# Big Data, Bigger Implementation Using Oracle And Hadoop

**Ronak Juneja[1], Rashmi Arora[2], Sadhana G.[3]**
[1,2,3]Department of Computer Science and Engineering, Dronacharya College Of Engineering , Farrukh Nagar, Gurgaon, India
[1]ronak_juneja18@yahoo.co.in, [2]miley.arora014@gmail.com,
[3]sadhna.bangalore@gmail.com

***Abstract*: Big data-means many different things to different people, but it's no longer confined to the technology only. Today it has become a business priority, giving its ability to profoundly affect the commerce and others. This paper includes enlightens study and implementation of big data and its various techniques to handle it. This paper provides discussion about big data, its dimensions and its significance in the upcoming world. There are various domains under which the study of big data in included. These domains basically divide the application big data all over. The oracle approach to handle the big data through oracle big data appliance x3-2 have been discussed in the paper. The Oracle Big Data Appliance is an engineered system optimized for acquiring, organizing and loading unstructured data into Oracle Database 11g. It combines optimised hardware components with new software solutions to deliver the most complete big data solution. The last section of the paper discusses about hadoop, which handles huge volumes of data across a large number of nodes. Hadoop operates by dividing a "task" into "sub-tasks" that it hands out redundantly to back-end servers, which all operate in parallel on a common data store.**

***Keywords:* Apache, SQL, Data ware house, Linux, parallelism, API**

## I .INTRODUCTION

Today, Obama Administration is announcing the "Big Data Research and Development Initiative"[1]. Big data, the name is itself bigger. Big data include the detailed study of defining, handling and updating bigger amount of data of any enterprise. The applications and implementation of it are increasing day by day. Today ever new comer in the companies is expected to be well equipped with the aspects of big data. This paper throws a limelight to these aspects about big data :its introduction, its dimensions, domains and what actually big data is.Oracle big data appliance is an engineered system that combines optimized hardware with the most comprehensive software stack featuring specialized solutions developed by Oracle to deliver a complete, easy-to-deploy solution for acquiring, organizing and loading big data into Oracle Database 11g. It is designed to deliver extreme analytics on all data types, with enterprise-class performance, availability, supportability and security. With Big Data Connectors, the solution is tightly integrated with Oracle Exadata and Oracle Database, so we can analyze all your data together with extreme performance. Now a day, in the enterprises the information keeps piling up, as data volumes are doubling annually. And roughly 80 percent of that captured data is unstructured and must be formatted using a batch-processing platform such as Hadoop in order to be minable for information. Hadoop stores its data on hard disks spread across the many nodes. Also, it is open source, which means we save money at the expense of time — it is a developer tool requiring client-side development, but growing and adapting.

## II.INTRODUCING BIG DATA

The operations including the capture, storage, search, sharing, transfer, visualization and analysis seems quite easier while dealing with a small set data .It becomes challenging when we are to deal with collection of data sets that are large and complex. And collection of such large set data sets is termed as big data. Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers"[2][3].We can say big data to be a traditional data of any enterprise.: It must be captured, stored, organized, and analysed, and the results of the analysis need to be integrated into established processes and influence how the business operates[4].But the slight difference between the two is that the companies doesn't deal with the unstructured data as the traditional data enterprise does. Therefore, much of the information available to enterprises isn't captured or stored for long-term analysis, and opportunities for gaining insight are missed [4].Thus there arises the need of various techniques to handle big data effectively and efficiently.

## III.THREE DIMENSIONS: VOLUME, VELOCITY AND VARIETY

1. Volume: Every enterprise has ever-growing data and records of all types. It reaches to terabytes and even petabytes of information.

2. Velocity: sometimes even 2 seconds late provided information hampers the effectiveness of our enterprise. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

3. Variety: Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.

Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile, and to answer questions that were previously considered beyond your reach. Until now, there was no practical way to harvest this opportunity [24] by using big data analysis.

## IV.DOMAINS OF BIG DATA

The study and implementation of big data have led its division under five big domains these are healthcare in the United States, the public sector in Europe, retail in the United States, and manufacturing and personal-location data globally.

