Big Data Engineer

(Classroom)

Career path description

The Big Data Engineer career path prepares students to use the Big Data platform and methodologies in order to collect and analyze large amounts of data from different sources. This will require skills in Big Data architecture, such as Apache Hadoop, Ambari, Spark, Big SQL, HDFS, YARN, MapReduce, ZooKeeper, Knox, Sqoop, and HBase.

ibm.com/training

General information

Delivery method

95% instructor led and 5% web-based

Version

2018

Product

HDP Open Source and IBM Watson Studio

Audience

Undergraduate senior students from IT related academic programs i.e. computer science, software engineering, information systems and similar others

Learning objectives

After completing this course, you should be able to understand the following topics:

- Big Data and Data Analytics
- Hortonworks Data Platform (HDP)
- Apache Ambari
- Hadoop and the Hadoop Distributed File System
- MapReduce and Yarn
- Apache Spark
- Storing and Quering data
- ZooKeeper, Slider, and Knox
- Loading data with Sqooq
- Dataplane Service
- Stream Computing
- Data Science essentials
- Drew Conway's Venn Diagram - and that of others
- The Scientific Process applied to Data Science
- The steps in running a Data Science project
- Languages used for Data Science (Python, R, Scala, Julia, ...)
- Survey of Data Science Notebooks
- Markdown language with notebooks
- Resources for Data Science, including GitHub
- Jupyter Notebook
- Essential packages: NumPy, SciPy, Pandas, Scikit-learn, NLTK, BeautifulSoup...
- Data visualizations: matplotlib, ..., PixieDust
- Using Jupyter "Magic" commands
- Using Big SQL to access HDFS data
- Creating Big SQL schemas and tables
- Querying Big SQL tables
- Managing the Big SQL Server
- Configuring Big SQL security
- Data federation with Big SQL
- IBM Watson Studio
- Analyzing data with Watson Studio

Prerequisites Skills

- Basic knowledge of Linux
- Basic SQL knowledge
- Working knowledge with big data and Hadoop technologies
- Have a basic understanding of notebook technologies for data science
- Students can attend free courses at www.bigdatauniversity.com to acquire the necessary requirements
- Exposure to the IBM Skills Academy Portal learning environment
- Exposure to the IBM Skills Academy Cloud hands-on labs platform

Duration

34.8 hours

Skill level

Basic – Intermediate

Hardware requirements

Classroom (ILT) setup requirements	
Processor	3 GHz or higher
GB RAM	8 GB
GB free disk space	80 GB
Network requirements	Yes
Other requirements	IBM ID

Notes

The following unit and exercise durations are estimates, and might not reflect every class experience. If the course is customized or abbreviated, the duration of unchanged units will probably increase.

Course Agenda

MODULE I – BIG DATA OVERVIEW

Course I – Introduction to the Big Data Ecosystem

Duration: 1.6 hours

Course introduction Duration: 5 minutes

Unit 1. Introduction to Big Data Duration: 90 minutes	
Overview	In this unit you will learn about Big Data and understand why it's important.
Learning objectives	After completing this unit, you should be able to: • Understand what Big Data is • Develop an understanding of the complete open-source Hadoop ecosystem and its near-term future directions • Understand the major challenges of data • Understand how the growth of interconnected devices helps big data • List some real life examples of Big Data • Learn the types of Big Data • Student some Big Data use cases

MODULE II – Prerequisites

This course does not have any prerequisites

MODULE III – Big Data Engineer

Course I – Introduction to the Big Data Ecosystem

Duration: 20.2 hours

Course introduction Duration: 5 minutes

Unit 1. Introduction to Big Data
Duration: 30 minutes

Overview

In this unit you will learn about Big Data and understand why it’s important.

Learning objectives

After completing this unit, you should be able to:

- Develop an understanding of the complete open-source Hadoop ecosystem and its near-term future directions
- Be able to compare and evaluate the major Hadoop distributions and their ecosystem components, both their strengths and their limitations
- Gain hands-on experience with key components of various big data ecosystem components and their roles in building a complete big data
- solution to common business problems
- Learning the tools that will enable you to continue your big data education after the course

Unit 2. Introduction to Hortonworks Data Platform (HDP)
Duration: 30 minutes

Overview

In this unit you will learn about the Hortonworks Data Platform (HDP).

