



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

BIG DATA ANALYTICS ON THE CLOUD

Dr. Tamanna Siddiqui, Mohammad Al Kadri
Department Of Computer Science
Aligarh Muslim University, Aligarh, INDIA

Abstract: The late advancements in innovations and construction modeling made the big data –which is information examination procedure – one of the sultriest points in our present studies. For that big data applications get to be more the most required programming in our life, in light of the fact that all our business, social, instructive and medicinal studies consider this idea as nightmare specially when we discuss storing and accessing client information in business clouds, or mining of social information. Big data needs a gigantic hardware equipment and processing assets, this needs made the little and medium estimated organizations be perplexed from these expenses.

Anyhow, cloud computing offers all these needs for big data implementation to those organizations. It gives services to internet by using assets of the figuring framework to give diverse services of the web. It makes customers and organizations have the capacity to control their applications without establishment and retrieve their own data at any computer through internet access. The significant concerns according to cloud computing are security and loss of control.

This paper presents an overview of big data concept, its famous interactive analysis tools (Apache Hive, BigQuery, CitusDB, ..), big data software (MATLAB, R, SAS, STATA, ...) and their comparisons on the basis of specific parameters. This paper also focuses on Cloud computing services and its four models (PaaS, SaaS, IaaS, HaaS). Further useful comparisons between NoSQL databases on the basis of design, indexing, integrity, distribution, and system have been presented. This review work is an effort to present an overall study of Big Data Analytics on the Cloud to facilitate practitioners in this domain.

Keywords: Big Data; Cloud Computing; Big Data analytics software; Hadoop; NoSQL.

I. Introduction

Big data is an information investigation strategy empowered by another era of innovations and structural planning which bolster high-speed information catch, storage, and analysis. Data sources augment past the traditional corporate database to incorporate email, cell phone output, sensor-created information, and social media yield. Information are no more confined to structured database records however incorporate unstructured data – information having no standard arranging. Big data obliges enormous measures of storage room. While the cost of capacity kept on declining, the assets expected to influence big data can in any case posture budgetary challenges for little to medium estimated organizations. The storage of commonplace big data and the infrastructure of its analysis will be in light of clustered network-attached storage (NAS) [2].

The thought behind the Cloud is that clients can utilize the service whenever, anyplace through the Web, straightforwardly through the utilization of browser. In cloud computing information is put away in virtual space as it uses the browsers to utilize network services. Since, networks are related so primary concern is of security of information [1].

For a little to medium estimated organizations, information storage utilizing cloud computing is a practical choice when they want to use big data analytics procedures. Cloud computing is on-interest system access to processing assets which are regularly given by an outside element and oblige little administration exertion by the business. Nowadays, for cloud computing there are various architectures and organization models exist, and these architectures and models are able to be manipulated with different design approaches and technologies [2].

Owners of little to medium estimated organizations who should not able manage the cost of selection of clustered NAS innovation can consider various cloud computing models to meet their big data needs. Little to medium measured entrepreneurs need to consider the right cloud computing so as to stay both aggressive and gainful.

II. Big Data

Big data is more than simply a typical word in today's business ideas, it is a phenomena that be come about because of the colossal measure of information data produced crosswise over society, and gathered by business and government associations.

Case in point, the humanity made 2 exabytes (one exabyte is one quintillion bytes, or one thousand petabytes, or quite simply one billion gigabytes) of information since the beginning of time, till the early 2000's. But comparably, in 2012 as example around 2.5 exabytes bytes of information was made consistently [3].

However, these days, the normal routines of human and machines (communicating, browsing, documenting, capturing content, sharing, searching, and .. etc) create upwards of 2.5 exabytes of data every day across the globe.

In the current Information and Communication Technology (ICT) era, the most difficult exploration issues get to be: how to successfully and ideally manage "Information explosions", which is created from various sorts of sources including cell phones, sensors, individual files, Internet of Things, ventures, cams, social media, programming logs, health information and so forth. The issue is otherwise called the 'Big Data' problem, which is characterized as the act of gathering complex information sets so extensive that it gets to be hard to analyze and interpret physically or utilizing available information administration applications.

These days there are a considerable measure of delegate applications incorporate clinical choice emotionally supportive networks, computerized horticulture, social networking analytics, high vitality physical science, earth perception, genomics, auto reproductions, therapeutic imaging, body range systems, translational medication, and so forth which control the big data issue, because we can discover - from the point of view of genuine applications – this issue has additionally turned into a typical marvel in space of science, prescription, building, and trade [3].

