Springboard Open-ended Capstone Overview

This project starts with finding a large data set, then builds a data pipeline to explore and extract valuable information from a small sample data set on a local Apache Spark system, then scale up and deploy it on to Microsoft Azure Databricks for production.

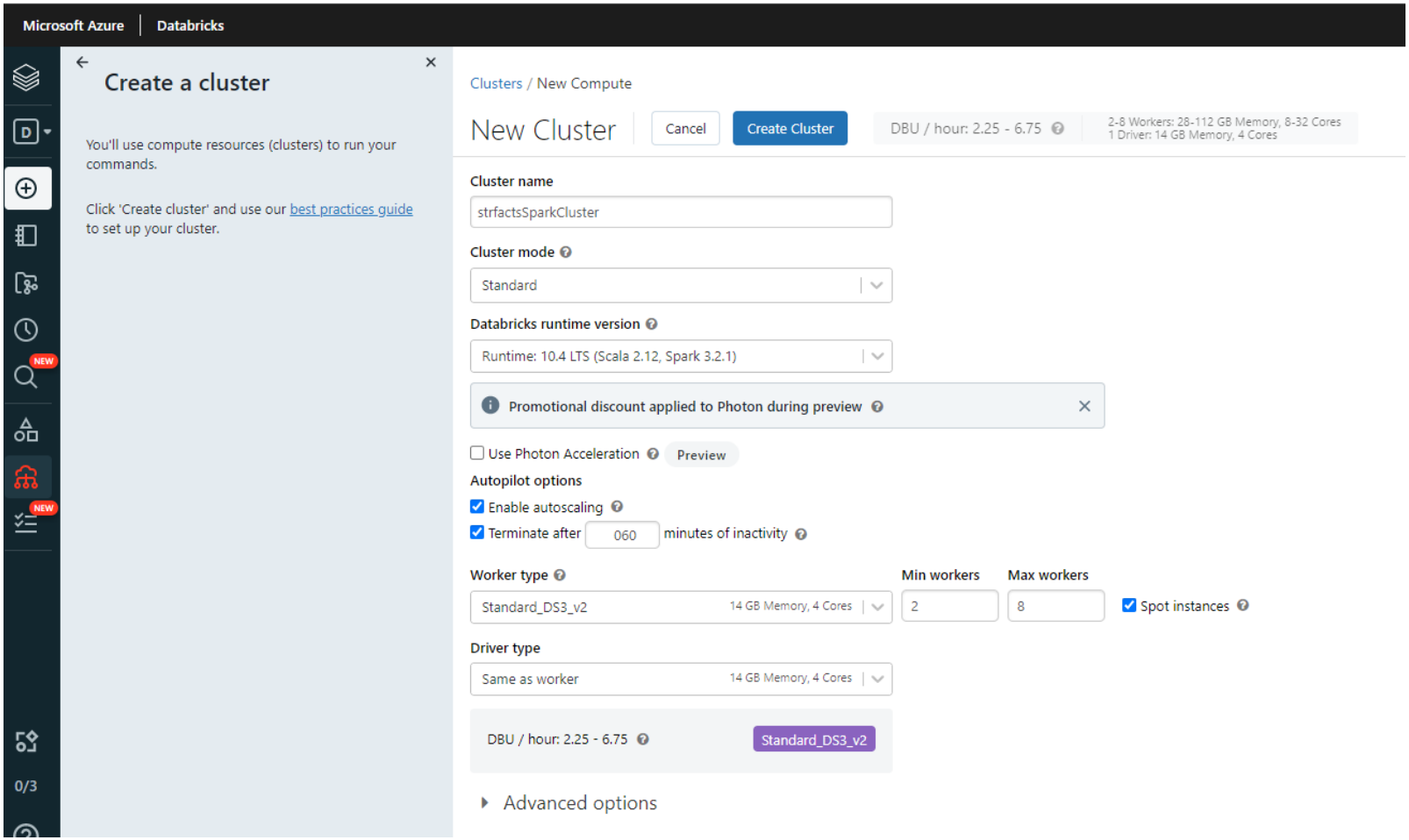
Short Term Rental Facts

- Short term rentals have exploded with the help of platforms such as airbnb.com, booking.com, vrbo.com.
- Both travellers and hosts need information to find or host a space.
- This project collects large historical, current and future data to extract valuable information.

