Big Data in Industry: A Study and Survey of Applications and Challenges

Tushar Chavan*, Mandar Jagtap, Aditya Kumbhar, Sahil Shelke
* Department of Mechanical Engineering, Walchand College of Engineering, Sangli, India

ABSTRACT
Now-a-days we rarely observe any company or any industry who don't have any database. Industries with huge amounts of data are finding it difficult to manage. They all are in search of some technology which can make their work easy and fast. The primary purpose of this paper is to provide an in-depth analysis of different platforms available for performing big data over local data and how they differ with each other. This paper surveys different hardware platforms available for big data and local data and assesses the advantages and drawbacks of each of these platforms.

KEYWORDS: Big data, Local data, HadoopBase, Clusterpoint, Mongodb, Couchbase, Database.

INTRODUCTION
This is an era of Big Data. Big Data is making radical changes in traditional data analysis platforms. To perform any kind of analysis on such huge and complex data, scaling up the hardware platforms becomes imminent and choosing the right hardware/software platforms becomes very important. In this research we are showing how big data has been improvising over the local databases and other technologies.

Present day, big data is making a huge turnaround in technological world and so to manage and access data there must be some kind of linking between big data and local data which is not done yet. This whole study is based on the relations between big data and local data and how big data is better than any other databases. In our study we are overcoming the problem of connectivity between big data and local data. The linking will be done by connecting to the server which will help both the databases to share their data.

This article is structured as follows: Section 2 consist of the purpose behind writing this paper and studying the corresponding topics, Section 3 present the background details about the study, Section 4 include the comparison of between different types of data and databases, Section 5 resides the conclusion laid from the study.

PURPOSE
The purpose behind this research is to show how big data is ruling the databases of the industrial giants, to put light on its different applications, how much feasible it is over local data or relational database. Big data today is even in the form of cluster also referred by HBase which works over Hadoop Distributed File System (HDFS), now here the data is stored in the form computational cluster, as Hbase is an unstructured database so if in the case of cloud disaster the data will be lost thus, the paper even suggests how this can be overcome. Paper provides a comparative study between big data and small data so that a user can easily establish one for motive.

BACKGROUND
Big Data
Big data is a term that describes the large volume of data – both structured and unstructured which is used to manage and access huge amounts of data. As far as the industrial giants are concerned they have huge amounts of data which they are finding it difficult to handle. According to our research we are showing the importance of big data how easily and quickly we can manage, access, retrieve data for future usage. Data will be processed with various tools which includes analytics and algorithms to provide meaningful information. Big data is utilized most by media. These media industries are moving away from using newspapers, magazines, television shows and instead they are showing interest with working with consumer mindset. Big data analytics have been very helpful in healthcare by providing medicines and perspective analytics.

Local Database
A local database is one that is local to our application only. In this study the local database uses an SDF file which is in SQL format. What this does is if any error occurs in the cloud then the data in the cloud can be stored in the local databases like Mongodb. This prevents loss of data and data will be safe.

Clusterpoint
Clusterpoint is a framework which is used for faster access of data. Suppose we are having a database which includes data then that data can be written in raw format. That raw data can perform many transformations.

Security is very important for keeping the data safe. Through clusterpoint a security code is generated for every company or an industry through which they can keep their data safe.

Clusterpoint is very strong technically and have many advantages like in it increases the speed of search and query, it even reduces the cost by scaling elastically on commodity hardware. So because the cost is low then the industries would love to purchase this technology so that their problems may be solved.

Hadoop base

Hadoop Base is an open-source framework which allows storing and processing of big data across clusters of computers which is in distributed environment with the use of normal programming models. In this there is a master node that manages the cluster of data that stores the portions of tables and performs the work. As the Hbase is very sensitive to loss of its master node, so in case of any cloud disaster the data of Hbase can be stored in the local database like mongoDB or any other local database. HBase allows for many attributes to be grouped together into what are known as column families, such that the data of a column family are all stored together. This is lot different from a row-family where all the columns of a given row are stored together so that the data can be seen at one place.

In a database which uses multiple machines the work is divided. All the data is accessed on one or more machines and all the data processing is done on another machine or server.

Now on a hadoop cluster, the data inside HDFS and other systems are accessed on every machine in the cluster. There are two benefits of this, firstly if any of the one machine in the cluster goes down it provides redundancy to the system and secondly it enable a data processing software in the machine where the data is actually stored and which further enhances the speed of retrieving the information.

