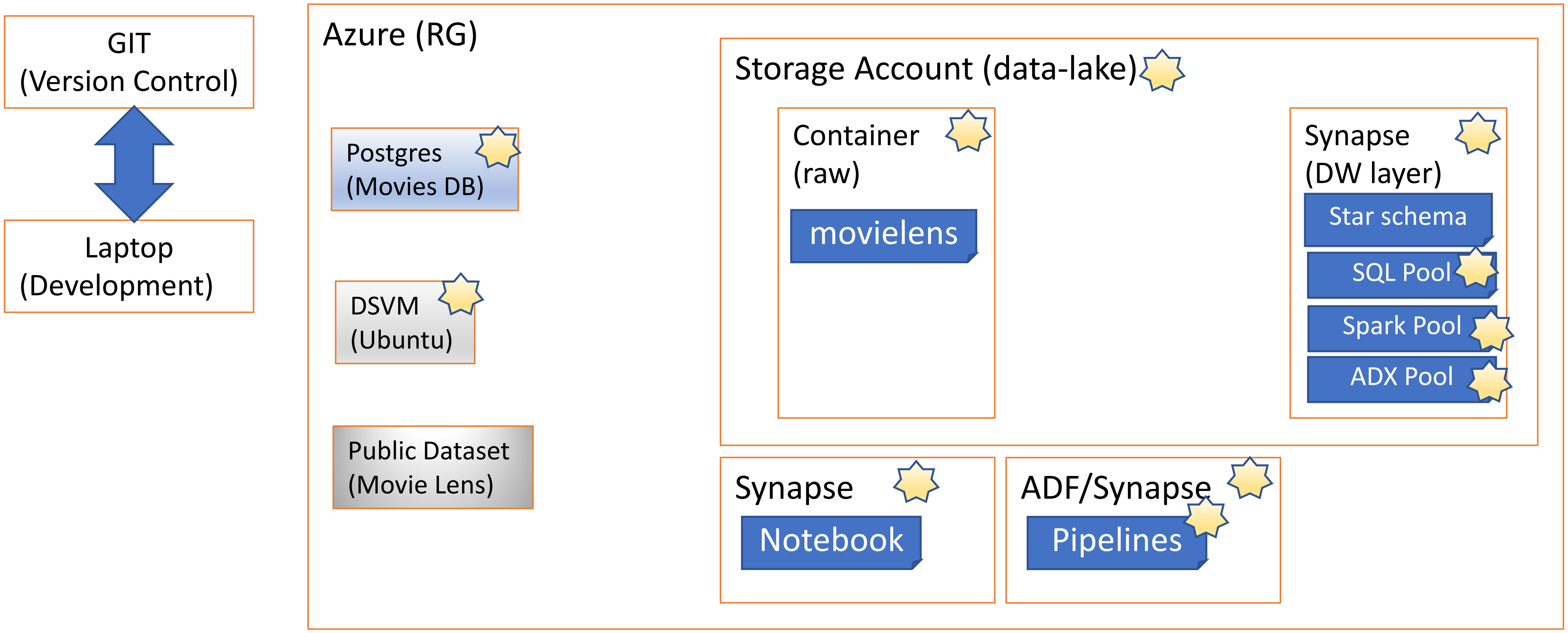
MovieLens Demo

Motivations

- Explore various capabilities offered by Azure to:
  - Provision a developer workspace / ADLS / Synapse / Jupyter notebook via IaC
  - Explore a non-trivial dataset (Movie Lens) following persona-based journeys.
    - Developer Journey – Cost conscious, VM deployment for Data Science development.
    - Data Analyst Journey – Use sandbox database with Kusto query language for insights.
    - Data Engineer Journey – Production ready from other journeys (Spark, Serverless, Dedicated Synapse Pools)
  - Share code-base with the team to allow for enhancements/experiments
    - https://github.com/SowmyaVenky/Azure-DP-203
    - YouTube follow-along https://www.youtube.com/channel/UCI5gdy3DaIITi_jTYXhLmVA
  - Cover concepts that can help with various Azure Certifications.

