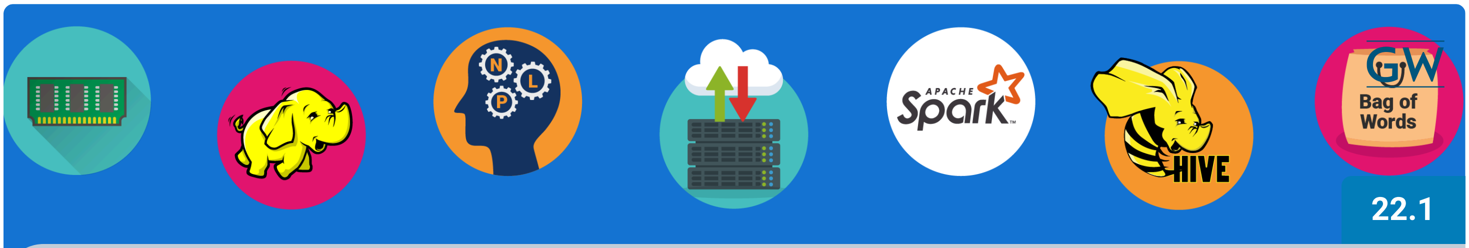

Introduction to Big Data

Overview

For today's lesson, you'll learn how to identify parts of the Apache Hadoop ecosystem. Then, you’ll learn how to write Python scripts that implement the Apache MapReduce programming model. Next, you'll learn the differences between the Apache Hadoop and Apache Spark environments. Finally, you'll learn how to create and filter DataFrames using PySpark.

What You'll Learn

By the end of this lesson, you will be able to:

- Identify the parts of the Hadoop ecosystem.
- Write a Python script that implements the MapReduce programming model.
- Identify the differences between the Hadoop and Spark environments.
- Create a DataFrame by using PySpark.
- Filter and order a DataFrame by using Spark.

22.1 Activity Files

Download the following files to prepare for today’s class:

22.1 Activity Files (https://static.bc-edx.com/data/dl-1-2/m22/lms/activities/Class_1_Activities.zip)

You don’t need to complete any of the activities before class, but feel free to review the material ahead of time to preview what the lesson will cover.