Create Deployment Architecture

Capstone Step 7

Decision Making Process

This process was fairly straightforward, as I now have experience using each of the tool required.

1. I did have to decide between using **Apache Airflow** vs **Azure Data Factory** for orchestration of ETL pipeline tasks

    I chose **Azure Data Factory** because I knew it would ease any pain encountered when authenticating connections between services

2. I successfully imported and queried the data using both:

    Databricks using a Pyspark kernel

    Azure Data Studio using a SQL kernel

    The speeds were roughly the same, but VM core quota issues with **Databricks** frustrated me, and I chose to serve the data with **Azure Data Studio**

3. I was unable to create a **Microsoft Power BI** resource without using an organization email account

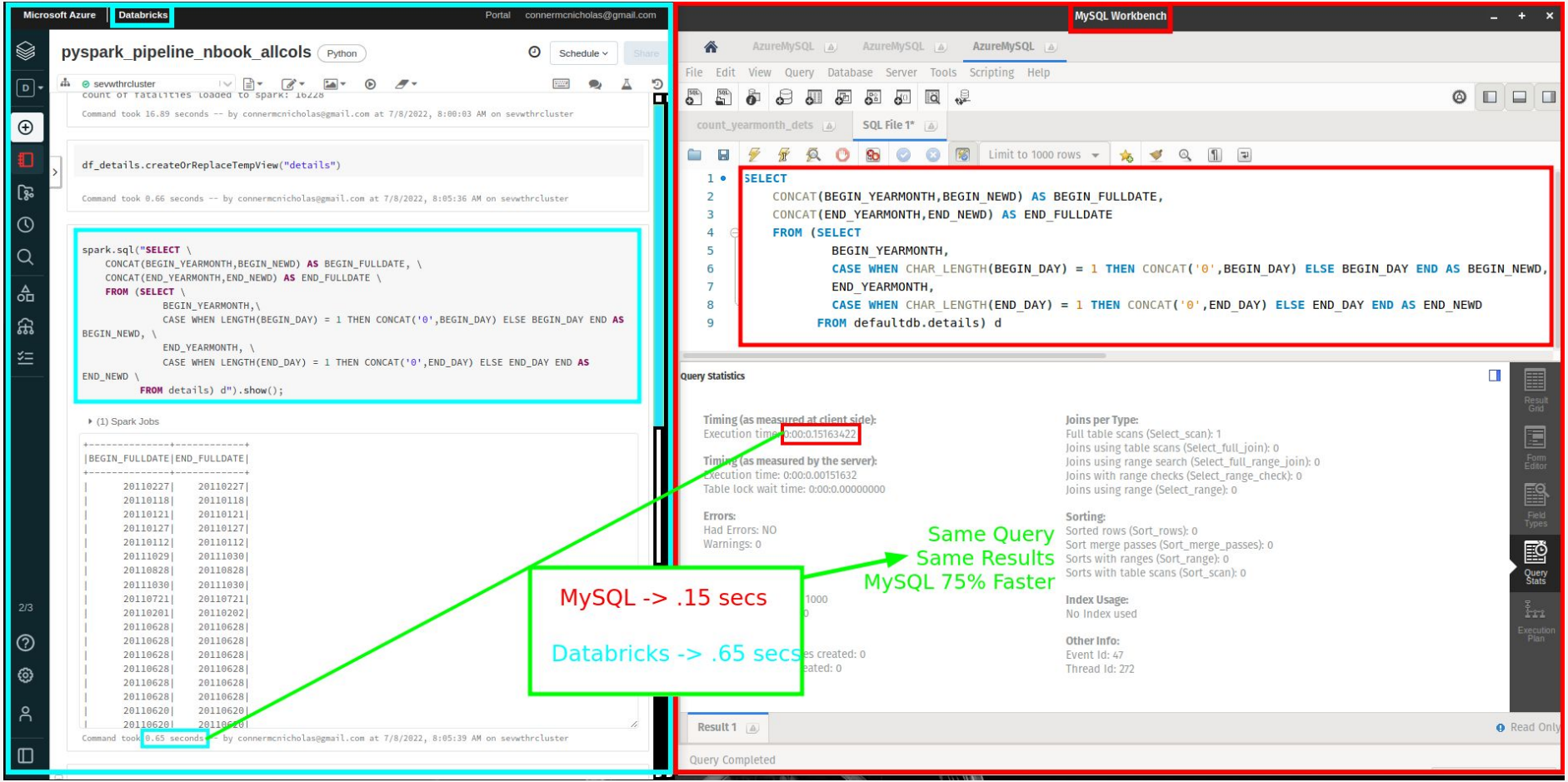
    Provisioning business intelligence software is overkill for my requirements, as **Azure Data Studio** extension **SandDance** accommodates plotting.

