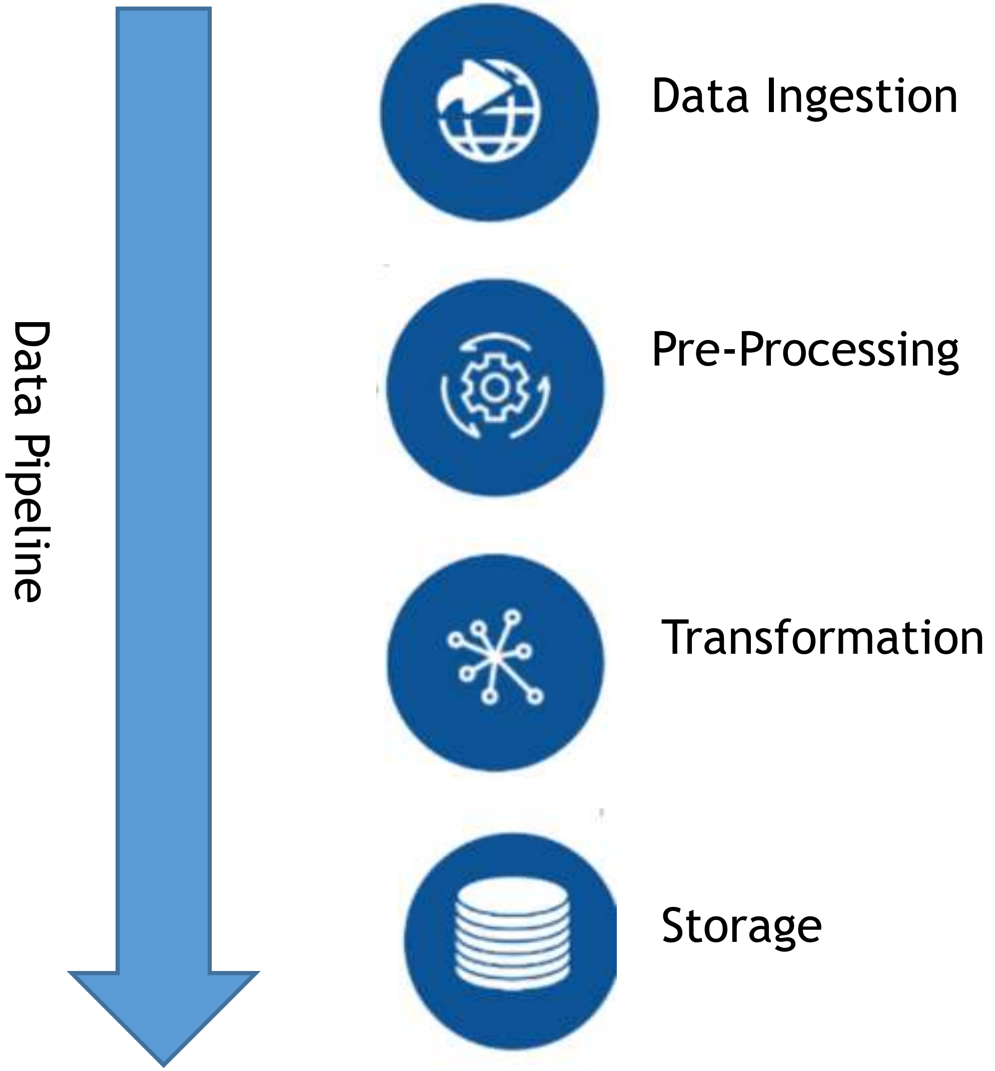
Project Flow



Data Pipeline

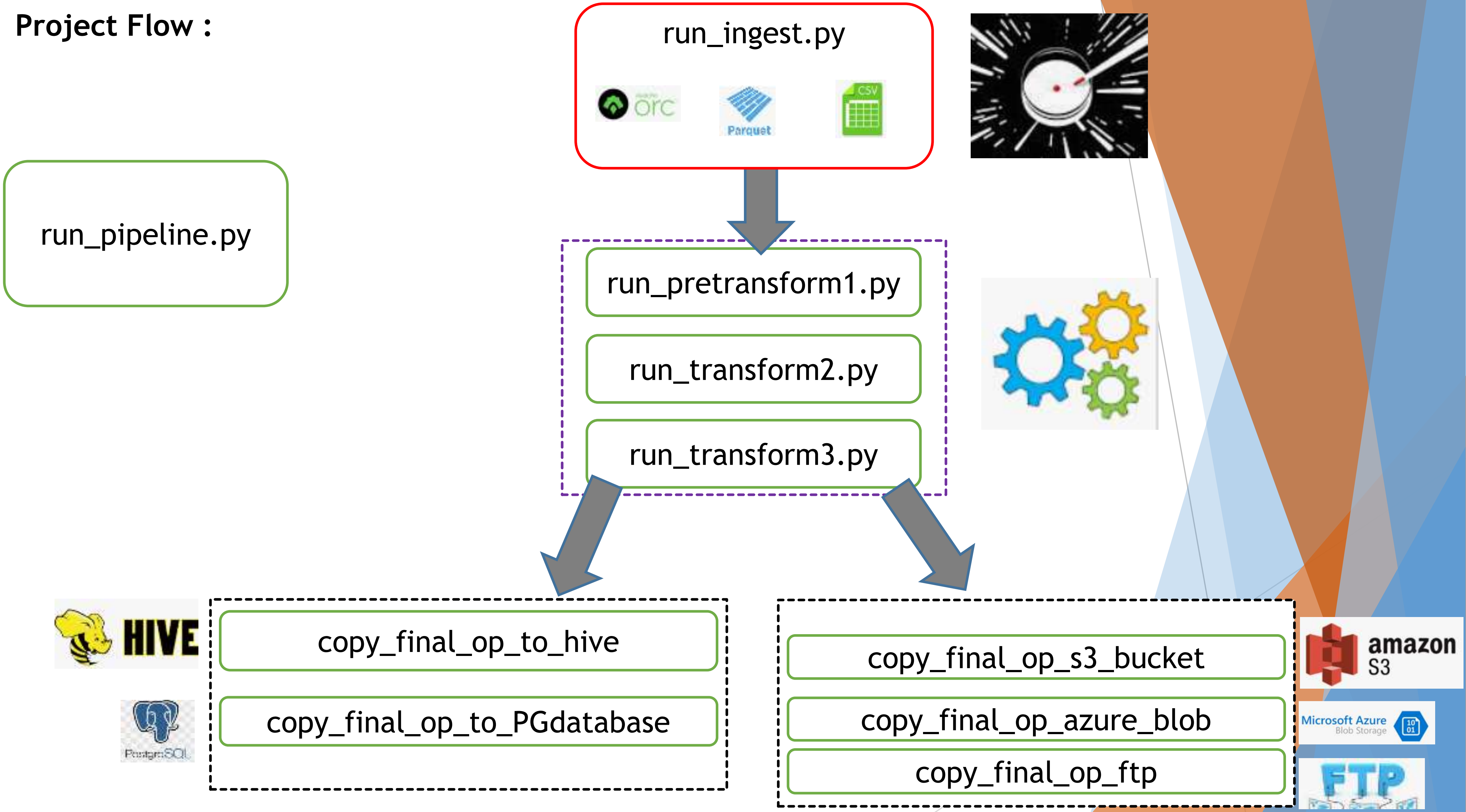
Data Ingestion

Pre-Processing

Transformation

Storage

Project Flow :



