Azure Data Engineering project

(ETL)
pipeline

Data source (kaggle) (uploaded on github)

↓

Data integration (Data factory) (Data flow)

↓

store data (storage acc Azure)
(storing in a container)

↓

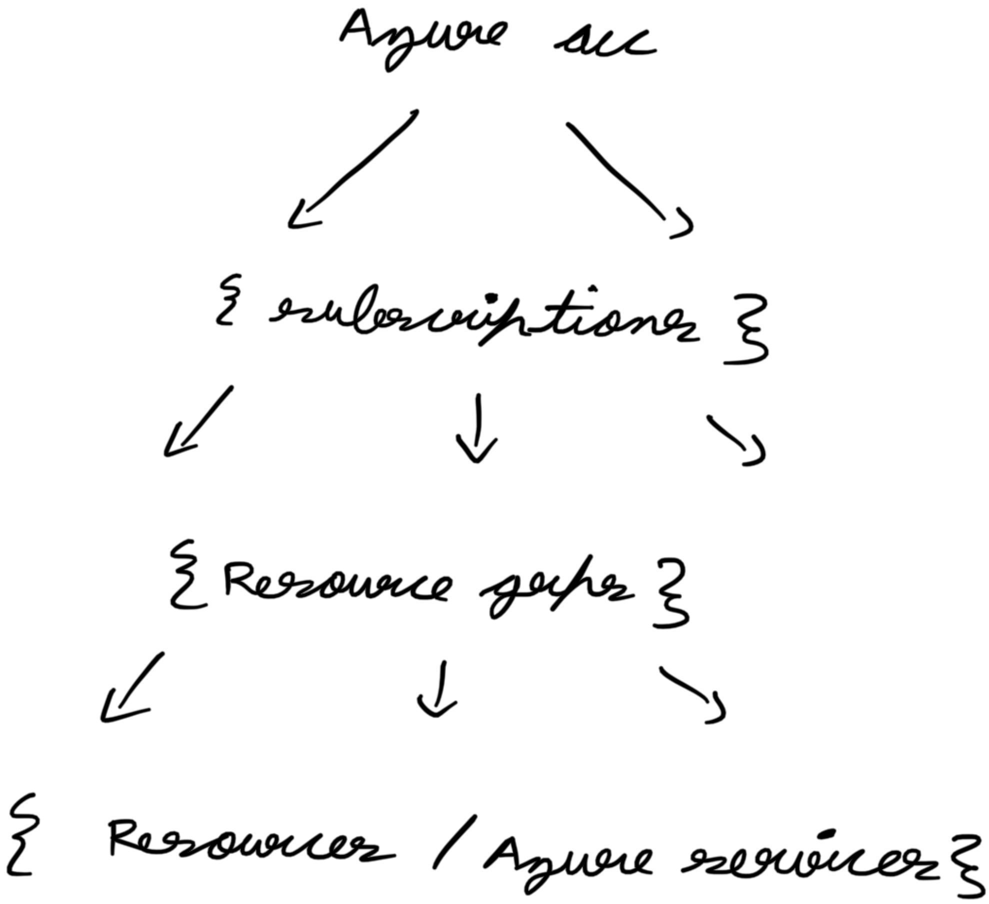
Transformation (Databricks, spark)

↓

store data

↓

Analytics (Azure synapse analytics / SQL /
Dashboard)


Azure acc
{ subscriptions }
{ Resource grps }
{ Resources / Azure services }

Store data

↳ container

- raw data
- transformed data

Datafactory (triggers → schedule the end to end pipeline execution)

↳ activities

- copy data (source → sink)

source	sink
- url from github (http)	- azure data lake storage Gen2

loaded data from source

data factory → storage container

Databrics

↳ new compute (to write spark code)

↳ new notebook (spark cluster)

raw data → Databricks → transformed data raw_data → databricks → transformed_data

have to connect databricks to ADLS

- use key valult to store & not expose the keys while mounting

- we are using the app to get the data (need to give permission using IAM)

- while writing the files apache stores them in a folder along with metadata

- if we have a very large file spark will divide the file into multiple files

Azure synapse

- after loading the tables to the DB

  (options
  - custom
  - from template
  - from Data Lake)

  ↳ SQL script
  ↳ notebook
  ↳ ML → train/predict

(ran SQL queries, MYSQL syntax)
(result -> table & chart)

Triggers in datafactory :

1) schedule (many to many)

2) Tumbling window (seperate files for every execution)

3) Event based (blob related events del/generation of blob)