International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 09 , September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Big Data and Cloud Computing

A.Zakiuddin Ahmed[1], R.Deepika[2] , P.Rizwan Ahmed[3]

[1]*Assistant Professor of Computer Science, Mazharul Uloom College, Ambur*

[2]*Research Scholar, Mazharul Uloom College, Ambur*

Assistant Professor & Head of Computer Applications, Mazharul Uloom College, Ambur

ABSTRACT

big data can bring huge benefits to businesses of all size. However, as with any business project, proper preparation and planning is essential, especially when it comes to infrastructure. Until recently it was hard for companies and organizations to get into big data without making heavy infrastructure investments (expensive data warehouses, software, analytics staff, etc.). But times have changed. Now cloud computing has opened up a door to the companies to avoid investing lot of money for infrastructure.

Index Terms— BigData, Hadoop, MapReduce, MongoDB, Casandra, variety, velocity and computing.

I. INTRODUCTION

Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine structured and unstructured data are important as they can help organizations gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web. The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimize the use of resources.

This paradigm is being popularly termed as Despite the popularity on analytics and Big Data, putting them into practice is still a complex and time consuming Endeavour. 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. An organization willing to use analytics technology frequently acquires expensive software licenses; employs large computing infrastructure; and pays for consulting hours of analysts who work with the organization to better understand its business, organize its data, and integrate it for analytics. This joint effort of organization and analysts often aims to help the organization understand its customers' needs, behaviors, and future demands for new products or marketing strategies. Such effort, however, is generally costly and often lacks flexibility. Cloud computing 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.

In order to get going with big data and turn it into insights and business value, it's likely you'll need to make investments in the following key infrastructure elements: data collection, data storage, data analysis, and data visualization/output. Let's look at each area in turn.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 09 , September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

II. DATA COLLECTION

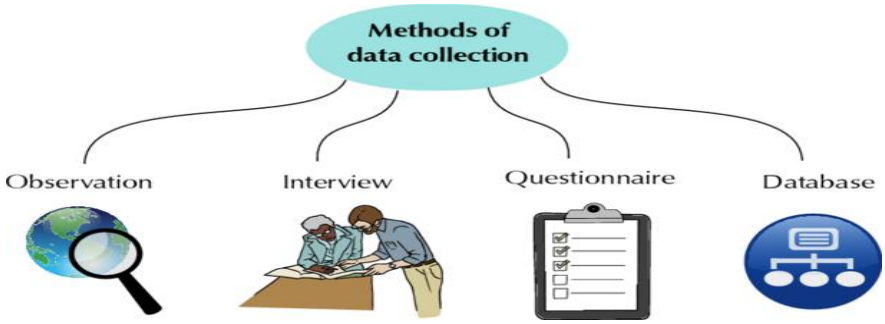

Figure-1 methods of data collection

This is where the data arrives in a company. It includes everything like sales records, customer database, feedback, social media channels, marketing lists, email archives and any data gleaned from monitoring or measuring aspects of the operations. You may already have the data you need, but chances are you need to source some or all of the data required.

If you do need to source new data, this may require new infrastructure investments. Infrastructure requirements for capturing data depend on the type or types of data required, but key options might include: sensors (that could sit in devices, machines, buildings, or on vehicles, packaging, or anywhere else you would like to capture data from); apps which generate user data (for example, a customer app which allows customers to order more easily); CCTV video; beacons (such as iBeacons from Apple which allow you to capture and transmit data to and from mobile phones); changes to your website that prompt customers for more information; and social media profiles.

Regular hard disks are available at very high capacities and for very little cost these days and, if you're a small business, this may be all you need. But when you start to deal with storing and analyzing a large amount of data, or if data is going to be a key part of your business going forward, a more sophisticated, distributed (usually cloud-based) system like Hadoop may be called for.

I think cloud-based storage is a brilliant option for most businesses. It's flexible, you don't need physical systems on-site and it reduces your data security burden. It's also considerably cheaper than investing in expensive dedicated systems and data warehouses.