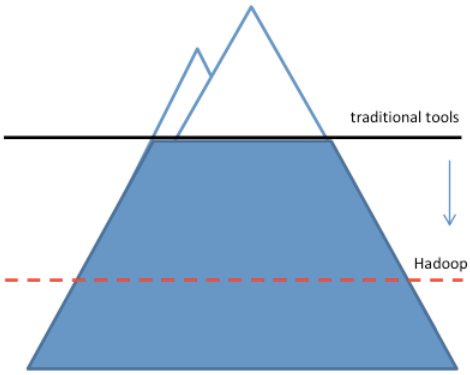

Fig 1: Hadoop Level

Couchbase

Couchbase is an open source software which is used to optimize various applications. We can create, retrieve, store, manipulate, and present the data in different ways. We can say it is a very precise and formal way of presenting the data. Couchbase

MongoDB

Mongodb is a platform which is a collection of various documents. In this research we are having a set of data or set of collections then one collection can hold different documents at one time. Mongodb is one the most widely used local databases for storing huge data of industries so that the data remains safe and can stay for further use. In big industrial giants they have huge data they have their data in documents. They have a collection of documents so it is very difficult to manage so many documents of data at one time. So to make this task easier and faster mongodb came into act.

This has a number of advantages over files stored in a file system. Unlike file system, the database supported by MongoDB can deal with thousands of objects without any problem. In addition to that we get the power of the database when we are accessing or dealing with this data. We can do advanced queries to find a file, using indexes, even can do clear stuff like achieve redundancy of the whole file set. Objects in MongoDB are stored in the binary form which is known as BSON. Big Data is a BSON data type for a binary byte array. MongoDB objects are all limited to 4MB in size and to achieve this bigger files are divided into smaller or into multiple objects of each size less than 4MB. This has the advantage of letting us efficiently retrieve a specific range of the file.

3.7.More about Big Data

- In the past years 5 Exabytes of data were created by human. But today this work is done in two days. Big data requires a huge revolutionary step forward to take the data analysis era to new heights. This depends three main components variety, volume and velocity of data.
- Big data analytics have a significant role of cloud which will grow as cloud is getting adopted by a big number of organizations. Service oriented computing is achieved by cloud computing and it provides three types of services at various IT levels: Infrastructure as a Service (IAAS), Platform as a Service (PAAS) and Software as a Service (SAAS).
- Big data has another important aspect Data mining which is used for the collection of things from different cases having different patterns permitted, which is recognized by simplification and summarizing of data.
- Big data have various issues; is the size of the data, big data is difficult to work with using relational databases and desktop statistics, requiring massive software's running simultaneously on tens, hundreds, or even thousands of servers. With the invention of new technologies and tools required to build big data solutions the availability of skills is a big challenge. A higher level of professionals in the data sciences are required to access big data solutions today because they are not yet familiar with the tools and still IT graduates are required to configure a big data system.
- There are various ways to leverage big data:
  - Big data can unlock significant values by making information transparent; as there are many information which are not yet snapshotted in digital form like any data on hardcopy or not easily searched.
  - As organisations create and store more transactional data in digital form, more information can be accurately collected.
  - Big Data allows narrower separation of customers and thus easily provide products and services.
  - Analytics can substantially improve decision-making, minimise risks, and unearth valuable insights that would otherwise remain hidden.

COMPARISONS

Between Big Data & Small Data

CATEGORY	BIG DATA	SMALL DATA
DATA SOURCES	This includes data generation from non-traditional data sources outside the enterprise, E.g.: • Device data • Log data • Social media • Sensor data • Images, videos, etc.	While in contrast, the small data is generated from traditional data sources like, • Customer Relationships Management (CRM) systems • Web transactions • General ledger data or Financial data • Enterprise resource planning transactional data
VOLUME	Big data has a wide range of volume measuring unit, E.g.: • Terabytes (10^{12}) • Petabytes (10^{15}) • Exabytes (10^{18}) • Zettabytes (10^{21})	Small data has only 2 measuring units for volume, • Gigabytes (10^{9}) • Terabytes (10^{12})
VELOCITY	❖ Data is often dealt in Real time. ❖ Data stream requires immediate response in case of change applied.	❖ Data is dealt in Batch or near Real time. ❖ Data stream does not constantly require immediate response.
VARIETY	Usually support 3 types of data: • Structured: Data in row column traditional tables. • Un-Structured: even includes multimedia content, text, etc. • Multi-Structured: resides in non-transactional systems and have many formats of application	Support only 2 types of data. • Structured • Un-Structured
VALUE	Supports complex, advanced, predictive business analysis and insights.	Supports business intelligence, analysis and reporting.

