# A Contrastive Study of Big Data Analytics Approaches and Tools with Simple Data analytics

Preeti Jakhar[1], Aijaz Ul Haq[2]

[1] Student, Computer Science Engineering, Satya College of Engineering and Technology , Haryana, India
[2] Assistant Professor, Computer Science Engineering, Satya College of Engineering and Technology , Haryana, India

## ABSTRACT

*This paper presents a contrastive study of traditional data and modern data . Data can exist in a variety of forms -- as numbers or text on pieces of paper, as bits and bytes stored in electronic memory, or as facts stored in a person's mind.. Moreover, data is present each and every place in the universe. There is no limit of the data present all over the universe. Since the day when human civilization starts, we always exist with some data. Our day to day life deals with some data with or without any reason. The data can be of any type like remembering someone's name or saving a playlist of thousands of songs. But, with the development of the human civilization the rapid growth of the data has already come into the existence. As we live in the twenty first century we exist with a new concept of data which is known as "BIG DATA". In last century the human civilization faced a new concept of manipulating data known as Data Analytics in order to make the business profitable and the life more simple that deals with large volume of data. In this paper we have discussed about different types of data, the problems in the traditional data analytic s process. After that we discussed about what is big data and different parameters of big data. How the problems in traditional data analytics is being resolved by big data analytics also discussed in this paper.*

**Keyword : -** *Big data, data analytics, data storage procedure, big data landscape, hadoop, Hadoop Distributed File System(HDFS), Business Intelligence (BI).*

## 1. INTRODUCTION

Big Data technologies are transforming the way data is used to be analyzed. One reason is the massive amount of data that is being generated from different sources such as social networks, sensors; search engines, banks, telecommunication and web, handling this massive amount of data take us in the era of Big Data.

According to the YouTube statistics 400 hours of video are being uploaded to YouTube [1] servers in every minute. Facebook is dealing with more than 500 terabytes of data daily [2], companies such as a Google and Yahoo are recording search engine results for analyzing the searching trends; crawling different web sources to analyze for any important events; gathering marketing data for analyzing the current and future trends which all results the generation of large data sets also referred to as Big Data.

Data is everywhere, from social sciences to physical science, business and commercial world, for example, digitizing the past fifty year's newspapers will results the massive amount of data, in astronomy storing billions of astronomical objects, in biology storing genes, proteins and small molecules results in mas sive amounts of data. In business world such as handling millions of call data records in telecommunication, handling millions of transactions in banking and handling millions of transactions for multinational grocery store results in large data sets. Analyzing these large datasets and getting out meaningful information from it is a challenging in itself.

### 1.1 Big Data

Big Data can be described in 3 V's such as variety, volume and velocity [3].

#### 1.1.1 Variety

Data has different variations, for example semi-structured or unstructured, such as data, generated from web sites, social networks, emails, sensors and web logs is unstructured. Structured data refers to as data generated in result of conversion from call data record to tabular format in order to calculate the monetary value out of it or banks transactions data or data generated from the airline ticketing system are different varieties in the data.

### 1.1.2 Volume

Volume refers to the amount of data or size of the data set. Nowadays figures are in Tera and Peta bytes. For instance Airbus can generate half of terabytes of data in one flight [4].

### 1.1.3 Velocity

Velocity refers to the speed of data generation which is very fast nowadays. For example weather sensors are kept on generating data as new updates comes, Twitter is generating data at 9100 tweets per second and on Facebook users is sending 3 million messages to each other every 20 minutes [5].

There are different technologies to analyze the Big Data. Hadoop [6] is one of the most popular among them. There are many others, such as Cloudera and Cassandra and technologies for warehousing such as Hive and HBase [6] which can be used in conjunction with Hadoop to ease the analysis and provides the abstraction over Hadoop platform.

