A Survey on Big data Analytics in Cloud Environment

Miss. Sohile Kent
M. Tech (CSE) Student
Assam Down Town University
Panikhaiti, Assam, India
sohilekent@gmail.com

Dr. Lakshmi Prasad Saikia
Professor & HOD, Computer Science& Engineering
Assam Down Town University
Panikhaiti, Assam, India
lp_saikia@yahoo.co.in

Abstract- The continuous and rapid growth in the volume of data captured by organizations, such as social media, Internet of Things (IoT), machines, multimedia, GPS has produced an overwhelming flow of data. Data creation is occurring at a record rate, referred to as big data, and has emerged as a widely recognized trend. To take advantage of big data, real-time analysis and reporting must be provided in tandem with the massive capacity needed to store and process the data. Big data is affecting organization such as Banking, Education, Government, Health care, Manufacturing, retails and eventually, the society. On the other hand, Cloud computing eliminates the need to maintain expensive computing hardware, dedicated space, and software. Cloud provides larger volume of space for the storage and different set of services for all kind of applications to the cloud customers. Therefore, all the companies are nowadays migrating their applications towards cloud environment, because of the huge reduce in the overall investment and greater flexibility provided by the cloud.

Keywords- *Big Data, Analytics, Cloud Computing, Hadoop, MapReduce*

*****

I. INTRODUCTION

Big Data Computing is a new paradigm which combines large scale compute, new data intensive techniques and mathematical models to build data analytics. The IDC [1] report predicted that there could be an increase of the digital data by 40 times from 2012 to 2020. Big Data organizes and extracts the valued information from the rapidly growing, large volumes, variety forms, and frequently changing data sets collected from multiple, and autonomous sources in the minimal possible time, using several statistical, and machine learning techniques. As Yu [2] points out, Big Data offers substantial value to organizations willing to adopt it, but at the same time poses a considerable number of challenges for the realization of such added value.

Cloud Computing is emerging today as a commercial infrastructure that eliminates the need for maintaining expensive computing hardware. It has been revolutionizing the IT industry by adding flexibility to the way IT is consumed, enabling organizations to pay only for the resources and services they use. Its infrastructure provides both computational and data processing application [3] [4]. The main reason for small to medium sized businesses to use cloud computing for big data technology implementation are hardware cost reduction, processing cost reduction, and ability to test the value of big data.

II. BIG DATA

Big data refers to a collection of large sets of data or large volume of data – either structured, semi structured or unstructured which can be stored, analyzed, manipulated to reveal or discover patterns or trends. Big Data is so large that it is difficult to store, handle, process and analyze using the traditional database technologies. Big data is being generated from numerous sources such as public web, sensor data, social media, mobile phones, cameras, microphone etc.

Big data are characterized by three aspects:

i. Data are numerous,
ii. Data cannot be categorized into regular relational databases, and
iii. Data are generated, captured, and processed rapidly.

A. Characteristics of Big Data

Big data is not only characterized by the three shared characteristics i.e Volume, Variety and Velocity but may also extend to five Vs, namely, volume, variety, velocity, value and veracity [5].

The terms volume, variety, and velocity were originally introduced by Gartner [6] to describe the elements of big data challenges. IDC also defined big data technologies as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis."

a) Volume – it refers to the amount of data that is being generated by organizations or individuals every second.

b) Variety – it refers to the ability to process data from different sources and formats,both structured and unstructured.

c) Velocity – the speed at which data is generated, captured and shared.