4. I then realized that if and when I want to move to a non-Azure infrastructure, I would want to be cloud-agnostic,

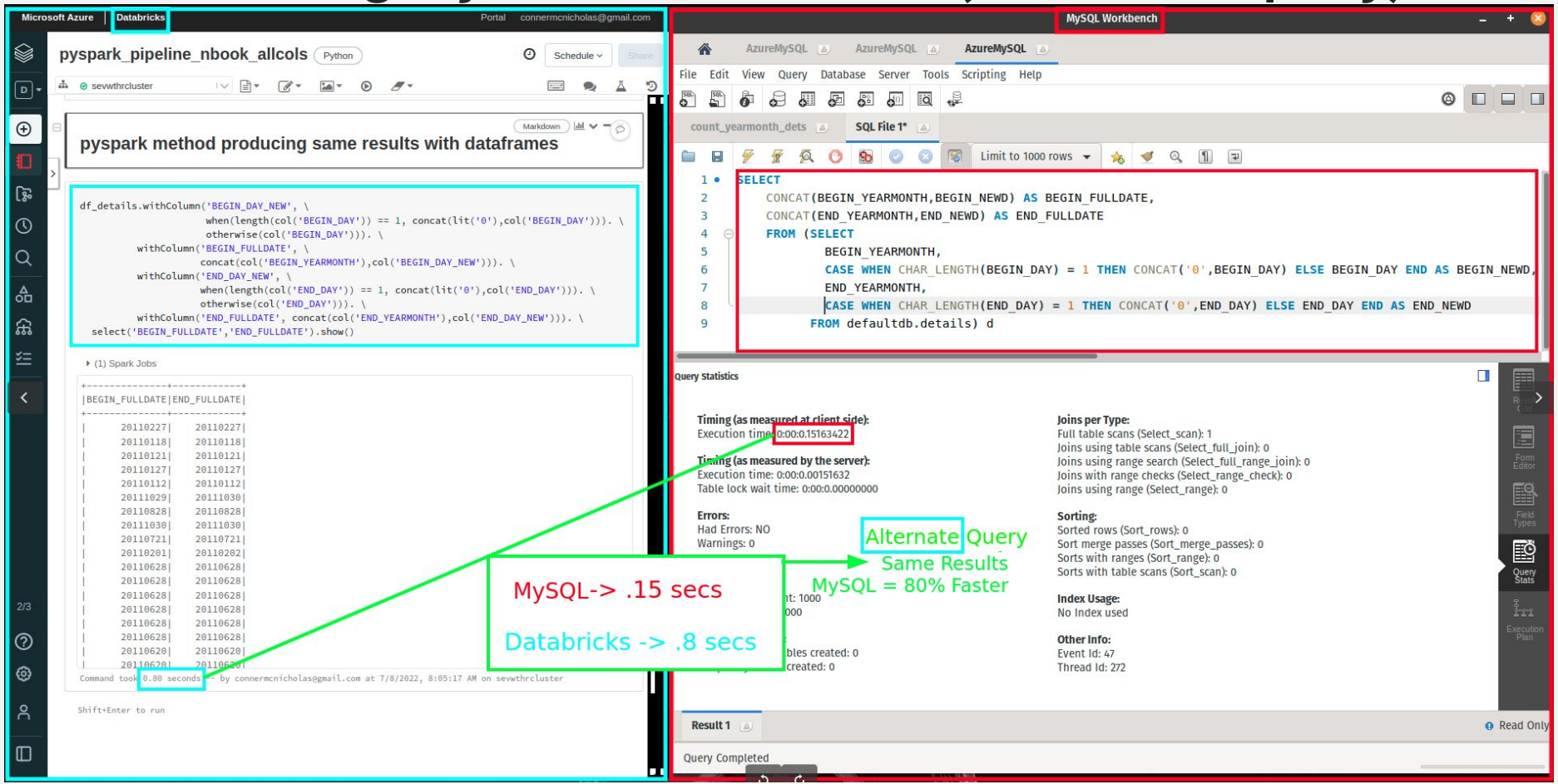
    so I converted **Azure Data Studio** + **SandDance** to **MySQL** + **Jupyter**

    Comparing speeds with **Databricks** again shows **MySQL** to be superior (at least for the resources I have configured):

