Big Data Analytics

Course 3572 – 40 Hours

Overview

Business success in the information age is predicated on the ability of organizations to convert massive amount of raw data coming from various sources into high-grade business information. Many organizations are overwhelmed by the sheer volume of information they have to process in order to stay competitive. Traditional database systems may become either prohibitively expensive to handle the exponential growth of data volumes or found unsuitable for the job.
Each organization has its own solution to deal with amount of data and the need to analyze the data and get insights.

This course is designed to introduce you to:

- Different Bigdata processing solutions
- How organizations handle Bigdata on premises and in the cloud
- How to analyze data in the era of Bigdata

Hands on exercises are included.

Who Should Attend

This course is mainly intended for Database Developers, Business Intelligence professionals, Data Analysts, Product Managers, and other roles responsible for analyzing high volumes of data.

Prerequisites

- Basic Knowledge of Python, or experience with other programming languages
- SQL
- Hands on experience with Databases

JOHN BRYCE
Leading in IT Education
a matrix company

matrix

Course Contents

BigData solutions overview

- What challenges BigData addresses
- BigData components in a nutshell
- How a typical BigData solution looks like
- How BigData platform handle data - Blocks, Partitions and distributed processing

Cloud and On-Premises solutions for big data

- What do different clouds offer for BigData Solutions
- BigData Solutions On-Premisses
- Hybryd approach
- Non-Cloud-Specific solutions

Introduction to Hadoop

- What is Hadoop and how it handles BigData
- Hadoop components
- What is the Hadoop Data File System and how it works
- Working with Hive
    - Creating tables (and their types - Managed / Extrnal)
    - Popuplating data in Hive
    - Storage Types (Avro, Parquet, ORC) and when to choose which

Intoductions to NoSQL

- What is NoSQL and when to use it
- NoSQL families
- HBase data modeling concepts
- MongoDB
    - MongoDB model concepts
    - MongoDB architecture (Clusters, Sharding)
    - Qurying and Updating Data with MongoDB
- Elasticsearch and Kibana
    - The ELK Stack
    - What is a lucene index
    - How to analyze data with Kibana
- Graph Databases and Neo4j
    - The Neo4j modeling concepts and technique
    - Querying and updating data with Neo4j

Data analytics tools and Methods

- How to approach data analytics
- Data analytics with R
- Data analytics with Python
- Tableau for data analysis and data science
- Data storytelling

Data Ingestion

- Types of sources
- Streaming with Kafka

Data preparation methods

- Handling missing data
- Identifying and handling outliers
- Preparaing data for machine learning

Data preparation tools

- Working with Python and pandas
- Working with Spark
- Overview of existing tools (Tableau Data Prep, Knime, Talend, Trifacta)

Working with PySpark

- Concepts and architecture
- Working with PySpark Core (RDD)
- Working with PySpark SQL (Dataframes)
- Working with Data Streaming
- Data preparation and data preparation for ML

Advanced SQL

- Why SQL is still the best tool for data analytics and data preparation
- Windows functions
- CTE
- Aggregations

Cloud databases (Synapse, RedShift, BigQuery, Snowflake, Presto)

- Overview of the different BigData DBs
- Synapse architecture and concepts
- RedShift architecture and concepts
- Snoflake concepts
- Presto (Starbase) architecture and conepts