BIG DATA INTRODUCTION

Reading sources:

1. Book - Understanding Big Data by IBM <http://www.ibmbigdatahub.com/whitepaper/understanding-big-data-e-book> or from<http://www.utdallas.edu/~axn112530/cs6350/Understanding_BigData.pdf>
2. Paper

Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, *19*(2), 171-209.

<http://mmlab.snu.ac.kr/~mchen/min_paper/BigDataSurvey2014.pdf>

or from the ACM Digital Library<http://dl.acm.org/citation.cfm?id=2843712>

# What is Big Data

Answer the following questions by reading the sources mentioned

## Section 1.1 of the paper:

* + 1. What does the term Big Data (BD) refer to? How is BD different from traditional datasets?

Bigdata usually means humongous data sets. This data sets can be structured, semi structured or unstructured. As compared to traditional datasets, big data includes masses of unstructured data that need more real-time analysis.

* + 1. What challenges have emerged because of the rise of BD?

With the vast generation of data everyday we are confronted with the main challenge of collecting and integrating massive data from widely distributed data sources.

With huge generation of heterogenous datasets the main problem arises of storing and managing datasets with moderate requirements on hardware and software infrastructure.

## Section 1.2 of the paper:

1. This section presents several definitions and features of BD. Write down in pointwise fashion the features of BD. Pay special attention to the 3V definition proposed by Laney and understand what each term means.

Here is a start:

* Datasets which could not be captured, managed, and processed by computers within a reasonable time frame [Hadoop]
* Big Data can be defined as data growth whose challenges and opportunities are three dimensional, i.e., increasing volume, velocity, and variety.
* Volume means with generation of data; data becomes increasingly big.
* Velocity means the timeliness of big data.
* Variety indicates the various types of data.

Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of data, by enabling the high-velocity capture, discovery, and/or analysis. [IDC Report]

Big Data can be redefined as four Vs: Volume -great volume, Variety -various modalities, Velocity rapid generation, and Value - huge value but very low density

* Big data means the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies [NIST]

## Characteristics of BD (Chapter 1 of book):

Read and understand the 3V character of BD. Answer the following questions:

1. What is meant by **volume** of BD. How has it changed over time?

Volume of data means the amount of data generated every day. Everyday huge amount of data is being generated.

1. How has **increased volume** created a "blind zone" for organizations?

As data is on the rise, the percent of data it can process, understand, and analyze is on the decline, hence creating the blind zone.

1. What is meant by **variety** of BD? What are the various types of data that large organizations acquire today?

Variety in Big Data refers to all the structured and unstructured data that has the possibility of getting generated either by humans or by machines. Various types of data that large organizations acquire today are raw, semi structured, and unstructured data from web pages, web log files, search indexes, social media forums, e-mail, documents, sensor data etc.

1. How is **velocity** of data applied to data in motion. What are the advantages of **streams computing**?

# What is the value of Big Data

## Section 1.3 of the paper and chapter 2 of the book

1. Read section 1.3 of the paper and chapter 2 of the book. They list several industries (e.g. US medical industry, retail industry, government operations, public health, etc) that can benefit enormously by using Big Data techniques. Choose any one such industry and do research about Big Data applications in that industry. Write a brief 2-3 paragraph report.

Big Data has many applications in various industries. One such application is in the US medical industry. In 2009, google was able to analyze big data and provide more valuable information than

The disease prevention centers. Hospitals were responsible for providing information to disease prevention centers. This took time for hospitals to send information as patients took time to report to the hospitals. Therefore, when the public became aware of new type of influenza, the disease might have already been spread. Google found out that entries in its search engine were frequently sought and use frequencies of the entries were correlated to the influenza spreading in both time and location. Google found out 45 search entry groups that were relevant to outbreak of influenza and created models which forecasted the spreading of influenza and predicted places where influenza spread from.

Big data is also used in medical industry to identify trends and develop better treatment plans or predict risk patterns. Healthcare analytics use big data to compare chronic disease and population growth in neighborhoods to identify problem areas. Hospitals can also use big data to see the trend of appointments patients at specific times of the day and schedule right number of staff during peak hours. One of the most widespread application of big data is Electronic Health Records (EHRs). Every patient has his own digital record which includes demographics, medical history, allergies, laboratory test results, etc. EHRs trigger warnings and reminders when a patient should get a new lab test or track prescriptions to see if a patient has been following doctors’ orders.