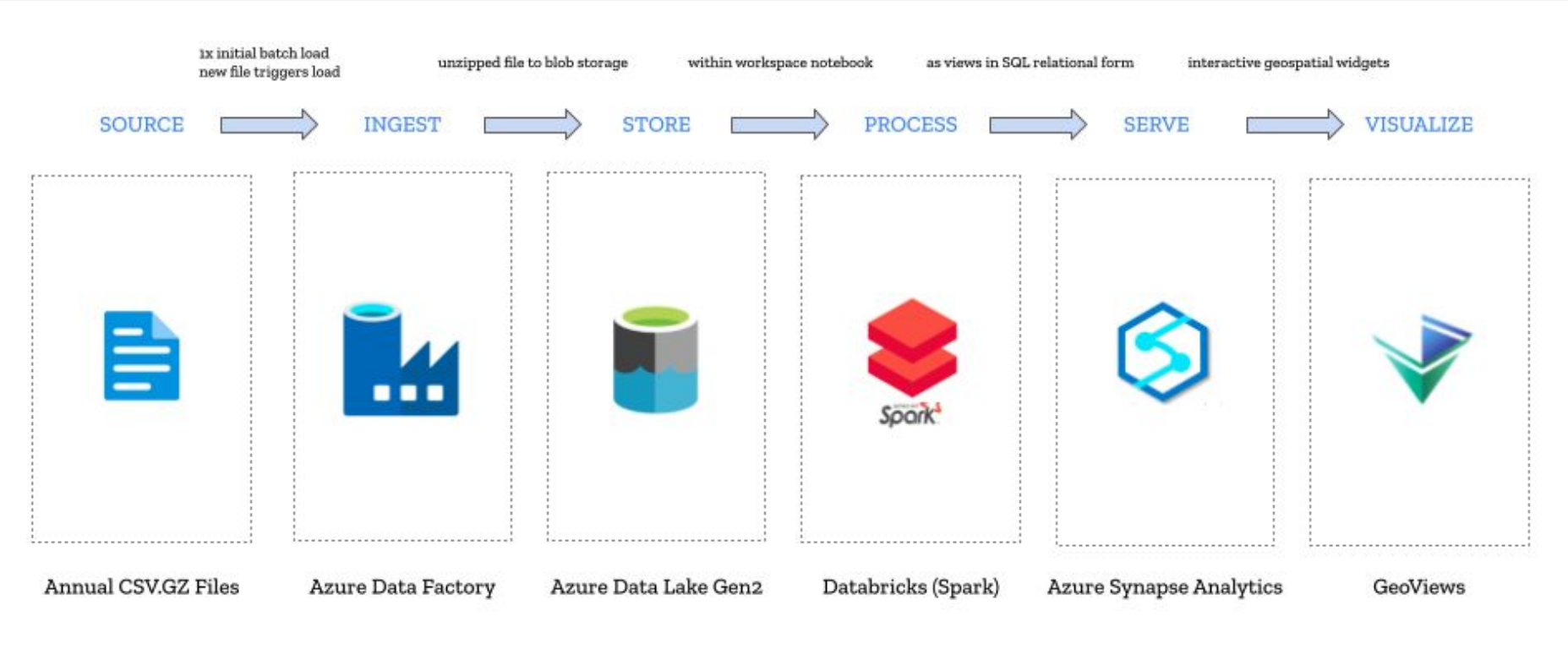
Benchmarking MySQL vs Databricks (pyspark.sql query)



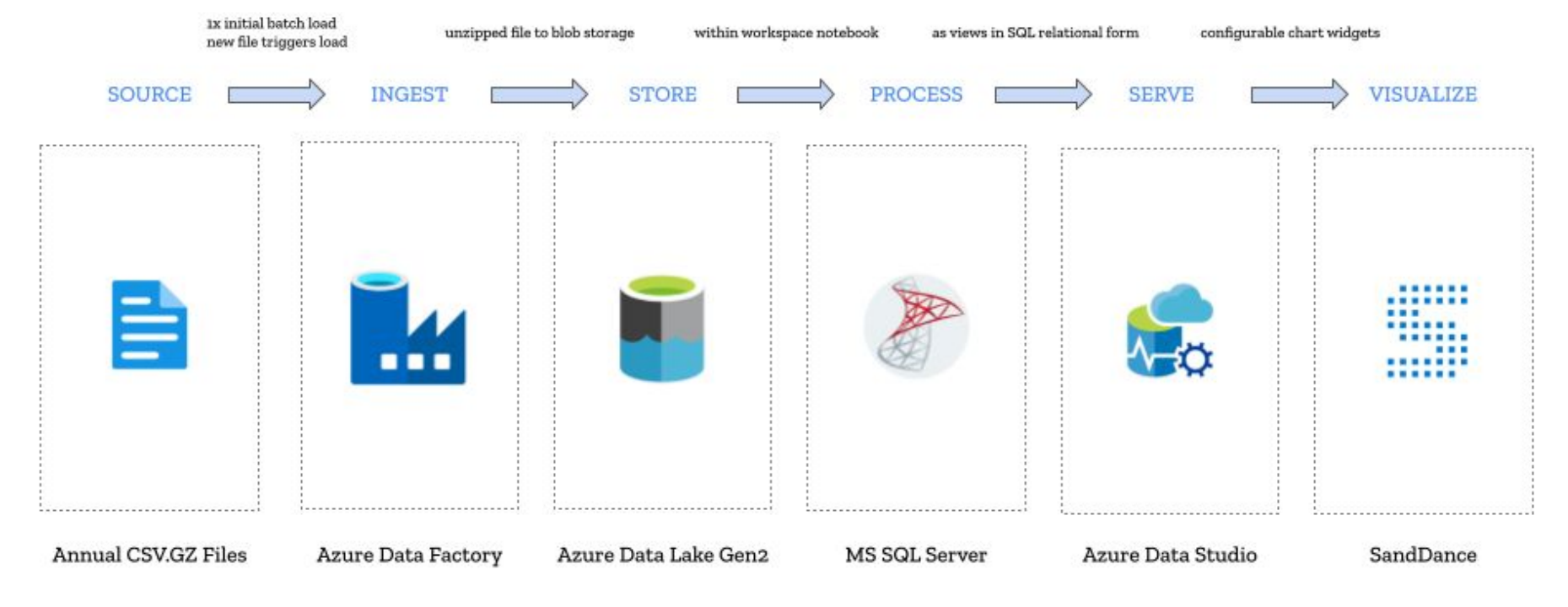
Benchmarking MySQL vs Databricks (dataframe query)



