Implementation of Big Data Technology using Hadoop Platform

*Mr. Tanay Gupta**
*Dr. Seema Gupta***

Abstract

Big data analytics is a set of advanced analytic techniques used against very large, miscellaneous data sets that include. Using advanced analytic techniques such as machine learning, Analytics like Text and Predictive, statistics, data mining, etc. Businesses can examine previously untouched data sources independently or together with their current enterprise data to gain new perceptions, resulting in faster and better decisions. In this research paper, the Hadoop platform provides an improved programming model that is used to create and run distributed systems quickly and efficiently to process high volumes of data. Big Data, i.e., terabytes to exabytes, consists of large datasets that cannot be managed efficiently by the common database management systems. Mobile phones, Credit cards, Radio Frequency Identification (RFID) devices, and social networking platforms create huge amounts of data that may reside unutilized on unknown servers for many years. However, with the evolution of Big Data, this data can be accessed and analysed regularly to generate useful information.

Keywords: Hadoop, mapreduce, JRE, YARN

1. Introduction:

A big data system consists of a setup that adheres to these seven Vs i.e., 1)Volume, 2)Velocity 3) Variety, 4) Veracit,y 5) Variability, 6) Value 7)Visualization, and provides a great infrastructure that can withstand the influx of huge datasets with high velocity. Meanwhile, providing an effective mechanism to process the datasets by cleansing, shaping, filtering, and sorting into meaningful information aimed towards making the data both user and machine-friendly. Beneath the complex system of architecture, sophisticated hardware, and methodologies working in conjunction with each other, lie the interfaces that are responsible for communicating with the hardware and user simultaneously, programmable applications or the tools that are the prime drivers of the efficiency of a typical big data system setup. The adoption of a contemporary technology like big data can enable the altering innovation that can bring a transition in the structure of a business, either with its services, products, or organizations. Big data frameworks are not push–button answers. For data analysis/analytics to offer value, corporations ought to have data management and the governance frameworks of big data. Complete well well-defined processes and ample skill sets for those who will be responsible for customizing, implementing, populating, and using big data solutions are also necessary. Additionally, the data quality aimed for big data-powered processing needs to be evaluated as well. The chunk of big data comes from three primary sources: machine **data, i.e.**, the data created from/by sensors installed in machinery and industrial equipment, and even logs that track user behavior. **Social data** comes from the tweets, likes, comments, retweets, video uploads, and the overall media that is shared on the world's most popular social media platforms and **transactional data** is the data generated from online and offline transactions occurring dail,y like invoices, storage records, payment orders and delivery receipts.

*** Mr. Tanay Gupta,** *Student, SRM University,*
****Dr. Seema Gupta,** *Associate Professor, IIMT, Delhi, seemagupta.iimt@gmail.com*

2. Methods and Practices for Big Data Implementation:

The market is flooded with corporations offering custom-made tools and frameworks for implementing big data and analytics. NOSQL database offers a provision for storage and extraction of the data modelled in tabular relations instead of relational databases to cater efficiently to real-time situations.

Data Management tools are available as solutions like Amazon Elastic MapReduce (EMR) that run underneath a customized version of Apache Hive, Pig, Spark, Couchbase, MapReduce, Hadoop, MongoDB, etc.

Data virtualization of multiple data sources into one helps in real-time extraction, fetching, and storage operations from multiple sources such as Hadoop and distributed data stores.

Search and knowledge finding tools and applications aid in self-service processes to extract information and new findings from humongous storage space consisting of structured/unstructured data residing in numerous sources such as databases, file systems, APIs, streams, other platforms, and applications. Stream analysis tools and applications can enrich, aggregate, filte,r and analyse a high data influx from multiple incongruent real-time data sources. Data memory composition tools provide faster access and processing of humongous data by spreading it across the dynamic RAM, SSD, or flash storage of a distributed computer system. Big data predictive analytics comprises hardware or tool-based solutions to let the organization discover, evaluate, deploy, and optimize predictive models by evaluating big data sources to better business performance and alleviate risks.

3. Applications of Big Data:

Big data's crucial role in transforming the most adverse situations for companies and organizations, or even smaller hotel chains, is no uncommon achievement for a model that is meant to failsafe you against the worst of the conditions. Big data comes with much deeper and broader applications. Big data is used to have a better understanding of a customer's behaviours, need,s and preferences. For example, a car dealership can predict when the next car is going to be sold. Walmart can predict the best-selling item at any point in time for a month in a year, or around any holiday season.

Big data is now seeping into those areas that were earlier prone to miscalculations and predictions, such as the stock inventory model where a retailer could not decide whether to stock up for the upcoming seasonal sales based on the factors around or not. Now, the same retailer can optimize their stock from the web search trends, social media data and weather forecast predictions. In supply chain or delivery route optimization, Big data is helping big time as well. Radio sensor,y along with route optimization based on the traffic data, road blockags or even live protest detectors ,are being actively used by many postal corporations. Science and research is currently under transformation by big data and its associated technique,s being actively used e.g., CERN ,the Large Hadron Collider nuclear physics world's most powerful and largest particle accelerator, is currently experimentally on the genesis of the universe in search of the elusive God particle .Big data governance is a crucial factor in dealing with the management of diverse datasets because many times, such data poses as risks, like unplanned costs, input and misleading data .Data governing tools actively deploy data pipelining technology, which enables sequential data processing where the output from one process works as an input for the next process. With these pipelines being linear or dynamic, the scale of data flexibility becomes high in data governance.

