Integrating Big Data in Cloud Environment: A Review

Richa Mathur
Ph.D., Computer Science and Application

Abstract-- Cloud Computing is a powerful model that is capable to process and store massive amount of data. Cloud Computing offers on-demand elastic, scalable and cost-effective resources with eliminating the need to maintain expensive hardware, software and costly space on cluster of servers. To efficiently manage and analyze Big Data is time-consuming and challenging task. This gives rise to incorporate Big Data in Cloud environment to take advantage of both technologies. This paper presents a review on the integration of Big Data in Cloud environment. Big Data and its characteristics, Cloud Computing with its features are introduced. The reason to integrate these two technologies is discussed with some examples. Further some challenges offered by Big Data in cloud are presented. Then some tools and techniques to manage Big Data in Cloud platform are discussed as a solution.

Index Terms-- Big Data, Cloud Computing, Big Data Cloud Provider, Challenges, Big Data Management tools

I. INTRODUCTION

The term "Big Data" describes the large amount of data sets that is so complex that it is not possible to process them via conventional methods and technologies. Such a large volume of data can be generated from various sources as: log files, mobiles, social networking sites, transactions, sensors, satellites etc. Where "Cloud" in Cloud Computing means "The Internet" which describes operations and services are provided using Internet with its benefits such as: unlimited, on-demand, elastic, computing and data storage resources. And Big Data environment requires multiple servers that are capable to process large data volume, high velocity and varied formats of Big Data. Some real-time applications that support Cloud Computing are: Google Calendar, Gmail, Google Docs etc. Many organizations like Amazon AWS, IBM Smart Cloud, and Windows Azure are now migrating their Big Data to Clouds to take their advantage.

A. What is Big Data?

Now-a-days Data is becoming more valuable but how to handle data and finding hidden facts from
it is more important. Big Data is a wide term for any voluminous and compound datasets i.e. too large, fast growing, and difficult to handle by using conventional tools

and techniques [2]. Big Data can be generated via various sources like mobile devices, sensors, audio and video inputs and social media, are all increasing the volume and variety of data [1]. Big Data have the potential to give valuable information after processing that can be discovered through deep analysis and efficient processing of data by decision makers. To extract valuable insights from such varied and rapid growing datasets, various tools and techniques of Big Data Analytics can be used that may lead to better decision making and strategic planning. Big Data technology relies on massively parallel processing, in-database execution, storage optimization and mixed workload management.
Big Data has leading five V's of Big Data as:

Volume: Refers to vast amount of generated and stored data not in Terabytes but Zettabytes or Yottabytes. The "size" indicates usefulness of data and its potential to be considered as "Big Data" or not.

Variety: Refers to different types of data that we use like structured, semi-structured, unstructured and raw data. Wide data variety requires different techniques and approaches to store all type of raw data.

Velocity: Refers to the "speed" at which new data generate (data in or out) and at which it moves around.

Variability: Refers to the presence of "variance" or inconsistency at the time of data analysis or at time of preparation of summaries. It is difficult to efficiently handle it by data analyst.

Veracity: Refers to "uncertainty" due to inconsistency, incompleteness and doubted data.

B. What is Cloud Computing?

Cloud Computing is a successful paradigm of service oriented computing, capable in providing on-demand, Pay-as-you-use, secure storage management, easy and agile development, and ubiquitous access of computing resources to its shared remote users for data storage and
processing [3]. The world of Cloud Computing is totally virtual to its users that require minimum effort from user to manage with features like: on-demand, scalability, reliability, maintenance, cost-effective, and flexibility. The services are delivered to users of it through the use of Internet and sharing of resources can be done using network of remote servers to store, manage and process data with distributed data processing system. Its service-oriented architecture supports "everything as a service", offers their "services" according to different models with infrastructure-, platform- and software-as-a-service.

II. WHY INTEGRATE BIG DATA AND CLOUD COMPUTING?

Both technologies are continually evolving and are complementary technologies. The benefits of using them in an integrated way are: scalability, agility and elastic on-demand availability of data [2]. Environment of Big Data need cluster of servers to maintain the tools capable to process large data volumes with high velocity and varied formats. This type of service is offered by cloud computing in a cost-effective way with deployment of cluster of servers, storage and networking resources that can scale us or down as needed. With the use of Cloud Computing a single server can serve multiple customers to retrieve and update their data without paying for different applications.

Fig i: Big Data and Cloud Computing integration

Cloud Computing technology delivers parallel-processing as a solution for handling Big Data technologies with advanced analytical application. Several benefits can be achieved using Infrastructure-as-a-Service (IaaS) in cloud such as: reduce data centre costs, improve utilization, secure and scalable data solution.

