Assignment 5 Report

Aditya Jain

1 Part A

1.1 Part A.1

- **Azure Data Lake**: Data lake is primarily a data storage service that ingests and stores large volumes of data in its original form. Due to its open open, scalable architecture, it can accommodate all types of data from any any source, structured or unstructured. The data can then be processed and used as a basis for a variety of analytical needs.
- **Azure Databricks**: Azure Databricks allows data analysis and building ML solutions using Apache Spark environment. It supports languages such as Python, R, Java, and Scala, and ML libraries such as TensorFlow, PyTorch, and scikit-learn.
- **Azure Data Factory**: Azure Data Factory is a cloud-based ETL and data integration service that allows us to create data-driven workflows for orchestrating data movement and transforming data at scale.
- **Azure Synapse Analytics**: Azure Synapse Analytics is an analytics engine, designed to process large amounts of data quickly. It accelerates time to insight across data warehouses and big data systems.
- **Azure Cosmos DB**: Azure Cosmos DB is a globally distributed multi-model database, offering single digit millisecond latencies at the 99th percentile anywhere in the world,. It is scheme-agnostic and automatically indexes all the data without requiring to deal with scheme and index management.

Following will be my choice of Azure components for the big data architecture:

1. **Ingest**: Azure Data Factory (ADF)
2. **Data Store**: Azure Data Lake, Azure Cosmos DB
3. **Prepare and transform data**: Azure Data Factory
4. **Model and serve data**: Azure Databricks, Azure Synapse Analytics

ADF is designed for ingesting data into the system and and transforming it using its ETL framework. Azure Data Lake and Cosmos DB are the fundamental storage options, while Databricks and Synapse Analytics can be used for doing analysis on top of the transformed data.

1.2 Part A.2

Stream Analytics in Azure is desgined to analyze and process large volumes of streaming data. There are three building blocks to it: **Ingest**, **Analyze**, and **Deliver**. In the Ingest stage, real-time data is fed from a variety of input sources such as edge devices, sensors, clickstreams, and social media feeds. In the next Analyze stage, the data is analyzed in real-time using services such as Azure ML. In the final Deliver stage, the found patterns in the data can be used for feeding information to a reporting tool such as Power BI, creating alerts and actions on Event Hubs, Service Bus, Azure Functions etc. or storing the transformed in Blob storage or Data Lake for later use.

1.3 Part A.3

Figure 1 shows the deployed resources for the stream analytics job and figure 2 shows the result.

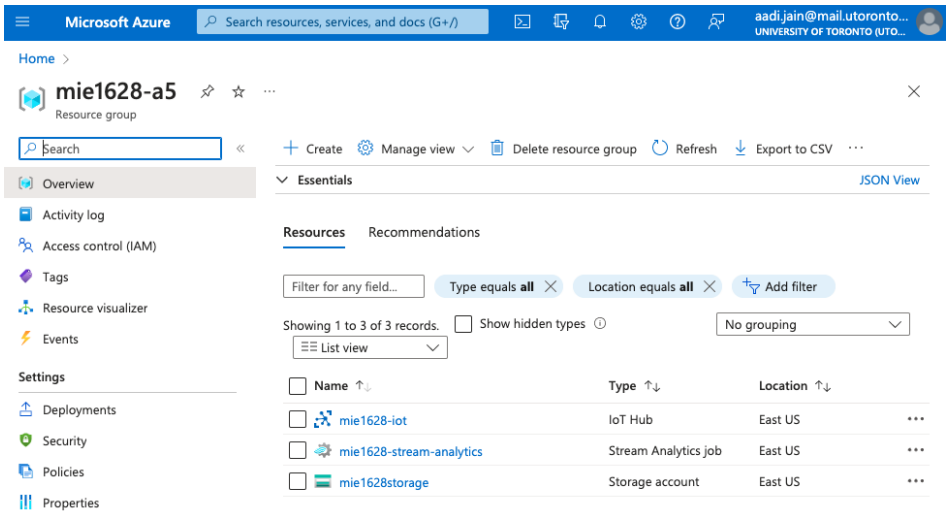

Figure 1: Deployed resources for Part A.3

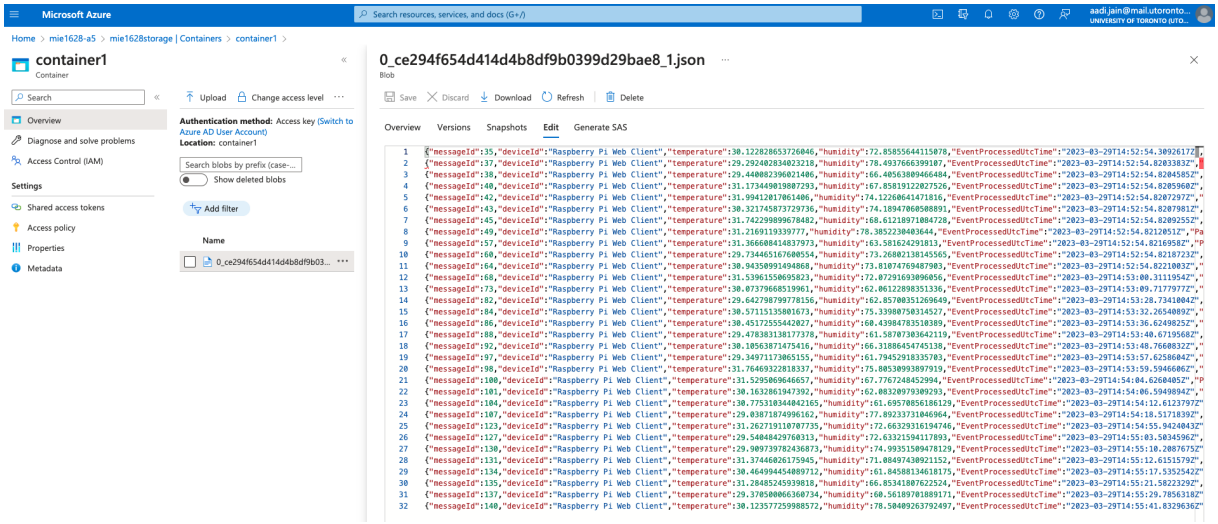

Figure 2: Result of Part A.3