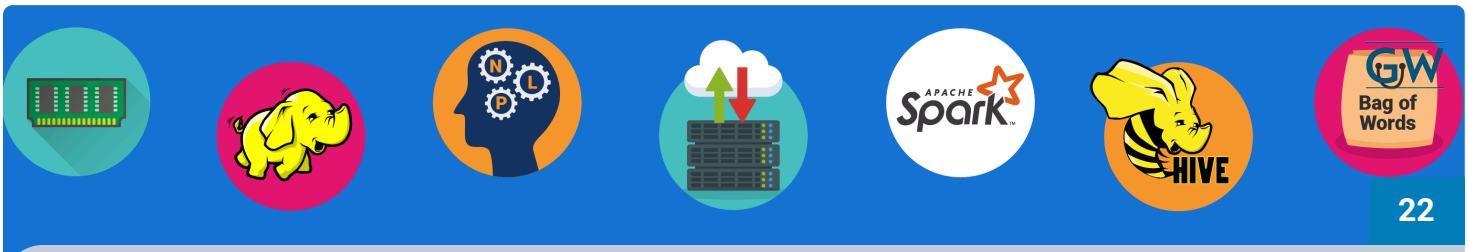

Review

Congratulations on completing Module 22!

After all of your learning and practice in this module, you should 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.
- Apply grouping and aggregation functions to a dataset by using Spark.
- Parse and format date timestamps by using Spark.
- Use temporary tables to prepare data for SQL.
- Combine PySpark and SQL to run queries.
- 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.

Take a few minutes to reflect on your learning:

- What new topics did you learn in this module?
- How has your understanding changed or evolved?
- What are you wondering about?
- What questions do you have?

You will continue to build your knowledge and skills throughout the boot camp. Reach out to your instructor team with any outstanding questions about the content in this module.