1. There is vast implementation of big data. First, big data can unlock significant value by making information transparent and usable at much higher frequency. Second, as organizations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything. Third, big data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services.


**Fig.1**:Big data and it's associated terms.

Fourth, sophisticate analytics can substantially improve decision-making.Finally,big data can be used to improve the development of the next generation of products and services.

2. Data have swept into every industry and business function and are now an important factor of production, alongside labour and capital.

3. From the standpoint of competitiveness and the potential capture of value, all companies need to take big data seriously. In most industries, established competitors and new entrants will have to be competent with its use, application and deeper knowledge.

4. The use of big data will underpin new waves of productivity Growth and consumer surplus. [5]

## V.FUTURE OF BIG DATA

Big Data is a sea change that, like nanotechnology and quantum computing, will shape the twenty-first century [6].Another perspective frames this new ability to unveil stylised facts from large datasets as "the fourth paradigm of science" [23].If the question is about development in coming 10 years, the answer wouldn't be straightforward. First, because the new types of data that people will produce in ten years is unknown. Secondly, because a great deal will depend on the future strategic decisions taken by a myriad of actors—chief of which are policymakers? If, however, we ask how Big Data for Development can fulfil its immense potential to enhance the greater good, then the answer is clearer. What is needed is both intent and capacity to be sustained and strengthened, on the basis of a full recognition of the opportunities and challenges. Specifically, its success hinges on two main factors. One is the level of institutional and financial support from public sector actors, and the willingness of private corporations and academic teams to collaborate with them, including by sharing data and technology and analytical tools. Two is the development and implementation of new norms and ontologies for the responsible use and sharing of Big Data for Development, backed by a new institutional architecture and new types of partnerships. [6]

## VI. THE ORACLE APPROACH

Oracle offers a broad portfolio of products including Oracle Big Data Appliance, Oracle Big Data Connectors, Oracle Exadata and Oracle Exalytics In-Memory Machine [4] which helps enterprise to manage, integrate and acquire big data in most economic, reliable and fastest means. Itdeliver sorganized, analysed and secure data with enterprise class performance. Basically oracle big data appliance includes integrated use of hardware and software together for its effective working.


Fig.2 Oracle's big data solution

## VII.About Oracle Big Data Appliance X3-2

Introduced in January 2012, Oracle's Big Data Appliance supports both Hadoop and Oracle's NoSQL database. [12].It is built using industry standard Hardware from sun and market leading cloudera distribution including Apache Hadoop.Using Cloudera's Distribution including Apache Hadoop (CDH) and Oracle NoSQL Database as data management capabilities, the Big Data Appliance runs on Oracle Linuxand Oracle HotSpotJVM.It includes a combination of Oracle Enterprise Manager and Cloudera Manager for both hardware and software cluster administration and monitoring. [13].It is designed tosupportpilot projects, and is flexible enough to grow with the needs of any enterprise. It integrates tightly with Oracle Exadata and Oracle Database using Oracle Big Data Connectors, which enables analysis of both structured and unstructured data in the enterprise.

### A. Hardware specifications:

It requires a full rack configuration of 18 Sun servers for a total storage capacity of 648TB. Every server in the rack has 2 CPUs, each having 6 cores for a total of 216 cores per full rack. Each server has 48GB memory; so that there is total of
864GB of memory per full rack.Further it features the latest 8-core Intel® Xeon E5-2600 series of processors.

### B.Software specifications:

1.Oracle Enterprise Linux 5.6
2.Oracle Hotspot JVM
3.Cloudera Distribution using Apache Hadoop 4
4.Cloudera Manager 4
5.Open Source Distribution of R
5.Oracle NOSQL Database
6.Oracle Big Data Appliance Enterprise Manager Plug-In

## VIII. WORKING OF ORACLE BIG DATA APPLIANCE

The major software components perform three basic tasks:
■ Acquire
■ Organize
■ Analyze and Visualize

