Xtage DE Assignment

Hello,
My name is Yashraj Jagtap and I once again thank you all for considering me for the Data engineering role at Xtage Technologies.
I have 2+ years of experience, I have been working on many different exciting projects in my current organisation
I have hands-on experience with python, SQL, spark and Azure Databricks.
Having keen interest in data engineering and like problem solving

This power-point presentation provides a brief overview of the solution. With every step briefly explained. Detailed solution can also be found in READ.MD on GitHub but you can reach me anytime for further explanation

---

Yashraj.jagtapcj@gmail.com
GitHub:https://github.com/YASHRAJML/Xtage_DE_Assignment.git

Ingestion

Step 1

Ingestion:

The pipeline starts with various data sources including CSV files, external and internal APIs, and a PostgreSQL database.

I have written a comprehensive python with error handling to import all the necessary component need to proceed

- Data Ingestion: CSV files are stored in S3 buckets
- APIs are accessed through API Gateway
- Lambda functions handle the initial data fetch
- RDS hosts the PostgreSQL databases

Standardization

Step 2

Standardization:

The standardize_data.py file is used to join and integrate all the CSV files into a single dataset. The script integrates and standardizes data from multiple sources (sales, products, transactions, customers, and exchange rates) into a unified format

Data Standardization: AWS Lambda functions process the ingested data

- Data from different sources is merged
- Data is converted into a standardized format

Data Preprocessing

- Step 3

Data Preprocessing

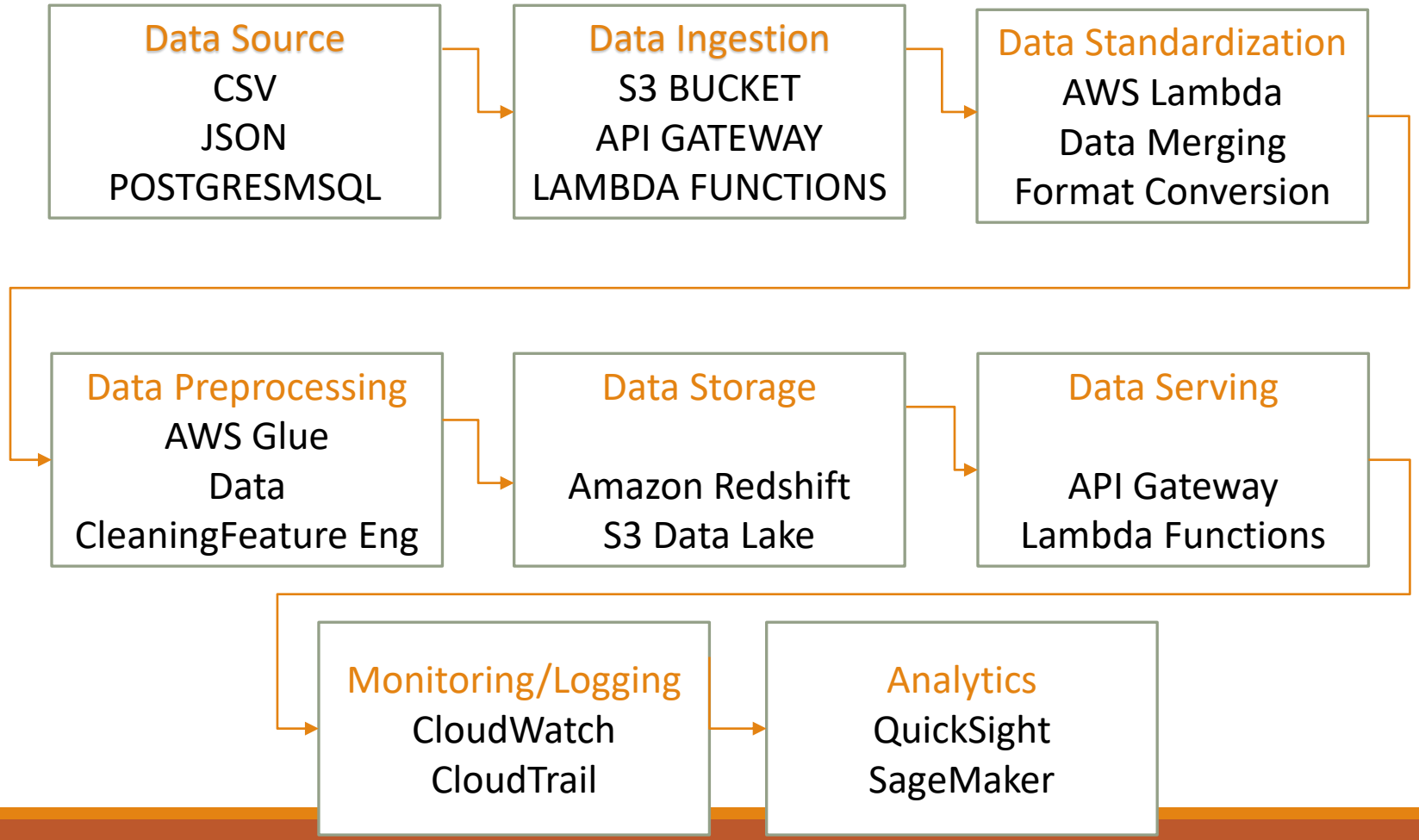
A python code is written to handle all data discrepancy and introduce functional engineering

AWS Glue is used for ETL jobs

This stage includes data cleaning, handling missing values, removing duplicates

Feature engineering is performed here

DIAGRAMATIC FLOW CHART OF AWS PIPELINE



DEPLOYMENT

We have chosen AWS to host and deploy our pipeline due to steady and cost effective nature.

The previous slide gives us the flowchart of how as to the cloud architecture would be designed and implemented

We can create a S3 BUCKET UPLOAD OUR FILES AND USE THE .PY FILES ON GITHUB TO EXCECUTE THE PIPELINE ATERNATIVELY WE CAN ALSO USE AWS INHOUSE SERVICE TO DEPLOY THE PIPELINE

Thanks a lot……