwidth issue.
Data Variety: The diverse nature of data, including raw, structured, semi-structured, and unstructured data, presents a challenge for existing traditional analytic systems. From an analytical standpoint, this variety poses one of the most significant obstacles to effectively utilize large volumes of data.

2. Objective of Research

Big Data Characteristics: The term 'big' in big data inherently denotes its volume. Currently, data exists in petabytes and is anticipated to grow to zettabytes in the near future. Data volume quantifies the amount of data available to an organization, which need not necessarily own all of it as long as access is feasible. Data Velocity in big data encompasses the speed at which data arrives from diverse sources. This characteristic isn't limited solely to the speed of incoming data but also encompasses the rapid flow and aggregation of data. Data variety signifies the richness of data representation, encompassing text, images, video, audio, etc. The produced data isn't confined to a single category; it includes traditional data as well as semi-structured data from various sources such as web pages, web log files, social media sites, emails, and documents. Data value gauges the usefulness of data in decision-making. While data science aids in exploring and understanding data, 'analytic science' encompasses the predictive capabilities of big data. Users can execute specific queries against stored data to extract crucial results, subsequently ranking them based on their required dimensions. These reports facilitate the identification of business trends, allowing for strategic adjustments. Data complexity assesses the degree of interconnectedness and interdependence within big data structures. Small changes in one or a few elements can trigger significant changes or cascades throughout the system, substantially affecting its behaviour, or sometimes no change at all.

Challenges in Big Data: Addressing the challenges in Big Data is crucial for real-time implementation, demanding immediate attention. Failure to navigate these hurdles during implementation might lead to technology failures and undesirable outcomes.
Privacy and Security Challenges: Among the most critical issues in Big Data are concerns regarding privacy and security. Big data, often sensitive in nature, carries conceptual, technical, and legal significance. Personal information within databases of merchants or social networking sites, when combined with external extensive datasets, can unveil new and potentially sensitive information about individuals. These revelations might contradict individuals' desires for data privacy. Data collected from various sources often aims to enhance an organization's business by creating insights into people's lives, sometimes unbeknownst to them. This process could inadvertently lead to social stratification, where technologically literate individuals leverage Big Data for predictive analysis. The utilization of Big Data by law enforcement could increase the risk of adverse consequences for specific individuals tagged within datasets, possibly without the ability to contest or even the awareness of discrimination. Accessing and sharing data within company information systems is crucial for timely and accurate decision-making. This necessity complicates data management and governance, requiring data to be open and available to government agencies in standardized formats with standardized APIs and metadata. Such practices aim to enhance business intelligence and productivity. However, sharing data between companies poses challenges due to the competitive nature of businesses. Sharing information about clients and operations threatens the traditional culture of secrecy and competitiveness, creating an awkward dynamic in data exchange.

3. Methodology

Big data poses significant analytical challenges in deriving actionable insights. The diverse nature of this data—be it unstructured, semi-structured, or structured—demands advanced analytical methodologies. Analysing such vast datasets requires a high level of expertise. The selection of analytical methods hinges on the desired outcomes, emphasizing the critical role of robust decision-making. Two primary approaches prevail: either integrating massive volumes of data into analysis or predefining the pertinent data for the desired insights. The workforce in the realm of big data is in its nascent stage, demanding a unique blend of skills and expertise. Attracting talent to navigate this emerging technology is crucial for organizations. Beyond technical prowess, skills in research, analysis, interpretation, and creativity are pivotal. Building these competencies necessitates tailored training programs within organizations. Additionally, universities must incorporate comprehensive big data curricula to groom a workforce adept in this field, meeting the growing demand for skilled professionals.

The primary challenging questions encompass:

- What happens when the data volume becomes extensive and diverse, and there's uncertainty about how to manage it?

- Is it necessary to store all available data, or is selective storage more prudent? Similarly, must all data undergo analysis?
- What methodologies can identify the truly crucial data points amidst the vast pool of available information?
- In what ways can data be optimally utilized to yield maximum benefits and advantages?

Technical Challenges

Fault Tolerance: With the advent of new technologies like cloud computing and big data, the focus is on ensuring that when failures occur, the impact remains within an acceptable threshold rather than restarting the entire task. Achieving fault-tolerant computing is a complex endeavour, involving intricate algorithms. It's practically unattainable to create machines or software that are 100% reliable in tolerating faults. The primary objective, therefore, is to minimize the probability of failure to a level deemed 'acceptable'. Regrettably, as we aim to reduce this probability, costs tend to increase. In the realm of big data, there are two methods that appear to enhance fault tolerance. The first involves breaking down the entire computation into tasks and distributing these tasks among different nodes for processing**.** Secondly, in a system where one node oversees the proper functioning of multiple nodes, it's common practice to assign specific nodes the task of monitoring others, with a restart mechanism in place should any issues arise. However, certain computations might not easily divide into independent tasks. Some tasks could be inherently recursive, where the output of one computation serves as the input for the next, making restarting the entire computation process arduous. To address this challenge, implementing checkpoints becomes crucial. These checkpoints capture the system's state at intervals in time. In the event of a failure, the computation can resume from the last recorded checkpoint, preventing the need to restart the entire process.