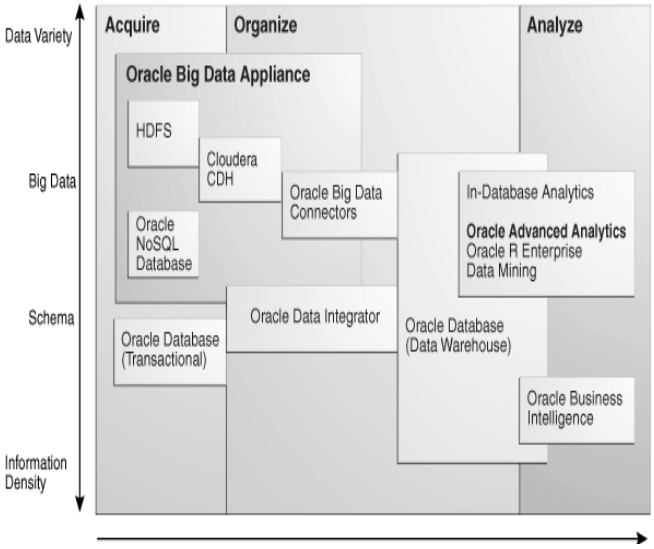


**Fig.3** Oracle Big Data Appliance Software Overview [14]

### A. Acquiring Data for Analysis

Oracle Big Data Appliance provides these facilities for capturing and storing big data:
■ Hadoop Distributed File System (HDFS)
■ Oracle NoSQL Database

**1) Hadoop Distributed File System:** Cloudera's Distribution including Apache Hadoop (CDH) on Oracle Big Data Appliance uses the Hadoop Distributed File System (HDFS). [14]. HDFS stores extremely large files having record-oriented data. It splits large data files into chunks of 64 MB, and replicates the chunk across three different nodes in the cluster HDFS with the help of chunking is able to store files that are larger than the physical storage of one server. It also allows the data to be processed in parallel across multiple machines having multiple processors, with all of them working on data that is stored locally

***2) Oracle NOSQL database:*** Whereas HDFS stores unstructured data in very large files, Oracle NoSQL Database indexes the data and supports transactions. But unlike Oracle Database, which stores highly structured data, Oracle NoSQL Database has relaxed consistency rules, no schema structure, and only modest support for joins, particularly across storage nodes. [14] Oracle NoSQL Database is mainly used for low latency data capture and fast querying of that data, typically by key lookup. I come with an easy to use Java API and a management framework.

***3) Cloudera Manager:*** Oracle Big Data Appliance also contains Cloudera Manager. It is an end-to-end management

application for CDH. Cloudera Manager gives a cluster-wide, real-time view of nodes and services running. It also provides a single, central place to enact configuration changes across the cluster; and incorporates a full range of reporting and diagnostic tools to help optimize cluster performance and utilization. [8]

**B. Organizing Big Data**

Oracle Big Data Appliance provides several ways of organizing, transforming, and reducing big data for analysis:

■Map Reduce
■Oracle R Support for Big Data
■Oracle Big Data Connectors

**1) Map Reduce:** Map Reduce uses a parallel programming model for processing data on a distributed system. [14] It can process very large amounts of data quickly and can scale it linearly. It is an effective mechanism for batch processing of data which can be unstructured and semi-structured. It consists of two phases, the Map phase and the Reduce phase. Each Map phase applies a transform function over each record in the input data, which then produce a set of records expressed as key-value pairs. The output from the Map phase becomes the input to the Reduce phase. In the Reduce phase the Map output records are sorted into key-value sets so that all records in a set have the same key value. [14]

**2)Oracle R Support for Big Data:** R is an open source language and environment for statistical analysis and graphing, providing linear and nonlinear modeling, standard statistical methods, time-series analysis, classification, clustering, and graphical data displays [14]. Using R on a PC limits the amount of data and the processing power available for analysis. Oracle eliminates this restriction by extending the platform of R to directly leverage Oracle Database and Oracle Big Data Appliance to obtain a fully scalable solution. This helps the analysts continue to work on their PCs using the familiar R user interface while manipulating huge amounts of data stored in an Oracle database or in HDFS.

**3) Oracle Big Data Connectors:** Oracle Big Data Connectors facilitate data access between data stored in CDH and Oracle Database. They are licensed separately from Oracle Big Data Appliance.

