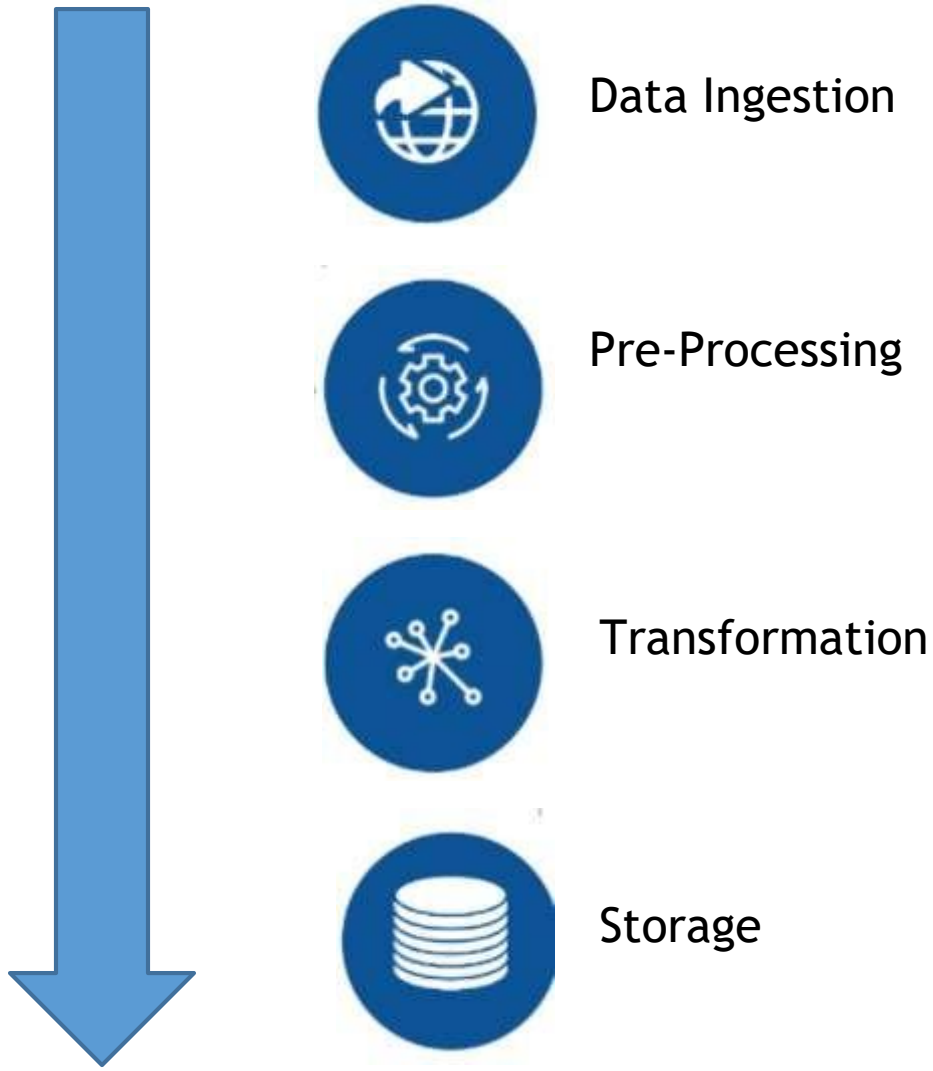
Project Flow


Data Pipeline
Data Ingestion
Pre-Processing
Transformation
Storage

Project Flow :