III.Data Analysis

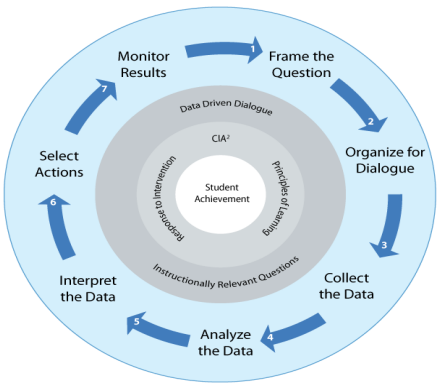

Figure-2 Data Analysis Chart

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

When you want to use the data you have stored to find out something useful, you will need to process and analyze it. So this layer is all about turning data into insights. This is where programming languages and platforms come into play. The analysis phase basically lets us to observe and gain knowledge about our subject. For example the figure given below:

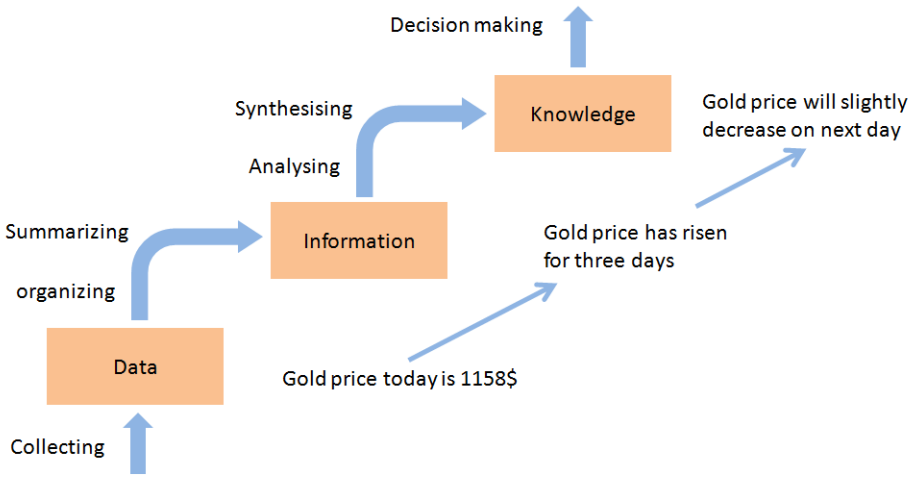

Figure -3 Example For Data Processing

There are three basic steps in this process:

1. Preparing the data (identifying, cleaning and formatting the data so it is ready for analysis)
2. Building the analytic model and
3. Drawing a conclusion from the insights gained.

Software exists from vendors such as IBM, Oracle and Google to help you do all of this: turning raw data into insights. Google has BigQuery, which is designed to let anyone with a bit of data science knowledge run queries against vast datasets. Other analytics options include Cloudera, Microsoft HDInsight and Amazon Web Services. And many startups are piling into the market, offering simple solutions which claim to let you feed it with all of your data, and sit back while it highlights the most important insights, and suggests actions for you to take.

IV. DATA VISUALIZATION / OUTPUT

This is how the insights gleaned from analyzing the data are passed on to the people who need them, i.e. the decision makers in your company. Clear and concise communication is essential, and this output can take the form of brief reports, charts, figures and key recommendations.

All too often I see businesses bury the real nuggets of information that could really impact strategy in a 50-page report or a complicated graphic that no one understands. It's clearly unrealistic to expect busy people to wade through mountains of data with endless spreadsheet appendices and extract the key messages. Remember: if the key insights aren't clearly presented, they won't result in action.

Key data output options include management dashboards, commercial data visualization platforms that make the data attractive and easy to understand, and simple graphics (like charts and graphs) that communicate insights. In my experience, for smaller businesses looking to improve their decision making, simple graphics or visualization tools like word clouds are more than enough to present insights from data.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Technologies not only support the collections of large amounts such data effectively. Transactions that are made all over the world in a Bank, Walmart customer transactions, and Facebook users generating social interaction data are few examples for big data usage.