run_ingest.py

run_pipeline.py

run_pretransform1.py

run_transform2.py

run_transform3.py

copy_final_op_to_hive

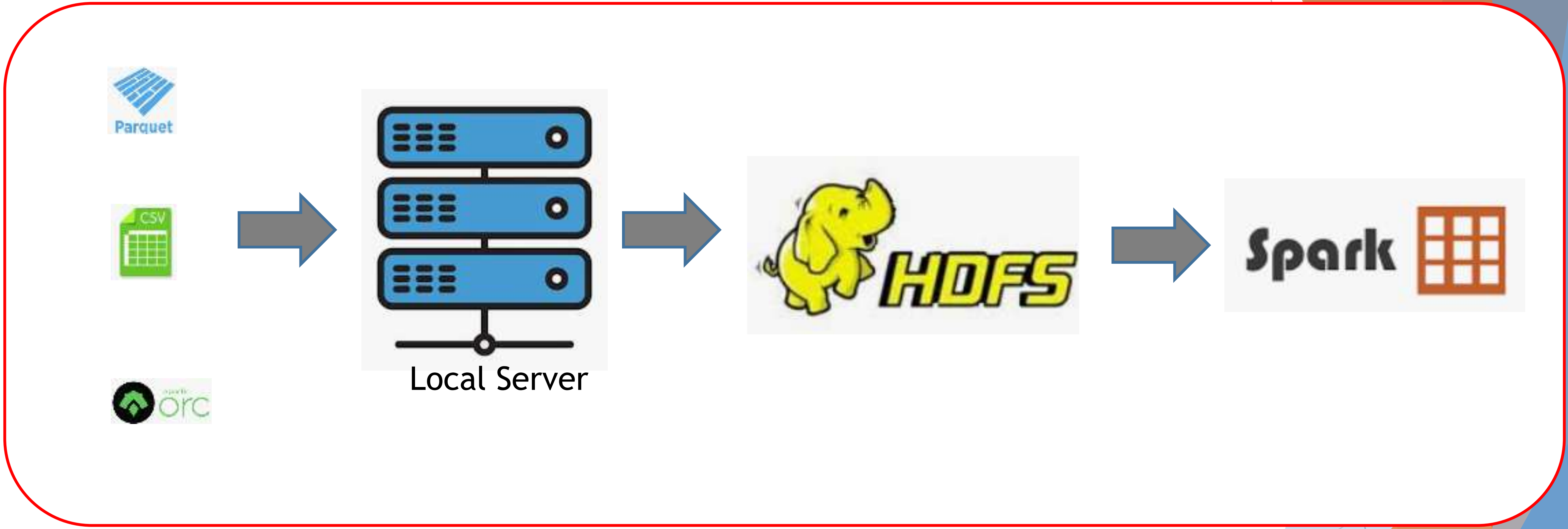
copy_final_op_to_PGdatabase

copy_final_op_s3_bucket

copy_final_op_azure_blob

copy_final_op_ftp

Data Ingestion:



Transformations

Pre-Transform 1 … Pre-Transform n Transform 1 … Transform n

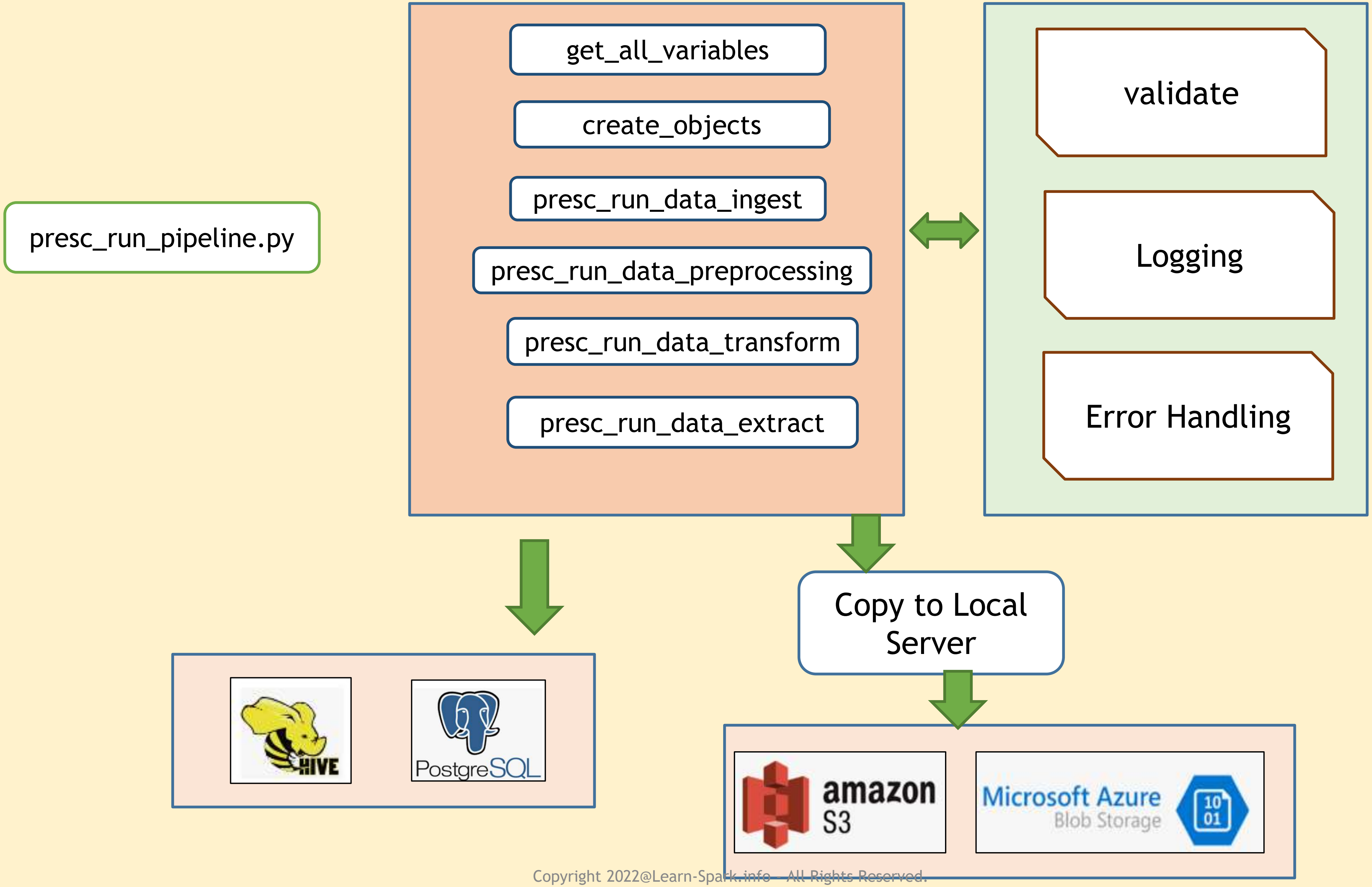
Storage

PostgreSQL

Transfer

Microsoft Azure
Blob Storage
10
01

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

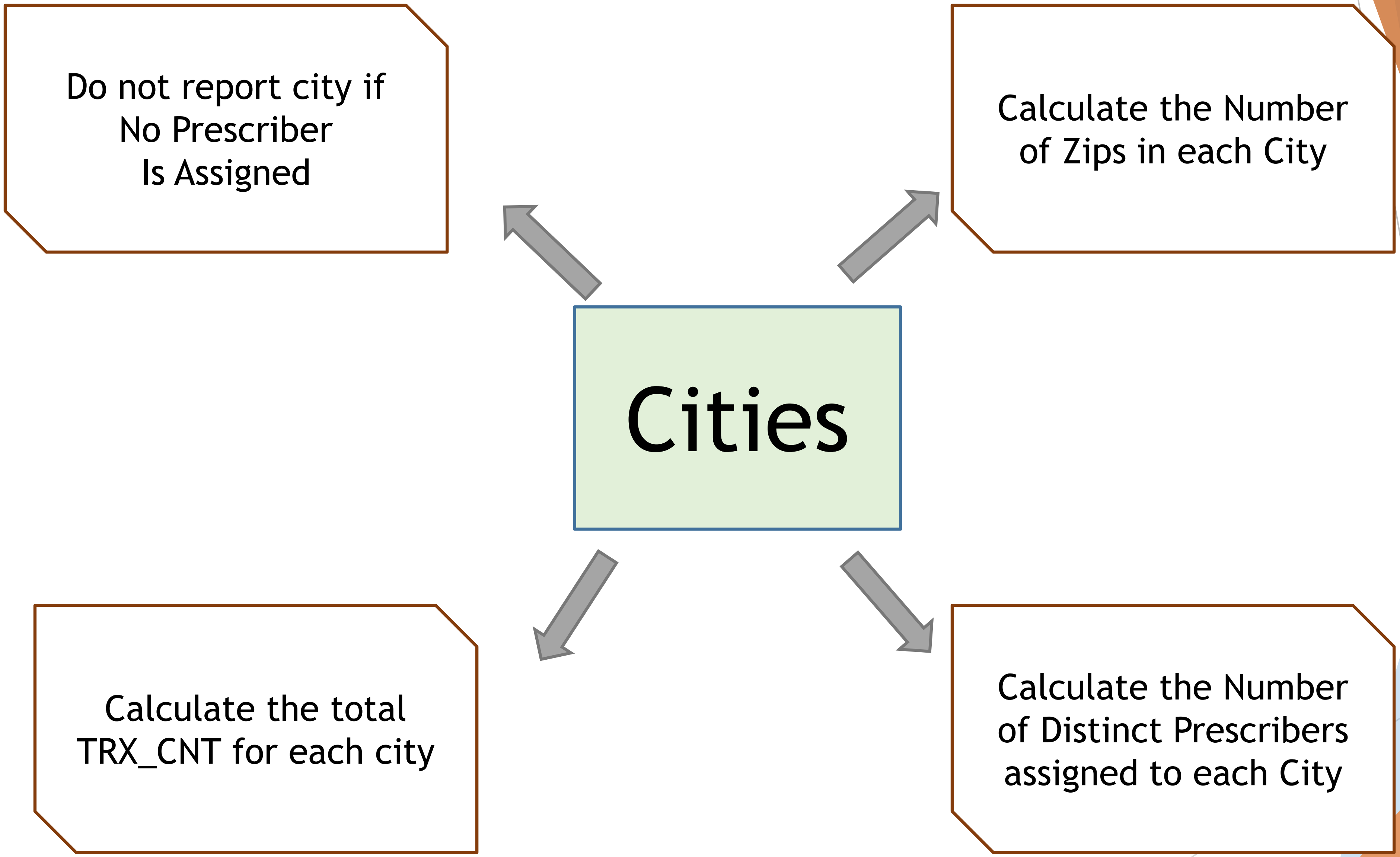
presc_run_data_extract

validate

