IBM - IBM Open Platform with Apache Hadoop

Code:	DW606G
Length:	2 days
URL:	View Online

IBM Open Platform (IOP) with Apache Hadoop is the first premiere collaborative platform to enable Big Data solutions to be developed on the common set of Apache Hadoop technologies. The Open Data Platform initiative (ODP) is a shared industry effort focused on promoting and advancing the state of Apache Hadoop and Big Data technologies for the enterprise. The current ecosystem is challenged and slowed by fragmented and duplicated efforts between different groups. The ODP Core will take the guesswork out of the process and accelerate many use cases by running on a common platform. It allows enterprises to focus on building business driven applications.

This module provides an in-depth introduction to the main components of the ODP core --namely Apache Hadoop (inclusive of HDFS, YARN, and MapReduce) and Apache Ambari -- as well as providing a treatment of the main open-source components that are generally made available with the ODP core in a production Hadoop cluster.

Skills Gained

- List and describe the major components of the open-source Apache Hadoop stack and the approach taken by the Open Data Foundation.
- Manage and monitor Hadoop clusters with Apache Ambari and related components
- Explore the Hadoop Distributed File System (HDFS) by running Hadoop commands.
- Understand the differences between Hadoop 1 (with MapReduce 1) and Hadoop 2 (with YARN and MapReduce 2).
- Create and run basic MapReduce jobs using command line.
- Explain how Spark integrates into the Hadoop ecosystem.
- Execute iterative algorithms using Spark's RDD.
- Explain the role of coordination, management, and governance in the Hadoop ecosystem using Apache Zookeeper, Apache Slider, and Apache Knox.
- Explore common methods for performing data movement
  - Configure Flume for data loading of log files
  - Move data into the HDFS from relational databases using Sqoop
- Understand when to use various data storage formats (flat files, CSV/delimited, Avro/Sequence files, Parquet, etc.).
- Review the differences between the available open-source programming languages typically used with Hadoop (Pig, Hive)

and for Data Science (Python, R)

- Query data from Hive.
- Perform random access on data stored in HBase.
- Explore advanced concepts, including Oozie and Solr

Who Can Benefit

This intermediate training course is for those who want a foundation of IBM BigInsights. This includes: Big data engineers, data scientist, developers or programmers, administrators who are interested in learning about IBM's Open Platform with Apache Hadoop.

Prerequisites

None, however, knowledge of Linux would be beneficial.

Course Details

Unit 1: IBM Open Platform with Apache Hadoop

- Exercise 1: Exploring the HDFS

Unit 2: Apache Ambari

- Exercise 2: Managing Hadoop clusters with Apache Ambari

Unit 3: Hadoop Distributed File System

- Exercise 3: File access and basic commands with HDFS

Unit 4: MapReduce and Yarn

- Topic 1: Introduction to MapReduce based on MR1
- Topic 2: Limitations of MR1
- Topic 3: YARN and MR2
- Exercise 4: Creating and coding a simple MapReduce job
- Possibly a more complex second Exercise

Unit 5: Apache Spark

- Exercise 5: Working with Spark's RDD to a Spark job

Unit 6: Coordination, management, and governance

- Exercise 6: Apache ZooKeeper, Apache Slider, Apache Knox

Unit 7: Data Movement

- Exercise 7: Moving data into Hadoop with Flume and Sqoop

Unit 8: Storing and Accessing Data

- Topic 1: Representing Data: CSV, XML, JSON, and YAML
- Topic 2: Open Source Programming Languages: Pig, Hive, and Other [R, Python, etc]
- Topic 3: NoSQL Concepts
- Topic 4: Accessing Hadoop data using Hive
- Exercise 8: Performing CRUD operations using the HBase shell
- Topic 5: Querying Hadoop data using Hive
- Exercise 9: Using Hive to Access Hadoop / HBase Data

Unit 9: Advanced Topics

- Topic 1: Controlling job workflows with Oozie
- Topic 2: Search using Apache Solr
- No lab exercises

ExitCertified® Corporation and iMVP® are registered trademarks of ExitCertified ULC and ExitCertified Corporation and Tech Data Corporation, respectively
Copyright ©2019 Tech Data Corporation and ExitCertified ULC & ExitCertified Corporation.
All Rights Reserved.