Learning objectives

After completing this unit, you should be able to:

- Describe the functions and features of HDP
- List the IBM value-add components
- Explain what IBM Watson Studio is
- Give a brief description of the purpose of each of the value-add components

Lab 1. Exploration of the lab environment
Duration: 1 hour

Overview

In this lab, you will explore the lab environment. You will access your lab environment and launch Apache Ambari. You will startup a variety of services by using the Ambari GUI. You will also explore some of the directory structure on the Linux system that you will be using.

Learning objectives

After completing this lab, you should be able to:

- Explore the lab environment
- Launch Apache Ambari
- Start a variety of services using Apache GUI
- Explore some of the directory structure on the Linux system

Unit 3. Apache Ambari
Duration: 30 minutes

Overview	In this section you will learn about Ambari, which is one of the operations tools that come with HDP.
Learning objectives	After completing this unit, you should be able to: • Understand the purpose of Apache Ambari in the HDP stack • Understand the overall architecture of Ambari, and Ambari’s relation to other services and components of a Hadoop cluster • List the functions of the main components of Ambari • Explain how to start and stop services from Ambari Web Console

Lab 1. Managing Hadoop clusters with Apache Ambari
Duration: 1 hour

Overview	In this lab you will explore the Apache Ambari web console and perform basic starting and stopping of services, giving you experience in using Apache Ambari to manage your Hadoop cluster.
Learning objectives	After completing this lab, you should be able to: • Managie Hadoop clusters with Apache Ambari o Start the Apache Ambari web console and perform basic start/stop services o Explore other aspects of the Ambari web server

Unit 4. Hadoop and HDFS
Duration: 1 hour

Overview	This unit will explain the underlying technologies that are important to solving the bigdata challenge and tell about BigInsights.
Learning objectives	After completing this unit, you should be able to: • Understand the basic need for a big data strategy in terms of parallel reading of large data files and internode network speed in a cluster • Describe the nature of the Hadoop Distributed File System (HDFS) • Explain the function of the NameNode and DataNodes in an Hadoop cluster • Explain how files are stored and blocks ("splits") are replicated

Lab 1. File access and basic commands with HDFS

Duration: 1 hour

Overview	This lab is intended to provide you with experience in using the Hadoop Distributed File System (HDFS). The basic HDFS file system commands learned here will be used throughout the remainder of the course. You will also be moving some data into HDFS that will be used in later units of this course. The files that you will need are stored in the Linux directory /home/labfiles.
Learning objectives	After completing this lab, you should be able to: • File access and basic commands with HDFS

Unit 5. MapReduce and YARN

Duration: 2 hours and 20 minutes

Overview	In this unit you will learn about MapReduce and YARN.
Learning objectives	After completing this unit, you should be able to: • Describe the MapReduce model v1 • List the limitations of Hadoop 1 and MapReduce 1 • Review the Java code required to handle the Mapper class, the • Reducer class, and the program driver needed to access MapReduce • Describe the YARN model • Compare Hadoop 2/YARN with Hadoop 1

Lab 1. Running MapReduce and YARN jobs

Duration: 1 hour

Overview	In this lab, you will run Java programs using Hadoop v2, YARN, and related technologies.
Learning objectives	After completing this lab, you should be able to: • Run MapResuce and YARN jobs

Lab 2. Creating and coding a simple MapReduce job

Duration: 1 hour

Overview	In this lab, you will compile and run a more complete version of WordCount that has been written specifically for MapReduce2.
Learning objectives	After completing this lab, you should be able to: • Create and code a simple MapReduce job

Unit 6. Apache Spark
Duration: 2 hours

Overview	In this unit you will learn about Apache Spark.
Learning objectives	After completing this unit, you should be able to: • Understand the nature and purpose of Apache Spark in the Hadoop ecosystem • List and describe the architecture and components of the Spark unified stack • Describe the role of a Resilient Distributed Dataset (RDD) • Understand the principles of Spark programming • List and describe the Spark libraries • Launch and use Spark's Scala and Python shells

