Ayan Ashkenov
CS779: Advanced Database Management
03/16/2023

Term Project Update #2

My project outline remains the same, with only a slight change in one of my data sources. Originally, I planned on using a local instance of SQL Server as one of my data sources for the Databricks Lakehouse. However, since MongoDB and Amazon S3 are cloud based, I decided it made more sense to adopt cloud architecture throughout. As a result, I will be using an instance of Azure SQL Server instead.

Current project plan:

1) Preliminary research on Databricks and which data sources to use with it (SQL Server, MongoDB, Amazon S3). **Completed**
2) Create an instance of each data source and test connections. **Completed**
3) Consult Apache Spark documentation for read and write methods. **Completed**
4) Get data from MongoDB (which is in JSON-like format) and save it as a SQL table in Databricks. Check if it queries and returns the expected output. **Completed**
5) Perform the same test for Azure SQL Server (extract data and save as SQL table in Databricks, using PySpark) **In progress**
6) Find data for final version use (currently looking at pseudo transaction data from a paid API and pseudo customers data from another API) **In progress**
7) Populate the data sources (MongoDB, Azure SQL Server, Amazon S3) with data from the API (Data pipelines in Python).
8) ETL inside Databricks using PySpark for each data source. Test if ingested data is available in Databricks for querying.
9) Connect Databricks to Tableau and create a sample dashboard.
10) Connect MLFlow to Databricks and create a sample ML pipeline.
11) *If time permits, add scheduling to all pipelines.

Data Sources:

https://www.mongodb.com/docs/atlas/
This is the official Atlas MongoDB documentation page. It provides information on how to work with a cloud MongoDB instance.

https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_df.html
This is the official Apache Spark (PySpark specifically) documentation. It provides information on how to work with Spark dataframes.

https://docs.databricks.com/getting-started/quick-start.html
This is the official Databricks documentation. It provides information about notebooks and common use cases.