Scalability: The scalability challenge posed by Big Data has propelled the rise of cloud computing. Presently, it consolidates numerous diverse workloads, each with distinct performance objectives, into extensive clusters. This necessitates substantial resource sharing, an expensive endeavour that introduces several hurdles. These hurdles include optimizing job execution to achieve cost-effective workload goals and efficiently managing frequent system failures within large clusters. The amalgamation of these factors raises concerns regarding program expression, especially for intricate machine learning tasks. Notably, there has been a significant technological shift: traditional Hard Disk Drives (HDD) are giving way to Solid State Drives and Phase Change technology. However, these newer options exhibit differing performance metrics between sequential and random data transfers. Consequently, determining the appropriate storage devices for data storage still remains a significant inquiry.

Quality of Data: Acquiring and storing extensive datasets incurs expenses. Utilizing larger volumes of data for decision-making or predictive analysis undoubtedly yields superior outcomes in business operations. Business leaders consistently pursue increased data storage, while IT leaders prioritize technical considerations before storing vast amounts of data. Big data fundamentally emphasizes quality data storage over amassing excessive and irrelevant information, aiming for more precise outcomes and informed conclusions.

4. Data Analysis

To navigate effectively, we must first define our interests and have an idea about it. Consider whether we aim to predict customer behaviour or analyse customer driving patterns for insurance assessments or perhaps the interest lies in seeking insights from system log data for anticipating potential issues. In the context of a church, employing big data techniques allows for extensive data storage, encompassing audio/video demand and old church records such as birth, marriage, and death indexes along with their scanned images. Big data analysis facilitates advanced search methods within this unstructured data, driven by the overarching problems we intend to solve. The nature of our high level problem dictates the analytics approach we adopt. When the specific business problem remains unclear, focusing on areas in the business needing improvement may be beneficial. Even an analytics driven strategy, targeted accurately, can yield valuable results with big data. Advanced data analysis involves complex techniques like predictive modelling and other pattern matching techniques, offering deeper data insights. Operational analytics integrate seamlessly into the business workflow, while monetized analytics directly contribute to revenue generation. Basic analytics serve as an exploratory tool for data when specific issues are yet undefined but hold potential value. This might encompass simple visualizations or basic statistics within the realm of big data.

Basic analysis is often used for handling extensive and diverse datasets. Advanced analytics offer intricate algorithms for complex analysis of either structured and unstructured data. This includes sophisticated statistical models, machine learning, neural networks, text analytics, and other advanced data-mining techniques. Advanced analytics serve various purposes, including identifying data patterns, predictive modelling, forecasting, and handling complex event processing. Operationalizing analytics involves integrating them into business processes. For instance, statisticians within an insurance firm might create a model that predicts the likelihood of a claim being fraudulent. This model, combined with decision rules, could be

incorporated into the company's claims-processing system to flag claims with a high probability of fraud. These flagged claims undergo further scrutiny by an investigation unit. In some scenarios, the model itself might not be as apparent to the end user. For instance, a predictive model can be designed to identify potential customers for upselling when they contact call centres. During the customer-agent interaction, the agent would receive prompts about specific additional products to sell and market them to customers, without knowing that a predictive model is driving these recommendations. Monetizing analytics proves instrumental in optimizing business strategies, fostering informed decisions, and enhancing both bottom-line and top-line revenue. Yet, big data analytics extends beyond the insights it provides just for a single department or company. It can compile and assemble a unique data set that is valuable to other companies, as well. For instance, credit card providers leverage aggregated data to offer value-added analytics products, akin to financial institutions. Telecommunications companies are venturing into selling location-based insights to retailers. The concept revolves around amalgamating diverse data sources—billing, location, text messaging, web browsing—to extrapolate customer behaviour patterns useful to retailers, either collectively or individually.

5. Conclusions and implication

Technologies and Project: Big data demands exemplary technologies capable of swiftly and efficiently processing massive data volumes within manageable timeframes. Technologies applied to big data encompass massively parallel processing (MPP) databases, data mining grids, distributed file systems, cloud computing platforms, and scalable storage systems. Real-time or near-real-time information delivery stands out as a defining trait of Big Data Analytics. A wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyse, and visualize big data. These techniques draw from several fields including statistics, computer science, applied mathematics, and economics. Thus, organizations aiming to derive value from big data must embrace a flexible, multidisciplinary approach. Leveraging the entirety of available information within large datasets, rather than just a subset of its data, provides a formidable edge over market competitors. Big Data unlocks insights and empowers better decision-making, offering unprecedented business advantages and enhanced service delivery. The average end user accesses myriad websites and employs a growing number of operating systems and apps daily utilizing a variety of mobile and desktop devices. This translates to an overwhelming and ever-increasing volume, velocity, and variety of data generated, shared, and propagated. Effective protection hinges on a right combination of methodologies, human insight, a comprehensive understanding of the threat landscape, and the efficient processing of big data to create actionable intelligence. According to the International Data Corporation (IDC), overall data is predicted to grow by 50 times by 2030, driven in large part by more embedded systems such as sensors in clothing, medical devices and structures like buildings and bridges. The study forecasts that unstructured data—such as files, emails, and videos—will constitute a staggering 90% of all data created in the next decade.

6. ACKNOWLEDGMENT:

1. http://www.futureofprivacy.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf
2. http://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.5_2014.pdf
3. http://www.ey.com/Publication/vwLUAssets/EY-Global-Forensic-Data-Analytics-Survey-
4. 2014/$FILE/EY-Global-Forensic-Data-Analytics-Survey-2014.pdf
5. http://www.cpni.gov.uk/advice/Personnel-security1/risk-assessment/
6. https://www.google.co.in/?gfe_rd=cr&ei=rMvqVaDxDLPG8Afdx7mIDA&gws_rd=ssl#q=security+risk
7. +assessment http://www.ejst.tuiasi.ro/