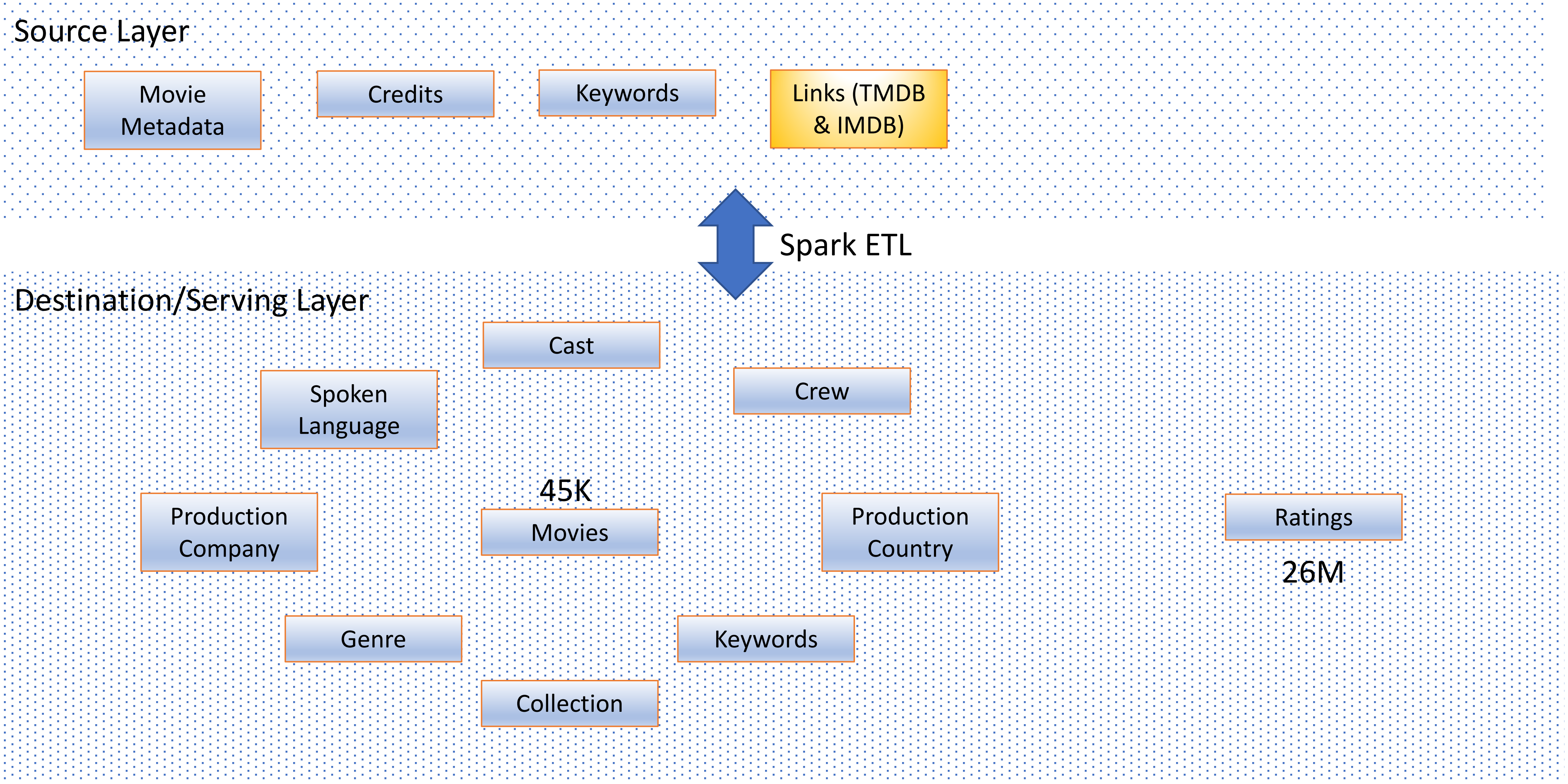
Demo Project (MovieLens)



☆ ARM Template

Dataset Introduction

- Movielens dataset from Kaggle. This is the basis for most of the demonstration.
- The dataset is in CSV format, but some columns are complex (Array of structures)



Source Layer

Movie Metadata

Credits

Keywords

Links (TMDB & IMDB)