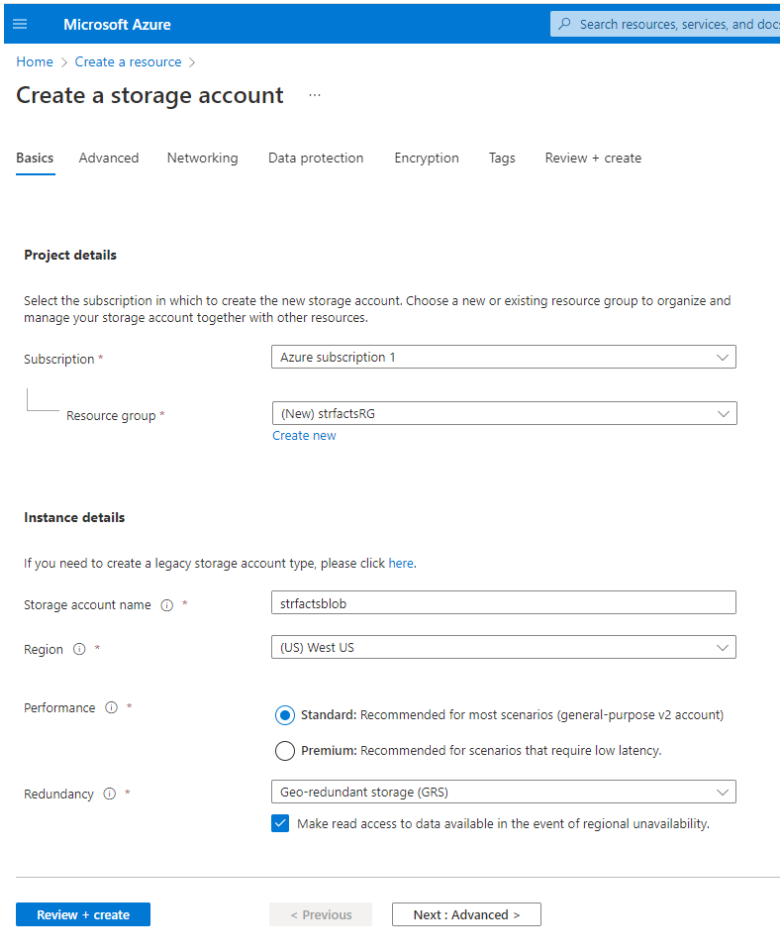
Data Pipeline Design

- The data pipeline should be run quarterly based on the chosen data set.
- Data acquisition: python program that downloads the data set from a web source.
- Data ingestion: the data set is read, cleaned, then written to storage in partitions.
- Data analytics: Focused on the reviews of each property. It extracts the most frequent used words from the reviews to give people some idea without reading the actual reviews.