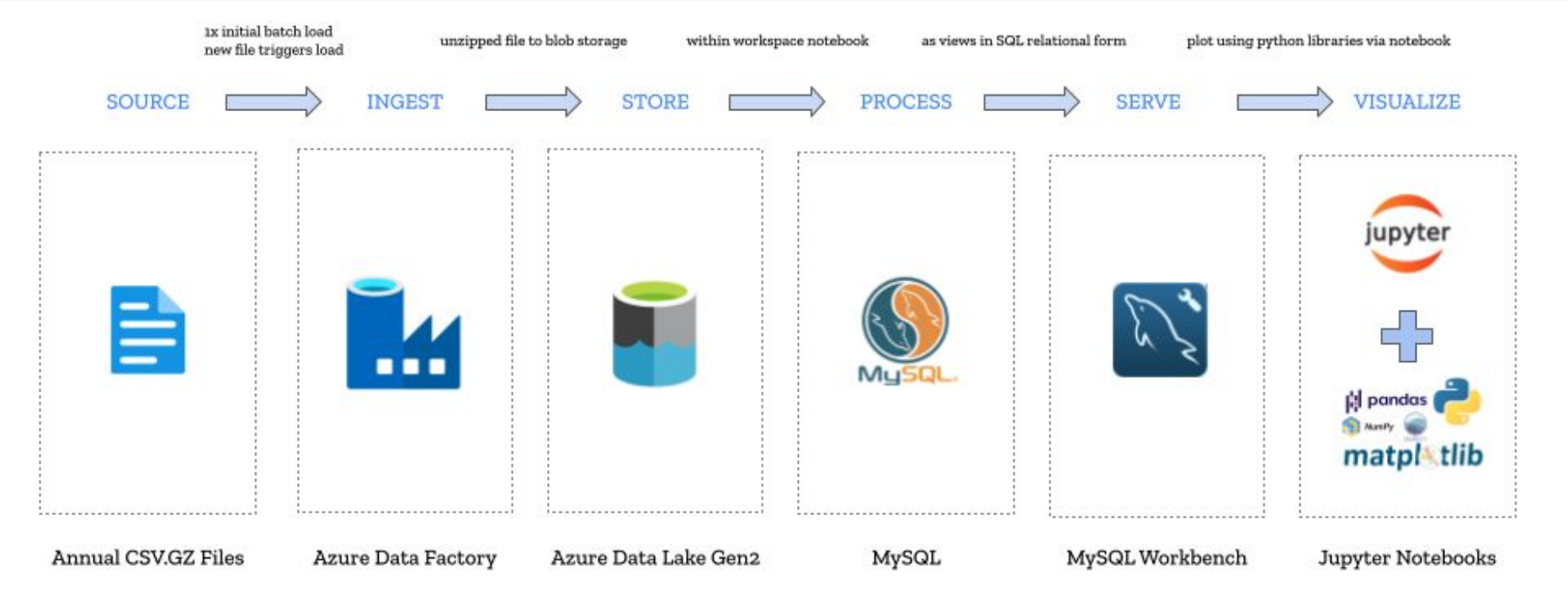
Initial Architecture



Intermediate Architecture



Final Decision Architecture



Resource Deployment Exports

The ability to deploy and export the resources was possible for all resources except **Azure Data Factory**, which lacks the feature.