Spark ETL

Destination/Serving Layer

Cast

Spoken Language

Crew

45K

Production Company

Movies

Production Country

Ratings

26M

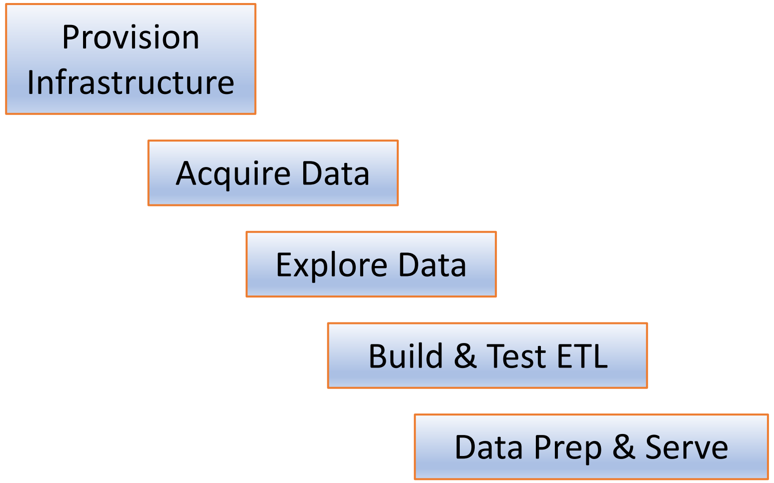
Genre

Keywords

Collection

Journey # 1 – Talented Developer who wants to experiment!

- To create a consistent playground, we are using the Microsoft Data Science Virtual Machine (DSVM). This has all the libraries pre-loaded and configured to work out of the box. Has Spark and Jupyter loaded and ready to go. No install and configuration pains!
- Has consistent paths to allow sharing between developers.
- Exploratory analytics on raw data via DSVM + Jupyter notebooks.
- Spark is used to read the CSVs and shred the data into a more relational format.
- A Postgres PaaS DB is created.
- All the required tables are created.
- Spark loads the data into the Postgres DB, and we issue some simple queries to demonstrate some fun facts.



