A Survey on Hadoop: Solution of Big Data Processing on Cloud

Janvi Patel[1] Nirali Mankad[2]
[1]P. G. Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]Noble Group of Institution, Gujarat, India

Abstract— Big data is an emerging paradigm applied to datasets whose size or complexity is beyond the ability of commonly used computer software and hardware tools. One of the key drivers of Big Data is Hadoop, combination of HDFS a distributed file system, MapReduce data processing and Resource Manager (YARN) for allocating resource on cluster for Job execution. Datasets are often from various sources (Variety) unstructured such as social media, sensors, scientific applications, surveillance, video and image archives, Internet texts and documents, Internet search indexing, medical records, business transactions and web logs and are of large size (Volume) with fast data in/out (Velocity). More importantly, big data helps in business decision making (Veracity) and gaining insight in real time which is hard to achieve using traditional system. As estimated that about 40% data globally would be touched with Cloud Computing. Cloud Computing provides strong storage, computation, network and distributed capability in support of Big Data processing. Every entity on cloud is a virtual entity created by underlying virtualization technique. On demand, ease of use, elastic, automated service, PAYG properties helps building proves beneficial for auto scaling the cluster in Hadoop. This paper presents the survey of big data, issues with big data, how Hadoop works.

Key words: Big Data, Hadoop, HDFS, Map Reduce, Cloud Computing

I. INTRODUCTION

Business Intelligence is useful in decision making. In the digital age, the previous day data is treated as archival data as the BI decisions are now based on data generated before minutes and seconds. There is a need to design a new platform or to restructure the existing one to analyze the huge amount of day generated per day. Big Data is the term which refers to the type (Variety), size (Volume), speed (Volume) of the data. Big data brings this technology change in terms of storing, retrieving, analyzing and visualizing the data.

Based on GFS, Hadoop an open source java implementation of GFS was developed. Hive, Sqoop, Zookeeper, HBase, Strom, Kafka, Yarn, and many others services are part of the Hadoop ecosystem. There is challenge at the hardware side which is unable to store the entire data on a single system leading to distributed storage. Handling of such cluster of hardware in terms of high availability, maintenance, security, fault tolerance is a huge challenge. Its downtime leads to unavailability of data to customer which results in loss of business. These challenges are addressed by cloud computing paradigm which serve scalability, on demand, orchestration and measured usage of resources.

II. BIG DATA

"Big Data" is buzzword used to describe the voluminous of structured, semi-structured and unstructured data coming from heterogeneous and independent sources. Big data is so large that is difficult to be captured and processed or analyzed within the time necessary to make them useful using traditional database or conventional tools and technologies. Despite these problems, big data has the potential to help companies improve operations and make faster, more intelligent decisions.

III. BIG DATA CHARACTERISTICS

Big Data as having three dimensions: volume, variety, and velocity. Thus, IDC defined it: "Big data technologies describe a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis." [2] Two other characteristics seem relevant: value and complexity. We can describe the big data characteristics using following five Vs [3]:

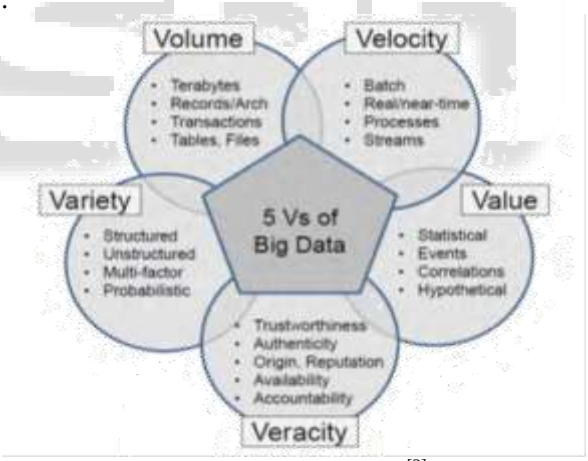

Fig. 1: 5vs of Big Data[3]

A. Volume

Volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. This volume presents the most immediate challenges to conventional IT structure

B. Velocity

Velocity refer to the speed with which new data is generated, streaming, and aggregation. So, the speed at which the data is stored processed and analyzed by traditional database.

C. Variety

Variety is a measure of the richness of the data representation. Big data comes from different variety of

formats, types, and structures. Big Data has in three types namely structured, unstructured and semi structured data.

D. Veracity

There will be dirty data when we are dealing with high volume, velocity and variety of data, so it is not possible that all the data is going to be 100% correct. As a result quality of data can vary greatly. The veracity of the source data upon which accuracy of analysis is depend.

E. Value

Value is the most important aspect in the big data. Big data is useless unless and until it turn into value.