Lab 1. Working with a Spark RDD with Scala
Duration: 1 hour

Overview	In this lab, you will learn to use some of the fundamental aspects of running Spark in the HDP environment.
Learning objectives	After completing this lab, you should be able to: • Work with Spark RDD with Scala

Unit 7. Storing and quering data
Duration: 2 hours

Overview	In this unit you will learn about storing and quering data.
Learning objectives	After completing this unit, you should be able to: • List the characteristics of representative data file formats, including flat/text files, CSV, XML, JSON, and YAML • List the characteristics of the four types of NoSQL datastores • Describe the storage used by HBase in some detail • Describe and compare the open source programming languages, Pig and Hive • List the characteristics of programming languages typically used by • Data Scientists: R and Python

Lab 1. Using Hive to access Hadoop/HBase data
Duration: 30 minutes

Overview	In this lab, you will use Hive to access Hadoop/HBase data.
Learning objectives	After completing this lab, you should be able to: • Use Hive to access Hadoop/HBase data

Unit 8. ZooKeeper, Slider, and Knox
Duration: 1 hour

Overview	In this unit you will learn about ZooKeeper, Slider and Knox.
Learning objectives	After completing this unit, you should be able to: • Understand the challenges posed by distributed applications and how ZooKeeper is designed to handle them • Explain the role of ZooKeeper within the Apache Hadoop infrastructure and the realm of Big Data management • Explore generic use cases and some real-world scenarios for ZooKeeper • Define the ZooKeeper services that are used to manage distributed systems • Explore and use the ZooKeeper CLI to interact with ZooKeeper services • Understand how Apache Slider works in conjunction with YARN to deploy distributed applications and to monitor them • Explain how Apache Knox provides peripheral security services to an Hadoop cluster

Lab 1. Explore Zookeeper
Duration: 30 minutes

Overview	In this lab, you will connect to ZooKeeper and explore the ZooKeeper files.
Learning objectives	After completing this exercise, you should be able to: • Connect to ZooKeeper and explore the ZooKeeper files

Unit 9. Loading data with Sqoop
Duration: 30 minutes

Overview	In this unit you will learn how to load data with Sqoop.
Learning objectives	After completing this unit, you should be able to: • List some of the load scenarios that are applicable to Hadoop • Understand how to load data at rest • Understand how to load data in motion • Understand how to load data from common sources such as a data warehouse, relational database, web server, or database logs • Explain what Sqoop is and how it works • Describe how Sqoop can be used to import data from relational systems into Hadoop and export data from Hadoop into relational systems • Brief introduction to what Flume is and how it works

Lab 1. Moving data into HDFS with Sqoop
Duration: 30 minutes

Overview	In this lab, you will learn how to move data into an HDFS cluster from a relational database.
Learning objectives	After completing this lab, you should be able to: • Move data into HDFS with Sqoop

Unit 10. Security and Governance
Duration: 1 hour and 15 minutes

Overview	In this unit you will learn about the need of data governance and the role of data security in it.
Learning objectives	After completing this unit, you should be able to: • Explain the need for data governance and the role of data security in this governance • List the Five Pillars of security and how they are implemented with HDP • Discuss the history of security with Hadoop • Identify the need for and the methods used to secure Personal & Sensitive Information • Describe the function of the Hortonworks DataPlane Service (DPS)

Unit 11. Stream Computing
Duration: 1 hour

Overview	In this unit you will learn about stream computing.
Learning objectives	After completing this unit, you should be able to: • Define streaming data • Describe IBM as a pioneer in streaming data - with System S ℗IBM Streams • Explain streaming data - concepts & terminology • Compare and contrast batch data vs streaming data • List and explain streaming components & Streaming Data Engines (SDEs)

Course II – Introduction to Data Science

Duration: 1.6 hours

Course introduction
Duration: 5 minutes

Unit 1. Data Science and Data Science Notebooks
Duration: 45 minutes

Overview	In this unit, you will learn about data science and data science notebooks.
Learning objectives	After completing this unit, you should be able to: • Have a better understanding of methodology “scientific approach” methods used & skills practiced by Data Scientists • Recognize the iterative nature of a data science project • Outline the benefits of using Data Science Notebooks • Describe the mechanisms and tools used with Data Science Notebooks • Compare and contrast the major Notebooks used by Data Scientists