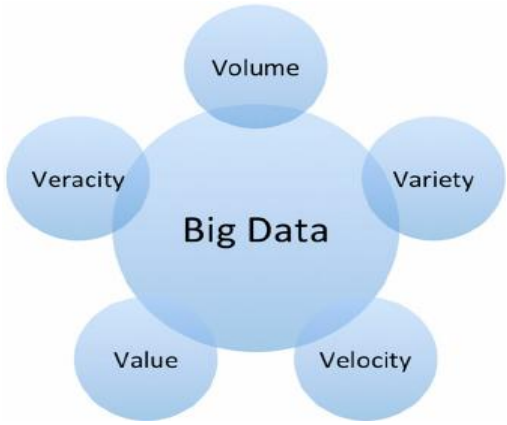

Figure 1. Five Vs of Big Data

d) Value – it refers to our abilityto turn our data into value. i.e the process of discovering huge hidden values from large datasets with various types and rapid generation.

e) Veracity - the correctness and accuracy of information

This 5V definition is widely recognized because it highlights the meaning and necessity of big data. Because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big data is creating a new generation of decision support data management.

III. BIG DATA ANALYTICS

Big data analytics is a process that works with large volumes of heterogeneous data. It uses sophisticated quantitative methods to explore the data and to discover valuable information. It is data which is huge in size (volume), variety, and generated at high velocity and this data may be structured and/or unstructured. According to NIST [7] big data analytics involve the following characteristics: Value: when the data is analyzed, veracity: which measures timeliness, accuracy and quality of data, latency between availability and measurement of data and cleanliness of data.

The term Analytics (including its Big Data form) is often used broadly to cover any data-driven decision making [8]. People are analyzing data to extract information and knowledge useful for making new discoveries or for improving the business by smart decision. This can be done by exploiting big data analytics techniques and tools. Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable.

The key to deriving value from big data is the use of analytics. Collecting and storing big data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value. Big data and analytics are intertwined. Big data analytics helps understand what customers want to buy and what they don't like about your products or services.

A. Techniques for analyzing big data

For storing and processing massive amount of data, big data requires different techniques like:

a. Association Rule Learning
b. Genetic algorithms
c. Machine learning
d. Regression analysis
e. Sentimental Analysis
f. Social network analysis.

1) Association Rule Learning

This rule is sometimes referred to as "Market Basket Analysis"; it consists of data present and then finding the relationship between the data present. i.e this rule is used for finding a set of items that occurs frequently in a data items. This rule is widely used in various areas such as marketing, catalog design, sale campaign analysis, web log analysis etc.

As per Techopedia, "Association rule is a procedure which is used to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories".

The most common example of this rule is "if a customer purchases <bread, egg> from a market, then they are likely to purchase <butter, milk>".

2) Genetic algorithms

Genetic Algorithms (GA) was first proposed by John Holland in 1975. GA [9] is used to obtain optimized solutions from a number of candidates. It uses the principles of selection and evolution to produce several solutions to a given problem. GAs is also being developed to optimize placement and routing of cell towers for best coverage and ease of switching.

3) Machine Learning

Machine learning is a science of finding patterns and making predictions from data based on already-identified trends and properties in the training data set. In Machine Learning, a system learns from past or present experiences and is able to build a model which would most likely be able to comprehend future instances. It is applied in manufacturing, retail, healthcare and life sciences, travel and hospitality, financial services etc to gain deeper insights and to improve decision making.

4) Regression Analysis

Regression analysis is a form of predictive modeling technique which estimates the relationship between dependent variable and independent variable. This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

5) Sentimental Analysis

It is also known as Opinion Mining. It refers to the use of natural language processing tools to study the emotions,

attitude, and opinion of people, customers to identify and extract valuable information [10]. It is especially useful in social media monitoring as it allows us to review how people feel about certain topics and how they are responding to it. E.g. suppose a new movie is released, it can be used to review how a customer feels about the new movie like – awesome, good, interesting, bad, dreadful, boring etc.

6) Social Network Analysis

A social network is an interconnection of nodes and links. The nodes represent people or organization (actors), and the links are relationships or interdependencies between the actors/ nodes.

Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other information/knowledge processing entities." (Valdis Krebs, 2002).

Social network analytics is one of the techniques that can be used to determine the influence of an individual amongst others. SNA provides both a visual and a mathematical analysis of human relationships. Example Facebook use basic elements of SNA to identify and recommend potential friends based on friend of friends.