We suggest there are three fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues.

IV. HADOOP: SOLUTION FOR BIG DATA

One of the best open source tools used in the market to utilize the distributed architecture in order to solve the data processing problems is Apache Hadoop. The Apache Hadoop is developed for data intensive jobs. Hadoop is programming framework use for writing and running distributed application that process very large dataset. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is projected to support parallel processing using the Google's MapReduce programming model. Here the concept is that the data is distributed across the cluster and the processing performed on the nodes where the data resides. MapReduce is the programming model used to process this distributed data.

Hadoop framework consists on two main layers: Hadoop Distributed file system (HDFS), MapReduce layer. It automatically handles data replication and node failure. Hadoop includes a Distributed File System that stores large amount of data on cluster called as Hadoop Distributed File System (HDFS) and MapReduce is the data processing component of Hadoop.

Hive, Sqoop, Zookeeper, HBase, Strom, Kafka, Yarn, and many others services are part of the Hadoop ecosystem.

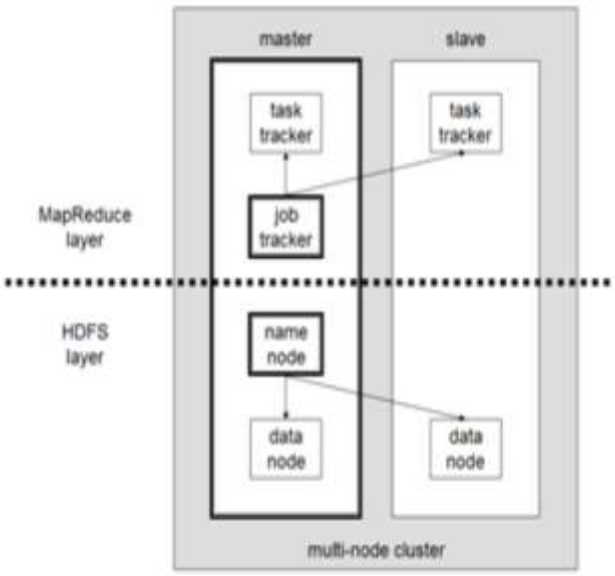

Fig. 2: Hadoop Architecture [6]

A. Hadoop Distributed File System (HDFS):

It is distributed file system which stores files into predefined block of chunks. These chunks are replicated based on replication factor (by default 3) on different nodes on the cluster. It supports rack awareness through which the nearest rack having the requested data is returned to the user. This file system is based on WORM (write once read many time) mechanism. The namenode process in the HDFS manages the metadata of the HDFS and the image of the file system. The datanode is the process running on the nodes where actual data resides on the HDFS. The datanodes sends heartbeats to namenode which is used by namenode to check the health of the node.

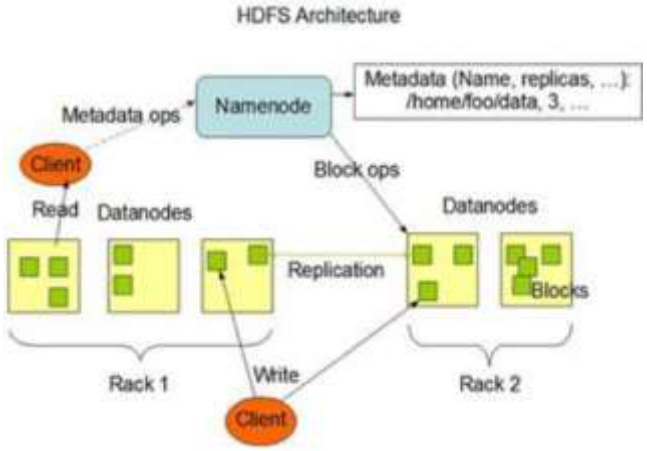

Fig. 3: HDFS Architecture [5]

B. MapReduce

It is a parallel processing programming model for large data sets. It has two functional parts Mapper and Reducer. The Mapper is the first part to execute & it transforms a piece of data into some number of <key, value> pairs, so the resultant output of the Mapper is the intermediate data which is provided as an input to the Reducer. Reduce function is used to merge the values into a single result.

