

20775 Performing Data Engineering on Microsoft HD Insight

Overview

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

Target Audience

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.

Course Objectives

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

Getting Started with HDInsight

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

Deploying HDInsight Clusters

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

[Register Online](#)

Schedule

Class Length: 5 Days

G2R = "Guaranteed to Run" | OLL = "Online LIVE"
ILT = "Instructor-Led-Training"

This course is not currently available on the public schedule. Please contact us using the information in the footer below to inquire about future dates or to schedule a private class.

Authorizing Users to Access Resources

Non-domain Joined clusters
Configuring domain-joined HDInsight clusters
Manage domain-joined HDInsight clusters
Lab : Authorizing Users to Access Resources

Loading data into HDInsight

Storing data for HDInsight processing
Using data loading tools
Maximising value from stored data
Lab : Loading Data into your Azure account

Troubleshooting HDInsight

Analyze HDInsight logs
YARN logs
Heap dumps
Operations management suite
Lab : Troubleshooting HDInsight

Implementing Batch Solutions

Apache Hive storage
HDInsight data queries using Hive and Pig
Operationalize HDInsight
Lab : Implement Batch Solutions

Design Batch ETL solutions for big data with Spark

What is Spark?
ETL with Spark
Spark performance
Lab : Design Batch ETL solutions for big data with Spark.

Analyze Data with Spark SQL

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

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
Lab : Analyze data with Hive and Phoenix

Stream Analytics

Stream analytics

Process streaming data from stream analytics

Managing stream analytics jobs

Lab : Implement Stream Analytics

Implementing Streaming Solutions with Kafka and HBase

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

Develop big data real-time processing solutions with Apache Storm

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

Create Spark Streaming Applications

Working with Spark Streaming

Creating Spark Structured Streaming Applications

Persistence and Visualization

Lab : Building a Spark Streaming Application