Unit 2. Data Science with Open Source Tools
Duration: 30 minutes

Overview	In this unit, we will concentrate on the Jupyter Notebook and Python
Learning objectives	After completing this unit, you should be able to: • Getting started with Jupyter Notebook • Data and notebooks in Jupyter • How notebooks help data scientists • Essential packages: NumPy, SciPy, Pandas, Scikit-learn, NLTK, BeautifulSoup, ... • Data visualizations: matplotlib, ..., PixieDust • Using Jupyter “Magic” commands

Lab 1. Introduction to Jupyter Notebooks
Duration: 15 minutes

Overview

In this lab you will be introduced to Jupyter Notebooks.

Learning objectives

After completing this exercise, you should be able to:

- Start Jupyter - it will open in a web browser
- Import the lab file (all Jupyter files have a.ipynb suffix) into your default workspace
  - This is now a copy of the provided lab file and you can do anything with it
  - If you mess it up, you can re-import again later
- Explore the component panels - some are markdown, some are code, some are results of running the code (output data, visualizations, ...)
- Learn how to run single panels - and then the whole script
  - You may need to adjust the provided script to locate the data files thataccompany the Jupyter.ipynb file
  - Add some additional panels, as described in the lab script

Course III – Big SQL

Duration: 8.83 hours

Course introduction
Duration: 5 minutes

Unit 1. Using Big SQL to access data residing in the HDFS
Duration: 45 minutes

Overview

In this unit, you will learn about Big SQL, and how to use it to access data residing in the HDFS

Learning objectives

After completing this unit, you should be able to:

- Overview of Big SQL
- Understand how Big SQL fits in the Hadoop architecture
- Start and stop Big SQL using Ambari and command line
- Connect to Big SQL using command line
- Connect to Big SQL using IBM Data Server Manager

Lab 1. Connecting to the IBM Big SQL Server
Duration: 30 minutes

Overview

In this lab you will connect to the Big SQL Server using multiple techniques.You will first explore the lab environment. You will then learn how to set up JSqsh and use it to connect to the Big SQL server. You will also explore the Big SQL service using the Data Server Manager (DSM) graphical web interface.

Learning objectives

After completing this exercise, you should be able to:

- Configure images
- Start Hadoop components
- Start up the Big SQL and DSM services
- Connect to Big SQL using JSqsh
- Execute basic Big SQL statements
- Explore Big SQL through Ambari using DSM

Unit 2. Creating Big SQL schemas and tables
Duration: 55 minutes

Overview

In this unit, you will learn how to create Big SQL schemas and tables

Learning objectives

After completing this unit, you should be able to:

- Describe and create Big SQL schemas and tables
- Describe and list the Big SQL data types
- Work with various Big SQL DDLs
- Load data into Big SQL tables using best practices

Lab 1. Creating and managing Big SQL schemas and tables
Duration: 35 minutes

Overview

In this lab you will start off by creating and dropping a simple Big SQL table. You then will create multiple Big SQL tables using a variety of data types and load the tables with data. You will also work with views, external tables, and other methods of creating Big SQL tables.

Learning objectives

After completing this exercise, you should be able to:

- Create and drop simple Big SQL table
- Create sample tables
- Move data into HDFS
- Load data into Big SQL tables
- Create and work with views
- Create external tables

Unit 3. File formats and querying Big SQL tables
Duration: 1 hour

Overview	In this unit, you will learn about file formats and querying Big SQL tables.
Learning objectives	After completing this unit, you should be able to: • Describe Big SQL supported file formats • Query Big SQL tables using various DMLs

Lab 1. Querying Big SQL tables
Duration: 30 minutes

Overview	In this lab you will experiment with some more advanced SQL queries. You will then explore Big SQL's ARRAY type. You will also create a user-defined function (UDF) and write queries that call the UDF. Finally, you will store data in an alternate file format (Parquet).
Learning objectives	After completing this exercise, you should be able to: • Connect to Big SQL • Query data with Big SQL • Work with the ARRAY type • Work with Big SQL functions • Store data in an alternate file format (Parquet)