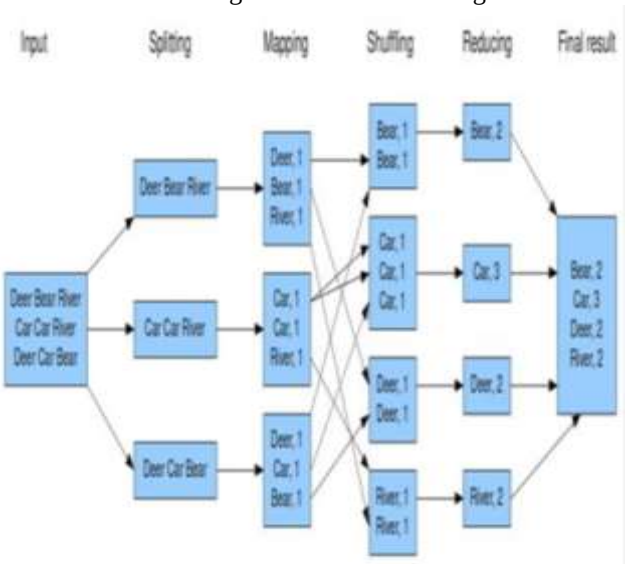

Fig. 4: MapReduce Architecture

V. CLOUD COMPUTING

Simply cloud computing provides a variety of computing resources, from servers and storage to enterprise applications, all delivered over the Internet. Cloud

computing is aimed at providing IT as a service to the cloud users on-demand basis with utility computing model. It takes away the burden of installing software's, and the burden of managing huge networks. Cloud computing is a computing paradigm shift where computing is moved away from personal computers or an individual server to a "cloud" of computers. This method of distributed computing achieved through pooling all computer resources together and being managed by software rather than a human and saving corporations money, time and resources.

Once a cloud is established, how its cloud computing services are delivered will depending on requirements. The primary service models are:

A. Software as a Service (SaaS)

Consumers purchase the ability to access and use an application or service that is hosted in the cloud. A benchmark example of this is Salesforce.com.

B. Platform as a Service (PaaS)

A cloud platform offers an environment on which developers create and deploy applications. Consumers purchase access to the platforms, enabling them to deploy their own software and applications in the cloud. The operating systems and network access are not managed by the consumer. A benchmark example of this is Google AppEngine and Microsoft Azure.

C. Infrastructure as a Service (IaaS)

Cloud Infrastructure offering virtualized resources (computation, storage, and communication) on demand is . Consumers control and manage the systems in terms of the operating systems, applications, storage, and network connectivity, but it does not itself control the cloud infrastructure. A benchmark example of this is Amazon Web Services.

VI. LITERATURE SURVEY

A. Big Data: Issues and Challenges Moving Forwar

This paper defines big data and it's important. Also initiates a collaborative research effort to begin examining big data issues and challenges. They have identified some of the major issues such as storage, management and processing. They have also identified some major challenges with big data as (1) design appropriate systems to handle the data effectively and (2) and analyze it to extract relevant meaning for decision making. [1]

B. Big Data Analysis Using Apache Hadoop

This paper describes the paradigm for processing huge data sets. As with the large amount of data they gather it is found that the data cannot be processed using any of the existing centralized architecture. Apart from the time issue there is also some other issues like efficiency, performance and infrastructure cost with centralized environment. These papers describe the paradigm shifted from centralized architecture to distributed architecture. Also stated that one of the best open source tools used in the market to harness the distributed architecture in order to solve the data processing problems is Apache Hadoop.[2]

C. A Review Paper on Big Data & Hadoop

This paper describes the concept of Big Data along with its characteristics such as volume, velocity and variety. This focus on problems like Heterogeneity and Incompleteness, Scale, Timeliness, Privacy and Human Collaboration. It also stated that term Big Data brings changes to techniques and technologies in terms of to capture, store, distribute, manage and larger-sized datasets with high-velocity and different structures. It requires new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

It describe Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers.[5]

D. Strategic alignment of Cloud-based Architectures for Big Data

This Paper describes the influence of Big Data on the enterprise architecture. It also describes different ways of integrating Big Data with existing enterprises. As with the advancement in cloud computing Big Data can be realized using cloud services, so, alternatives for implementing Big Data applications using cloud-services. New data sources can be integrated quickly and with an extremely low effort. Data and information bases are available on demand and on low price. The data supply is scalable: additional information can be obtained easily.[7]

E. Big Data computing and clouds: Trends and future directions

This paper discusses approaches and environments for carrying out analytics on Clouds for Big Data applications. As the large amount of data currently generated by the various activities of the society and even it is being generated in increasing speed.. This Big Data trend is being seen by industries as a way of obtaining advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more costumers, increase the revenue per customer, optimize its operation, and reduce its costs. Nevertheless, Big Data analytics is still a challenging and time consuming task that requires expensive software, large computational infrastructure, and effort. Cloud computing helps in alleviating these problems by providing resources on-demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand.[8]

