Module Title : 20775A: Performing Data Engineering on Microsoft HD Insight
Duration : 5 days

About this course

The main purpose of the course is to give students the ability plan and implement big data workflows on HDInsight.

Audience profile

The primary audience for this course is data engineers, data architects, data scientists, and data developers who plan to implement big data engineering workflows on HDInsight.

At course completion

After completing this course, students will be able to:

- Deploy HDInsight Clusters.
- Authorizing Users to Access Resources.
- Loading Data into HDInsight.
- Troubleshooting HDInsight.
- Implement Batch Solutions.
- Design Batch ETL Solutions for Big Data with Spark
- Analyze Data with Spark SQL.
- Analyze Data with Hive and Phoenix.
- Describe Stream Analytics.
- Implement Spark Streaming Using the DStream API.
- Develop Big Data Real-Time Processing Solutions with Apache Storm.
- Build Solutions that use Kafka and HBase.

Course Outline

Module 1: Getting Started with HDInsight

This module introduces Hadoop, the MapReduce paradigm, and HDInsight.

Lessons

- What is Big Data?
- Introduction to Hadoop
- Working with MapReduce Function
- Introducing HDInsight

Lab : Working with HDInsight

- Provision an HDInsight cluster and run MapReduce jobs

After completing this module, students will be able to:

- Describe Hadoop, MapReduce and HDInsight.
- Use scripts to provision an HDInsight Cluster.
- Run a word-counting MapReduce program using PowerShell.

Module 2: Deploying HDInsight Clusters

This module provides an overview of the Microsoft Azure HDInsight cluster types, in addition to the creation and maintenance of the HDInsight clusters. The module also demonstrates how to customize clusters by using script actions through the Azure Portal, Azure PowerShell, and the Azure command-line interface (CLI). This module includes labs that provide the steps to deploy and manage the clusters.

Lessons

- Identifying HDInsight cluster types
- Managing HDInsight clusters by using the Azure portal
- Managing HDInsight Clusters by using Azure PowerShell

Lab : Managing HDInsight clusters with the Azure Portal

- Create an HDInsight cluster that uses Data Lake Store storage
- Customize HDInsight by using script actions
- Delete an HDInsight cluster

After completing this module, students will be able to:

- Identify HDInsight cluster types
- Manage HDInsight clusters by using the Azure Portal.
- Manage HDInsight clusters by using Azure PowerShell.

Module 3: Authorizing Users to Access Resources

This module provides an overview of non-domain and domain-joined Microsoft HDInsight clusters, in addition to the creation and configuration of domain-joined HDInsight clusters. The module also demonstrates how to manage domain-joined clusters using the Ambari management UI and the Ranger Admin UI. This module includes the labs that will provide the steps to create and manage domain-joined clusters.

Lessons

- Non-domain Joined clusters
- Configuring domain-joined HDInsight clusters
- Manage domain-joined HDInsight clusters

Lab : Authorizing Users to Access Resources

- Prepare the Lab Environment
- Manage a non-domain joined cluster

After completing this module, students will be able to:

- Identify the characteristics of non-domain and domain-joined HDInsight clusters.
- Create and configure domain-joined HDInsight clusters through the Azure PowerShell.
- Manage the domain-joined cluster using the Ambari management UI and the Ranger Admin UI.
- Create Hive policies and manage user permissions.

Module 4: Loading data into HDInsight

This module provides an introduction to loading data into Microsoft Azure Blob storage and Microsoft Azure Data Lake storage. At the end of this lesson, you will know how to use multiple tools to transfer data to an HDInsight cluster. You will also learn how to load and transform data to decrease your query run time.

Lessons

- Storing data for HDInsight processing
- Using data loading tools
- Maximising value from stored data

Lab : Loading Data into your Azure account

- Load data for use with HDInsight

After completing this module, students will be able to:

- Discuss the architecture of key HDInsight storage solutions.
- Use tools to upload data to HDInsight clusters.
- Compress and serialize uploaded data for decreased processing time.