run_pipeline.py

run_ingest.py

run_pretransform1.py

run_transform2.py

run_transform3.py

HIVE

copy_final_op_to_hive

copy_final_op_to_PGdatabase

copy_final_op_s3_bucket

copy_final_op_azure_blob

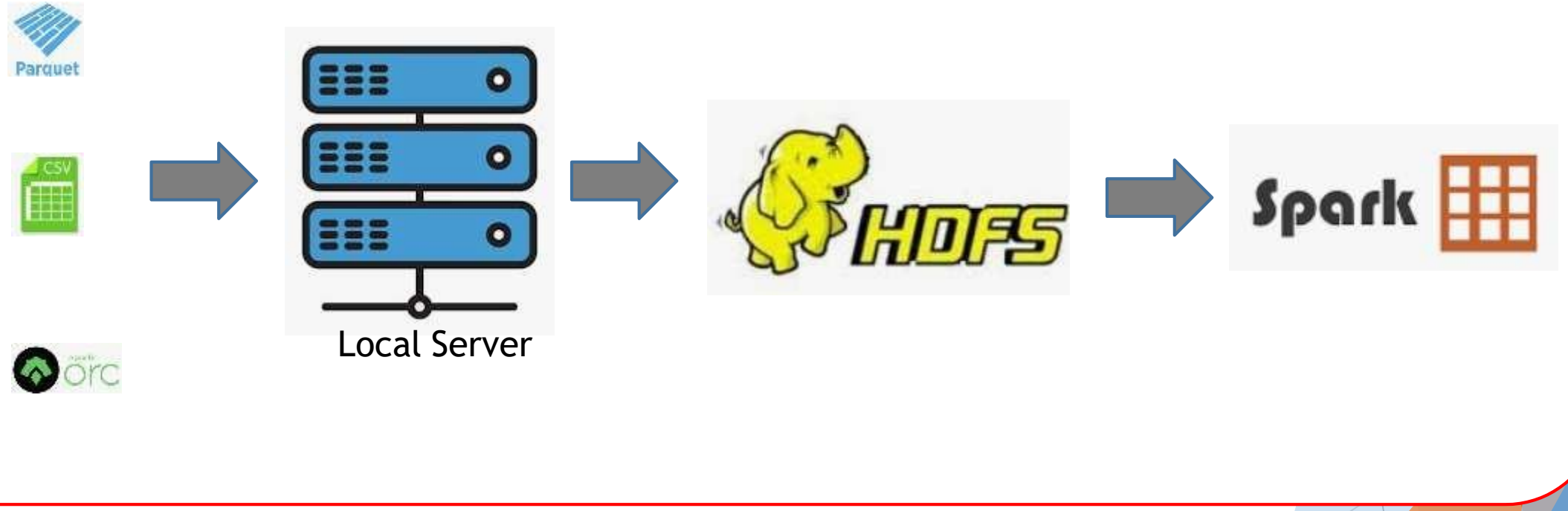
copy_final_op_ftp

amazon S3

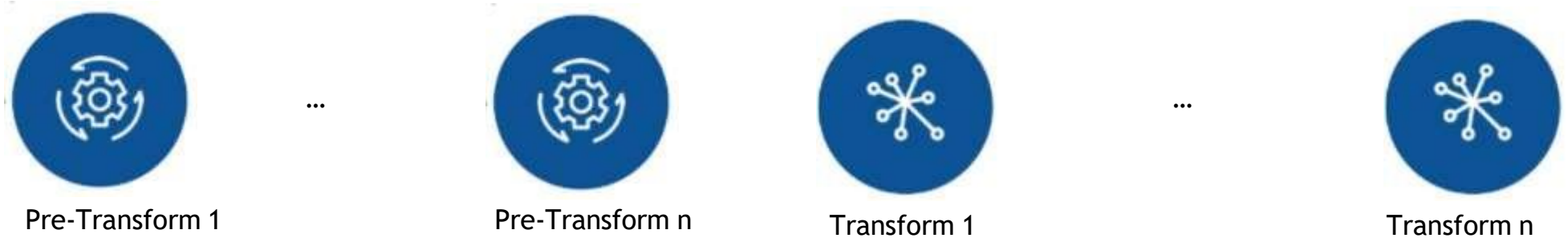
Microsoft Azure Blob Storage

FTP

Data Ingestion:



Transformations



Storage

Transfer

Project Flow (Code Level)

