INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

BIG DATA: THE CHALLENGES, SOLUTIONS AND FUTURE SCOPE

MS. DEEPASHRI S. KHAWASE[1], MR. H. R. DESHMUKH[2], S. H. KUCHE[3], A. S. MAHALLE[3]

1. M.E I Year, IBSS COE, Amravati.
2. Prof and Head, CSE, IBSS COE Amravati.
3. Asst. Prof., CSE, IBSS COE Amravati.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: Big data is the voluminous amount of unstructured and semi-structured data including data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data is referred to huge quantities of data, social media analytics, real-time data and next generation data management capabilities. Big Data today is also a business priority, having ability to affect commerce in the global integrated economy profoundly . Along with providing solutions to long-standing business challenges, big data is inspiring various ways to transform processes, organizations and even entire industries. Software frameworks such as Hadoop and MapReduce, which support distributed processing applications across relatively inexpensive commodity hardware, make it possible to mix and match data from many disparate sources. This paper includes the overview of the new challenges, existing solutions and future scope of Big Data in Business Intelligence.

Keywords: Big data, Business Intelligence, New challenges, Apache Hadoop, Map Reduce

PAPER-QR CODE

Corresponding Author: MS. DEEPASHRI S. KHAWASE

Access Online On:

www.ijpret.com

How to Cite This Article:

Deepashri Khawase, IJPRET, 2014; Volume 2 (9): 679-686

INTRODUCTION

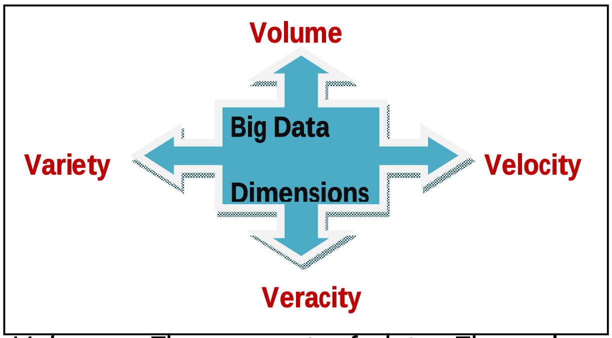
Big Data is emerging as a discipline. Notwithstanding the usual scalability, performance, and modeling concerns, Big Data signifies a radical shift in the way we can leverage data in both science and business. Big Data has the potential to revolutionize not just research, but also education .Big Data Analytics explain the efficient use of a simple model applied to volumes of huge data that would be too big for the traditional analytical environment.

Solutions are needed for big data analysis ,that is from acquiring the data and to discover new insights to make repeatable decisions and scale the associated information systems .The challenge that companies are facing today is how to analyze this massive amount of data to find those critical pieces of information that provide a competitive edge. Different techniques applied to analyze such a mass data include analyzing raw data, unlocking Big Data, Mining and Predicting and decision management. Hadoop provides the framework to handle massive amounts of data to either transform it to a more usable structure and format or analyze and extract valuable analytics from it. Big data spans four dimensions: Volume, Velocity Variety and Veracity.

I. Materials and Methods:

*The 4V's of Big Data***:**

The convergence of these four dimensions helps both and distinguish big data:



Volume : The amount of data. The volume refers to the large quantities of data the organization are trying to harness to improve the decision-making across the enterprises. Data volumes are increasing at an unprecedented rate.

Variety: Variety is managing the complexity of multiple data types that include structured, semi-structured and unstructured data. Organizations feel the need to integrate and analyze data from a complex array of the traditional and non-traditional information sources. In awash of technology advent like sensors, smart devices and social collaboration technologies, data is thus being generated in countless forms, that include: text, data from web, tweets ,sensor data, audio data, video data, click streams, log files and much more.

Velocity: Data in motion. The speed at which data is being created, processed and analyzed is continuously accelerating. Contributing to higher velocity is the real-time nature of data creation, along with the need to incorporate streaming data into business processes and decision making.

