Big Data and Data Science Seminar

8 hours

Course Overview:

The continued rise in the volume, velocity and diversity of available data presents both opportunities and challenges for businesses to handle information in an efficient and timely manner.

Big Data:

Big Data technologies enable us to build highly scalable, available and capable applications that can deal with the increasing rise of data which is being generated. It also provides various technologies to process, analyze and visualize these huge amounts of data to gain valuable insights. Leveraging Big Data technologies can enable organizations to target markets, engage with new prospects, compete more effectively and grow. This seminar will provide you with an introduction to the diverse world of Big Data technologies and then focus on the Hadoop ecosystem and Spark.

Data Science:

A lot of organizations experienced a 17-49% increase in productivity when they increased data usability by 10%. Given the positive impact that professionals can have on their organizations' performance by employing data science, we created this seminar to help both business and nonbusiness professionals acquire foundational data science understanding that will help them get right decisions and analyze data better.

This seminar offers an overview of foundational data science and machine learning methods and how they can be used to solve real-world problems.

Who Should Attend:

BI experts, developers, architects, DBAs, Devops experts, IT managers, product managers, decision makers with strong technical background.

Required Skills:

Knowledge and experience with RDBMS and Information systems

Course Contents:

Introduction to Big Data

- History of Database Systems
- Internet of Things and Data Explosion
- Birth of clustered computing systems
- Big Data technologies – overview
    - Apache Hadoop
    - NoSQL
    - Search Engines
    - NewSQL

Hadoop – a closer look

- Introduction to Hadoop
- The Hadoop Eco-System – High level overview of the main tools and technologies
    - Hadoop core – HDFS and YARN
    - Processing frameworks – Map/Reduce and Spark (Including Spark Streaming)
    - Using Hive to enable SQL queries on unstructured data.
    - Selected Hadoop eco-system technologies and concepts.
    - Data ingestion tools, including:
        - Using Sqoop for relational databases integration
        - Using Flume to ingest data from various sources in real time.
- Hadoop and Data Science
    - Leveraging advanced analytics and Machine Learning in Hadoop
    - Clustering, Classification and Regression
    - Supervised vs Non-Supervised learning
- Spark – A closer look
    - Apache Spark basics
    - Spark Data Model and Operations: RDDs, Actions, Transformations
    - Spark Execution Models Overview: YARN, Spark Standalone
    - Spark APIs – working with DataFrames and DataSets
    - Additional Spark capabilities:
        - Spark SQL
        - Spark ML / MLlib (Machine Learning)
        - Spark Streaming (DStreams) and Structured Streaming
        - GraphX
- Hadoop on the Cloud
    - Virtualization and the Private Cloud
    - The Public Cloud – Amazon AWS, Microsoft Azure and Google Cloud
- Real world implementations of Hadoop

NoSQL

- Introduction to NoSQL
- Main characteristics of NoSQL solutions
- The four types of NoSQL solutions, leading technologies and real world use cases:
    - **Key/Value Stores** (DynamoDB, Redis)
    - **Document Stores** (MongoDB)
    - **Wide-Column Stores** (HBase, Cassandra)

- **Graph Databases** (Neo4j)

Search Engines

- Apache Lucene
- Indexing, storing and analyzing data in Search Engines
- Leading Search engines today - Elastic search and SOLR
- Real world use-cases of search engines today

NewSQL databases - Overview

- Characteristics of NewSQL databases
- Leading NewSQL databases today: Vertica, VoltDB and MemSQL.

Data Science:

- Background
  - What is data science?
  - Problems and models
    - Regression
    - Classification
    - Clustering
  - Main challenges
    - Overfitting
    - Feature engineering
- Beyond background
  - The working environment
  - Model selection - cross validation & grid search
  - Big data science with Spark ML
  - Demos