presc_run_pipeline.py

get_all_variables

create_objects

presc_run_data_ingest

presc_run_data_preprocessing

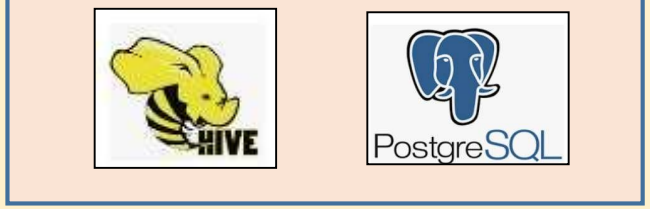
presc_run_data_transform

presc_run_data_extract

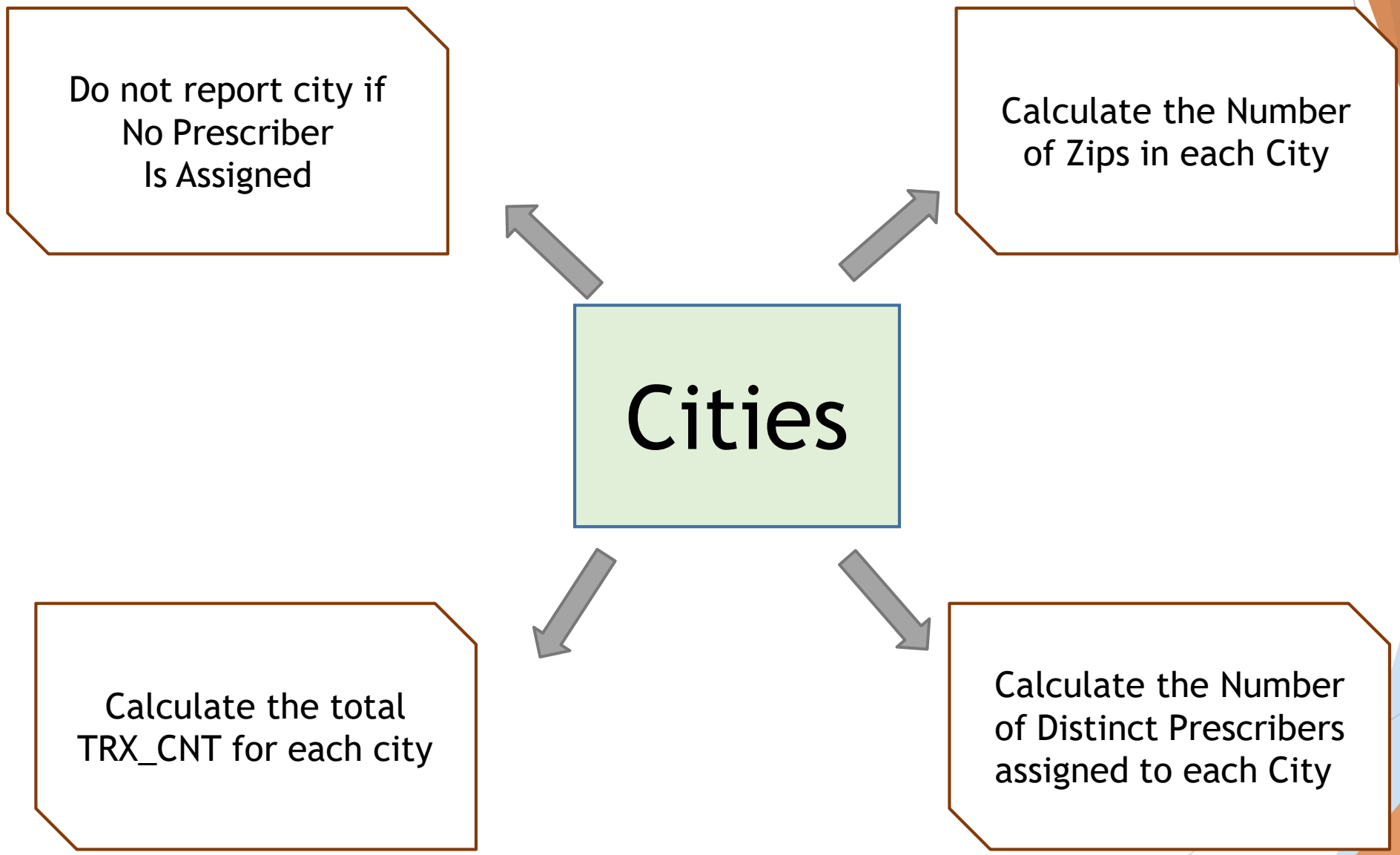
validate

Logging

Error Handling



Copy to Local Server


Do not report city if
No Prescriber
Is Assigned
Calculate the Number
of Zips in each City
Cities
Calculate the total
TRX_CNT for each city
Calculate the Number
of Distinct Prescribers
assigned to each City

Prescriber Report

Apply a filter to consider the prescribers only from 20 to 50 years of experience

Rank the Prescribers based on their TRX_CNT for each state.

Select top 5 prescribers from each state.