Veracity : Data uncertainty. Veracity is referred to the level of reliability that associates with certain kind of data. Striving for high data quality is an important big data requirement and challenge, but even the best data cleansing methods cannot remove the inherent unpredictability of some data, like the weather, the economy, or a customer's actual future buying decisions.

III. Techniques for Analyzing Big Data –

Big data analysis is making "sense" of large volumes of diverse data that in its raw form lacks a data model to define what each element means in the context of the others. The figure 2 illustrates the techniques to analyze Big Data.

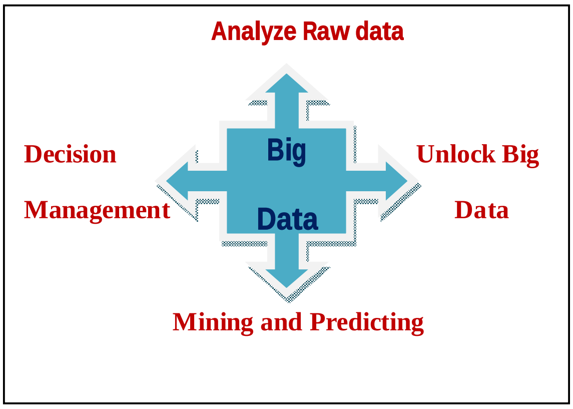


Figure 2: *Big Data Analyzing Techniques*

The Techniques that are used for the Big Data analysis as described as below:

1. *Analyze Raw Data*– In many cases you don't really know what you have and how different data sets relate to each other. You must figure it out through a process of exploration and discovery.

2. *Unlock Big Data* – Because the actual relationships are not always known in advance, uncovering insight is often an Iterative process as you finds the answers that you seek. The nature of iteration is that it sometimes leads you down a path that turns out to be a dead end.

3.*Mining and Predicting* – Big data analysis is not black and white. You don't always know how the various data elements relate to each other. As you mine the data to discover patterns and relationships, predictive analytics can yield the insights that you seek.

4. *Decision Management* – Consider the transaction volume and velocity. If you are using big data analytics to drive many operational decisions then you need to consider how to automate and optimize the implementation of all those actions.

IV. Big Data –Existing Solutions:

The solutions the organizations have adopted to handle Big Data include:

1. Apache Hadoop and MapReduce
2. IBM Big Insights
3. Teradata Aster Data
4. Oracle Big Data
5. Greenplum HD(EMC)
6. SAP Hana

V. Introduction and Working Of Hadoop

Hadoop is the popular open source implementation of MapReduce, a powerful tool designed for deep analysis and transformation of very large data sets.

In a Hadoop cluster, data is distributed to all the nodes of the cluster as it is being loaded in. The Hadoop Distributed File System (HDFS) as shown in (figure 3) will split large data files into chunks which are managed by different nodes in the cluster.

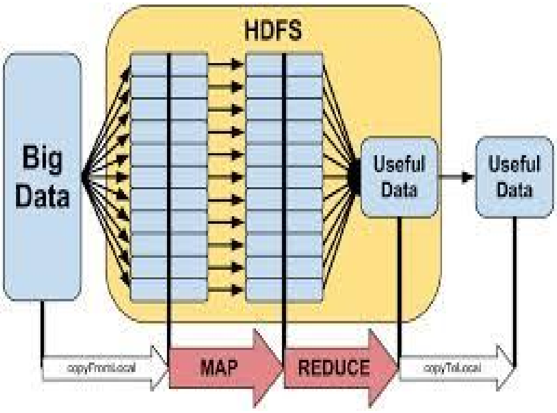


Figure 3: *The Hadoop Distributed File System (HDFS)*

Hadoop enables you to explore complex data, using custom analyses tailored to your information and questions, provides storage capability through a distributed, shared-nothing file system, and analysis capability through Map Reduce. Data is conceptually **record-oriented** in the Hadoop programming framework. Individual input files are broken into lines or into other formats specific to the application logic. Each process running on a node in the cluster then processes a subset of these records. The Hadoop framework then schedules these processes in proximity to the location of data/records using knowledge from the distributed file system. Hadoop will not run just any program and distribute it across a cluster. Programs must be written to conform to a particular programming model, named "MapReduce “.

