Beam/ Dataflow setup:

- Wiki section of repo -> beam setup guide
- Enable the API: Only one person in the group needs to enable the API
- Service Account: Allows beam to access resources from Big Query and the storage buckets.
- Google Cloud Storage bucket: Name should be globally unique
    - This should be done by each person in the group individually
- A VM instance needs to be created.
    - This should also be done per person
- SSH into the VM
- pwd - tells home directory (present working directory)
- ls -> list all the files in the directory
- ls -la - > shows all the hidden files in the directory

Apache Beam:

- Unique feature of the system (Dataflow) - can process batch and streaming data
- Batch data is bounded while streaming data is continuous
- Three aspects:
    - Pipeline
    - PCollection
    - Transform
- Python SDK will be used in class

Beam Pipeline:

- A DAG
- Takes in PCollections -> transformed using PTransform -> outputs PCollections
- We will be running in batch mode
- PCollections (Parallel collection)
    - a collection of data elements
    - Elements can be made up of strings, integers, arrays as long as the schema is consistent between all the PCollections
    - immutable
- PTransform (Parallel Transform)
    - Transformation of PCollection
    - Element wise: maps elements to one, zero, many output elements
    - Generates a new PCollection (as they are immutable)
    - Properties:
        - Serializable: Converted to byte stream to transfer over the network
        - Parallelizable: Many instances will be running it as subsets of the data will be using it
        - Idempotent: safe to apply multiple times leading to similar results
- ParDo
    - maps input to 1, 0, many elements
    - Formatting, parsing, cleansing the data