

# Managing Distributed Large System with Big Data



**Sourav Chandra**

**180309**

**Cognizant**

**12/22/2015**

## Contents

1.	Introduction .....	2
2.	Facts and Statistics till Jan 2015.....	2
3.	Big Returns from Big Data.....	3
4.	What is Big Data? .....	4
5.	Big Data is not new, Origin comes from Google.....	5
6.	Why Hadoop?.....	6
7.	What's the big deal? .....	7
8.	Big Data Processing for Better Decisions.....	8
9.	Living on the edge of Analytics .....	8
10.	Big Differences - Conventional RDBMS and Hadoop.....	9
11.	Conclusion: Before you think Hadoop. ....	9

## 1. Introduction

### How far you can depend on the machines/Servers?

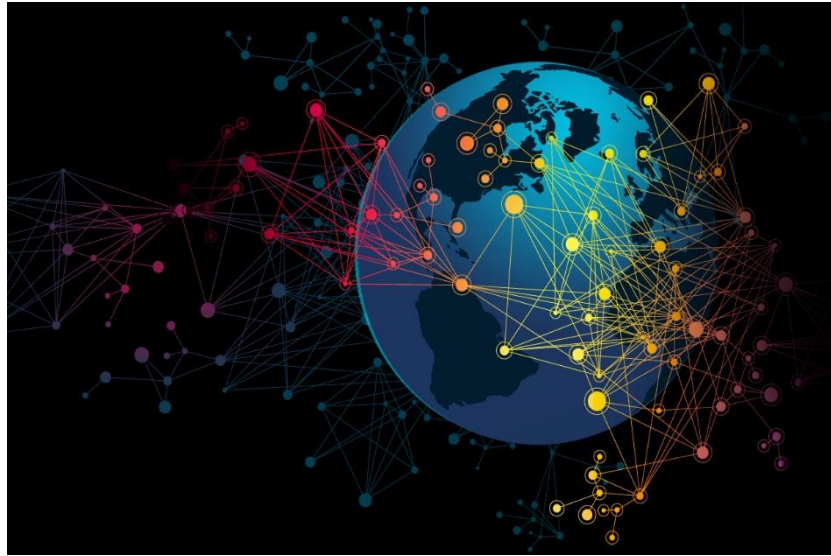
Machines/servers are expected to fail but we can reduce the failure points and maximize the scalability and availability as per business needs. And we always want that, any system which is up and running in Production should never fail completely and it should be optimum in terms of availability and performance.



## 2. Facts and Statistics till Jan 2015

- ✚ The World's Data is Doubling Every 1.2 Years
- ✚ There are 7 Billion people in the world and 5.1 Billion of them own a Cell Phone
- ✚ Each Day we send over 11 billion Texts & Watch Over 2.8 Billion Youtube videos
- ✚ Perform almost 5 Billion Google Searches & Generate 2.5 Quintillion Bytes of Weblogs (10<sup>18</sup>) 1 Quintillion Bytes = 1000 Petabytes
- ✚ For example Consumer Transactions, Communication Device, Online Behavior, Streaming Services etc.
- ✚ In 2015 the World's information Totaled over 5 Zetabytes
- ✚ By 2020 we will need 10x more servers, 50x more Data Management systems and 75x more file to handle it all
- ✚ 80% of this new data is Unstructured which is too large to manage.

### 3. Big Returns from Big Data



Data in the hands of right people is the most important asset an organization has. But, the more data it has to contend with, the bigger the scale of the challenge. The raise in cloud, mobile and social computing means that organizations are faced with ever increasing volumes of new type of Data. There was no Global recession in the growth of Digital data. Between 2005 and 2010 digital Data grew from 130 1227 Exabyte. In today's world as per the current trend Organizations need right kind of people with right kind of tools and technologies to understand the underlying patterns that generate a return on Data. Return on Data can be calculated with the below equation.

$$\text{Return on Data} = \text{Value of Data} / \text{Cost of Data}$$

The difference between winning and losing in the Data driven world will be the ability to reduce the ongoing costs of managing increasing volumes of data with the ability to extract value from it. In this is document some facts and examples are explained to realize big returns from Big Data.

## What makes Data Valuable to any business?

Data with the below qualities can convert any business into Big returns. But most organizations struggling to make this a reality.

