SCHOOL OF DATA
by Springboard

Data Engineering Career Track
Azure Data Engineering Hands-on Labs

Now that you're wrapping up the Azure section of the course, it's important for you to get some hands-on experience creating Azure data pipelines and utalizing all the various Azure components that you've been learning about.

Please complete the following labs, in the order they appear below. Each lab builds iteratively upon the lab that came before it. In the first lab you'll work with Spark to prepare your data in a data warehouse. In the second lab, you'll transform data using one of two Azure tools. In the third lab, you'll optimize your SQL queries. Be sure to answer the reflection questions on the second page and turn them into your mentor.

When you've finished the labs, consider reading some of the optional blog posts on the second page.

Lab 1 - Explore, transform, and load data into the Data Warehouse using Apache Spark:

- Please follow the lab setup instructions for this module.
- Follow the steps as laid out here in order to complete the lab.

Lab 2 - Transform data with Azure Data Factory or Azure Synapse Pipelines:

- The lab setup for lab 1 will be used for this lab as well.
- Follow the steps as laid out here in order to complete the lab.

Lab 3 - Optimise query performance with dedicated SQL pools in Azure Synapse:

- The lab setup for lab 1 will be used for this lab as well.

- Follow the steps as laid out here in order to complete the lab.

Reflection Questions - To Be Submitted To Your Mentor

- Why should one use Azure Key Vault when working in the Azure environment? What are the alternatives to using Azure Key Vault? What are the pros and cons of using Azure Key Vault?
- How do you achieve the loop functionality within an Azure Data Factory pipeline? Why would you need to use this functionality in a data pipeline?
- What are expressions in Azure Data Factory? How are they helpful when designing a data pipeline (please explain with an example)?
- What are the pros and cons of parametrizing a dataset in Azure Data Factory pipeline's activity?
- What are the different supported file formats and compression codecs in Azure Data Factory? When will you use a Parquet file over an ORC file? Why would you choose an AVRO file format over a Parquet file format?

OPTIONAL: Now that you're done with the labs, consider cementing your knowledge with the following blog posts:

- Azure Synapse vs Databricks - Differences and features (13 min read)
- Pyspark — Parallel read from database (2 min read)
- How to Optimise Azure SQL Database (7 min read)