Logging

Error Handling

Copy to Local Server



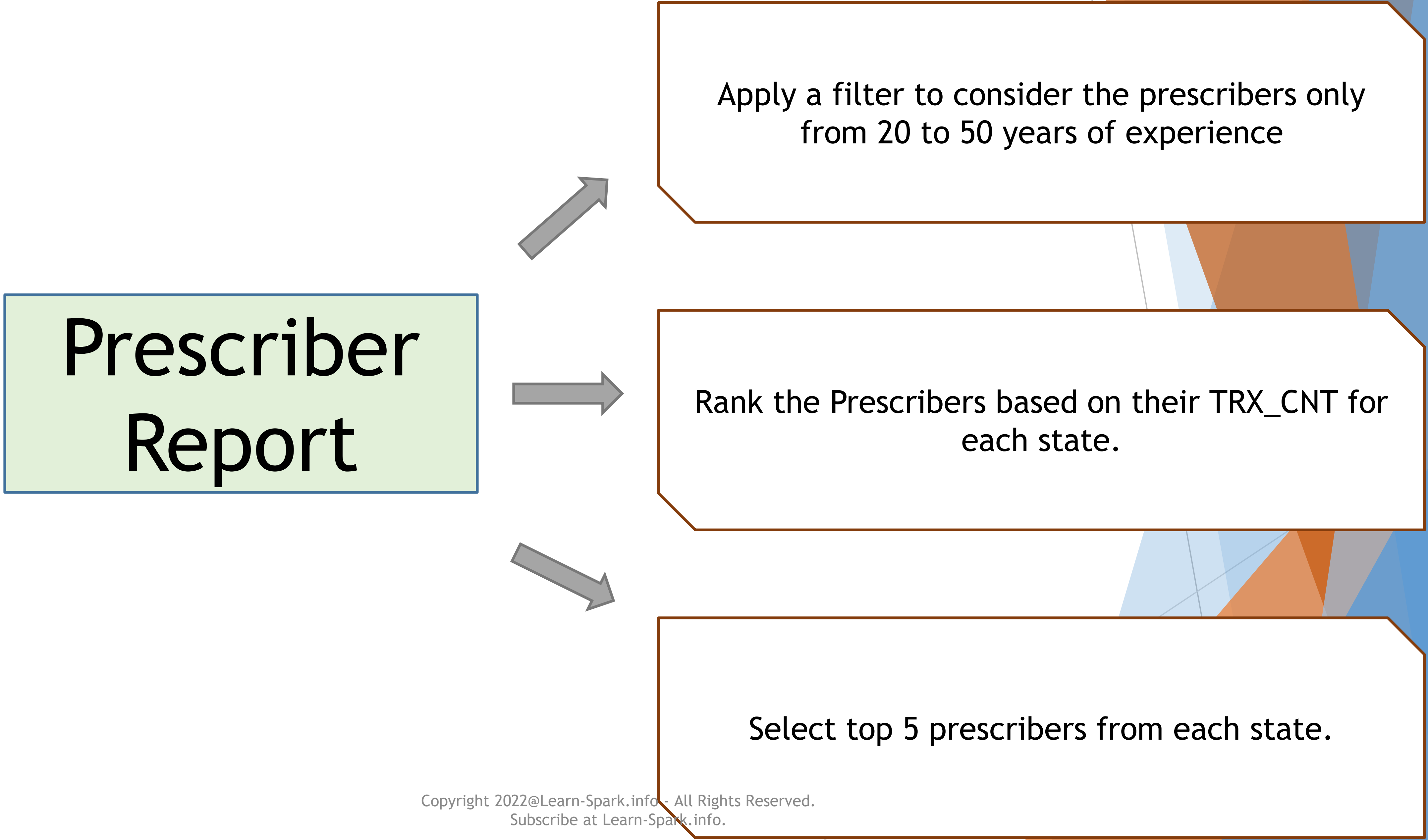
Cities

Do not report city if
No Prescriber
Is Assigned

Calculate the Number
of Zips in each City

Calculate the total
TRX_CNT for each city

Calculate the Number
of Distinct Prescribers
assigned to each City