Part 1

presc_run_pipeline.py

Copy Input files to HDFS

get_all_variables

create_objects

presc_run_data_ingest

presc_run_data_preprocessing

presc_run_data_transform

presc_run_data_extract

validate

Logging

Error Handling

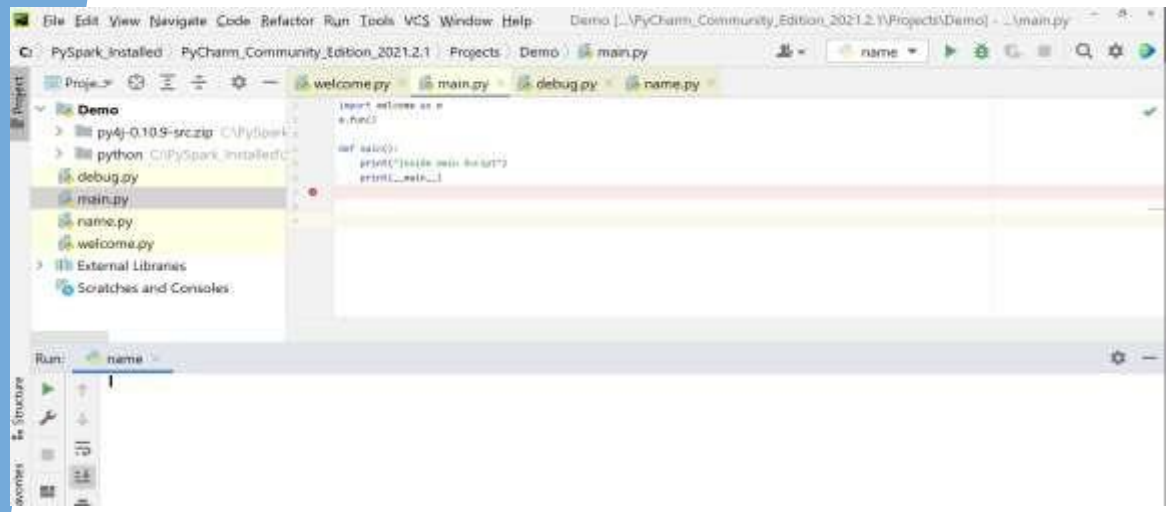
Part 3

Part 2

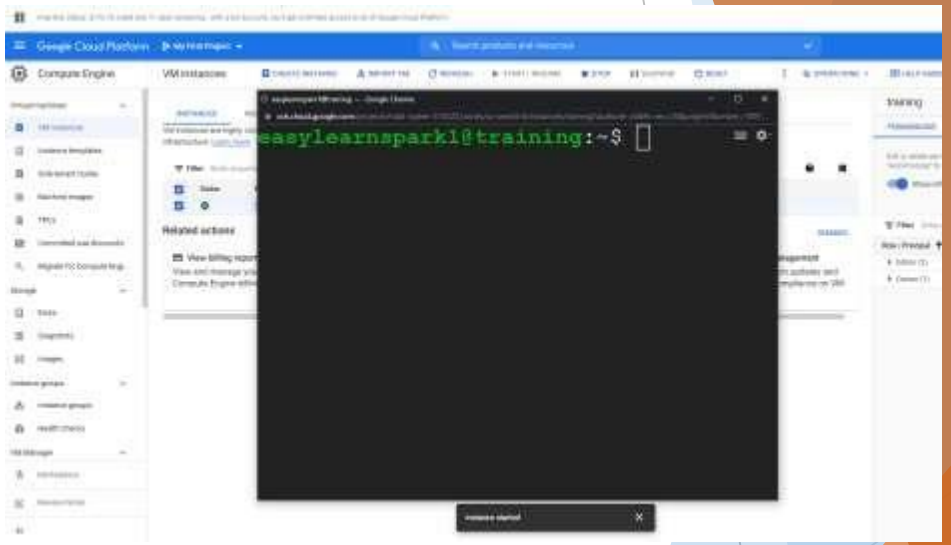
Copy to Local Server

Project Approach

Test and Debug at Windows PyCharm



Deployment at Single Node Cluster



Download City Dimension File at below Link:

https://prescpipeline.blob.core.windows.net/input-vendor-data/city/us_cities_dimension.parquet?st=2022-04-21T14:19:25Z&se=2022-12-31T22:19:25Z&si=read&spr=https&sv=2020-08-04&sr=c&sig=wjY0KtPvyy%2BblpopBqMKAGmmSHsSvLhqL0n%2BBGFVXOQ%3D

Download Prescriber Fact File at below Link:

https://prescpipeline.blob.core.windows.net/input-vendor-data/presc/USA_Presc_Medicare_Data_2021.csv?st=2022-04-21T14:19:25Z&se=2022-12-31T22:19:25Z&si=read&spr=https&sv=2020-08-04&sr=c&sig=wjY0KtPvyy%2BblpopBqMKAGmmSHsSvLhqL0n%2BBGFVXOQ%3D