Bigdata Analysis with Power BI

Objective: To pre-process and analyze the large data set and represent it in graphical format via Power BI Tools

Pre-requisites: To learn Bigdata Analytics, you must have a basic knowledge about the Linux commands, Basic SQL Queries and Logic Building

Session 1 (Introduction to Traditional Databases)

- Introduction to database
- Data Models
- ER Diagram
- 3-Tier Architecture
- Entity Relational Model

Session 2 (SQL – Structured Query Language)

- Create Database, Drop Database
- Queries
- Update and Delete Queries
- Create Table and Insert Values
- Logical Operators (AND, OR, NOT)
- LIKE and TOP Clause

Session 3 (Introduction to SQL with Advanced Clause)

- Order By, Distinct Keyword
- Using Joins, Union Clause
- Using Alias and Truncate
- Group By
- SQL Constraints
- UNION
- Having Clause
- Subqueries

Session 4 (Data Backup Process)

- Backup Entire Database
- Backup Single Table
- Backup Single Database

Session 5 (Fundamentals of Linux Operating System)

- Introduction to Linux Operating System
- Types of Linux Versions and Their Uses
- Linux Installation
- Linux Basic Commands and Advanced Commands

Session 6 (VMWare Installation)

- Introduction to VMWare
- Explaining the Packages
- Installing VMWare

Session 7 (Introduction to HADOOP)

- Hadoop Architecture
- Hadoop Distributed File System
- MapReduce
- Environment Setup
- Difference Between Hadoop1, Hadoop2 and Hadoop3
- Man in the middle of attack

Session 8 (HDFS Overview)

- HDFS Architecture
- Importing Data into HDFS
- Data Node
- HDFS Commands
- Name Node
- MapReduce Job Management

Session 9 (Single Node Cluster Configuration)

- Hadoop cluster setup with one VM
- Hadoop Installation and Configuration

Session 10 (Multi Node Cluster Configuration)

- Master and Slave Concept
- Handshaking
- SSH 256
- Hadoop Installation and Configuration

Session 11 (Cluster Maintenance)

- Checking HDFS Status
- Breaking The Cluster
- Adding and Removing Cluster Nodes
- Copying Data between Cluster
- Cluster Upgrading
- Rebalancing The Cluster

Session 12 (Hadoop Administration)

- Hadoop High Availability
- Hadoop Multi Network
- Advanced Concepts of Administrator

Session 13 (Hadoop Ecosystem - Hive)

- Introduction to Data Warehouse
- Hive Architecture
- Installing Hive
- Data Management using Hive
- Hive Partitioning
- Hive Bucketing
- Hive Serde
- HQL
- HIVE Script
- JSON and XML data in HIVE

Session 14 (Hadoop Ecosystem - Pig)

- Pig Overview
- The Need of Apache Pig
- Apache Pig Architecture
- Downloading and Installing Pig
- Pig Latin Basics
- Latin Built In Functions and Data Management
- Difference Between Pig and Hive
- Pig UDF
- Pig Eval Function
- Pig Scripting

Session 15 (Cluster Monitoring, Troubleshooting and Optimization)

- Checking HDFS with fsck
- Breaking the Cluster
- Checking HDFS with fsck
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Clusters Self-Healing Feature

Session 16 (Hadoop Ecosystem - Sqoop)

- Introduction to Sqoop
- Downloading and Installing Package
- Server Installation
- Client Installation
- Upgrading Server
- Sqoop Jobs
- Sqoop Incremental Append
- Sqoop Advanced

Session 17 (Hadoop Ecosystem - Flume)

- The need of Apache Flume
- Data Management using Flume
- Downloading and Installing Flume

Session 18 (Restoring Data)

- Process Understanding
- Pre-requisites for data restore
- Data Restoring

Session 19 (Troubleshooting Cluster)

- Validate Environment Information
- Validate Hadoop Cluster Health
- Troubleshooting HDFS
- Troubleshooting HIVE

Session 20 (Spark, Kafka, Presto, Samza)

- Introduction to Spark
- Introduction to Kafka
- Introduction to Presto
- Introduction to Samza

Session 21 (Power BI Tools)

- Introduction to Power BI
- Connecting with HDFS
- Visualizing data with Power BI

Session 22 (Project Work)