Module Title : Data Engineering Workshop
Duration : 5 days

Overview

For over 10 years, there has been an intense focus by companies to extract business value from their data. Out of this activity, a role called the data scientist emerged. However, it quickly became obvious that a majority of a data scientist's time was spent on data preparation or moving analytical models into production environments. Thus, the **data engineer** has emerged as a highly desirable and indispensable member of an analytics project team. This instructor-led workshop covers the content and hands-on lab exercises provided in the following courses:

- Data Warehousing with SQL and NoSQL
- ETL Offload with Hadoop and Spark
- Data Governance, Security and Privacy for Big Data
- Processing Streaming and IoT Data
- Building Data Pipelines with Python

This training prepares the learner for a major portion of the Dell Technologies Proven Professional data engineering specialist-level certification exam (DES-7DE1). Review the exam description document to understand all the related data engineering training and consumption options.

Audience

This course is intended for data engineers, data scientists, data architects, data analysts or anyone else who wants to learn and apply data engineering principles and tools. Possible workshop participants include:

- Current business and data analysts looking to add data engineering to their skillset
- Database professionals looking to expand their Big Data skills
- Managers of teams of business intelligence, analytics, and big data professionals

Prerequisite Knowledge/Skills

To complete this course successfully and gain the maximum benefits from it, a student should have the following knowledge and skill sets:

- Experience with a programming language such as Java, R, or Python
- Familiarity with the non-statistical aspects of the Data Science and Big Data Analytics v2 content

- Understanding of the data engineer role provided in the Introduction to Data Engineering (course id #: ES731OCMIDENG)

Course Objectives

Upon successful completion of this course, participants should be able to:

Data Warehousing with SQL and NoSQL

- Provide an overview of data warehouses
- Explain the purposes of databases and their various types
- Describe various SQL and NoSQL tools

ETL Offload with Hadoop and Spark

- Identify business challenges with ETL (Extract-Transform-Load)
- Explain ELT and ETL processes
- Describe the Hadoop ecosystem as an ETL offload solution

Data Governance, Security and Privacy for Big Data

- Describe data governance, roles, and responsibilities
- Discuss data governance models
- Describe metadata, metadata types and uses
- Explain master data, framework, and purpose
- Explain Hadoop security controls
- Discuss data governance tools Apache Atlas, Ranger and Knox
- Describe cloud security consideration
- Explain GDPR and data ethics

Processing Streaming and IoT Data

- Describe streaming and IoT data environments
- Explain Kafka messaging system with examples
- Explain the key features, architecture and various use cases of stream processing tools such as Storm, Spark Streaming, and Flink
- Explain various IoT related projects such as Project Nautilus, Pravega, and EdgeX Foundry

Building Data Pipelines with Python

- Write Python scripts to perform key data processing activities
- Describe data pipelines and tools
- Build data pipelines using Python

Course Topics

The content of this course is designed to support the course objectives.

Data Warehousing with SQL and NoSQL

- Data warehouses
- Relational databases
    - SQL operations
    - Transactional vs. analytical
    - Design and performance considerations
- NoSQL
    - SQL vs. NoSQL
    - NoSQL database types and examples
- Redis
- Apache Cassandra
- Apache CouchDB
- Data Lakes

ETL Offload with Hadoop and Spark

- Hadoop ecosystem
- Hadoop Distributed File System (HDFS)
- Data ingestion tools
    - Apache Flume
    - Apache Sqoop
- Apache Spark
- ETL schedulers
    - Apache Oozie
    - Apache Airflow
- ETL offload implementation considerations

Data Governance, Security and Privacy for Big Data

- Data Governance Overview
- Data Governance Roles
- Data Governance Models
- Metadata
- Master Data Management
- Security Controls in Hadoop Ecosystem
- Apache Atlas
- Apache Ranger
- Apache Knox
- Security Considerations in the Cloud
- General Data Protection Regulation (GDPR)
- Data Ethics: Avoiding Hidden Biases

Processing Streaming and IoT Data

- Processing Streaming and IoT Data Overview
- Streaming and IoT Data Processing Tools Framework
- Apache Storm
- Apache Kafka
- Apache Spark Streaming
- Apache Flink
- Pravega
- Project Nautilus
- EdgeX Foundry

Building Data Pipelines with Python

- Introduction to data pipelines
- Introduction to Python
    - Features of Python
    - Basic Syntax of Python
    - Data Types, Operators, and Conditional Statements
    - User-defined Functions and Classes
- Python libraries
- Data structures in Python
- Data pipeline best practices