These are the connectors

■ Oracle Direct Connector for Hadoop Distributed File System
■ Oracle Loader for Hadoop
■ Oracle Data Integrator Application Adapter for Hadoop
■ Oracle R connector for Hadoop.

**Oracle Direct Connector for Hadoop Distributed File System**

Oracle Direct Connector for Hadoop Distributed File System is a high speed connector for accessing data on HDFS directly from Oracle Database.[8] External table are created in Oracle Database, which enable direct SQL access on data stored in HDFS. The data stored in HDFS can then is queried via SQL, joined with data stored or loaded in Oracle It gives the flexibility of querying data from HDFS at any time, whenever needed.

**Oracle Loader for Hadoop**

Oracle Loader for Hadoop (OLH) enables users to use Hadoop Map Reduce processing to create optimized data sets for efficient loading and analysis in Oracle Database 11g.[8].OLH is added as the last step in the Map Reduce transformations and is a separate map – partition – reduce step. Using the CPUs in the Hadoop cluster, this last step format the data into Oracle understood formats, allowing for a lower CPU load on the Oracle cluster and higher data rates because the data is already formatted for Oracle Database. Once loaded, this data is permanently available in the database providing very fast access to it.

**Oracle Data Integrator Application Adapter for Hadoop**

Oracle Data Integrator Application Adapter for Hadoop simplifies data integration from Hadoop and an Oracle Database through Oracle Data Integrator's easy to use interface. Once the data is accessible in the database, end users can use SQL and Oracle BI Enterprise Edition to access data. [8].

**Oracle R connector for Hadoop**

Oracle R Connector for Hadoop is an R package that provides transparent access to Hadoop and to data stored in HDFS. [8]R Connector for Hadoop provides users of the open-source statistical environment R to analyze data stored in HDFS. Also it helps to run R models at scale against large volumes of data leveraging Map Reduce processing – without requiring R users to learn yet another API or language. The advantage is that the end users can leverage over 3500 open source R packages to analyze data stored in HDFS, while administrators do not need to learn R to schedule R Map Reduce models in production environments.

## IX.BIG DATA USING HADOOP

The most popular "spearhead" of Big Data, right now, appears to be Hadoop. As noted, it provides a intensive applications (including Hadoop Common that divides nodes into a master coordinator and slave tsk executors for file-data access, and Hadoop Distributed File System [HDFS] for clustering multiple machines), and therefore allows parallel scaling of transactions against rich-text data such as some social media data. As it turns out, there are limits even to Hadoop's eventual-consistency type of parallelism. In

particular, it now appears that the metadata which supports recombination of the results of "sub-tasks" must itself be "federated" across multiple nodes for both availability and scalability purposes. In fact, Pervasive Software notes that its own investigations show that using multiple-core "scale-up" nodes for the sub-tasks improves performance compared to proliferating yet more distributed single-processor scale-out servers. In other words, the most scalable system, even in Big Data territory, is one that combines strict and eventual consistency, parallelism and concurrency, distributed and scale-up single-system architectures, and NoSQL and relational technologies. Solutions like Hadoop are effectively out there "in the cloud" and therefore outside the usual walls of enterprise data centres. Thus, there are fixed and probably permanent physical and organizational boundaries between IT's data stores and those serviced by Hadoop. Moreover, it should be apparent from the above that existing business intelligence and analytics systems will not suddenly convert to Hadoop files and access mechanisms, nor will "mini-Hadoops" suddenly spring up inside the corporate firewall and create havoc with enterprise data governance. The use cases are simply too different. The whole idea of Big Data brings with it its own special tools and frameworks that are needed to manage the truly enormous mountains of data that are generated, analyzed, and correlated. One of the frameworks that have found success in Big Data is Hadoop, which is managed by the Apache Foundation. Hadoop is used by a wide variety of organizations to manage and process large quantities of data across computer clusters using simple programming models.Hadoop can handle all types of data from disparate systems: structured, unstructured, log files, pictures, audio files, communications records, email– just about anything you can think of, regardless of its native format. Even when different types of data have been stored in unrelated systems, you can dump it all into your Hadoop cluster with no prior need for a schema. By making all of your data useable, not just what's in your databases, Hadoop lets us see relationships that were hidden before and reveal answers that have always been just out of reach. You can start making more decisions based on hard data instead of hunches and look at complete data sets, not just samples.

