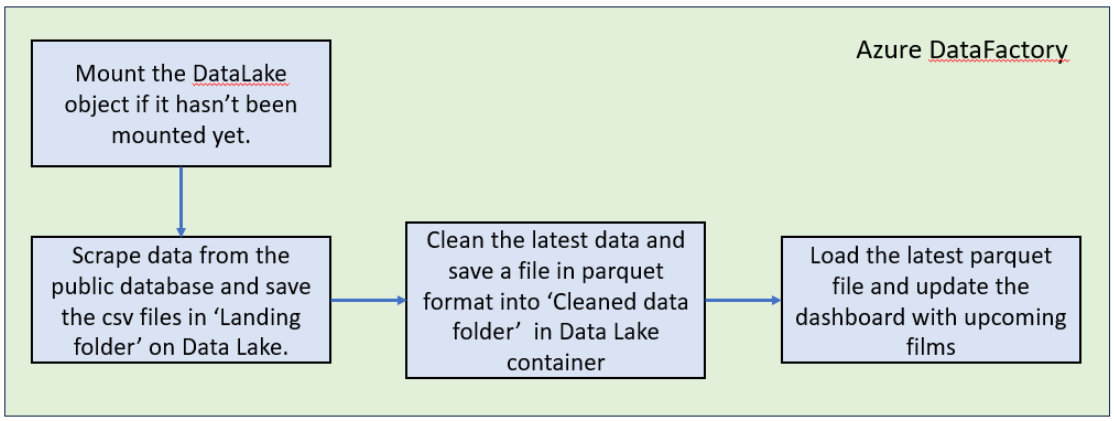
Azure pipeline for upcoming films:

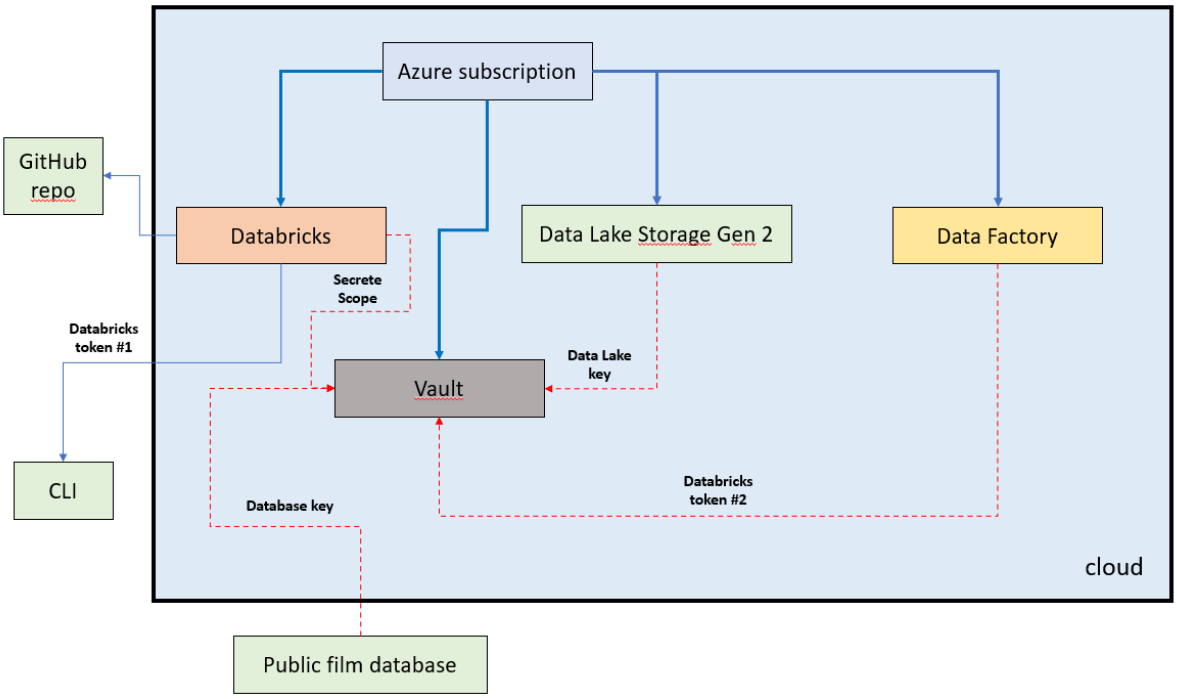
Goal:

- create a pipeline which scrapes data from public film database, process the data and create dashboard on the monthly basis

Flowchart:



Azure Resources and their connections:



Create a subscription in Azure:

- choose ‘Pay-As-You-Go’ subscription

Vault:

- create a vault resource group

- set up the role for viewing the secrets in the vault

Data Lake:

- create a resource selecting ‘Storage account’

- put the access key in the vault as a secrete

- create a container for landing the data

Databricks:

- in Azure create Databricks resource group

- during the creation of workspace – choose ‘Premium tier’ – allows creation of Secrete scope

- set up the cluster (choose single node)

- create a secrete scope backed by ‘Azure Key Vault’

- verify the creation of scope installing the CLI in VSC (need to generate a token in Databricks for that)

- mount the cloud object (Data Lake) using the vault

- create a repo in Databricks and connect it with your repo in GitHub (all commits in Databricks will appear in the GitHub)

- write a notebook for:

    - mounting the Data Lake

    - scraping the data and saving into ‘Landing folder’ on cloud

    - cleaning the data and saving into ‘Cleaned data folder’

    - read cleaned data and create a dashboard in Databricks with filtering widgets

Data Factory:

- in Azure create a resource group

- create a token in Databricks – add it in the vault and grant access to it

- link the Databricks and Data Lake and publish the changes

- create a pipeline using the Databricks notebooks and set the trigger for the pipeline