Big Data Outlines:

1. What is Big Data?

**Big data** is an all-encompassing term for any collection of [data sets](http://en.wikipedia.org/wiki/Data_set) so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.

1. What data types that big data consists of?
   1. Structured data: Predefined data type (Fixed schema).

This type of data normally can be stored into tables with columns and rows.

* 1. Unstructured data: Non-predefined data model.

This includes audio, video and picture files

* 1. Semi-structured data: Structured data embedded with some unstructured data such as email and XML.

1. Why we have big data?
   1. Evolution of Technology
   2. IoTs (Internet of Things)
   3. Social Media
   4. Other factors: Transportation, Retail, Banking & Finance, Media & Entertainment, Healthcare, Education, Government also contributing large amount of data.
2. Problems with handling big data:
   1. Problem 1: growing huge datasets.
   2. Problem 2: data in complex structures
   3. Problem 3: processing data fast
3. What is Hadoop?

**Apache Hadoop** is a framework that allows us to store and process large data sets in parallel and distributed fashion. (Open-source framework)

1. What are components of Hadoop?
   1. **HDFS (Hadoop Distributed File System)** (Storage) – allows to dump any kind of data across the cluster.
   2. **MapReduce** (Processing) – allows parallel processing of the data stored in HDFS.

**Hadoop**

**HDFS (Storage)**: allows to dump any kind of data across the cluster

**MapReduce (Processing)**: allows parallel processing of the data stored in HDFS

1. What is HDFS (Hadoop Distributed File System)?

A distributed file system that provides high-throughput access to application data.

1. What is HDFS consisted of?
   1. **NameNode (Master)**: is the main node that maintains and manages DataNode
   2. **DataNodes (slaves):** are commodity hardware in the distributed environment, stores actual data, serves read/write requests from the clients
   3. **Secondary NameNode**: is not a backup of NameNode, whose main function is to take checkpoints of the file system metadata present on NameNode.

**NameNode**

**(Master)**

**(Slaves)**

**DataNode**

**DataNode**

**DataNode**

**DataNode**

**Secondary**

**NameNode**

1. What is MapReduce?

**MapReduce**: is a **programming framework** that allows us to perform **distributed** and **parallel** processing on large data sets in a distributed environment.

**Map()**

**Map()**

**Map()**

**HDFS**

**Input Data**

**Output Data**

**Reduce()**

**Reduce()**

Aggregated data

**Map Tasks**

**Reduce Tasks**

1. Solutions to big data problems?
   1. Problem 1: Storing exponentially growing huge datasets.

Solution: Hadoop HDFS (easy to add DataNode)

* 1. Problem 2: Storing unstructured data.

Solution: Hadoop HDFS (store any kind of data)

* 1. Problem 3: Processing data fast.

Solution: Hadoop MapReduce (parallel processing of data)

1. What is Apache Spark?

Apache Spark is a unified analytics engine for large-scale data processing (Big Data), with built-in modules for streaming, SQL, machine learning and graph processing.