**A. Apache hadoop has two main subprojects:**

*1)Map Reduce :* The framework that understands and assigns work to the nodes in a cluster.

*2)HDFS :*A file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system. HDFS assumes nodes will fail, so it achieves reliability by replicating data across multiple nodes.

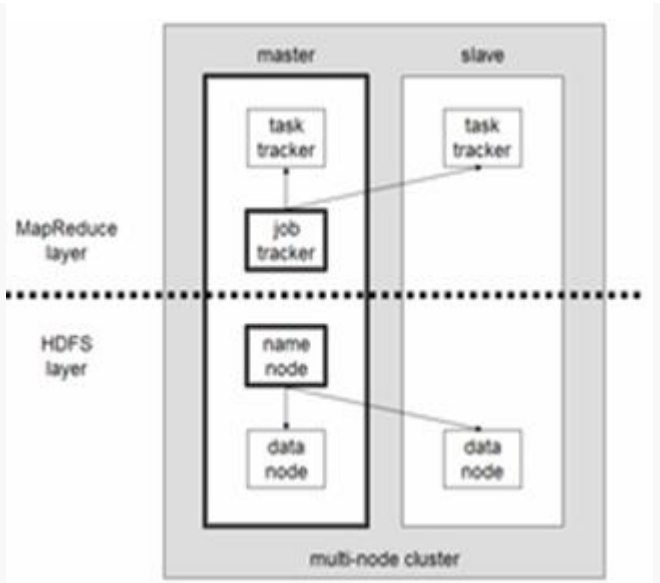


Fig 4:' *How the final multi-node cluster will look like*

**B. Hadoop enables a computing solution that is:**

**Scalable**– New nodes can be added as needed and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top.

**Cost effective**– Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.

**Flexible**– Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.

**Fault tolerant**– When you lose a node, the system redirects work to another location of the data and continues processing

**C. The Benefits of Hadoop are as follows: [22]**

- Reduced computational processing times by providing a much faster way of handling data processing in large data volume scenarios. The HDFS and Map Reduce applications allow the processing of large data sets in much faster time. This benefit is seen in response times for internet search engines and the websites of large online traders where a query is computed in a short time.

- It provides redundancies for data and applications. By spreading an application and the associated file systems it creates an environment where data and application services are backed up. This prevents entire system failure in case of an error in any one of the various nodes that make up the system.

## X.CONCLUSION

This paper is about the definition and implementation of big data. The big data analysis and furthermore its application by using oracle and hadoop.Big data basically is dealing with larger amount of records or data of any enterprise. Such data basically could not be dealt with traditional methods rather need some higher approach or technique to handle it. This paper have successfully introduced with the big data terminology and its definition. Big data also refer with the management, updating and approaching the data. This paper has dealt with the various aspects or say dimensions of big data. The various domains where it's needed are also thrown into view. To derive real business value from big data, we need the right tools to capture and organize a wide variety of data types from different sources, and to be able to easily analyze it within the context of all your enterprise data. By using the Oracle Big Data Appliance and Oracle Big Data Connectors in conjunction with Oracle Exadata, enterprises can acquire, organize and analyze all their enterprise data – including both structured and unstructured. Hadoop, comprised at its core of the Hadoop File System and Map Reduce, is very well designed to handle huge volumes of data across a large number of nodes. At a high level, Hadoop leverages parallel processing across many commodity servers to re the key difference is, rather than only looking at parallel computing, it looks at parallelizing the data access. Spond to client applications. Scientific applications which require high degree of parallelism or need to operate on large volumes of data also benefit from Map Reduce and Hadoop. Scientific applications are mostly used by companies in the Bioinformatics and Healthcare verticals.

## REFERENCES

**[1]** Post by Tom Kalil, "office of science and technology policy"

**[2]** Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue.

**[3 ]** Bid data ,from Wikipedia ,the big encyclopaedia .

**[4]** Oracle-white paper ,integrate for insight.