Performing Data Engineering on Microsoft HD Insight - MOC On Demand (MS-20775)

Modality: On Demand

Duration: 2 Days

SATV Value: 2

About this course:

This course is an on-Demand course of Microsoft Official that is available for ninety days from the course demand date if you have a yearly subscription, or buy the course separately. The access to this course will terminate after 90 days of course registration.

MS Azure HDInsight gives a pay-as-you-go way out for the batch processing of Hadoop-based big information that financially benefit as you don't have to focus on configuring and installing on-premises infrastructure. The fundamental motivation behind the course is to give learners the implement big data workflows and ability plan on HDInsight. This course likewise helps the learners in the groundwork for Microsoft 70-775: Perform Data Engineering on the certification exam of MS Azure HDInsight.

The normal compensation for Data Engineer is $84,143 annually.

Course Objective:

At the time of course completion, students will be competent to:

- HDInsight Clusters Deployment.
- Data Analysis with Spark SQL.
- Uploading information into HDInsight.
- Users Authorization to Access Resources.
- Define Stream Analytics.
- Batch Solutions Implementation.
- HDInsight Troubleshooting.
- Plan Solutions for Batch ETL for Big Data with Spark
- Solutions for Big Data Real-Time Processing Development with Apache Storm.
- Data Analysis with Phoenix and Hive.
- Create Solutions that utilize HBase and Kafka.
- Spark Streaming Implementation with the help of the DStream API.

Audience:

This course is designed for:

Data engineers, data developers, data architects, and data scientists who design to perform workflows for big data engineering on HDInsight.

Prerequisites:

Along with their professional experience, learners who appear this course must have:

- Experience of Programming using R, and awareness of common packages of R.
- A Simple understanding of the operating system of MS Windows and its core operations.
- Information about common methods of statistics and best practices of data analysis.
- Relational databases working knowledge.

Proposed prerequisites courses:

Databases in Azure

Course Outline:

Module 1: Getting Started with HDInsight

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

Lessons

- Big Data
- Hadoop
- MapReduce
- HDInsight

Lab : Querying Big Data

- Query data with Hive
- Visualize data with Excel

After completing this module, students will be able to:

- Describe Big data.
- Describe Hadoop.
- Describe MapReduce.
- Describe HDInsight.

Module 2: Deploying HDInsight Clusters

At the end of this module the student will be able to deploy HDInsight clusters.

Lessons

- HDInsight cluster types
- Managing HDInsight Clusters
- Managing HDInsight Clusters with PowerShell

Lab : Managing HDInsight clusters with the Azure Portal

- Create an HDInsight Hadoop Cluster
- Customise HDInsight using a script action
- Customize HDInsight using Bootstrap
- Delete an HDInsight cluster

After completing this module, students will be able to:

- Describe HDInsight cluster types.
- Describe the creation, management, and deletion of HDInsight clusters with the Azure portal.
- Describe the creation, management, and deletion of HDInsight clusters with PowerShell.

Module 3: Authorizing Users to Access Resources

This module covers permissions and the assignment of permissions.

Lessons

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

Lab : Authorizing Users to Access Resources

- Configure a domain-joined HDInsight cluster
- Configure Hive policies

After completing this module, students will be able to:

- Describe how to authorize user access to objects.
- Describe how to authorize users to execute code.
- Describe how to manage domain-joined HDInsight clusters.

Module 4: Loading data into HDInsight

This module covers loading data into HDInsight.

Lessons

- HDInsight Storage
- Data loading tools
- Performance and reliability

Lab : Loading Data into HDInsight

- Loading data using Sqoop
- Loading data using AZcopy

- Loading data using ADLcopy
- Use HDInsight to compress data

After completing this module, students will be able to:

- Describe HDInsight storage configurations and architectures.
- Describe options for loading data into HDInsight.
- Describe benefits of compression and pre-processing in HDInsight.

Module 5: Troubleshooting HDInsight

This module describes how to troubleshoot HDInsight.

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:

- Analyze HDInsight logs.
- Analyze YARN logs.
- Analyze Heap dumps.
- Use the operations management suite to monitor resources.

Module 6: Implementing Batch Solutions

This module describes how to implement batch solutions.

Lessons

- Apache Hive storage
- Querying with Hive and Pig
- Operationalize HDInsight

Lab : Backing Up SQL Server Databases

- Load data into a hive table
- Query data with Hive and Pig

After completing this module, students will be able to:

- Describe Apache Hive storage.
- Query data using Hive and Pig.
- Operationalize HDInsight.

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

This module describes how to design batch ETL solutions for big data with Spark.

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 Spark and when to use it.
- Describe the use of ETL with Spark.
- Analyze Spark performance.

Module 8: Analyze Data with Spark SQL

This module describes how to analyze data with Spark SQL.

Lessons

- Implement interactive queries
- Perform exploratory data analysis

Lab : Analyze data with Spark SQL

- Implement interactive queries
- Perform exploratory data analysis

After completing this module, students will be able to:

- Implement interactive queries.
- Perform exploratory data analysis.

Module 9: Analyze Data with Hive and Phoenix

This module describes how to analyze data with Hive and Phoenix.

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

This module introduces Azure Stream Analytics.

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 it?s capabilities.
- Process streaming data with stream analytics.
- Manage stream analytics jobs.

Module 11: Spark Streaming using the DStream API

This module introduces the Dstream API and describes how to create Spark structured streaming applications.

Lessons

- Dstream
- Create Spark structured streaming applications
- Persistence and visualization

Lab : Spark streaming applications using DStream API

- Creating Spark streaming applications using the DStream API
- Creating Spark structured streaming applications

After completing this module, students will be able to:

- Explain DStream.
- Create Spark structured streaming applications.
- Describe persistence and visualization.

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: Analyze Data with Spark SQL**

This module describes how to analyze data with Spark SQL.

Lessons

- Implement interactive queries
- Perform exploratory data analysis

Lab : Analyze data with Spark SQL

- Implement interactive queries
- Perform exploratory data analysis

After completing this module, students will be able to:

- Implement interactive queries.
- Perform exploratory data analysis.