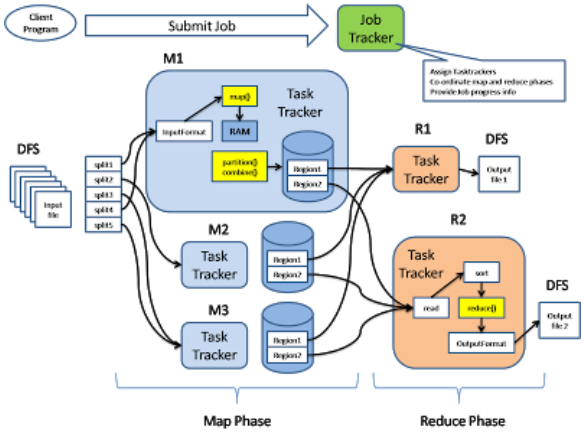

Figure -4 Big Data Tools Working Architecture

Task trackers are responsible for running the tasks that the job tracker assigns them Job trackers has two primary responsibilities which are managing the cluster resources and scheduling all user jobs.

Data engine consists of all the information about the processing the data Fetch manager helps to fetch.

V. MAPREDUCE

MapReduce framework is used to write apps that process a large amounts of data in a reliable and fault tolerant way. The application is initially divided into individual chunks which are processed by individual map jobs in parallel. The output of map sorted by a framework and then sent to the reduce tasks. The monitoring is taken care by the framework. The input data is divided into individual chunks and are provided for processing by the map task. These map task process the data in parallel and the result from the map task is then provided to the reduce task where the results that are generated in parallel by the map task are consolidated and the reduced report is given as output.

A. Big Data Applications

In the current age of data explosion, parallel processing is very much essential for performing a massive volume of data in a timely manner. Parallelization techniques and algorithms are used to achieve better scalability and performance for processing big data. Map reduce is a very popularly used tool or model used in industry and academics.

The two major advantages of map reduce are encapsulation of data storage, distribution, replication details. It is very simple for use by the programmers to code for the map reduce task. Since the map reduce is schema free and index free, it requires parsing of each records at the reading point. Map reduce has received a lot of attentiveness in the fields of data mining, information
retrieval, image retrieval etc.

The computation becomes difficult to be handled by traditional data processing which triggers the development of big data apps[8]. Big data provides an infrastructure for maintaining transparency in manufacturing industry,

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

which has been having the ability to unrevealing uncertainties that exists in the component performance and availability. Another application of the big data is the field of bioinformatics [9] which requires large scale data analysis.

B. Advantages of Big Data

The big data allows an individual to analyze the threats he/she faces internally by noosing onto the entire data landscape over the company using the rich set of tools that the software supporting the big data provides. This is an important advantage of big data since it allows the user to make the data safe and secure. The speed, capacity and scalability of cloud storage provide a mere advantage for the company and organization. Big data even allows the end users to visualize the data and companies can find new business opportunities. Data analytics is one more notable.

VI.MONGO DB

MongoDB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

C. Database

Database is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases.

D. Collection

Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose.

E. Document

A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data.

The following table shows the relationship of RDBMS terminology with MongoDB.

RDBMS	MongoDB
DataBase	DataBase
Table	Collection
Tuple/Row	Document
Column	Field
Table Join	Embedded Documents

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Primary Key	Primary Key (Default key _id provided by mongodb itself)
DataBase Server and Client	
MySql/Oracle	Mongo
MySql/SqlPlus	Mongo

VII. APACHE CASSANDRA

Cassandra is a distributed database from Apache that is highly scalable and designed to manage very large amounts of structured data. It provides high availability with no single point of failure.

Apache Cassandra is a highly scalable, high-performance distributed database designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. It is a type of NoSQL database. Let us first understand what a NoSQL database does.

F. NoSQLDataBase

NoSQL database (sometimes called as Not Only SQL) is a database that provides a mechanism to store and retrieve data other than the tabular relations used in relational databases. These databases are schema-free, support easy replication, have simple API, eventually consistent, and can handle huge amounts of data.