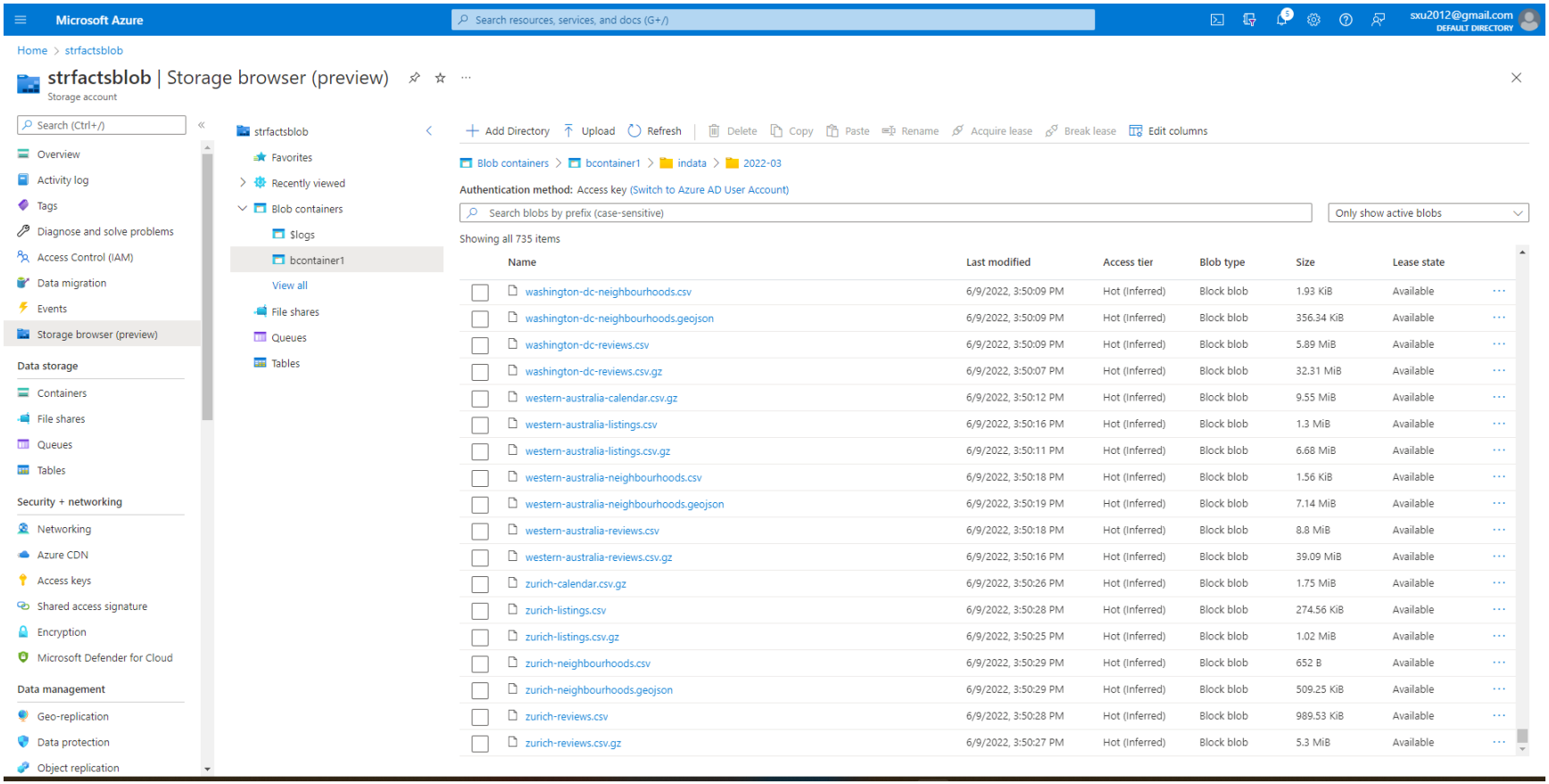
Asure Databricks Spark Cluster



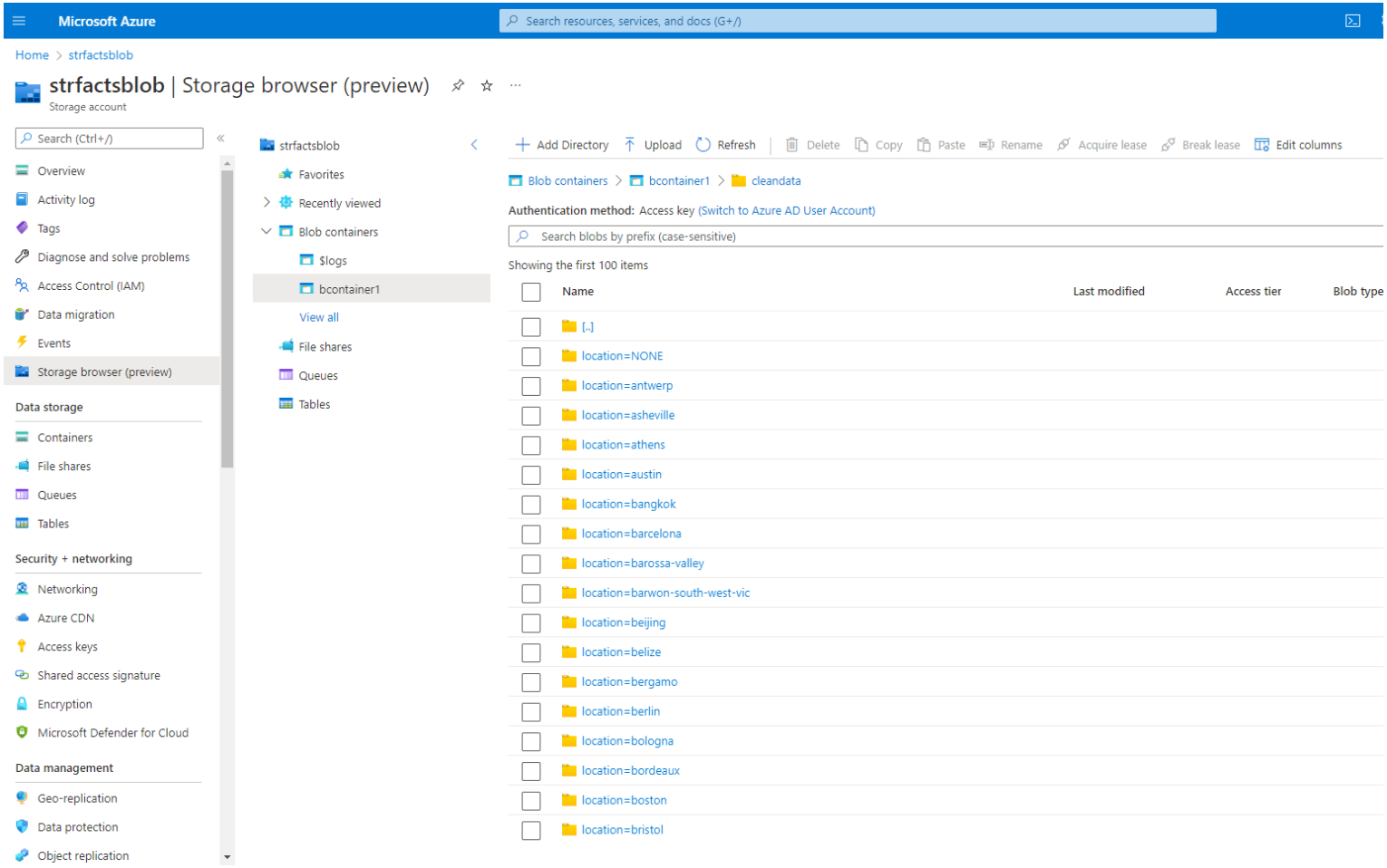