VI. MapReduce:

The MapReduce component of Hadoop is a framework for processing huge data sets on the Hadoop cluster (Figure 4). MapReduce workloads can be divided into two distinct phases:

Map phase: The submitted workload is divided into smaller sub workloads and assigned to mapper tasks. Each mapper processes one block of the input file. The output of the mapper is a sorted list of key-and-value pairs.

Reduce phase: The input for the reduce phase is the list of key-value pairs received from mappers. The job of a reducer task is to analyze, condense, and merge the input to produce the final output. The final output is written to a file in HDFS.

Individual node failures can be worked around by restarting tasks on other machines.

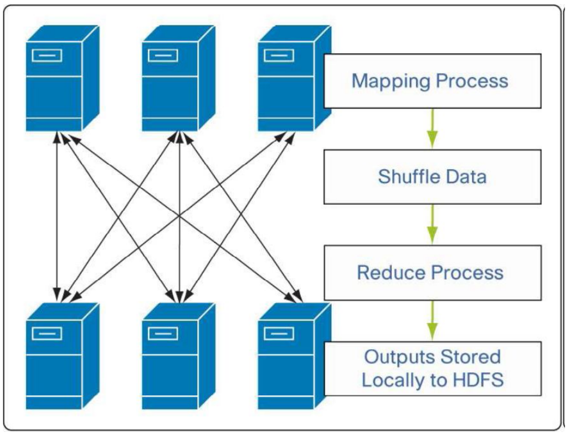


Figure 4: *Hadoop MapReduce*

The other workers continue to operate, leaving the challenging aspects of partially restarting the program to the underlying Hadoop layer.

VII. Future Scope and Challenges:

Future Scope: Owing to the ever growing Internet and media world Big Data have a vast scope and opportunities in near future in the following areas:

1. Non-traditional forms of media
2. Real-time information
3. Data influx from new technologies
4. Social media data

Future Challenges:

Big data introduces a great challenge for database and data analytics. The challenge is the ability to do something meaningful with that data, cost effectively. The Challenges are:

1) Architecture: The challenge is - knowing how an optimal architecture of an analytics system should be to deal with historic and real-time data at the same time.

2) Distributed mining: Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and to detect change first.

3)Security Challenges: Data tampering at the public cloud is predictable and is a big concern. To find a robust security mechanism for making use of the public cloud like Cloud Computing Storage System (CCSS) is a challenging problem.

VIII. CONCLUSION:

The value from analysis on structured, transactional data is well understood and much of its value has been realized. Forward-looking models and other analysis that benefit from larger, more unstructured data sets not as well understood, yet experts suggest that this new frontier of analytics holds untapped promise.

If the enterprise has an unmet business need for strategic decision making ,Analytics and Hadoop combination offers significant opportunity to gain advantage. Big data is already a fact of life for enterprises, but the sheer volume and massive complexity of big data can feel overwhelming. Companies suddenly must struggle with making sense of and creating opportunities from both data at rest and data in motion, from structured, unstructured, and multi-structured data.

IX. REFERENCES:

1. Sam Madden Massachusetts Institute of Technology From Database to Big Data,IEEE,2012

2. P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis,Understanding big data – Analytics for enterprise class Hadoop andstreaming data, McGraw-Hill, 2012

3. Xindong Wu,,Xingquan Zhu, Gong-Qing Wu, Wei Ding, "DataMining with Big Data", 1041-4347/13, 2013 IEEE

4. Wei Fan, Albert Bifet, " Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2

5. Mrs. Deepali Kishor Jadhav, "Big Data: The New Challenges in Data Mining ",International Journal of Innovative Research in Computer Science & Technology (IJIRCST),ISSN: 2347- 5552, Volume-1, Issue-2, September, 2013.