4. Technology Infrastructure Requirement

Big data is simply a large data repository, as the prime driver of an organization must be robust,scalable ,ductile, and fail-safe for unplanned situations. Another driving force behind the successful implementation of big data is the software analytical and infrastructure. Primary infrastructure is called Hadoop. Hadoop platform is an improved programming model that is used to create and run distributed systems that provide analytical technologies and computational power required to work with such a large volume of data. Hadoop is changing the conventions of big data management, especially with unstructured data. Multiple nodes in distributed environments may not always cooperate with each other through a communication system, leaving a lot of scope for errors. Apache Hadoop streamlines the excess data for any distributed processing system across computer clusters using simple programming-based models. Instead of having a hardware dependency to provide the uptime, the library is built with features at the application layer to detect and handle breakdowns, providing a reliable and always available service along with a computer cluster, since both versions may be prone to failures. Basically, The Hadoop community package consists of: – OS level and file system abstractions, the Hadoop Distributed File System which is used for storage, the MapReduce or YARN (Yet another resource negotiator) Parallel Processing of large datasets, Java Archive files (JAR), Scripts needed to start Hadoop, documentation and source code, and a contribution section.

One of the world's leading networking organizations that has transformed the way people connect, communicate and collaborate. Cisco IT developed a Hadoop Platform using Cisco@UCS Common Platform Architecture(CPA) for Big Data, which needs to use big data analytics for business advantage, including high performance,scalability, and ease of management. Cisco IT uses MapR distribution for Apache Hadoop and code written in advanced C++. Hadoop complements rather than replaces Cisco IT's traditional data processing tools. such as Oracle and Teradata. Its unique value is to process unstructured data and very large data sets far more quickly and at far less cost. The HDFS system splits the data into smaller chunks for further processing and performing ETL(Extract, Transform, and Load) operations.

Results:

The main result of transforming the business using big data by Cisco IT is that the company has introduced multiple big data analytics programs, which are based on CPA for Big Data. Many employees who work as knowledge workers in Cisco used to take a lot of time to search for the content on websites throughout the day, as most of the content was not tagged with keywords.But now Cisco IT has replaced the static and manual tagging process with dynamic tagging on the basis of user feedback. This process uses machine learning techniques to examine usage patterns adopted by users and also acts on user suggestions given for searching by new tags.

Moreover, the Hadoop platform analyses log data of collaboration tools, such as Cisco Unified Communications, email, Cisco Telepresence, Cisco Webex, Cisco Webex Social, and cisco jabber to reveal commonly used communication methods and organizational dynamics.

5. Conclusion:

Big data analytics helps corporations to utilize their data and use it in identifying new opportunities, which further leads to more efficient operations, smarter and well-calculated business moves, happier clients, and higher revenues. Big data technologies like Hadoop bring cost reduction when it comes to the storage of large data and to recognizing more efficient ways of doing business. With the evolving new age technologies and memory analytics, coupled with the ability to analyze new data sources, corporations are able to immediately analyze the information and make better and faster decisions based on the learning they derive. With the clarity to read the customer's needs and analytical satisfaction enables the power to give consumers what they want- even to the level of tailoring the solution according to the requirements of each customer individually. More such technological prowess has enabled and opened up further potential arenas of customer servicing.

6. Limitation & Future Scope of Work:

The research paper gained from the arrival of big data science laid the way for further contemporary big data projects, like weather prediction, supercollider data analytics, and other physics-based research ,astronomical science,s and data collection like planetary image detection, medical research, and others .Big data has become such a dynamic force that it doesn't apply only to sciences anymore; many businesses have got their critical data-based services hooked onto its methodologies, techniques, and objectives too which has allowed the businesses to unleash the data value that might have gone unnoticed earlier.

References:

Beyond the Hype: Big Data Concepts, Methods, and Analytics." International Journal of Information Management 35, no. 2 (2015): 13

Big Data Technologies: A Survey." Journal of King Saud University - Computer and Information Sciences 30, no. 4 (2018): 431–48.

Castineira,R.,& Metzger,A.(2018,April).The transforming transport project-Mobility meets big data. http://doi.org/10.5281/zenodo.1484954

Cavanillas, J.M., Curry, E., & Wahlster, W. (Eds.) (2016).New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe. https://doi.org/10.1007/978-3-319-21569-3

Curry, E. (2020).Real-time linked data spaces. https://doi.org/10.1007/978-3-030-29665-0

Curry, E., & Ojo, A. (2020).Enabling knowledge flows in an intelligent systems data ecosystem. In Real-time Linked Data spaces (pp.15-43). http://doi.org/10.1007/978-3-030-29665-0 2

Erl.T.,Khattak,W.,& Buhler,P.(2016),Big data fundamentals: concepts,drivers & techniques.Boston:Prentice Hall

Metzger, A., Thornton J., Valverde, F., Lopez, J.F.G., & Rublova, D. (2019a).Transforming Transport. Predictive analytics and predictive maintenance innovation via big data: The case of Transforming Transport. In the 13th intelligent Transport Systems-European Congress (ITS Europe),Brainport-Eindhoven,The Netherlands, June 3-6.

Palm,A.,Metzger,A., & Pohl,K.(2020).Online reinforcement learning for self-adaptive Information systems.In S. Dustdar,E. Yu,C.Salinesi,D. Rieu, & V.Pant (Eds.),Advanced Information systems engineering (pp.169-184).Cham:Springer.

Wadkar, S., Siddalingaiah, M., &Venner, J. (2014).Pro Apache Hadoop. Berkeley, CA: Apress Williams,S.(2016).Business Intelligence strategy and big data analytics: a general management perspective.Cambridge,MA:Morgan Kaufmann.