III. PROVIDERS IN CLOUD FOR BIG DATA

An ideal computing environment can be created using cloud and offering many different products for Big Data users. IaaS requires more investment of IT resources in implementing Big Data analytics with installation of software like: Hadoop framework, NoSQL database as Cassendra, MongoDB etc. Some examples of such type of cloud providers with IaaS for Big Data include: Amazon.com, AT&T and IBM. Some examples of using Cloud with Big Data:

A. IaaS in Public Cloud

Using IaaS provider can be capable to create on-demand virtual machines with unlimited storage and large processing power. The infrastructure of public cloud provider would be used for Big Data services as anybody doesn't want to use their infrastructure. An example: Amazon Elastic Compute Cloud (Amazon EC2) service to run real-time predictive model, requires parallel-processing of massively distributed data in a scalable manner [4].

B. PaaS in a Private Cloud

Paas provides tools and libraries to its developers in cloud to fast develop, run and deploy applications in a private or public cloud without worry about maintaining complexities of Hadoop like implementation environment. PaaS integrated with Big Data is a fully packaged infrastructure that includes Big Data software, infrastructure, tools and managed services. Using PaaS enterprises can rapidly develop secure tools and techniques to Big Data analytics applications. PaaS developers are moving to enhance capabilities of Hadoop and MapReduce like Big Data analytics applications [8]. An example: Google Cloud Engine offers cloud based capabilities for virtual machine computing with secure and flexible environment.

C. SaaS in Hybrid Cloud

SaaS can be provided as a standalone application or a solution to its developers by cloud. SaaS provides specific cloud-based application to its customers as required or analyzed by them. Applications that are required by business users are merged into SaaS and these applications are provisioned to its users in a Pay-as-you-go fashion. An example: Amazon Elastic MapReduce that provides a Hadoop framework with easy, fast, and cost-effective processing of vast amount of data across dynamically scalable Amazon Cloud Computing instances [7].

IV. CHALLENGES IN BIG DATA CLOUD ENVIRONMENTS

Although Cloud Computing is widely accepted by many organizations but several challenges are introduced when using Big Data in cloud environments. Some of them are:

A. Collection

The major problem of Big Data is its size that is rapidly growing at an exponential rate. Getting data or uploading data to cloud is a major problem as the data size is too large. Therefore it needs new tools and techniques with efficient data movement into cloud.

B. Storage

Handling large amount of growing datasets in an suitable manner, needs an appropriate scalable distributed data storage system in cloud infrastructure. Traditional RDBMS systems are not suitable to take advantage of cloud scalability. The use of NoSQL database to store and retrieve large volume of distributed datasets provide scalable, easy replication support, schema-free, consistent and flexible modes system.

C. Analysis

Big Data Analytics can discover valuable information, interrelationships and patterns as hidden facts from overwhelming amount of datasets that are capable to provide us insights and knowledge that lead to better decision making and strategic planning. An example of it is: Apache Hadoop framework that provides distributed

applications running on cluster of servers for data storage and processing with fault tolerance. Existing tools and technique requires enhancing their capabilities in processing or incorporating large datasets with ease of customization.

D. Security and Privacy

As the volume of Big Data is increasing and gives us valuable insights and information, the privacy and Security concern related to accessing and analyzing personal data is also increasing. In case of Cloud Computing resources are distributed to its clients by its service provider, there exists a possibility of unauthorized person gaining access of private user information or sensitive information related to data administrator. They can misuse such sort of private information by showing them a legitimate user. So privacy and Security of such type of data becomes biggest challenge when mining and analyzing data related personal and location-based information. Therefore it require strong and complex data encryption techniques that is high in processing time and, bandwidth and hard to crack.

V. BIG DATA MANAGEMENT TOOLS IN CLOUD

Big Data produces big challenge to manage massive amount of structured and unstructured data to handle. Cloud Computing offers scalable solutions to manage such a large amount of data in cloud environment to take advantage of both technology. To effectively incorporate and manage Big Data in cloud environment it is important to understand tools and services offered by them. Some vendors like Amazon Web Service (AWS), Google, Microsoft and IBM offers Cloud based Hadoop and NoSQL database platforms that are supporting Big Data applications. In addition to, many cloud providers offer their own Big Data services as: AWS's Elastic MapReduce, Google's BigQuery etc. Most of the cloud service provider offers Hadoop framework that scale automatically on-demand of customers for data processing.

