Data Engineering

* + 1. Introduction to google cloud platform
* Google cloud platform is a cloud computing services offered by google
* Includes range of hosted services for compute, storage and application development that run on google hardware
* Same hardware on which google runs its services
* There are different categories for this, we will be mainly focusing on big data and storage database.
* How to navigate into the space of google cloud platform?
* Google cloud platform mainly works into form of projects
* Create a new project or use existing projects
* Storage >>>>>>> cloud storage
* Create bucket, it is similar as AWS S3

1.2.1 Introduction to Docker

* Data pipeline gets data and produces more data, it can have python scripts with some csv files or data
* Host computer (Windows)

On this we run the containers (**Data pipeline**, python 3.9, pandas, write the results to postgres connection library

* We can have multiple containers and a host machine
* We run everything in a container, all the containers can run independently without knowing anything about each other.
* We can run pgAdmin as a container and communicate with postgres for sql queries
* Docker image contains all the information that needs to be done and we can run it in a different environment.
* Why should we care about Dockers?
  + - * + Reproducibility
        + Local experiments/tests
        + Integrations tests(CI/CD)
        + Running pipelines on the cloud (AWS Batch, Kubernetes jobs)
        + Spark
        + Serverless (AWS lambda, Google functions)

9.22 minutes done. Check how to install git bash on windows and proceed with the course.