Unit 4. Managing the Big SQL Server
Duration: 1 hour

Overview	In this unit, you will learn how to manage the Big SQL server
Learning objectives	After completing this unit, you should be able to: • Configure the Big SQL Server • Configure the Big SQL Scheduler • List the registries for compiler and runtime performance improvement • Backup and restore Big SQL

Lab 1. Managing the Big SQL Server
Duration: 30 minutes

Overview	In this lab you will investigate and explore Big SQL configurations for resource allocation as well as for compiler and runtime performance improvements.
Learning objectives	After completing this exercise, you should be able to: • Update the database resource percentage for the Big SQL database instance • Inspect the Big SQL scheduler configuration file • View the registries for the compiler and runtime performance improvement

Unit 5. Configuring Big SQL security
Duration: 1 hour

Overview	In this unit, you will learn about how to configure Big SQL security
Learning objectives	After completing this unit, you should be able to: • Configure authentication for Big SQL • Manage security with Apache Ranger • Enable SSL encryption • Configure authorization of Big SQL objects • Configure impersonation in Big SQL

Lab 1. Configuring Big SQL security
Duration: 30 minutes

Overview	In this lab you will work with Big SQL authorization techniques.
Learning objectives	After completing this lab, you should be able to: • Use column masking and row based access control to restrict access to your data

Lab 2. Configuring impersonation in Big SQL
Duration: 30 minutes

Overview	In this lab you will enable and configure impersonation with Big SQL
Learning objectives	After completing this lab, you should be able to: • Configure impersonation in Big SQL

Unit 6. Data federation with Big SQL
Duration: 45 minutes

Overview	In this unit, you will learn data federation with Big SQL
Learning objectives	After completing this unit, you should be able to: • Understand the concept of Big SQL federation • List the supported data sources • Set up and configure a federation server to use different data sources

Lab 1. Using Fluid Query with Big SQL
Duration: 15 minutes

Overview	In this lab you will configure Fluid Query with Big SQL
Learning objectives	After completing this lab, you should be able to: • Configure Fluid Query with Big SQL

Course IV – IBM Watson Studio

Duration: 2.6 hours

Course introduction
Duration: 5 minutes

Unit 1. Introduction to IBM Watson Studio
Duration: 30 minutes

Overview	In this unit, you will learn about Watson Studio.
Learning objectives	After completing this unit, you should be able to: • What is Watson Studio? • Setting up a project • Working with collaborators • Managing data assets

Lab 1. Getting started with Watson Studio
Duration: 1 hour

Overview	In this lab, you will create and manage a project, add collaborators, and load a data set to the object store.
Learning objectives	After completing this lab, you should be able to: • Sign up for a Watson Studio account • Create a new project • Manage a project • Add collaborators • Load data • Manage the object storage

Unit 2. Analyzing data with Watson Studio
Duration: 30 minutes

Overview	In this unit, you will learn how to analyze data with Watson Studio.
Learning objectives	After completing this unit, you should be able to: • Overview of Jupyter notebooks • Creating notebooks • Coding and running notebooks • Sharing and publishing notebooks

Lab 1. Analyzing data with Watson Studio
Duration: 30 minutes

Overview	In this lab, you will run through a sample notebook in Watson Studio and use PixieDust for data visualization
Learning objectives	After completing this lab, you should be able to: • Create a notebook • Use notebooks • Work with external data

IBM Official Badges and Associated Job Roles	
IBM Official Badges	Big Data Engineer 2018: Explorer | Mastery Award
Associated Job Roles	• Business Intelligence Analyst • Artificial Intelligence Analyst

For more information

To learn more about this career path and others, see ibm.biz/ibmskillsacademy

To learn more about validating your technical skills with IBM Open Badges, see www.youracclaim.com

To stay informed about the IBM Skills Academy, see the following sites:

Facebook: www.facebook.com/ibmskillsacademy