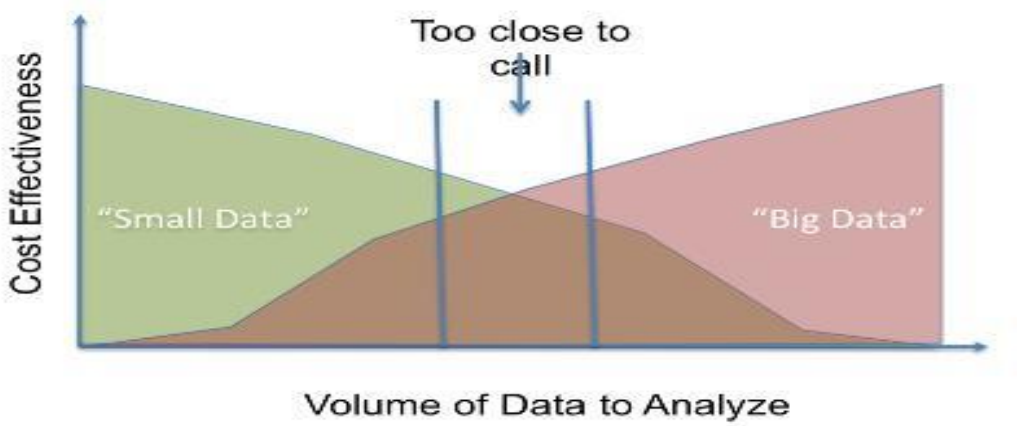

Fig 2: Big Data v/s Small Data

4.2. Between Big Database & Local Database

BIG DATABASE	LOCAL DATABASE
Data can be in form be it tables, multimedia, cloud, etc.	Data is in the form of traditional tables.
They have Hadoop, Hbase, Cluster and Cloud database management systems.	These are managed by Relational Database Management System and have oracle database IBM db2, MySql.
They support a large number of data, huge in size and formats.	They are not capable of storing and are feasible for smaller data with a fixed format.
They support all three types of data structured, unstructured and multi structured.	Supports only structured data.
Have a wide range of applications in government fields, international development, media, education, manufacturing, health care, Information technology and many more.	Its applications are in education, projection operations, relational calculus or algebra.

CONCLUSION

Now a days data managing is becoming very easy because of the introduction of Hadoop and cloud based analytics, this is becoming cost effective comparing between big data techniques and traditional architecture is tedious because of different in functionality nut order of magnitude can also be suggested by difference in price. Big data supports fast and better decision making, introducing new products and services which are usually ready for prime time. In older days systems were used to Extract, Transfer and Load data through systems into big warehouses supporting business intelligence initially data was stored into a data base where the reports used to run and everyone had access to it, thus database technologies simply couldn't handle multiple, continuous stream of data or a big volume of data while in contrast big data is secure, timely accessible as no manual time is wasted on managing data, it is relative authoritive and actionable fulfilling every demand of the user. In this paper we put light on some bid data and local database concept, HadoopBase, Clusterpoint, Couchbase, MongoDB, big data analysis and mining and how the insights into real time error by keeping up the customer trends on high point.

Where at one place big data is capable of storing huge amount of data it might lack somewhere or the other, in case of Cloud based big data or Hbase which is based on cloud if any cloud disaster occurs, by any chance if the cloud gets corrupts because of inefficiency of real time handling the data in the cloud might halt and might get distorted. We can manage this condition efficiently by establishing connection between big database and local database, by doing so we will still have our data in the basic formats of local database like MongoDB. This can be achieved by creating a reference or a copy of data every time it is stored in the cloud or Hbase further this reference can be linked to the local database or the copy can be provided to the local database which in turn will handle the data according to its module, accepting the fact that creating reference or copy of huge amount of data will be inefficient or how a local database will be able to handle such a big data is secondary as at least we will have a source of retrieving the data which got distorted or corrupted cause of external affairs.

At last, a system can be gingerly designed so that the unstructured data can be connected through their composite relationships forming valuable patterns and the improvement in the data capacity and association should assist patterns to guess the tendency and succeeding.

REFERENCES

[1] Understanding MongoDB, http://blog.mlab.com/2014/01/how-big-is-your-mongodb/.
[2] CouchBase Server, https://en.wikipedia.org/wiki/CouchBase_Server.
[3] Big Data and Open Data, https://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government.
[4] Small data vs. Big Data, https://www.linkedin.com/pulse/20140703195144-246665791-small-data-vs-big-data-back-to-the-basics.
[5] Company use CouchBase, http://www.slideshare.net/ddborkar/how-companies-use-nosql-and-couchbase.
[6] What is Big Data, http://www.sas.com/en_us/insights/big-data/what-is-big-data.html.
[7] Startups using CouchBase, https://www.quora.com/Which-startups-use-Couchbase.
[8] The Local databse, https://docs.mongodb.com/manual/reference/local-database/.
[9] Connecting Data in local Database, https://msdn.microsoft.com/en-us/library/ms171890.aspx.
[10] ClusterPoint No SQL database, http://www.itbaltic.com/e-health/clusterpoint-nosql-database/.
[11] Munesh Kataria, Pooja Mittal, "Big data- a review", Haryana, India, 2014.
[12] Research that change the world of big Data, http://bigdata-madesimple.com/research-papers-that-changed-the-world-of-big-data/.
[13] Three Big Benefits of big data, https://www.sas.com/en_ca/news/sascom/2014q3/Big-data-davenport.html.
[14] What is big data, https://marketingtechblog.com/benefits-of-big-data/.
[15] Why big data is the new competitive advantage, http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/.
[16] Hadoop, http://readwrite.com/2013/05/23/hadoop-what-it-is-and-how-it-works/.
[17] Storage in mongoDB, http://blog.mongodb.org/post/183689081/storing-large-objects-and-files-in-mongodb.