Big data generally indicates to datasets that are not liable to analysis by the relational database instruments, statistical analysis tools and visualization aids that have ended up recognizable in the past twenty years since the begin of the quick increment in digitized sensor information. Rather, it obliges 'hugely parallel programming running on tens, hundreds, or even a thousands of servers in some extreme cases' [1].

We can summarize the differences between Tradition Data and Big Data in this table:

Table I traditional Data vs. Big Data

Attribute	Traditional Data	Big Data
Volume	Gigabytes to terabytes	Petabytes to zettabytes
Organization	Centralized	Distributed
Structure	Structured	Semi-structured & unstructured
Data model	Strict schema based	Flat schema
Data relationship	Complex interrelationships	Almost flat with few relationships

Important characteristics of big data include (3Vs):

- **Volume:** It incorporates the mass information of big business.
- **Variety:** It manages the multifaceted nature of numerous information sorts, including structured and unstructured data.
- **Velocity:** It is the speed at which information is dispersed furthermore the rate at which it changes [1].

Information savants are likewise concentrating on extra V's in characterizing big data. One such is verification. The contention is that with big data comes such a various scope of information that nature of information and well as security get to be enormous central focuses for verification of the information. [4]

Variable Veracity also cited sometimes as an additional V. It manages the level of unwavering quality connected with specific sorts of information. It utilizes the standardization impact to investigate information at these tremendously diverse requests of size [1].

Big data can be broken into two areas:

1. Big Data Transaction Processing (also known as Big transactions).
2. Big Data Analytics.

In big data we store our data in two kind of stores (relational DBMS or file systems), the big data transaction processing should deal with both kind of stores, while it'll update that extreme. Typically, many NoSQL DBMSs don't focus on "ACID properties" transactional concept which all relational DBMSs take care of it very well. That problem will effect if it is unsatisfactory to lose a transaction e.g. a Banking deposit.

Organizations consider the most common types of data that they want to analyze include:

- **Web data:** such as e-commerce logs, web logs and social network data.
- **Industry specific big transaction data:** like Telco call data records (CDRs), retail transaction data and geo-location data.
- **Machine generated/sensor data:** to monitor everything from movement, temperature, vibration, location, airflow, light, liquid flow and pressure.
- **Text:** such as archived documents, customer interaction data or external content sources [1].

Let us take a Bigtable as an example which is a distributed storage system for managing structured data at Google [5]. It distinguishes with all these features together (high availability, wide applicability, high performance, and scalability).

Bigtable gives an alternate interface than Parallel databases and primary memory databases. But the likeness of Bigtable with database comes because of offering of usage databases' methodologies. Bigtable regards information as confused strings, despite the fact that customers regularly serialize different forms of structured and semi-structured data into these strings. For that information in Bigtable is indexed utilizing column and row

names that can be random strings. Customers can control the territory of their information through cautious decisions in their diagrams. At last, Bigtable pattern parameters let customers alertly control whether to serve information out of memory or from disk [5].

We can also mention a list of interactive analysis tools for big data for ad-hoc query, with some comparison between them from the data side (shapes, sources, etc.):

Table II Comparison of Big Data Interactive Analysis Tools

Tools	Low-latency	Data shapes	Data sources	Query languages	Columnar storage
Apache Drill	Yes	Nested, tabular	Extensible, incl. HDFS, HBase, Cassandra, MongoDB, RDBMS, etc.	Extensible, incl. SQL 2003, MongoQL, DSL, etc.	Yes
Apache Hive	No	Nested, tabular	HDFS, HBase	HiveQL	Possible
BigQuery	Yes	Nested, tabular	N/A	SQL subset	Yes
CitusDB	Yes	Nested, tabular	PostgreSQL, MongoDB, HDFS	SQL	No
Hadapt	Yes	Tabular	HDFS/RDBMS	SQL subset	No
HAWQ	Yes	Tabular	HDFS, HBase	SQL subset	Yes
Impala	Yes	Tabular	HDFS, HBase	SQL/HiveQL subset	Yes
Phoenix	Yes	Tabular	HDFS, HBase	SQL subset	No

Also if we consider healthcare business or crosswise over different commercial ventures as well, information can be arranged as structured information (eg, research center information), unstructured information (eg, post-operation notes and patient release summaries), imaging information (eg, figured tomography imaging studies), and streaming information (e.g. electroencephalography information). New innovation has developed that finds, lists, seeks, and explores different wellsprings of information. We have huge measures of databases and information stores that feed into information distribution centers (data warehouses) [4].