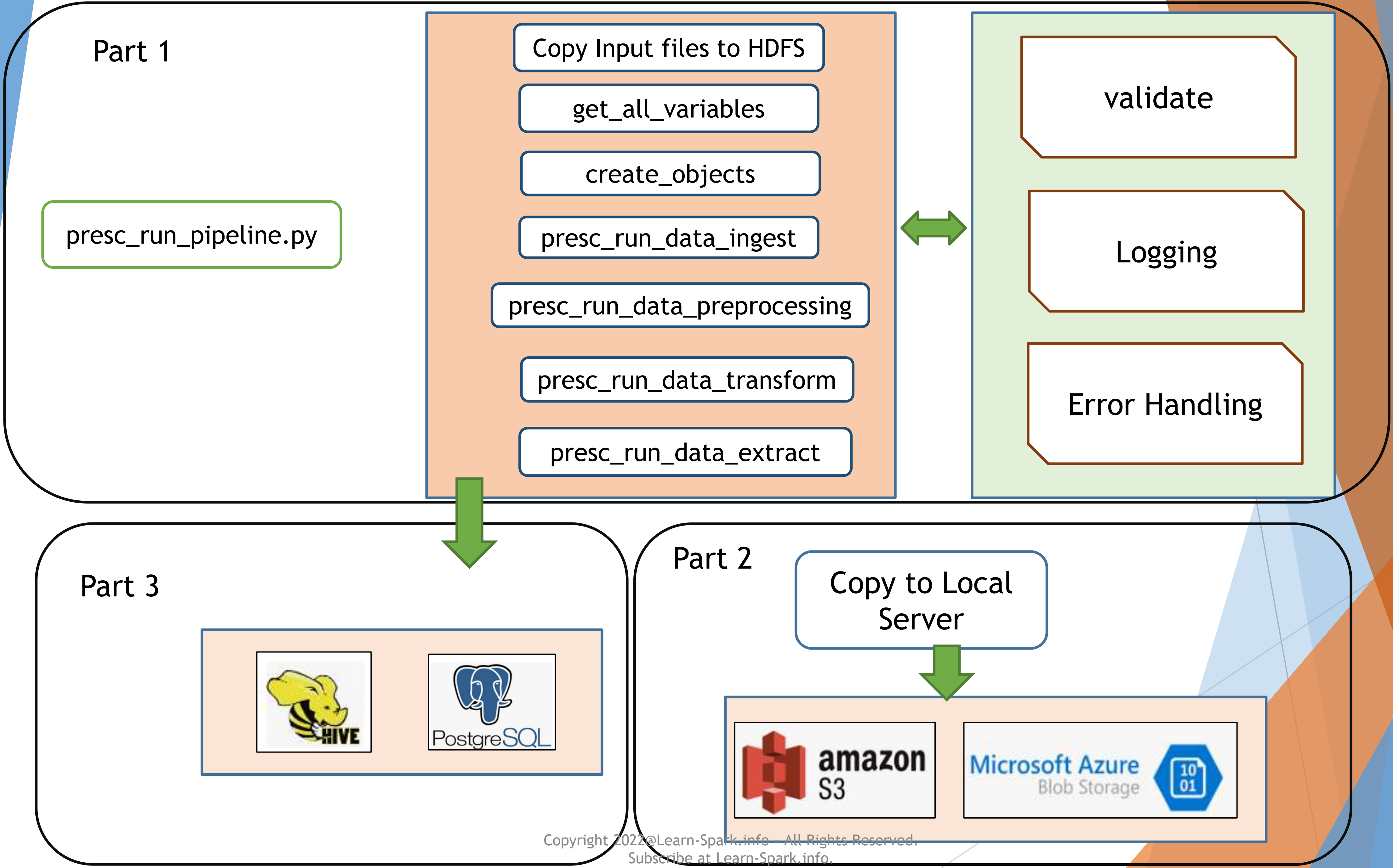


Prescriber Report

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

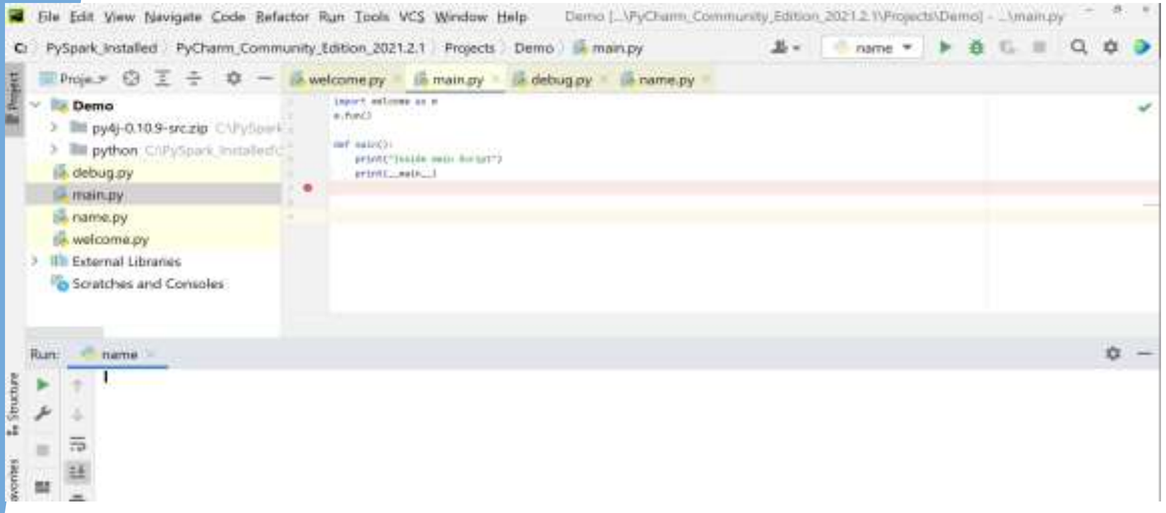
Part 2

Copy to Local Server

Part 3

Project Approach

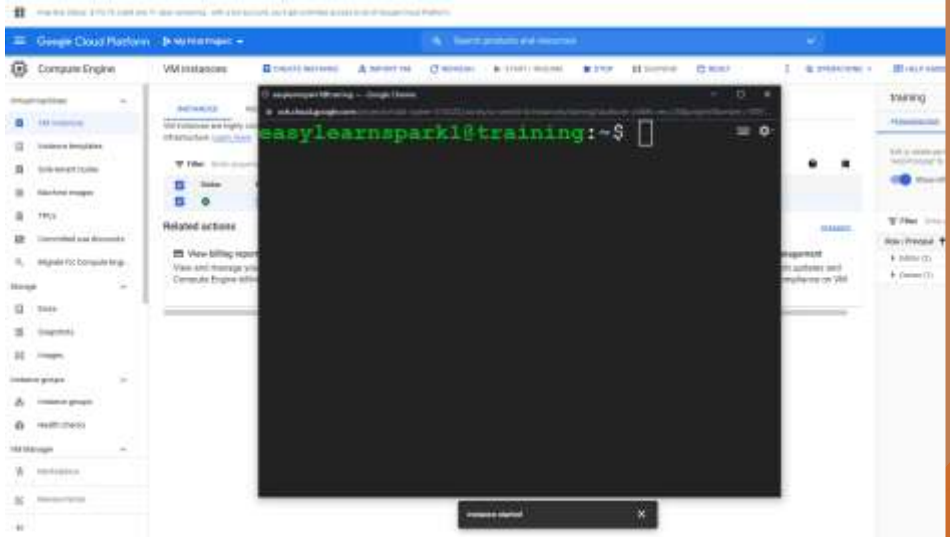
Test and Debug at Windows PyCharm


File Edit View Navigate Code Refactor Run Tools VCS Window Help
Demo [...\PyCharm_Community_Edition_2021.2.1\Projects\Demo] - ...\main.py
PySpark_Installed
PyCharm_Community_Edition_2021.2.1
Projects
Demo
main.py
name
Proje...
welcome.py
main.py
debug.py
name.py
Demo
py4j-0.10.9-src.zip C:\PySpark
python C:\PySpark_Installed\
debug.py
main.py
name.py
welcome.py
External Libraries
Scratches and Consoles
Run:
name
Project
Structure
Favorites

```
import welcome as m
m.fun()

def main():
    print("Inside main Script")
    print(__main__)
```

Deployment at Single Node Cluster



```
easylearnspark1@training:~$
```

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