Big data analytics used for healthcare could change the way images are read. Algorithms developed analyzing hundreds of thousands of images could identify specific patterns in the pixels and convert it into a number to help the physician with the diagnosis. By diving down into insights such as medication type, symptoms, and the frequency of medical visits, it’s possible for healthcare institutions to provide accurate preventative care and, ultimately, reduce hospital admissions.

# Challenges of Big Data

## Section 1.5 of the paper

* + 1. Read section 1.5 of the paper and summarize in your own words the challenges of developing and managing Big Data applications.

Some of the challenges that arose because of big data are as follows:

Data representation- data representation aims at making heterogenous data meaningful for computer analysis and user interpretation. Improper data representation will reduce the value of the original data thus affecting data analysis.

Redundancy reduction and data compression- datasets comes with high redundancy. Redundancy reduction and data compression is effective to reduce the indirect cost of the system.

Data life cycle management-One of the challenges is that the current storage system do not support s massive data.

Data confidentiality- data service providers find it difficult to maintain and analyze such huge datasets.

Energy management: main frame computing consumes high energy. With the increase of data volume and analytical demands, the processing, storage, and transmission of big data will inevitably consume more and more electric energy.

# Storage for Big Data

We will spend a significant amount of time discussing the storage mechanism of Big Data, so it's good to be familiar with the storage mechanism for Big Data.

## Section 4.2 of the paper

1. What factors should you take into account when using distributed storage for Big Data?

Factors taken in account using distributed storage for big data are as follows

Consistency: Consistency refers to assuring that multiple copies of the same data are identical. A distributed storage system requires multiple servers to store data.

Availability: a distributed storage system operates in multiple sets of servers ensuring that even if one server fails, it does not take down the entire system.

Partition Tolerance: There are multiple nodes in a network. The distributed system should have a certain level of tolerance to problems caused by network failures

## Chapter 4 of the book

One of the most popular distributed storage mechanisms for Big Data is Hadoop. Chapter 4 of the book presents a very good introduction to it.

Fill in the blanks / Short answer questions:

1. Hadoop is top level Apache project written in Java programming language.
2. Hadoop was inspired by Google’s work on its Google (distributed) File System (GPS) and Map Reduce programming paradigm. .
3. Hadoop is different from transactional systems in the following ways:

Hadoop is designed to scan through large data sets to produce its results through a highly scalable, distributed batch processing system. Hadoop is about discovery and making the once near-impossible possible from a scalability and analysis perspective.

1. Two parts of Hadoop are:

Two parts of Hadoop are: a file system (the Hadoop Distributed File System) and a programming paradigm (MapReduce).

1. Why is redundancy built into Hadoop environment?

Data is redundantly stored in multiple places across the cluster. Data and its associated programming across a very large cluster of commodity components. This redundancy provides fault tolerance and a capability for the Hadoop cluster to heal itself. This allows Hadoop to scale out workloads across large clusters.

## Components of Hadoop:

1. The three pieces of Hadoop project are:

The three pieces of Hadoop project are: Hadoop Distributed File System (HDFS), the Hadoop MapReduce model, and Hadoop Common.

## Hadoop Distributed File System:

1. How is it possible to scale Hadoop cluster to hundreds of nodes?

Data in the cluster are broken into smaller pieces and are distributed throughout cluster. The map and reduce functions are executed on smaller subsets thus providing scalability that is needed for big data processing.

1. Each server in a Hadoop cluster uses inexpensive internal (inexpensive / expensive) disk drives.
2. What is data locality. What does it achieve?

Hadoop makes use of available servers in a large cluster These have inexpensive disk drives. MapReduce tries to assign workloads to these servers where the data to be processed is stored. This is known as data locality.

1. What are the benefits of breaking a file into blocks and storing these blocks with redundancy?

Benefits of breaking a file into blocks and storing these blocks with redundancy is higher availability. Small Hadoop cluster chunks run jobs on all servers for better scalability. There is also benefit of data locality.

1. The default size of a block in HDFS is 64 MB.
2. What are the advantages of large block sizes in HDFS?

Working with large block size data is possible in Hadoop, hence each server can work on a larger chunk of data at the same time. Ability to process large chunks of work locally without sending data to other nodes which helps improve both performance and the overhead to real work ratio.

1. What is a NameNode in HDFS? What are its functions?

Hadoop’s data placement logic is managed by a special server called NameNode.

This NameNode server keeps track of all the data files in HDFS, such as

Where the blocks are stored.

Manages File system namespace.

 Assign work to DataNode.

Stores metadata of actual data.

1. All of NameNode's information is stored in (disk / memory).