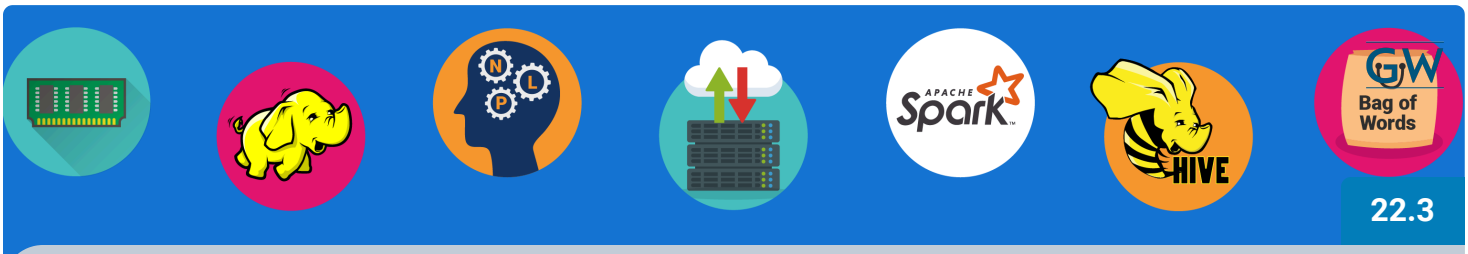

Optimizing Spark: Storage, Partitioning, and Caching

Overview

For today's lesson, you'll learn how to store data in parquet format and partition the parquet data to optimize query execution times. Then you'll practice caching data and determine optimal query execution times between partitioned and cached data.

What You'll Learn

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

- Compare the file storage types (other than tabular) that work the best for Spark.
- Understand how partitioning affects Spark performance.
- Explain the cause of shuffling and limit it when possible.
- Identify when caching is the best option.
- Explain how to broadcast a lookup table, and force it when it doesn't happen automatically.
- Set the shuffle partitions to an appropriate value and demonstrate how to cache data.

22.3 Activity Files

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

22.3 Activity Files (https://static.bc-edx.com/data/dl-1-2/m22/lms/activities/Class_3_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.