F. A study of Big Data solution using Hadoop to Process Connected Vehicle's Diagnostics Data

Recently, with period things are connected to the Internet, then why vehicle leaved back. Without doubt connected vehicles will generate a huge amount of vehicle's diagnostic data which will be sent to remotely servers or cloud providers. As the amount of data increases, automotive system will encounter difficulties to perform analysis of data gathered from connected vehicle for further services. This paper describe Apache Hadoop framework and its systems particularly Hive, sqoop have been deployed to process vehicle diagnostic data and generate useful information that

may be used to provide services to car owners. But here in this, Hadoop is not implemented on Cloud.[9]

VII. CONCLUSION

In this paper we presented the brief overview of big data and its issues such storage, processing and management. Also described how emergence of Hadoop and cloud solved Big Data storage and processing issues. . In this paper we have also presented survey of progress in Hadoop as a solution of Big Data processing on cloud.

The progress can extend by focusing on implementing Hadoop framework for Big Data processing on cloud.

REFERENCES

[1] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", Hawaii International Conference on System Sciences, IEEE, 2012
[2] Gantz, J. and E. Reinsel. 2011. "Extracting Value from Chaos", IDC's Digital Universe Study, sponsored by EMC
[3] Ishwarappa, Anuradha, "A Brief Introduction of Big Data 5Vs Characteristics and Hadoop Technology", International Conference on Intelligent Computing, Communication & Convergence, 2015
[4] Krunal Dav1, Mr.Jignesh Vania, " A Survey on Big Data Processing using Hadoop Components" International Journal for Scientific Research & Development , Volume 03– No.12, 2015.
[5] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, "A Review Paper on Big Data & Hadoop" International Journal of Scientific and Research Publications (IJSRP), 4 (10) , 2014
[6] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach ", IEEE, 2012
[7] Rainer Schmidt, Michael Möhring, "Strategic alignment of Cloud-based Architectures for Big Data", IEEE , 2013.
[8] Marcos D. Assunção , Rodrigo N. Calheiros , Silvia Bianchi , Marco A.S. Netto ,Rajkumar Buyya, "Big Data computing and clouds: Trends and future directions", Elsevier, 2014.
[9] Lionel Nkenyereye, Jong-Wook Jang, "A study of Big Data solution using Hadoop to process connected Vehicle's Diagnostics data", Springer, 2015.
[10] Muhammad Adnan, Muhammad Afzal , Muhammad Aslam , Roohl Jan, Martmez-Enrlquez A.M. , "Minimizing Big Data Problems using Cloud Computing Based on Hadoop Architecture" ,IEEE, 2015.
[11] S. Vikram Phaneendra, E. Madhusudhana Reddy, BigData - Solutions for RDBMS Problems – A Survey", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) , 2(9), 2013.
[12] Sae Song, "Storing Big Data—The Rise of the Cloud", 2012.
[13] Trapti Sharma, " Modelling Cloud Services for Big Data using Hadoop" , International Journal of Computer Science and Information Technologies (IJCSIT), 6 (2) , 2015.
[14] Noman Islam, Aqeel-ur-Rehman, "A comparative study of major service providers for cloud computing", (2013).
[15] A. Li, X. Yang, S. Kandula, and M. Zhang, "CloudCmp: Comparing Public Cloud Providers",presented at 10th ACM SIGCOMM conference on Internet measurement, Melbourne, Australia, 2010.
[16] Karthik Kambatla, Abhinav Pathak Saumitra Vaidya, Himabindu Pucha, "Towards Optimizing Hadoop Provisioning in the Cloud".
[17] Sanjay P. Ahuja, Bryan Moore, "State of Big Data Analysis in the Cloud", Network and Communication Technologies, 2(1), 2013.
[18] Saurabh Kumar Garg Steve Versteeg , Rajkumar Buyya, "A framework for ranking of cloud computing services", Elsevier, 29 (2013).
[19] Ivanilton Polato, Reginaldo Ré, Alfredo Goldman, Fabio Kon,"A comprehensive view of Hadoop research—A systematic literature review ", Elsevier, 2014.
[20] Ishwarappa, Anuradha, "A Brief Introduction of, Big Data 5Vs Characteristics and Hadoop Technology", International Conference on Intelligent Computing, Communication & Convergence, 2015.
[21] V. Nappinna lakshmi, N. Revathi, "DATA MINING OVER LARGE DATASETS USING HADOOP IN CLOUD ENVIRONMENT ", International Journal of Computer Science & Communication Networks, 3(2), 2015, 73-78.
[22] Krishna Kumar Rathore, Rakesh Patel, Girdhari Patel, "Agenda of the cloud computing ", IJSRD - International Journal for Scientific Research & Development, 2(10), 2014.

