Rumos valorizamos profissionais!

Performing Data Engineering on Microsoft HD Insight (20775)

- **Formato do curso:** Presencial
- **Localidade:** Porto
- **Data:** 01 Abr. 2019 a 05 Abr. 2019
- **Preço:** 1740€
- **Horário:** Laboral - das 09h30 às 17h30
- **Duração:** 35 horas

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

Destinatários

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.

Pré-requisitos

- 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

Objectivos

- 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

Programa

Getting Started with HDInsight

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

Deploying HDInsight Clusters

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

Authorizing Users to Access Resources

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

Loading data into HDInsight

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

Troubleshooting HDInsight

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

Implementing Batch Solutions

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

Design Batch ETL solutions for big data with Spark

- What is Spark?
- ETL with Spark
- Spark performance

Analyze Data with Spark SQL

- Implementing iterative and interactive queries
- Perform exploratory data analysis

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

Stream Analytics

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

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

Develop big data real-time processing solutions with Apache Storm

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

Create Spark Streaming Applications

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