III. Cloud Computing

Cloud computing is on-interest network access to computing resources, gave by an outside element. Common deployment models for cloud computing incorporate platform as a service (**PaaS**), software as a service (**SaaS**), infrastructure as a service (**IaaS**), and hardware as a service (**HaaS**) [6].

We can explain each of these four common models for cloud computing as:

- **Platform as a service (PaaS):** It is supplying platforms to develop and use of custom application via cloud computing, and consist of several tools such as :
 - Application testing
 - Deployment and hosting.
 - Application design and development tools
 - Integration.
 - Versioning.
 - State management.

Other preferences of utilizing PaaS incorporate bringing down risks by utilizing pretested innovations, advancing shared services, enhancing programming security, and bringing down ability necessities required for new frameworks advancement. As identified with huge information, PaaS gives organizations a stage to creating and utilizing custom applications expected to examine extensive amounts of unstructured information requiring little to no effort and generally safe in a protected domain [2].

- **Software as a service (SaaS):** gives organizations applications that are put away and run on virtual servers – in the cloud. Advantages of utilizing SaaS are:
 - Less demanding software organization
 - Programmed updates and patch administration
 - Programming similarity over the business
 - Simpler cooperation
 - Worldwide openness.

The primary focal point of SaaS is that the arrangement permits organizations to move the dangers connected with programming obtaining while moving IT from being responsive to proactive. Programming as an Administration gives organizations investigating enormous information demonstrated programming answers for information investigation. The distinction in the middle of SaaS and PaaS for this situation is that SaaS is not going to give a redid arrangement while PaaS will permit the organization to add to an answer custom-made to the organization's necessities [7].

- **Infrastructure as a service (IaaS):** In the IaaS model, a customer business will pay on an every utilization premise for utilization of hardware to bolster computing operations including storage, equipment, servers, and networking equipment. Infrastructure as a service is the cloud computing model getting the most consideration from the business sector, with a desire of 25% of endeavors wanting to receive a service supplier for IaaS [2].

Services available to businesses through the IaaS model - which is affording crest load capacity for variable processes – include:

- Compute as a service
- Data center as a service
- Storage as a service
- Disaster recovery
- Cloud bursting.
- Virtual desktop infrastructure
- **Hardware as a service (HaaS):** It allows the customer to license the hardware directly from the service provider which alleviates the associated costs [6].

There are three sorts of cloud computing:

- **The private cloud:** It's private data center (internal) for a business, not available to the general public but based on the structure of cloud.
- **The public cloud:** are the pay-as-you-go services.
- **The hybrid cloud:** both public cloud and private cloud are used in this cloud [2].

IV. Big Data and Cloud Computing

Big data refers to enormous, heterogeneous, and frequently unstructured advanced substance that is hard to process utilizing traditional information management instruments and systems. The term envelops the unpredictability and mixture of information and information sorts, ongoing information accumulation and handling needs, and the quality that can be gotten by smart analytics. Advanced data mining systems and related tools can help separate data from substantial, complex datasets that is helpful in settling on educated choices in numerous business and scientific applications including tax gathering, biosciences, market deals, social studies, and high energy material science. Consolidating big data analytics and knowledge discovery methods with adaptable computing frameworks will deliver new bits of knowledge in a shorter time. Albeit few cloud-based analytics platforms are accessible today, momentum exploration work foresees that they will get to be common within a couple of years. Some present arrangements are taking into account open source frameworks, for example, Apache Hadoop and SciDB, while others are proprietary solutions gave by organizations, for example, Google, IBM, Splunk Storm, Kognitio, InsightsOne, BigML, and EMC. As all the more such platforms rise, specialists will port progressively intense data mining programming tools and methodologies to the cloud to accomplish complex and adaptable programming models, for example, the dispersed workflow paradigm. The developing utilization of service-oriented computing could quicken this pattern [2].

Three main purposes behind little to medium estimated organizations to utilize cloud computing for big data innovation implementation are:

- Reducing hardware equipment costs.
- Decrease the preparing expense.
- Capacity to test the estimation of big data [2].