B. Tools and Technologies for Big Data

Lots of tools and technologies have emerged to enable large amount of unstructured data. Some of these are discussed below:

1) Hadoop

Hadoop is written in Java and is a top-level Apache project created *and started* by Doug Cutting and Mike Cafarella in 2006. It is a open source framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is designed in such a way that it can scale up from single servers to a larger number of machines each of them offering local computation and storage. Currently it is used by Google, Facebook, Yahoo!, Twitter etc. The main component of Hadoop is:

a. Hadoop Distributed File System (HDFS)
b. Yarn
c. MapReduce
d. Hadoop Common

2) Hadoop Distributed File System (HDFS):

HDFS is a distributed file system that provides big data storage solution through high-throughput access to application data. It is designed for storing very large files that are hundreds of TB or PB. It also implements write-once, read-many-times pattern, designed more for batch processing rather than interactive use by users. Compared to other distributed file system, HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications that have large data sets.

It uses Master/ Salves architecture. It breaks the data into smaller parts which are provided as input and then distributes it to various nodes in the cluster which allow for the parallel processing of data. Each piece of data is copied multiple times by the file system and then the data is distributed to individual nodes and also placing at least one copy on different server rack from all others.

3) MapReduce

It is a software framework with which we can write an application to process large amount of data, parallel on large clusters of commodity hardware in a very reliable manner. Many real world tasks are expressible in this model [11]. The computation of MapReduce takes a set of input key/value pairs, and produces a set of output key/value pairs. The two functions are: Map function and Reduce function. Map, is written by the user, takes an input pair and produces a set of intermediate key/value pairs. The Reduce function is also written by the user, accepts all values associated with the intermediate key and a set of values for that key. It merges together these values to form a possibly smaller set of values. Typically just zero or one output value is produced per Reduce invocation. This allows us to handle lists of values that are too large to fit in memory.

MapReduce is built on the concept of divide and conquer, it's much faster to break a massive task into smaller chunks and process them in parallel. This technique works with both structured and unstructured data

4) Sqoop

Sqoop is a command-line interface which is used for transferring data between relational databases and Hadoop. Sqoop got the name from sql-to-hadoop. It is used for import SQL-based data or tables to Hadoop, it also generates class that allows us to interact with the imported data etc. It is an open source framework provided by Apache.

5) Hive:

It is an SQL-based data warehousing application for summarization, querying and analyzing large datasets stored in Hadoop HDFS. Developed at Facebook. It runs SQL like queries called HQL (Hive query language) which gets internally converted to map reduce jobs.

6) Pig

It is a high level scripting language that is used with Hadoop for analyzing large data sets. It was developed at Yahoo! It consist of two components.

a. The Scripting Language called PigLatin
b. The runtime Environment for executing PigLatin program.

The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Every task which can be achieved using PIG can also be achieved using java used in Map reduce.

7) HBase

Hbase is an open source and non-relational (NoSQL) database that runs on top of HDFS and written in java. It is column oriented and horizontally scalable. It provides real-time read/write access to the users data in Hadoop. Like the traditional database, it consist of sets of table, each table containg rows and columns.

8) ZooKeeper:

It is an application programming interface (API) that provides centralized service for maintaining configuration across nodes, synchronizing process execution and implementing reliable messaging etc for distributed systems.

IV. CLOUD COMPUTING

Cloud computing is the delivery of computing services—servers, storage, databases, networking, software, analytics and more—over the Internet ("the cloud") [12]. The need to store, process, and analyze large amounts of datasets has driven many organizations and individuals to adopt cloud computing. A large number of scientific applications for extensive experiments are currently deployed in the cloud and may continue to increase because of the lack of available computing facilities in local servers, reduced capital costs, and increasing volume of data produced and consumed by the experiments. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available.

