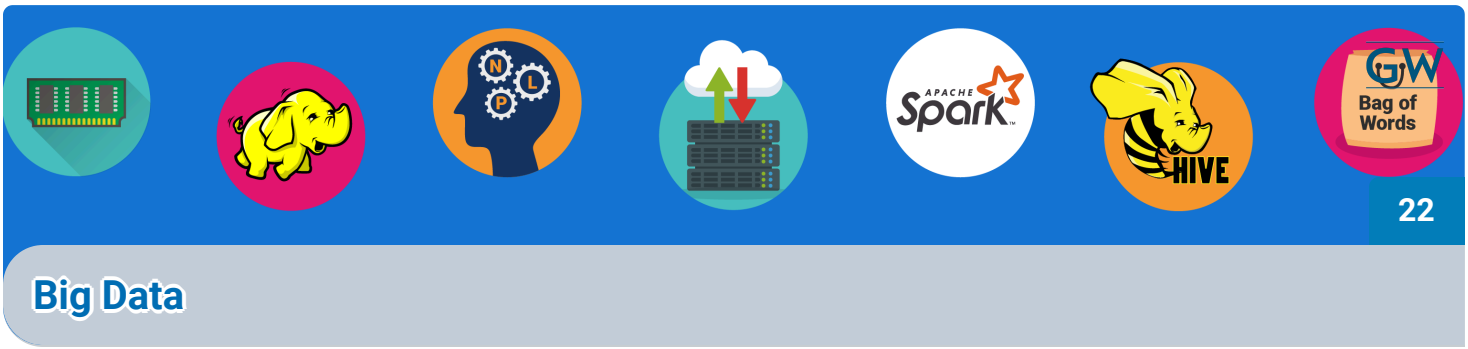

Big Data

This week we'll demystify the term “big data” and give you some hands-on experience working with it. We'll start by reviewing Hadoop and its ecosystem. Within this context, we'll cover MapReduce and how it has improved the process for handling big data. We'll then move on to PySpark, the leading technology for handling big data. After diving into some of the technologies, we'll learn about advanced features of PySpark and how to optimize query execution times with large datasets by storing data in parquet format and partitioning and caching data.

Preview This Week's Challenge

For this module’s Challenge, you will put your extract, transform, and load process (ETL) skills to the test by creating DataFrames to match production-ready tables from two big Amazon customer review datasets. You’ll also have the optional task of analyzing whether reviews from Amazon's Vine program are trustworthy.