### 1.2 Problem Description

There are different technologies to deal with Big Data analysis, but most of them are complex and requires expertise to deal with them. Especially for non-computer scientists such as social scientists, they require good programming skills and knowledge of configuring and maintaining the infrastructure which almost makes it impossible for them to explore the large data sets or to perform ad hoc analytics on it. For example, how a social scientist can explore the data to find an event in 1975 by having the previous fifty years of newspaper data? Or how a social scientist can predict the human behavior by analyzing its previous 5 years of data gathered from different sources such as cell phone records with GPS tracking, search engine queries, internet transaction data, consumer behavior or its social network activity? [8]

There are plenty of Big Data analysis platforms or frameworks are available nowadays in the market, but the problem for non-computer scientists is to master them because of the complexity involves with them and where necessary to take training in order to use them for exploring Big Data in adhoc manners and doing analytics on large data sets. These systems inherit the problem of maintaining them as well, which might include at application or infrastructure level.

## 2. TRADITIONAL DATA STORAGE PROCEDURE

In traditional data storage procedure the data model defines some properties. The structured data needs to satisfy all the properties defined by the data model in order to be stored in the database. So, the data will only be accepted if and only it satisfies all the properties defined by the data model. If the data doesn't satisfy at least a single property then the data will be rejected. The data basically stored in the row column format in relational databases. SQL is used to handle the data and to process it.

## 3. PROBLEMS IN TRADITIONAL DATA ANALYTICS PROCEDURE

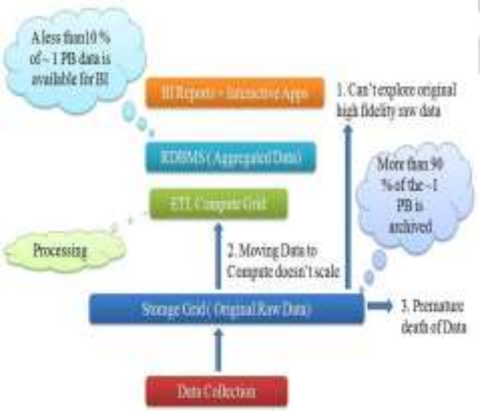
A less than 10 % of ~ 1 PB data is available for BI
BI Reports + Interactive Apps
1. Can't explore original high fidelity raw data
RDBMS ( Aggregated Data)
More than 90 % of the ~1 PB is archived
ETL Compute Grid
Processing
2. Moving Data to Compute doesn't scale
Storage Grid ( Original Raw Data)
3. Premature death of Data
Data Collection

**Fig 1: Block diagram showing Problems in Traditional Data Analytics Procedure**

## 3. INTRODUCTION TO BIG DATA

In a general meaning Big data is the huge amount of data. But the only parameter i.e., amount can't express the definition of big data completely. So, basically it is identified according to the large-volume, high-velocity and wide-variety of information.

### 3.1 V's of Big Data:

**i. Volume:**

In this case, the amount of data is taken into consideration. If large-volume of the data comes, then it will be difficult for the RDBMS to store and manage that amount of data. The name itself indicates that, the size of the data is the major parameter.

**ii. Velocity:**

It this case, the rate of speed at which the data is captured is the major concern. If large-volume of data comes at fraction of time, then that will be very difficult for the RDBMS to capture and process the data.

**iii. Variety:**

In this case, the type of the data becomes the major parameter. the RDBMS can process only structured the data. So, if the data is unstructured, then that can't be captured and process by the RDBMS

## 4. SOLVING THE PROBLEMS OF TRADITIONAL DATA ANALYTICS WITH BIG DATA

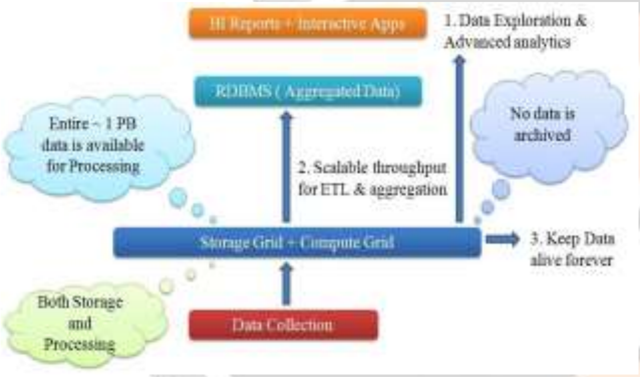
BI Reports + Interactive Apps
1. Data Exploration & Advanced analytics
RDBMS ( Aggregated Data)
Entire ~ 1 PB data is available for Processing
No data is archived
2. Scalable throughput for ETL & aggregation
Storage Grid + Compute Grid
3. Keep Data alive forever
Both Storage and Processing
Data Collection