The primary objective of a NoSQL database is to have

1. simplicity of design,
2. horizontal scaling, and
3. finer control over availability.

NoSql databases use different data structures compared to relational databases. It makes some operations faster in NoSQL. The suitability of a given NoSQL database depends on the problem it must solve.

G. NoSQL VS Relational DataBase

The following table lists the points that differentiate a relational database from a NoSQL database.

Relational Database	NoSql Database
Supports powerful query language.	Supports very simple query language.
It has a fixed schema.	No fixed schema.
Follows ACID (Atomicity, Consistency, Isolation, and Durability).	It is only "eventually consistent".
Supports transactions.	Does not support transactions.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Besides Cassandra, we have the following NoSQL databases that are quite popular:

- **Apache HBase**

HBase is an open source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as a part of Apache Hadoop project and runs on top of HDFS, providing BigTable-like capabilities for Hadoop.

- **MongoDB**

MongoDB is a cross-platform document-oriented database system that avoids using the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas making the integration of data in certain types of applications easier and faster.

Listed below are some of the notable points of Apache Cassandra:

- It is scalable, fault-tolerant, and consistent.
- It is a column-oriented database.
- Its distribution design is based on Amazon's Dynamo and its data model on Google's Bigtable.
- Created at Facebook, it differs sharply from relational database management systems.
- Cassandra implements a Dynamo-style replication model with no single point of failure, but adds a more powerful "column family" data model.
- Cassandra is being used by some of the biggest companies such as Facebook, Twitter, Cisco, Rackspace, ebay, Twitter, Netflix, and more.

H. Features of Cassandra

Cassandra has become so popular because of its outstanding technical features. Given below are some of the features of Cassandra:

- **Elastic scalability**

Cassandra is highly scalable; it allows to add more hardware to accommodate more customers and more data as per requirement.

- **Always on architecture**

Cassandra has no single point of failure and it is continuously available for business-critical applications that cannot afford a failure.

- **Fast linear-scale performance**

Cassandra is linearly scalable, i.e., it increases your throughput as you increase the number of nodes in the cluster. Therefore it maintains a quick response time.

- **Flexible data storage**

Cassandra accommodates all possible data formats including: structured, semi-structured, and unstructured. It can dynamically accommodate changes to your data structures according to your need.

- **Easy data distribution**

Cassandra provides the flexibility to distribute data where you need by replicating data across multiple data centers.

- ***Transaction support***

Cassandra supports properties like Atomicity, Consistency, Isolation, and Durability (ACID).

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

- **Fast writes**

Cassandra was designed to run on cheap commodity hardware. It performs blazingly fast writes and can store hundreds of terabytes of data, without sacrificing the read efficiency.

I. History of Cassandra

- Cassandra was developed at Facebook for inbox search.
- It was open-sourced by Facebook in July 2008.
- Cassandra was accepted into Apache Incubator in March 2009.
- It was made an Apache top-level project since February 2010.

VIII .CHALLENGES & DISCUSSIONS

We live in the period of the big data where we can gather more information from daily life of human being. So far, researchers are unable to unify the features that are more essential to big data, many think that big data is something which we cannot process using existing technology, theory or any methods of such kind. However the world has become helpless since enormous amount of data is being generated by science, business and even society. Big data has posed many challenges to the IT industry.

IX. BIG DATA MANAGEMENT

The needs of the big data are not being satisfied by the current technologies and the speed of increasing storage capacity is much less compared to the data. Thus a revolution reconstruction of information framework is needed very much. For this we need to design a hierarchical architecture for storage. The heterogeneous data are not efficiently handled by the efficient Algorithms that exist now and thus we need to even design a very efficient algorithm for the effective handling of the heterogeneous data.

J. Necessity of Security

The big data is used by many of the business but they may not have assets from perspective of the security. If any security threat occurs to big data, it may come out with even more serious issue. Nowadays, companies use this technology to store data of peta byte range regarding to the company, business and customers. This result in severe criticality for classification of information to secures the data we either need to encrypt, log or use honey pot techniques. The challenge of detecting threats and malicious intruders must be solved using big data style analysis.