Hadoop

Hadoop provides an open-source software framework for distributed storage and processing applications on very large datasets, written in java. Hadoop platform includes higher level declarative languages for writing queries and data analysis pipelines. Hadoop is used by approximately 63% of organizations to manage and analyze huge number of unstructured logs and events (Sys. Con Media, 2011). Hadoop is composed of many components but in Big Data two mostly components Hadoop Distributed File System (HDFS) and MapReduce are used. The other components provide complementary services and higher-level of abstraction.

MapReduce MapReduce system is the main part in Hadoop framework that is used for processing and generating large datasets on a cluster with distributed or parallel algorithm. It is a programming paradigm used to process large volume of data by dividing the work into various independent nodes. A MapReduce program corresponds to two jobs, A Map() method which include obtaining, filtering & sorting datasets and A Reduce() method which include finding out summaries and generate final result. MapReduce system arranges distributed servers, manage all communications, parallel data transfers, also provide redundancy and fault tolerance.

Hadoop Distributed File System (HDFS) HDFS is used to store large data files that are too much to store on a single machine typically in gigabyte to terabyte. HDFS is a distributed, scalable and portable file system written in java for Hadoop framework [7]. It maintains reliability by replicating data across multiple hosts to facilitate parallel processing, for that it split a file into blocks that will stored across multiple machines. The cluster of HDFS has master-slave relationship with single namenode and multiple datanode.

Cassandra and HBase Both are open-source, non-relational, distributed DBMS written in java, supports structured data storage for large tables and runs on top of HDFS. It is columnar data model with features like compression, in-memory operations and provides fault tolerance way of storing large quantities of sparse data.

Hive It is a warehouse infrastructure by facebook providing data summarization, adhoc querying and analysis. It provides SQL like language (HiveQL) to make powerful queries and get results in real time.

Pig It is a high-level data flow language (PigLatin) and execution framework for parallel computation.

Zookeeper It is a high performance coordination service for distributed application that can store configuration information and have master-salve node.

VI. OPEN PROBLEMS

As the amount of data is continuously growing everyday that produces big challenge to handle such an increasing amount of data to collect, transfer, store, process, and analyze securely within real-time. Many organizations are using Big Data analytics and cloud for parallel processing of distributed data. Big data and cloud computing are two sides of same issue so we can analyze and forecast data services accurately of each other. Existing tools and technologies requires incorporating advanced data acquisition methods, data management and analysis tools. Further, it is necessary to improve security challenges by innovative solutions and improve profitability of many enterprises with effective tools, techniques and strategic planning.

VII. CONCLUSION

This paper presented a description of a survey on incorporating Big Data in Cloud environment. As the data is rapidly growing, presents challenge related to transporting data to cloud or across different networks requires more time than actual processing .integrating Big Data in Cloud presents some potential features like: elasticity, scalability, deployment time, and reliability with cost-effective delivery model. Existing tools and technologies are not fully adequate to face all challenges presented by both technologies. Privacy and security preservation of sensitive information shared on cloud is a big challenge that should be improved with strong and efficient encryption algorithms techniques for data concealment in cloud environment.

ACKNOWLEDGMENT

With thanks to Prof (Dr.) Vibhakar Pathak and Dr. Ripu Ranjan Sinha for their advice, support and encouragement.

REFERENCES

[1] Rohit Chandrashekar,et.al, “Integration of Big Data in Cloud computing environments for enhanced data processing capabilities”, 2015
[2] Charlotte Castelino, et.al, “Integration of Big Data and Cloud Computing”, 2014
[3] K. Kala Bharathi, “Converging Technologies of Cloud and Big Data”, 2016
[4] Elmustafa Sayed Ali Ahmed1 and Rashid A.Saeed, “A Survey of Big Data Cloud Computing Security”, 2014
[5] Sanjay P. Ahuja & Bryan Moore, “State of Big Data Analysis in the Cloud”, 2013
[6] Judith Hurwitz,et.al, “How to Make Use of the Cloud for Big Data”
[7] Intel IT Center ,“Big Data in the cloud: Converging Technologies”
[8] Judith Hurwitz,et.al, “Big data Cloud Providers”
[9] Venkatesh et.al, “A Study on Use of Big Data in Cloud Computing Environment “, 2015
[10] Bo Li . 2013,” Survey of Recent Research Progress and Issues in Big Data”
[11] Divyakant Agrawal, et Al., Big Data White Paper, 2012, “Challenges and Opportunities with Big Data”
[12] Dr. Willie E. May , 2015, “NIST Big data Interoperability”, Draft Version 1
[13] Santosh Kumar Majhi, Gyanaranjan Shial, 2015, “Challenges in Big Data Cloud Computing and Future Research Prospects: A Review”