In these days we can find developers and researchers when they want to achieve big data analytics in the cloud, they embrace the software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) models in their solutions.

Analyzing big data is done by using MapReduce concept. The main idea in MapReduce paradigm is mapping the inquiry (query) related data in key values representation manner, then collect them to claim the query outcome result. The MapReduce paradigm obliges that tremendous measures of information be analyzed. The mapping is done simultaneously by every different NAS device; the mapping obliges parallel processing. The parallel processing needs of MapReduce are exorbitant, and require the arrangement noted beforehand for capacity. The processing needs can be met by cloud-service suppliers.

Developers can achieve big data analytics services by using each of these three models:

- **Data analytics software as a service:** provides a ready to- use knowledge discovery tool or well-defined data mining algorithm as an Internet service to final users, that able them to access it precisely via a Web browser.
- **Data analytics platform as a service:** provides a supporting platform that make developers don't care about the underlying infrastructure or distributed computing issues when they use it to create their own data analytics solutions or continue ones which are already existed.
- **Data analytics infrastructure as a service:** developers can build their own data mining applications or achieve their data analytics systems from scratch by using a set of virtualized resources which are provided in this case as a computing infrastructure [7].

V. Big Data Analytics Software

Designing and developing an innovation frame work to manipulate and carry Big Data Analytics needs will depend on one or some of the following technologies:

- Cloud computing infrastructures (e.g., Amazon, Google App Engine, Microsoft Azure, IBM Bluemix and .. etc) in conjunction with fast communication networks
- Data-intensive programming paradigms (MapReduce, distributed storage system, etc.)
- Semantic web.
- Machine learning algorithms [3].

And the recently software tools which we use for big data analytics (advanced analytics) depend on of the following techniques:

- Predictive analytics
- Data mining
- Text analytics
- Statistical analysis.
- Mainstream BI software
- Data visualization tools

For each of these analysis techniques we can find a lot of solutions and software that can help us and analyze our data in its way. If we take statistical analysis as example, we can find a huge number of software and each of them has some special features distinguish it from others, we compare here a short list of these software:

Table III Big Data Analytics Software and Tools

Software	Focus	Platforms	License
D3 (DATA DRIVER DOCUMENTS)	Visualization	Web-based	Open source
IBM SPSS	Statistics, data mining, and academic research	Microsoft Windows, Linux, and Mac OS X	Proprietary
MATLAB	Numerical computation	Microsoft Windows, Linux, and Mac OS X	Proprietary
MATHEMATICA	Mathematics (analytical computation)	Microsoft Windows, Linux, and Mac OS X	Proprietary
MS EXCEL	General purpose for office use	Microsoft Windows and Mac OS X	Proprietary
PYTHON	General-purpose language for extensive scientific libraries	Many	Open source
R	Statistics, data mining, and academic research	Many	Open source
SAS	Statistics, data mining, and academic research	Server tier on various platforms; client tier on Microsoft Windows	Proprietary
STATA	Statistics, data mining, and academic research	Microsoft Windows, Linux, and Mac OS X	Proprietary
WEKA	Machine learning	Microsoft Windows, Linux, and Mac OS X	Open source

But not all our data is structured data (it's just 20%) and almost (80%) of our data is not structured, thus the needs for manipulating and treating a semi-structured and unstructured data made the traditional data warehouses based on relational databases become not sufficient, which turned many organizations to an open source software framework which use a clustered systems to process a large and diverse data sets [8].

The core of these frameworks based on:

- **Hadoop and its tools:**

Hadoop is an open-source; Java-based software framework that supports the processing of distributed storage of large data sets in a distributed computing environment. It is part of the Apache project and the core of this framework consists of:

- Hadoop Common
- Hadoop Distributed File System (HDFS)
- Hadoop YARN
- Hadoop MapReduce

Hadoop comes with a lot of tools also such as Spark, Hive and Pig [9].

- **NoSQL databases:**

A **NoSQL** (often interpreted as **Not only SQL**) database affords a mechanism for storage and retrieve a data, not with tabular relations in relational database management system (RDBMS) but indeed via modeling that data by means. It uses a different data structure than which be used in relational databases, and it is distinguished by simplicity of design, finer control over availability, and horizontal scaling, for that it is progressively used in big data and real-time web applications.

