Data Engineering and Analytics Utilizing Snowflake

Chloe Israel

Mentor: Dr. Sarah Heckman

Fall 2025 (second 8 weeks)

Independent Development

Prerequisites

- CSC 316 - Data Structures and Algorithms

Description

Understanding how to work with data is essential for any aspiring data engineer or analyst. Raw data needs to be processed extensively before it can be usable. Processes like these are often collaborative, which makes knowing how to utilize data-sharing platforms equally important. For this independent study, I want to learn and demonstrate essential data engineering and analytics concepts and skills, utilizing the platform Snowflake. Snowflake is a cloud-based data warehousing platform that has recently become popular in industries seeking solutions for data management, engineering, and analytics. Platforms such as Snowflake provide a single, centralized place for data to be stored, managed, and analyzed, which is extremely useful for companies that work with and share large amounts of data. The company provides a free online service called Snowflake Academia, which includes free access to the platform, resources, and on-demand courses. I have selected specific courses on Snowflake Academia to help me learn topics such as data warehousing, advanced databases, data wrangling with unstructured and semi-structured data, SQL querying, exploratory data analysis, data application building, and AI/ML modeling.

I will spend six weeks completing the online courses, documenting the tasks I complete, my results (as many of the online courses provide grades at the end), the capabilities and limitations of Snowflake, and my assessment of the platform. For the final two weeks, I will complete a final project utilizing the skills I learned from the courses I completed. The final project

deliverable will involve creating a database or application using a dataset of my choice. The main requirements include showing the transformations from raw to clean/processed data, demonstrating SQL queries, providing a minimum of three relevant insights based on data analysis, and providing a minimum of two relevant insights based on AI/ML modeling techniques. The final report deliverable will be a summary and discussion of what I learned and the results of the final project. The details and deliverables for the independent study will be documented and stored on GitHub.

Proposed Timeline and Milestones

- **Week 1 (~16 - 20 hrs)**: Account and Environment Setup + Data Warehousing Workshop + Level Up: First Concepts + Level Up: Snowflake Key Concepts and Architecture + Documentation and Reflection + Weekly Check-in
- **Week 2 (~14 - 18 hrs):** Collab, Marketplace & Cost Estimation Workshop + Level Up: Data Loading + Level Up: Snowflake Ecosystem + Documentation and Reflection + Weekly Check-in
- **Week 3 (~14 - 18 hrs):** Data Application Builders Workshop + Level Up: Object Hierarchy + Level Up: Backup & Recovery + Documentation and Reflection + Weekly Check-in
- **Week 4 (~14 - 18 hrs):** Data Lake Workshop + Level Up: Getting Started with Snowpark for Python + Level Up: Connect to Snowflake with Snowpark + Documentation and Reflection + Weekly Check-in
- **Week 5 (~14 - 18 hrs):** Data Engineering Workshop + Level Up: DataFrames in Snowpark + Level Up: Snowpark in Python Worksheets + Documentation and Reflection + Weekly Check-in

- **Week 6 (~14 - 18 hrs):** Data Science Workshop + Level Up: Python User-Defined Functions in Snowpark + Level Up: Native App Development for Beginners + Documentation and Reflection + Weekly Check-in
- **Week 7 (~18 - 20 hrs):** Final Project + Weekly Check-in
- **Week 8 (~18 - 20 hrs):** Final Project and Final Report + Weekly Check-in

Deliverables and Grading

The minimum deliverables for this project are the Final Report and Final Project. An incomplete would be appropriate if one or both of the deliverables were not completed and/or determined to be incomplete (i.e, did not meet the outlined requirements in the description).