- ✚ Timely
- ✚ Accessible
- ✚ Holistic
- ✚ Trustworthy
- ✚ Relevant
- ✚ Secure
- ✚ Actionable

## 4. What is Big Data?

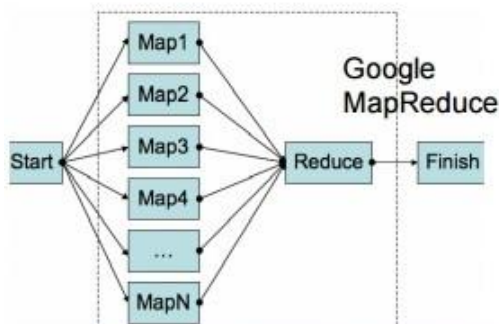
Big Data is not always about the size, size is of the factors but it also depends on the format, manageability of the Data, scalability and availability of the data. Here are few examples of Big Data to understand

- ✚ 40 MB Power point presentation is Big Data if someone cannot share it via mail with your client or customers.
- ✚ 1 TB of healthcare image is Big Data when someone can't share it accurately in a remote screen in real time for doctor to use in consultation with patient.
- ✚ 1 PB of movie/video is Big Data when someone can't edit it within the time constraints of Business.
- ✚ Big data is a new way of thinking about data.
- ✚ Not only the volume it also refers storing, distributed processing, searching, presentation and find value from that stored Data (Structured or Unstructured).
- ✚ Designed for high availability
- ✚ No single point failure and Guarantee consistent read on any node
- ✚ Very high data load rates and Executed as a distributed job
- ✚ Database remains available for query during load

## 5. Big Data is not new, Origin comes from Google

The Apache Hadoop platform is an open source version of the Google File System and Google MapReduce technology. It was developed by the search engine giant to manage and process huge volumes of data on commodity hardware. MapReduce is parallel programming model to process data on multiple systems, developers write Java applications that run in parallel in all the nodes. The key is computation runs on the same node as the data so you don't need to move the data around.

Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.



- ✚ Hadoop is 100% open source MapReduce Platform
- ✚ New and unique way of storing and processing data
- ✚ Enables distributed parallel processing

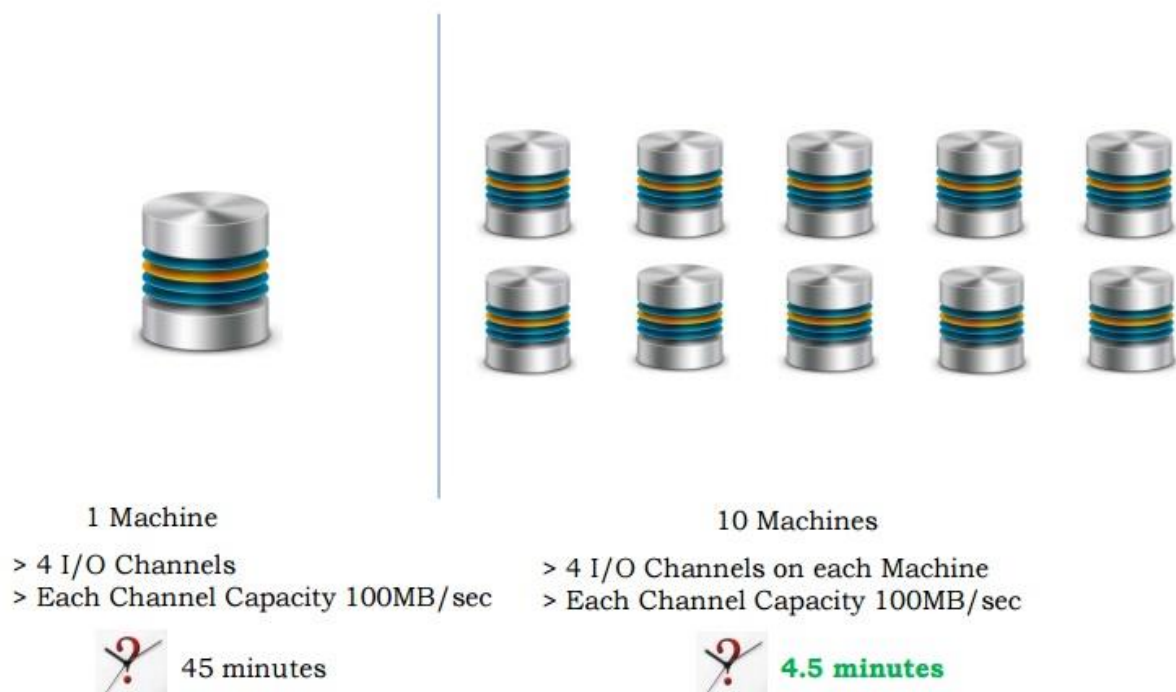


- ✚ With inexpensive and industry-standard servers
- ✚ Scale without limits
- ✚ All types of data