Module 5: Troubleshooting HDInsight

In this module, you will learn how to interpret logs associated with the various services of Microsoft Azure HDInsight cluster to troubleshoot any issues you might have with these services. You will also learn about Operations Management Suite (OMS) and its capabilities.

Lessons

- Analyze HDInsight logs
- YARN logs
- Heap dumps
- Operations management suite

Lab : Troubleshooting HDInsight

- Analyze HDInsight logs
- Analyze YARN logs
- Monitor resources with Operations Management Suite

After completing this module, students will be able to:

- Locate and analyze HDInsight logs.
- Use YARN logs for application troubleshooting.
- Understand and enable heap dumps.
- Describe how the OMS can be used with Azure resources.

Module 6: Implementing Batch Solutions

In this module, you will look at implementing batch solutions in Microsoft Azure HDInsight by using Hive and Pig. You will also discuss the approaches for data pipeline operationalization that are available for big data workloads on an HDInsight stack.

Lessons

- Apache Hive storage
- HDInsight data queries using Hive and Pig
- Operationalize HDInsight

Lab : Implement Batch Solutions

- Deploy HDInsight cluster and data storage
- Use data transfers with HDInsight clusters
- Query HDInsight cluster data

After completing this module, students will be able to:

- Understand Apache Hive and the scenarios where it can be used.
- Run batch jobs using Apache Hive and Apache Pig.
- Explain the capabilities of the Microsoft Azure Data Factory and Apache Oozie—and how they can orchestrate and automate big data workflows.

Module 7: Design Batch ETL solutions for big data with Spark

This module provides an overview of Apache Spark, describing its main characteristics and key features. Before you start, it's helpful to understand the basic architecture of Apache Spark and the different components that are available. The module also explains how to design batch Extract, Transform, Load (ETL) solutions for big data with Spark on HDInsight. The final lesson includes some guidelines to improve Spark performance.

Lessons

- What is Spark?

- ETL with Spark
- Spark performance

Lab : Design Batch ETL solutions for big data with Spark.

- Create a HDInsight Cluster with access to Data Lake Store
- Use HDInsight Spark cluster to analyze data in Data Lake Store
- Analyzing website logs using a custom library with Apache Spark cluster on HDInsight
- Managing resources for Apache Spark cluster on Azure HDInsight

After completing this module, students will be able to:

- Describe the architecture of Spark on HDInsight.
- Describe the different components required for a Spark application on HDInsight.
- Identify the benefits of using Spark for ETL processes.
- Create Python and Scala code in a Spark program to ingest or process data.
- Identify cluster settings for optimal performance.
- Track and debug jobs running on an Apache Spark cluster in HDInsight.

Module 8: Analyze Data with Spark SQL

This module describes how to analyze data by using Spark SQL. In it, you will be able to explain the differences between RDD, Datasets and Dataframes, identify the uses cases between Iterative and Interactive queries, and describe best practices for Caching, Partitioning and Persistence. You will also look at how to use Apache Zeppelin and Jupyter notebooks, carry out exploratory data analysis, then submit Spark jobs remotely to a Spark cluster.

Lessons

- Implementing iterative and interactive queries
- Perform exploratory data analysis

Lab : Performing exploratory data analysis by using iterative and interactive queries

- Build a machine learning application
- Use zeppelin for interactive data analysis
- View and manage Spark sessions by using Livy

After completing this module, students will be able to:

- Implement interactive queries.
- Perform exploratory data analysis.

Module 9: Analyze Data with Hive and Phoenix

In this module, you will learn about running interactive queries using Interactive Hive (also known as Hive LLAP or Live Long and Process) and Apache Phoenix. You will also learn about the various aspects of running interactive queries using Apache Phoenix with HBase as the underlying query engine.