Analysis and computation of big data: Speed is the main thing when we look up for querying in the big data. However the process may be time consuming only because of the reason that it cannot traverse all related data in the whole database in a short time. While the big data is getting complicated, the indices in the big data are aiming at the simple type of the data. The traditional serial algorithm is inefficient for this big data.

X. PROPOSED APPROACH FOR SECURITY OF BIG DATA IN CLOUD COMPUTING ENVIRONMENT

Here we present few security measures that can be used to improve the cloud computing environment.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

K. Encryption

Since the data in any system will be present in a cluster, a hacker can easily steal the data from the system. This may become a serious issue for any company or organization to safeguard their data. To avoid this, we may go for encrypting the data. Different encryption mechanisms can be used on different systems and the keys generated should be stored secretly behind firewalls. By choosing this method the data of the user may be kept securely.

L. Nodes Authentication

The node must be authenticated whenever it joins the cluster. If the node turns out to be a malicious cluster then such nodes must not be authenticated.

M. HoneyPot Nodes

The *HoneyPot* nodes appears to be like a regular node but is
a trap. It automatically traps the hackers and will not allow any damage to happen to the data.

N. Access Control

The differential privacy and access control in the distributed environment will be a good measure of security. To prevent the information from leaking we use a SELinux. The Security Enhanced Linux is a feature that provides the mechanism for supporting access control security policy through the use of Linux Security modules in Linux kernels.

XI.CONCLUSION

This paper gives a clear description of a systematic flow of survey of the big data in the environment of cloud computing. We discussed about the applications, advantages and challenges faced by big data when used over a cloud computing environment. We proposed few solutions to safeguard the data in the cloud computing environment. In future, the challenges are yet to overcome and make way for the even more efficient use of the big data by the user on a cloud computing environment.

It is very much needed that the computer scholars and IT professionals to cooperate and make a successful and long term use of cloud computing and explore new idea for the
usage of the big data over cloud environment. The better encryption techniques if applied the better secure the data will be over the cloud environment.

REFERENCES

[1] D.J. Abadi, Data management in the cloud: Limitations and opportunities, IEEE Data Engineering Bulletin 32 (1) (2009) 3–12.

[2] Amazon redshift, http://aws.amazon.com/redshift/.

[3] Amazon data pipeline, http://aws.amazon.com/datapipeline/.

[4] Amazon Elastic Map Reduce (EMR), http://aws.amazon.com/elasticmapreduce/.

[5] Amazon Kinesis, http://aws.amazon.com/kinesis/developer-resources/.

[6] R. Ananthanarayanan, K. Gupta, P. Pandey, H. Pucha, P. Sarkar, M. Shah, R. Tewari, Cloud Analytics: Do We Really Need to Reinvent the Storage Stack? in: Proceedings of the Conference on Hot Topics in Cloud

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Computing (HotCloud 2009), USENIX Association, Berkeley, USA, 2009. movement data, SIGKDD Explor. Newsl. 9 (2) (2007) 38–46.

[7] D. Borthakur, "The hadoop distributed file system: Architecture and design," Hadoop Project Website, vol. 11, 2007.

[8] The Apache Hadoop Project. http://hadoop.apache.org/core/, 2009.

[9] A. Abouzeid, K. B. Pawlikowski, D. J. Abadi, A. Rasin, and A.Silberschatz. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. *PVLDB*,2(1):922–933, 2009.

[10] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony,H.Liu, P. Wyckoff, and R. Murthy. Hive - A Warehousing SolutionOver a Map-Reduce Framework. *PVLDB*, 2(2):1626–1629, 2009.

[11] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[12] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.

[13] Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.

[14] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Busines on big dataapplications." Big Data, 2013 IEEE International *Conference*, Silicon Valley, CA, Oct 6-9, 2013, pp.32 - 37.

[15] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." *Open Cirrus Summit (OCS), 2012.*