Big Data Topics

Big Data Overview

- What's Big Data?
- Big Data: 3V's
- Explosion of Data
- What's driving Big Data
- Applications for Big Data Analytics
- **Big Data Use Cases**
- Benefits of Big Data

Hadoop(HDFS)

- History of Hadoop
- Distributed File System
- What is Hadoop
- Characteristics of Hadoop
- RDBMS Vs Hadoop(Hive)
- ETL vs ELT
- Hadoop Generations
- Components of Hadoop
- HDFS Blocks and Replication
- How Files Are Stored
- HDFS Commands
- Hadoop Daemons

Types of Data

Structure – tabular data
Semi Structure – JASON, XML, EMAIL
Unstructured – Logs, Image, Video
Data Frequency
Real Time(Streaming)
Near Real Time
Batch
Type of Files
Fixed Width
Delimited
Mainframe files(EBCDIC)

AVRO/ORC/PARQUET
Compression Techniques
Gzip
File level and Block level compression
Partitioning of Data
Random/Hash partitioning

Hadoop 2.0 & YARN

- Difference between Hadoop 1.0 and 2.0
- New Components in Hadoop 2.x
- YARN/MRv2
- Configuration Files in Hadoop 2.x
- Major Hadoop Distributors/Vendors
- Cluster Management & Monitoring
- Hadoop Downloads

Map Reduce

- What is distributed computing
- Introduction to Map Reduce
- Map Reduce components
- How MapReduce works
- Word Count execution
- Suitable & unsuitable use cases for MapReduce

Sqoop

- Architecture
- Basic Syntax
- Import data from a table in a relational database into HDFS
- import the results of a query from a relational database into HDFS
- Import a table from a relational database into a new or existing Hive table
- Insert or update data from HDFS into a table in a relational database

Flume

- Given a Flume configuration file, start a Flume agent
- Given a configured sink and source, configure a Flume memory channel with a specified Capacity

Hive Programming overview