==> BigData: Big data is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database system tools or traditional data processing applications.

5 V's of Big Data:
-Volume
-Variety - Different kind of data is being generated from different sources.
-Velocity - data is being generated at alarmic rate
-Value - mechanism to bring the correct meaning out of data
-Veracity - Uncertainty and inconsistencies in the data.

Big data Analytics:

Big data analytics examines large and different types of data to uncover the hidden patterns, correlations and other insights.

Stages in big data analytics:

- Identifying Problem
- Designing Data requirments
- Pre Processing data
- Performing analytics over data
- visualizing data

Types of Big data analytics:

- Discriptive Analytics -> what is happening now based on incoming data
- Predictive Analytics -> what might happen in future
- Perscriptive Analytics -> what action should be taken
- Diagnostic Analytics -> why did it happen

Hadoop:

Hadoop is a framework that allows us to store and process large data sets in parallel and distributed fashion.

```
                              Hadoop
                                |
                                |
               HDFS    ------------> (-) <----------   MapReduce
                                |
-------------------------------------------------------------------------------------
                                |
            Storage:                |      Processing:
```

```
         Distributed file       |     Allows Parallel and
           System               |     Distributed processing
                              |
                              |
                              |
```

Hadoop follows Master-Slave architecture:

```
                    MasterNode
                       /|\
                      / | \
                     /  |  \
             SlaveNode   |  SlaveNode
                        |
                     SlaveNode
```

Hadoop Core Components:

--HDFS:

```
                    MasterNode(NameNode) <----> SecondaryNameNode
                       /|\
                      / | \
                     /  |  \
         (SlaveNode)DataNode   |  DataNode(SlaveNode)
                        |
                  DataNode(SlaveNode)
```

--NameNode: -Maintains and Manages DataNodes.

-Records metadata i.e. information about datablocks e.g. location of blocks stored, the size of files, permissions, heirarchy etc.

-Recieves a heartbeat and block report from all the DataNodes.

--DataNodes: -Slave Daemons.

-Stores actual data.

-Serves read and write requests from the clients.

--SecondaryNameNode and checkpointing:

-checkpointing is a process of combining edit logs with Fsimage.

-SecondaryNameNode takes over the responsibility of checkpointing, therefore making NameNode more available.

-allow faster failover as it prevents edit logs from getting too huge.

-checkpointing happens periodically(default: 1 hour)

--HDFS Data Blocks: -Each file is stored on HDFS as blocks.

-the default size of each block is 128MB in Hadoop2.x(64MB in Hadoop1.x).

Hadoop Daemons:

```
                     nodes
                       |
         ----------------------------------------------
         |                                      |
    Master Node                              Slave Node
         |                                      |
         |--ResourceManager <== YARN ==>   Node Manager--|
         |                                      |
         |-- NameNode      <== HDFS ==>      DataNode--|
```