**Overview of BigData & Rise of Distributed File Systems**

**Hadoop Distributed File System**

* MapReduce
* Master & Slave Architecture (5 Daemons)
* Blocks
* InputSplit
* RecordReader
* InputFormat
* Job & Job Submission
* Scheduler Mechanism
* Partitioner
* Combiner
* LocalRunner
* ToolRunner
* Custom (Partitioner, Combiner, RecordReader)
* Joins
  + MapSide
  + ReduceSide
* SecondarySort
* Compressions
  + Codec Snappy, LZO, LZ2, GZip, etc
* YARN Architecture

**Hadoop  Administration**

**Pig**

* Datatypes
* Statements
* UDF
* Joins
* Store

**Hive**

* Datatypes
* Internal/External Tables
* Data loading
* Partition
  + Static/Dynamic
* Bucketing
* UDF
* Performance Tuning

Flume

* Agent
  + Source
  + Channel
  + Sink

Oozie

* Workflow
* Co-ordinator

**Zookeeper**

**Kafka**

* Consumer
* Producer
* Topics

**Storm**

* Bolts
* Spouts

**Spark**

* RDD
* Streaming
* MLib
* Core

**NoSQL Databases**

**CAP Theorem**

**Types of NoSQL**

* 1. **Key-Value Store** – It has a Big Hash Table of keys & values {Example- Riak, Amazon S3 (Dynamo)}
     + DynamoDB, ElasticSearch, Solr,
  2. **Document-based** **Store-** It stores documents made up of tagged elements. {Example- CouchDB}
     + MongoDB, CouchDB
  3. **Column-based Store-**Each storage block contains data from only one column, {Example- HBase, Cassandra}
     + HBase, Cassandra
  4. **Graph-based**- Neo4j
     + Neo4j

Data Science

Machine Learning

Data Mining Tools

     R

     SAS

     Matlab

     Weka

Python

Linux

Internet of Things (IOT)

DevOps