**Fig 2: Block diagram showing the Solution of the problem in Traditional Data Analytics Procedure**

## 5. BIG DATA LANDSCAPE

### 5.1. Big data infrastructures:

**Hadoop-** Hadoop is a tool which provides an ecosystem to store, process and analyze the data. It works with distributed file system which breaks a large file into small files and distributes the data across different nodes in order to process the data faster.

**NoSQL** - It Stands for Not-Only-SQL. It is used to process large volumes of unstructured or semi-structured data. Hbase is one of the popular NoSQL database which can work with Hadoop.

**Massively Parallel Processing Database**- This database distributes the data in different segments across multiple nodes, and then process them in parallel by using SQL. Basically it runs on expensive hardware. The difference between MPP and Hadoop is that, MPP runs on expensive hardware whereas hadoop runs on cheaper commodity hardware.

### 5.2. BIG data analytics technology:

Some of the big data infrastructure technologies provide the data analysis in some manner. The big data analytics technologies specifically provide the facility for analysis. The followings are some of the sub-categories.

**Analytical Platforms**- It integrates and analyses data to discover some new knowledge, which helps organizations for better decision making. The main focus in this case is to provide the solution of the users need in timely manner.

**Visualizations** - The main objective is to visualize the data. It takes the data and presents the data in some visual forms in order to extract the information from it.

**Business Intelligence** - It is specifically used to provide business solutions. It collects, integrates, and analyses data that is required for a business to get more profitable solutions. It enables users to build applications that help organizations to learn and understand their business,

**5. 3 . Applications:**

Applications are generally used to analyze big data and offer optimized insights to the end-users. Following are some of the application fields :

**Ad Optimization** - MediaMath is the first demand-side platform (DSP), changing the way digital media is purchased, and creating a new, more efficient way for advertisers to reach consumers, individually, at scale.

**Publisher Tools**- Visual Revenue is a real-time predictive analytics platform developing a suite of tools providing decision support for editors content.

**Energ**y- AutoGrid takes the data from smart meters, voltage regulators, thermostats to assist customers track the amount of power used, scale back waste, balance the grid, increase the system operations and forecasts the future consumption.

## 6. OPEN SOURCE TOOLS FOR BIG DATA

### 6.1. Big Data Analysis Platforms and Tools:

***6.1.1. Hadoop:***

It is developed by Apache Foundation. As discussed earlier A whole ecosystem of technologies designed for the storing, processing and analyzing the data. To extend the capabilities of hadoop, the Apache organization also provides some related technologies and projects. Some technology providers provide support for Hadoop and supported technologies.

Platforms : Windows, Linux, OS X.

**6.1.2. MapReduce:**

It is originally developed by Google. It is described as a programming model which processes huge amounts of data on large clusters of computer nodes very quickly. Mostly this software framework is used by Hadoop. It is also used by different applications which are used for data processing.

Platforms : OS Independent.

**6.1.3. Storm:**

It is a product of Twitter. It provides distributed real-time computation facilities and is described as the "Hadoop of realtime". It is highly scalable, robust and works with a lot of programing languages.

Platforms : Linux.

### 6.2. Database/ Data warehouse:

**6.2.1. Cassandra:**

It was a product of Facebook. But, now it is now maintained by the Apache Organization. Many organizations like Netflix, Twitter, Urban Airship, Reddit etc. use this.

Platforms : OS Independent.

**6.2.2. HBase:**

It is a product of Apache organizations. It is a non-relational data storage for Hadoop. Some of the features include modular scalability, consistent read and write, fail-over support etc.

**Platforms: OS Independent.**

**6.2.3. MongoDB:**

It supports wide range of databases. It is a NoSQL database and it provides document oriented data storage, duplication, support for full index and highly available, etc.

Platforms: Linux, Windows, Solaris, OS X.

**6.2.4. Neo4j:**

Now it is a leading graph database in the world. The performance of Neo4j is around 1000 more than the relational database.