There are a lot of various approaches with different categories for NoSQL databases, but based on data model we can classify NoSQL databases to these following categories (with examples) [10]:

Table IV NoSQL Database Categories

Category	Matching Database
Column	Cassandra, HBase, Druid, Accumulo, Vertica
Document	MongoDB, Lotus Notes, Clusterpoint, Couchbase, MarkLogic
Key-value	Dynamo, CouchDB, Riak, MemcacheDB, Redis, MUMPS
Graph	Neo4J, Allegro, Virtuoso, InfiniteGraph, Star dog
Multi-model	OrientDB, Alchemy Database, FoundationDB, ArangoDB, CortexDB

It's useful to make small comparison between these categories by focusing on one example from each category, in this comparison we consider some elements from designing features, indexing, integrity, distribution, and system aspects. Next table contain this comparison information:

Table V Comparison of some NoSQL databases' features

Features	MongoDB [11]	HBase [12]	Riak [13]	Neo4j [14]
Design				
Data Storage	Volatile memory file system	HDFS	Bitcask & eLevelDB	Volatile memory file system
Query Language	JSON-like query language	REST, Thrift, Avro, Hive, and Pig.	REST, HTTP and Erlang	Cypher
Integrity				
Atomicity	Conditional	Yes	No	Yes
Consistency	Yes	Yes	No	Yes
Isolation	No	No	Yes	Yes
Durability	Yes	Yes	-	Yes
Indexing				
Secondary Indexes	Yes	Yes	Yes	-
Full Text Search	No	No	Yes	Yes
Graph Support	No	No	Yes	Yes
Distribution				
Replication	Yes	Yes	Yes	Yes
Replication Mode	Master Slave Replication	Master Slave Replication	Multi-Master Replication	-
Sharding	Yes	Yes	Yes	Yes
System				
Operating System	Cross-Platform	Cross-Platform	Cross-Platform	Cross-Platform
Programming Language	C++	Java	Erlang	Java

VI. Conclusion

In this paper, we have highlighted the phenomena of big data and its interactive tools, cloud computing services and its four models (PaaS, SaaS, IaaS, HaaS), we have made comparison between traditional data and big data, also we have mentioned some of the software and techniques which data scientists usually used in their analytics on big data specially distributed frameworks and statistical solutions. NoSQL databases have been reviewed in detail and we found that **Neo4j** is better than other in respect of various features. We hope that we can improve the big data analytics in the future work by using a suitable combination between hadoop's languages, statistical software API, and data mining concepts.

VII. References

- [1] N. Arora and R. K. Bawa, "A Review on Cloud to Handle and Process Big Data," *International Journal of Innovations & Advancement in Computer Science*, 2014. .
- [2] B. Purcell, "Big data using cloud computing," *J. Technol. Res.*, vol. 5, pp. 1–8, 2013.
- [3] L. Wang and A. Zomaya, "Software Tools and Techniques for Big Data Computing in Healthcare Clouds," 2012.
- [4] Rasu B. Shrestha, "ENTERPRISE IMAGING Big data and cloud computing," *Big data could Comput.*, no. March, pp. 3–5, 2014.
- [5] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. a. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *7th Symp. Oper. Syst. Des. Implement. (OSDI '06)*, Novemb. 6–8, Seattle, WA, USA, pp. 205–218, 2006.
- [6] P. Kalagiakos and P. Karampelas, "Cloud computing learning," *2011 5th Int. Conf. Appl. Inf. Commun. Technol. AICT 2011*, 2011.
- [7] D. Talia, "Clouds for scalable big data analytics," *Computer (Long. Beach. Calif.)*, vol. 46, no. 5, pp. 98–101, 2013.
- [8] "What is big data analytics? - Definition from WhatIs.com." .
- [9] "Welcome to Apache™ Hadoop®!" .
- [10] <http://en.wikipedia.org>, "NoSQL," <http://en.wikipedia.org>. [Online]. Available: <http://en.wikipedia.org/wiki/NoSQL>.
- [11] K. Chodorow, *Mongo DB: The Definitive Guide*. 2013.
- [12] L. George, *HBase: The Definitive Guide (Google eBook)*. 2011.
- [13] E. Redmond and J. R. Wilson, *Seven Databases in Seven Weeks*. 2012.
- [14] I. Robinson, J. Webber, and E. Eifrem, *GraphDatabases*. 2013.