Figure 2. Cloud Computing Network

A. *Cloud Deployment Model*

Cloud Deployment models typically consist of PaaS, SaaS, and IaaS [5].

a) PaaS (Plateform as a service) - In the PaaS models, cloud providers deliver a computing platform, typically including operating system, programming-language execution environment, database, and web server. Example of PaaS provider is Google's Apps Engine, and Microsoft Azure.

b) Saas (Software as a service) - users gain access to application software and databases. Example of SaaS, provider are Google Docs, Gmail, Amazon* Elastic MapReduce.

c) IaaS (Infrastructure as a service) - refers to online services that abstract the user from the details of infrastructure like physical computing resources, location, data partitioning, scaling, security, backup etc. Example of IaaS provider is Citrix* CloudPlatform and Amazon's EC2.

V. BIG DATA IN THE CLOUD

Big data is mostly associated with the storage of huge loads of data, it also concerns ways to process and extract knowledge from it (Hashem et al., 2014). The cloud has helped make big data more efficient and effective.

Cloud computing offers a cost-effective way to support big data technologies and the analytics that drive business value. Cloud computing and big data are complementary because it delivers scalability, fault tolerance and availability through hardware virtualization. Cloud's ability to virtualized resources allows abstracting hardware, requiring little interaction with cloud service providers and enabling users to access terabytes of storage, high processing power, and high availability.

Since the cloud virtualizes resources in an on demand fashion, it is the most suitable and compliant framework for big data processing, which through hardware virtualization creates a high processing power environment for big data. Taking big data to the cloud offers up a number of advantages, including improved performance, targeted cloud optimizations, more reliability, and greater value.

Big data in the cloud gives businesses the type of organizational scale many are searching for. This allows many users, sometimes in the hundreds, to query data while only being overseen by a single administrator. That means little supervision is required.

Big data in the cloud also allows organizations to scale quickly and easily. This scaling is done according to the customer's workload. If more clusters are needed, the cloud can give them the extra boost. During times of less activity, everything can be scaled down. This added flexibility is particularly valuable for companies that experience varying peak times.

Big data utilizes distributed storage technology based on cloud computing rather than local storage attached to a computer or electronic device.

VI. CONCLUSION

This paper gave a brief description about big data analytics and the different tools and techniques use for processing and analyzing big data. Big Data is an umbrella term for a high velocity, high volume, high variety and veracity of data that is

difficult to manage by traditional solutions. There are various benefits in moving to cloud resources from dedicated resources for data management because cloud provides cost saving in hardware and processing etc.

REFERENCES

[1] IDC, The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, www.emc.com/leadership/digital-universe/index.htm.

[2] P.S. Yu, On mining big data, in: J. Wang, H. Xiong, Y. Ishikawa, J. Xu, J. Zhou (Eds.), Web-AgeInformation Management, in: Lecture Notes in Computer Science, vol. 7923, Springer-Verlag, Berlin, Heidelberg, 2013, p. XIV.

[3] D. Agarwal, S. Das, and A. E. Abbadi. Big data and cloud computing: New wine or just new bottles? PVLDB, 3(2):1647-1648

[4] Tharam Dillon et. al. "Cloud Computing: Issues and Challenges", 2010 24th IEEE International conference on Advanced Information Networking and Application, 2010 IEEE DOI 10.1109/AINA.2010.187. pp-27-33

[5] De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2016). "A Formal definition of Big Data based on its essential Features". *Library Review*. **65**: 122–135. doi:10.1108/LR-06-2015-0061

[6] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.

[7] NIST, http://www.nist.gov/itl/ssd/is/upload/NIST-stonebraker.pdf, accessed on September 2015.

[8] D.Fisher, R.Deline, M.Czerwinski and S. Drucker,"Interaction with big data analytics", Volume 19, No.3, May 2012.

[9] https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/?cdn=disable

[10] Kim S-M, Hovy E (2004) Determining the sentiment of opinions In: Proceedings of the 20th international conference on Computational Linguistics, page 1367.. Association for Computational Linguistics, Stroudsburg, PA, USA.

[11] http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

[12] Cloud Computing Portability and Interoperability : Cloud Computing Portability and Interoperability http://www.opengroup.org/cloud/cloud/cloud_iop/cloud_port.htm