Lessons

- Implement interactive queries for big data with interactive hive.
- Perform exploratory data analysis by using Hive
- Perform interactive processing by using Apache Phoenix

Lab : Analyze data with Hive and Phoenix

- Implement interactive queries for big data with interactive Hive
- Perform exploratory data analysis by using Hive
- Perform interactive processing by using Apache Phoenix

After completing this module, students will be able to:

- Implement interactive queries with interactive Hive.
- Perform exploratory data analysis using Hive.
- Perform interactive processing by using Apache Phoenix.

Module 10: Stream Analytics

The Microsoft Azure Stream Analytics service has some built-in features and capabilities that make it as easy to use as a flexible stream processing service in the cloud. You will see that there are a number of advantages to using Stream Analytics for your streaming solutions, which you will discuss in more detail. You will also compare features of Stream Analytics to other services available within the Microsoft Azure HDInsight stack, such as Apache Storm. You will learn how to deploy a Stream Analytics job, connect it to the Microsoft Azure Event Hub to ingest real-time data, and execute a Stream Analytics query to gain low-latency insights. After that, you will learn how Stream Analytics jobs can be monitored when deployed and used in production settings.

Lessons

- Stream analytics
- Process streaming data from stream analytics
- Managing stream analytics jobs

Lab : Implement Stream Analytics

- Process streaming data with stream analytics
- Managing stream analytics jobs

After completing this module, students will be able to:

- Describe stream analytics and its capabilities.
- Process streaming data with stream analytics.

- Manage stream analytics jobs.

Module 11: Implementing Streaming Solutions with Kafka and HBase

In this module, you will learn how to use Kafka to build streaming solutions. You will also see how to use Kafka to persist data to HDFS by using Apache HBase, and then query this data.

Lessons

- Building and Deploying a Kafka Cluster
- Publishing, Consuming, and Processing data using the Kafka Cluster
- Using HBase to store and Query Data

Lab : Implementing Streaming Solutions with Kafka and HBase

- Create a virtual network and gateway
- Create a storm cluster for Kafka
- Create a Kafka producer
- Create a streaming processor client topology
- Create a Power BI dashboard and streaming dataset
- Create an HBase cluster
- Create a streaming processor to write to HBase

After completing this module, students will be able to:

- Build and deploy a Kafka Cluster.
- Publish data to a Kafka Cluster, consume data from a Kafka Cluster, and perform stream processing using the Kafka Cluster.
- Save streamed data to HBase, and perform queries using the HBase API.

Module 12: Develop big data real-time processing solutions with Apache Storm

This module explains how to develop big data real-time processing solutions with Apache Storm.

Lessons

- Persist long term data
- Stream data with Storm
- Create Storm topologies
- Configure Apache Storm

Lab : Developing big data real-time processing solutions with Apache Storm

- Stream data with Storm
- Create Storm Topologies

After completing this module, students will be able to:

- Persist long term data.
- Stream data with Storm.
- Create Storm topologies.
- Configure Apache Storm.

Module 13: Create Spark Streaming Applications

This module describes Spark Streaming; explains how to use discretized streams (DStreams); and explains how to apply the concepts to develop Spark Streaming applications.

Lessons

- Working with Spark Streaming
- Creating Spark Structured Streaming Applications
- Persistence and Visualization

Lab : Building a Spark Streaming Application

- Installing Required Software
- Building the Azure Infrastructure
- Building a Spark Streaming Pipeline

After completing this module, students will be able to:

- Describe Spark Streaming and how it works.
- Use discretized streams (DStreams).
- Work with sliding window operations.
- Apply the concepts to develop Spark Streaming applications.
- Describe Structured Streaming.

Prerequisites

In addition to their professional experience, students who attend this course should have:

- Programming experience using R, and familiarity with common R packages
- Knowledge of common statistical methods and data analysis best practices.
- Basic knowledge of the Microsoft Windows operating system and its core functionality.
- Working knowledge of relational databases.