Platforms: Linux, Windows.

**6.2.5. CouchDB:**

It is basically developed for the Web. In this the data is stored in the JSON documents in order to access from the web or using JavaScript query.

Platforms: Linux, Android, OS X, Windows.

#### 6.2.6. Hive:

It was initially developed by Facebook. But now it is used and developed by other companies. It is known as the data warehouse for Hadoop. It uses HiveQL for queries.

Platforms: OS Independent.

#### 6.2.7. Hypertable:

It is developed by Zvents Inc. It is a NoSQL database that provides efficiency and performs faster which results in cost savings.

Platforms: Linux, OS X.

#### 6.2.8. FlockDB:

It is developed by Twitter. It is also popularly known as database of. Social graphs are stored here. i.e., the information of followings and followers and blocked users etc.

Platforms: OS Independent.

#### 6.2.9. Hibari:

Many of the telecom industries use this. It's an ordered key-value, storage of big data and guarantees high bandwidth and reliable.

Platforms: OS Independent.

### 6.3 Business Intelligence:

#### 6.3.1. Knime:

It is known as Konstanz Information Miner, or KNIME. It offers very user-friendly data integration, processing, analysis, and exploration.

Platforms: Windows, Linux, OS X.

#### 6.3.2. Palo BI Suite:

It includes an OLAP Server, Palo Web, Palo ETL Server and Palo for Excel.

Platforms: OS Independent.

#### 6.3.3. BIRT:

It is the short form of "Business Intelligence and Reporting Tools". It is an Eclipse-based tool that adds reporting features to Java applications.

Platforms: OS Independent.

#### 6.3.4. Pentaho:

It is used by more than thousands of companies. It offers BI tools and big data analytics tools with data mining, reporting and dashboard capabilities.

Platforms: Windows, Linux, OS X.

### 6.4. Data Mining:

#### 6.4.1. Rapid Miner / Rapid Analytics:

It is the world-leading open-source system for data and text mining. RapidAnalytics is a server version of RapidMiner.

Platforms: OS Independent.

#### 6.4.2. Mahout:

It is a project of Apache foundation and it offers algorithms for clustering, classification and batch-based collaborative filtering that runs on the top of Hadoop. The is used to build scalable machine learning libraries.

Platforms: OS Independent.

#### 6.4.3. SPMF:

It is a Java-based data mining framework. It is basically used in sequential pattern mining, but also includes tools for association rule mining, sequential rule mining and frequent item set mining.

Platforms: OS Independent.

#### 6.4.4. Weka:

It stands for "Waikato Environment for Knowledge Analysis." It offers a set of algorithms for data mining that can be applied directly on data or can be used in another Java application. It is sponsored by Pentaho.

Platforms: Windows, Linux, OS X.

### 6.5. File Systems:

#### *6.5.1. Gluster:*

It was developed by Red Hat. It offers object storage for very large datasets. It can be used to extend the capabilities of Hadoop beyond the limitations of HDFS.
Platforms: Linux.

#### 6.5.2. Hadoop Distributed File System:

It is popularly known as HDFS. It is the primary storage system for Hadoop. It quickly distributes the data into several nodes in a cluster and also replicates the data. It provides reliable, fast performance.
Platforms: Windows, Linux, OS X.

### 6.6. Programming Languages

#### 6.6.1. Pig:

It is an Apache Big Data project. It is a data analysis platform that uses a textual language called Pig Latin and it produces sequences of Map-Reduce programs.
Platforms: OS Independent.

#### 6.6.2. R:

R is developed by R core team. R is a programming language and an environment for statistical computing and graphics and is provided by R foundation for statistical computing. It provides a set of tools that make it easier to manipulate data, perform calculations and generate charts and graphs.
Platforms: Windows, Linux, OS X.

### 6.7. Data Aggregation and Transfer

#### 6.7.1. Sqoop:

It is developed by Apache Software Foundation. It is used to transfer the data between Hadoop and RDBMS. It imports individual tables or entire databases to HDFS. It provides the ability to import from SQL databases straight into the hive data warehouse.
Platforms: OS Independent.