Provision Infrastructure

Acquire Data

Explore Data

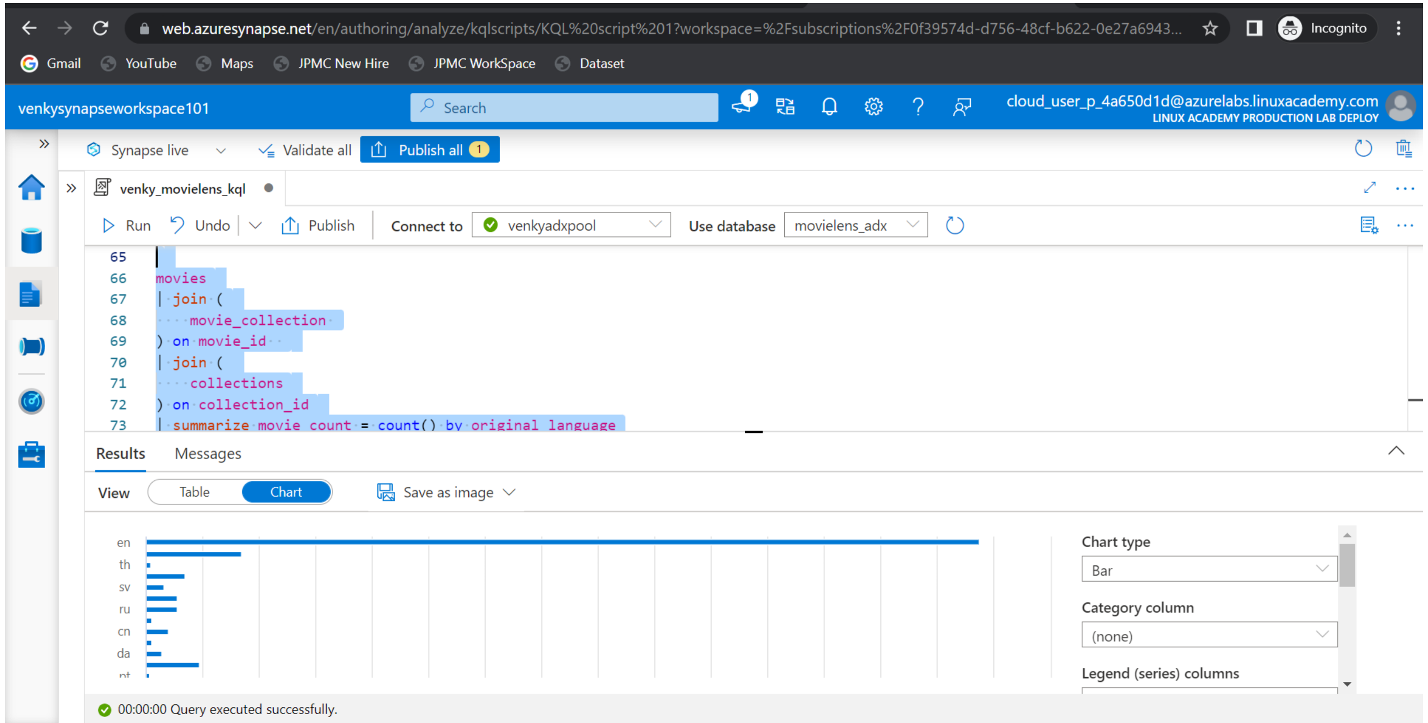
Build & Test ETL

Data Prep & Serve

Video walkthrough
https://www.youtube.com/watch?v=-ba02-rVJdo
https://www.youtube.com/watch?v=Z1eG9kHW-tk
https://www.youtube.com/watch?v=zj3pyBIvmcM
https://www.youtube.com/watch?v=ZDEwmcnKuUo

Journey # 2 – Data Analyst exploring data from a central lake inside own sandbox

- Understand the layers of the existing data lake implementation.
- Log into the Synapse ADX (Azure Data Explorer) workspace.
- Create personal ADX database.
- Ingest data from central data lake into ADX database.
- Perform explorations using ADX Kusto Query Language.


web.azuresynapse.net/en/authoring/analyze/kqlscripts/KQL%20script%201?workspace=%2Fsubscriptions%2F0f39574d-d756-48cf-b622-0e27a6943...
Incognito
Gmail
YouTube
Maps
JPMC New Hire
JPMC WorkSpace
Dataset
venkysynapseworkspace101
Search
cloud_user_p_4a650d1d@azurelabs.linuxacademy.com
LINUX ACADEMY PRODUCTION LAB DEPLOY
Synapse live
Validate all
Publish all 1
venky_movielens_kql
Run
Undo
Publish
Results
Messages
View
Table
Chart
Save as image
en
th
sv
ru
cn
da
pt

Connect to venkyadxpool

Use database movielens_adx

```
movies
| join (
    movie_collection
) on movie_id
| join (
    collections
) on collection_id
| summarize movie_count = count() by original_language
```

Chart type: Bar

Category column: (none)

Legend (series) columns

00:00:00 Query executed successfully.

Video walkthrough
Coming soon – YouTube limitations ☹

Journey # 3 – Pro Data Engineer - Logical Data Warehouse, Spark Analytics, and Dedicated DW

- Understand the layers of the existing data lake implementation.
- Create a pre-prod DW using Synapse Serverless pools. This gives the flexibility of analyzing how a DW would look without the added costs (since it is serverless)
- Create external tables, file-formats, and explore data in the data-lake. All analysis is SQL based.
- If you want more flexibility and are good in Spark, use that as a tool to explore, ETL, and create tables in the logical warehouse.
- Once satisfied, convert the logical DW to a dedicated pool DW and performance tune for production usage.

Video walkthrough
https://www.youtube.com/watch?v=LaoNNY8JtZE
https://www.youtube.com/watch?v=ges-hCIMd24
https://www.youtube.com/watch?v=9OUEigKAqyY
https://www.youtube.com/watch?v=cwFwsSfDMqw
https://www.youtube.com/watch?v=Zo3fq4FL0DA

Recap

- Azure has a lot of power/flexibility to fit usages from various personas.
- Lot of mundane tasks can be automated as IaC to allow precise reproduction many times.
- Data Engineering tasks are made easier with cutting edge, inter-operable tools.
- Capability to store large quantities of data and optimize it to serve business needs.

- What we did not look at:
  - Integration to other Azure tools for Data classification, Retention, and governance activities.
  - Machine learning integrations to data in ADLS or Synapse databases.
  - Data security features like Dynamic Data Masking, Always Encrypted and Role based access to data.
  - And much much more!

Code Repo: https://github.com/SowmyaVenky/Azure-DP-203