Azure Blob Storage



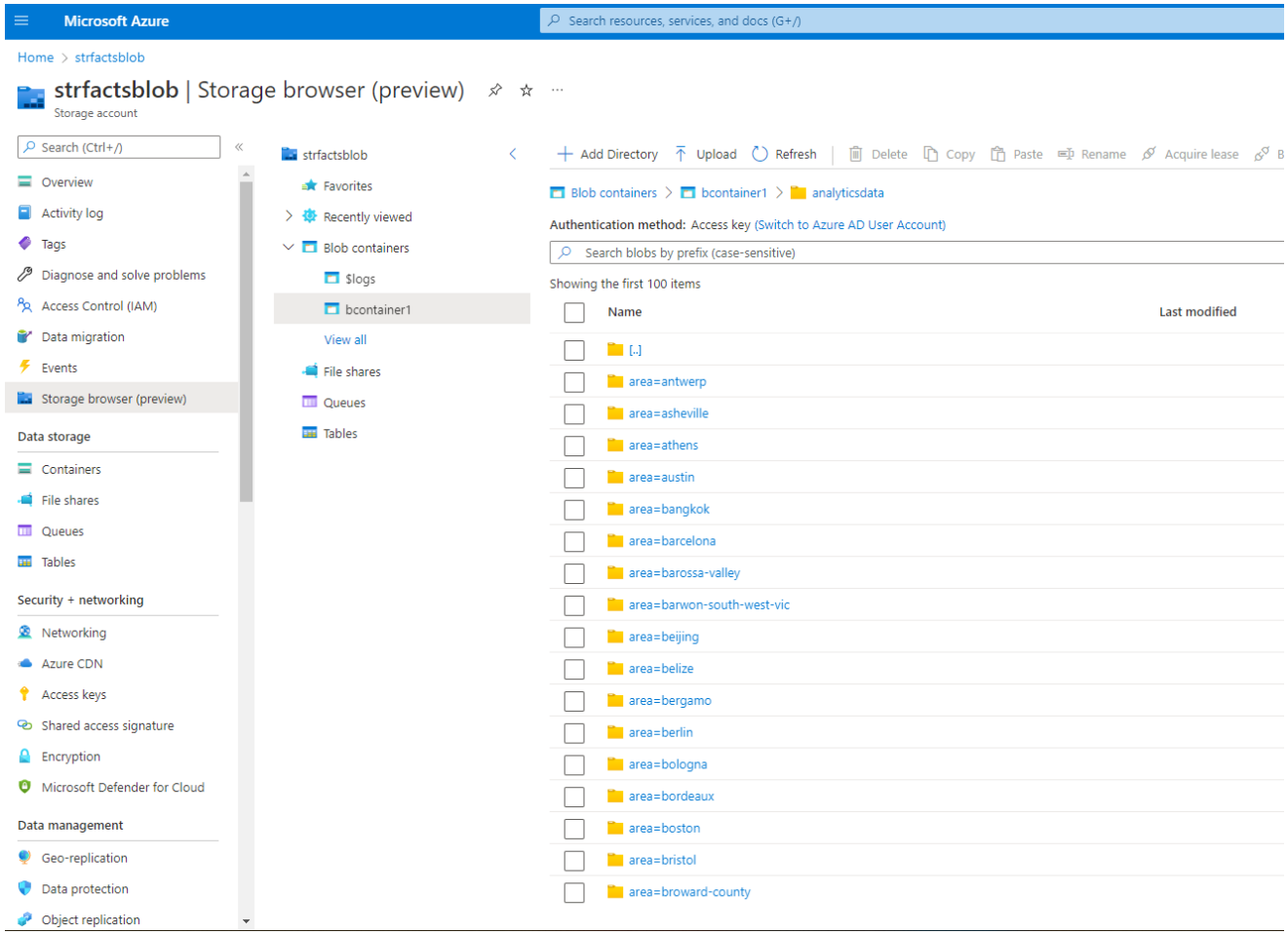
Blob After Data Acquisition