#### 6.7.2. Flume:

It is a project by Apache organization. It collects, aggregates and transfers large amount of log data from applications to HDFS. It is written in java and robust and fault-tolerant.
Platforms: Windows, Linux, OS X.

## 7. MAJOR INDUSTRIES USING BIG DATA

| Industry | Area where big data is used |
|---|---|
| Retail | Supply Chain Analysis, Dynamic Pricing, Sentiment Analysis |
| Banking | Modelling true risks,Fraud Detection, Threat Analysis, Trade Surveillance, Credit Scoring and Analysis |
| Advertising | Ad targeting, Recommendation Engine, Click Fraud Detection |
| Telecommunication | Customer churn Prevention, Network Optimization, Calling data record analysis |
| Healthcare | Gene Sequencing, Bioinformatics, Pharmaceutical Research, Prediction of diseases |
| Manufacturing | Product Research, Engineering Analysis, Quality Analysis |

**Table 1: Type of industries that use big data in different fields**

## 7. RAPID GROWTH OF THE GLOBAL DATA

According to a survey conducted by CMC, the production of the data is expanding at an astonishing pace. It predicts there will be 4300 % increase in annual data generation by 2020. According to that survey, data production will be 44 times greater in 2020 than it was in 2009.

## 8. FUTURE WORK

- Detecting the problems in Big data analytics
- Solution of the detected problem
- Defining a universal model for all types of data analytics

## 9. CONCLUSION

After this survey we got that data is growing rapidly irrespective of the type and size. So, we can conclude that in the near future we may deal with some new data definition having more advanced characteristics and there will be some more advanced tools which can solve the problems caused by that new type of data.

## 10.REFERENCES

[1]www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

[2]https ://research.facebook.com/blog/facebook-s-top-open-data-problems/

[3]DBg Database Group MIT Computer science by Michael Stonnekar

[4]https://blogs.msdn.microsoft.com/shishirs/2014/11/30/big-data-internet-of-things-and airlines/

[5]https ://research.facebook.com/blog/facebook-s-top-open-data-problems/

[6] S. Sruthika, N. Tajunisha, A Study On Evolution Of Data Analytics To Big Data Analytics and Its Research Scope. in: Paper presented in IEEE Sponsored International Conference on Innovations in Information Embedded and Communication Systems ICIIECS, 2015.

[7] Y. Demchenko, P. Membrey (2014), Defining Architecture Components of the Big Data Ecosystem. in: Paper published in Collaboration Technologies and systems (CTS),pp-104-112.doi:10.1109/CTS.2014.68675

[8] D. Singh, CK. Reddy (2014), A survey on platforms for big data analytics. In: paper published in Journal of Big Data. doi : 10.1186/s40537-014-0008-6.

[9] H. Hu, Y. Wen, TS. Chua, X. Li (2014), Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. In: Access, IEEE (Volume:2 ), PP- 652 - 687. doi : 10.1109/ACCESS.2014.2332453.

[10] RJT. Morris, BJ. Truskowski (2003), The Evolution of Storage Systems. In : IBM Systems Journal (Volume:42 , Issue: 2 ). PP - 205-217. doi: 10.1147/sj.422.0205.

[11] RIDER, FREMONT. The Scholar and the Future of the Research Library. 236 pp. New York, Hadham Press, 1944.

[12] Automatic data compression, published in Communications of the ACM, Volume 10 Issue 11, Nov.

1967 Pages 711-715, doi:10.1145/363790.363813.

[13] Y. Chen , S. Alspaugh, and R. Katz, Interactive analytical processing in big data systems: A cross-industry study of map reduce workloads, Proc.VLDB Endowment, vol. 5, no. 12, pp. 1802_1813, 2012.

[14] J. Gantz and D. Reinsel, The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, In : Proc. IDC I View, IDC Anal. Future, 2012.

[15] B. Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams With Advanced Analytics, vol. 56. New York, NY, USA:Wiley, 2012.

[16] N. Tatbul, Streaming data integration: Challenges and opportunities, In : Proc. IEEE 26th Int. Conf. Data Eng. Workshops (ICDEW), Mar. 2010, pp. 155_158.