## 6. Why Hadoop?

- ✚ For any organization there are different sources of data. However, since 80% of this data is "unstructured", it must be formatted (or structured) so that it can be processed and stored. And the stored data can be used for Analytics and Reporting.
- ✚ Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes.

### How to read 1TB data (with and without distributed/parallel processing)







This is a simple example (How to read 1TB data) to show how distributed/parallel processing is necessary for any large system when it comes to data volume and processing time. In this example, it clearly shows when data is processed with 1 machine (no scalability and very less availability) it takes more time. And same

amount of data processed 10 times faster with a distributed architecture where the scalability and availability is also optimized.

That's the main driving factor for the most of the organizations to rethink the way they are processing data. And as per the current industry trend commodity and cheap hardware machines can be leveraged to implement this kind of distributed architecture with Hadoop. So Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

## 7. What's the big deal?

Hadoop changes the economics and the dynamics of large scale computing. Its impact can be boiled down to four salient characteristics.

-  **Cost** – 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.
-  **Availability** – When you lose a node, the system redirects work to another location of the data and continues processing without missing a beat. That's how it deals with fault tolerance
-  **Scalability** – 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



## 8. Big Data Processing for Better Decisions



- ✚ Organizations are now looking at different types of sources to collect data and find value out of it. And 80% of the data coming out of these sources are unstructured. With the help of HDFS (Hadoop distributed file system) all these unstructured data can be analyzed, stored and some can also perform queries to answer lot of analytical questions (never been asked before)
- ✚ Both Hive and Pig are very well thought-out tools that enable the lay engineer to quickly being productive with Hadoop. After all, Hive and Pig are two tools that are used to translate analytics queries in common SQL or text into Java Map/Reduce jobs that can be deployed in a Hadoop environment.
- ✚ Hadoop is very popular because within an hour, an engineer can download, install, and issue a simple query. It's also an open source project, so there are no software costs, which makes it a very attractive alternative to Oracle and Teradata.
- ✚ Hadoop is great at analyzing large amounts of data and summarizing or “data pipelining” to transform the raw data into something more useful for another application (like search or text mining) – that’s what’s it’s built for.

## 9. Living on the edge of Analytics

- ✚ Browse, Model and Potential Relations
- ✚ MapReduce Transform, Organize ,Aggregate
- ✚ Identify Patterns across Datasets
- ✚ Data Mining
- ✚ in Database Analytics

## 10. Big Differences - Conventional RDBMS and Hadoop

Attributes	RDBMS	Hadoop
Transaction Per Second	1000's	N/A
Concurrent Queries	100's	10's
Update Pattern	Read / Write	Append Only
Join Complexity	100's of Table	Attribute Keys (Not Predefined)
Schema Complexity	Structured	Structured or Unstructured
Total Data Volume	1000's TB	10's PB
Per Job Data Volume	10's TB	10's PB
Processing Freedom	SQL	MapReduce, Streaming, Hive, Pig etc
Hardware Profile	High End Servers	Commodity / Utility Hardware
Cost Structure	Upfront Licensing + Annual Maintenance	Free Software + Support (Optional)

## 11. Conclusion: Before you think Hadoop.

- ✚ Hadoop is designed for large files, not large quantities of small files, so if you have millions of 50 Kb documents, that is not Hadoop's sweet spot.
- ✚ Hadoop stores its data on hard disks spread across the many nodes. This is opposite to the industry standard of storing the data on a single (or a few) file servers, NAS or SAN. So if you already have big data, then moving to a Hadoop

system will require time and resources to re-architect. Managing Large

Distributed Systems with Big Data Page 10

- ✚ Even though Hadoop leverages many servers, each one requires a significant amount of memory (more than your typical desktop), and if the name node runs out of memory, you are looking at a crash.
- ✚ For harder analytics problems, Hadoop quickly falls flat and requires you to directly develop Map/Reduce code directly.
- ✚ Ironically, it's also a framework that requires a lot of programming effort to get those questions answered.