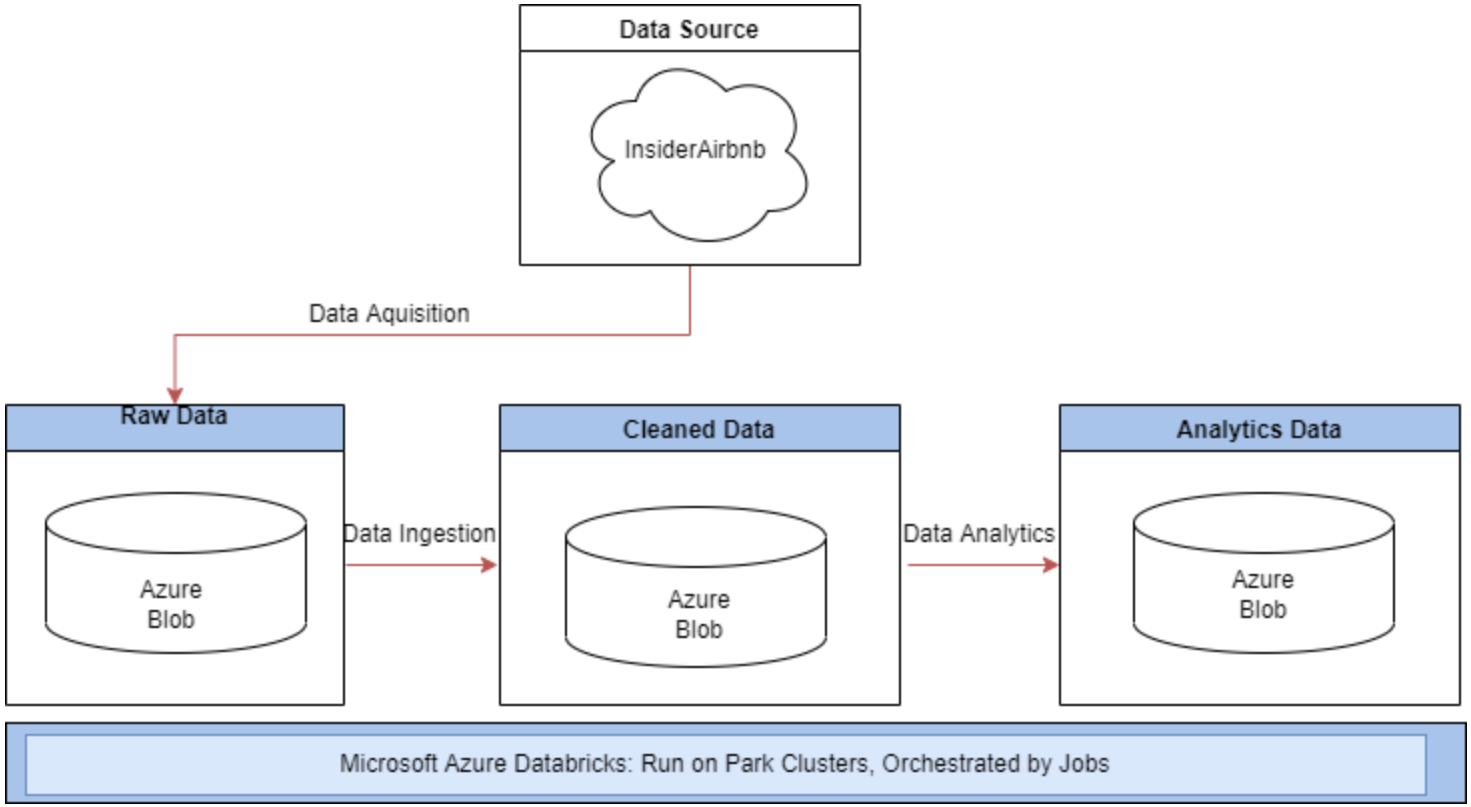
Blob After Data Ingestion



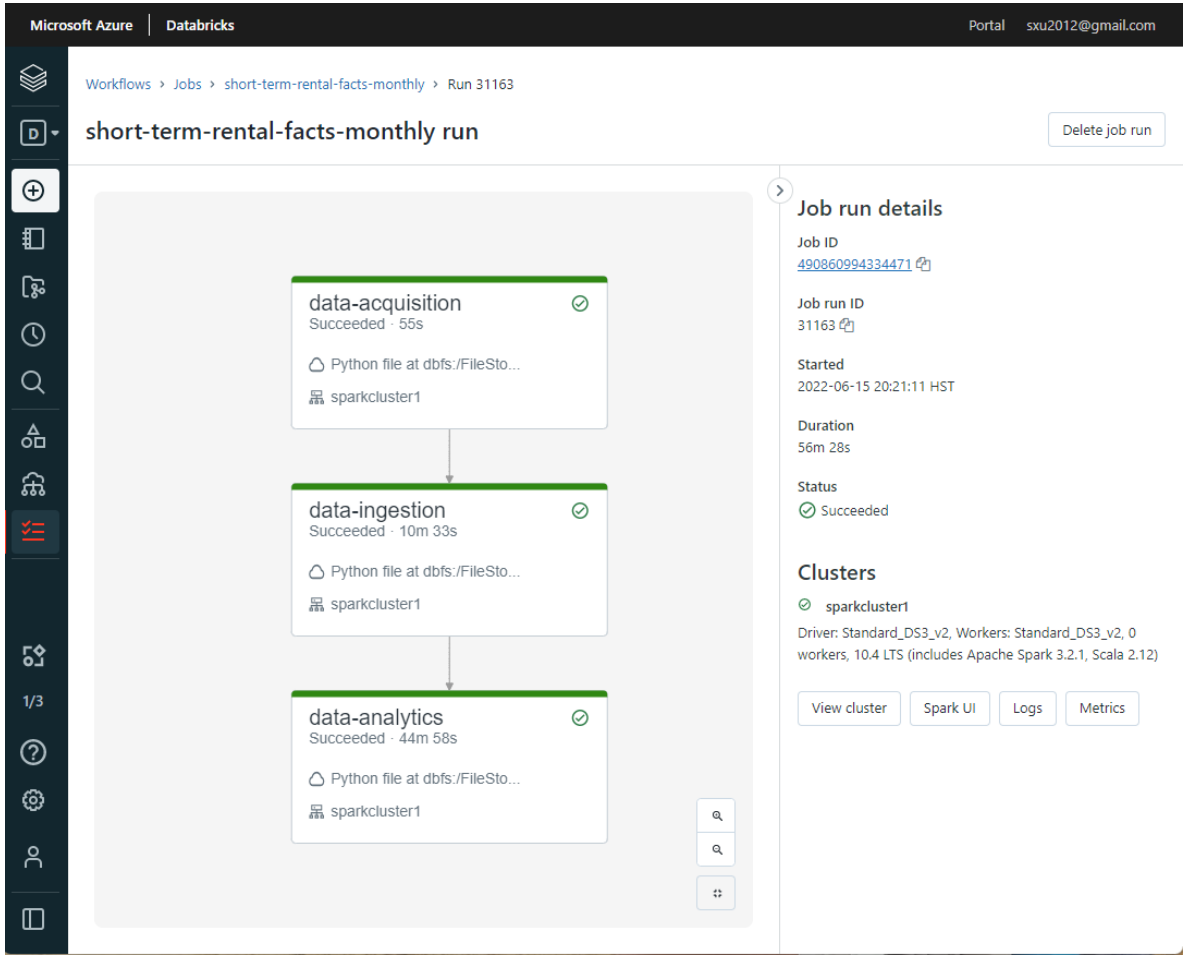
Blob After Data Analytics



Deployment Architecture



Pipeline Orchestration



Pipeline Monitoring