GOOD READS

ETL PIPELINE

JUSTICE OHENE AMOFA

07 **ADAPTABILITY**

Adaptability to other approaches and results.

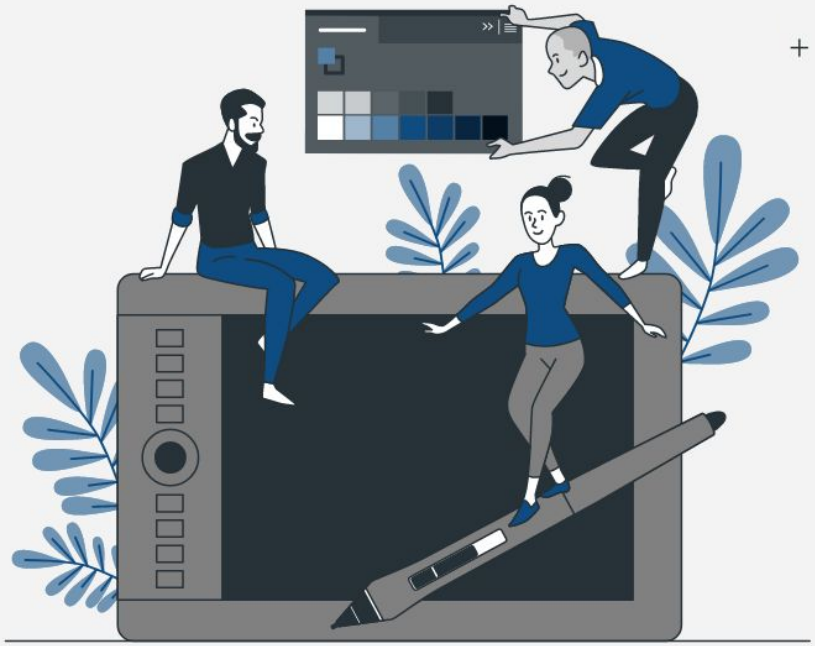
08 **TECHNICAL TOOLS AND TECHNOLOGY**

Tools and mediums used for the project

09 **CLOSING STATEMENT**

10 **REFERENCES**

Project Overview



Goal : To Develop a real-time ETL pipeline for Goodreads data, enabling efficient data capture, storage, transformations, and analysis

Focus : Implementing an automated, scalable solution using Apache Spark, Apache Airflow, AWS S3 and Redshift

My Roles

01

Designing and Implementing the ETL pipeline.

02

Setting up and configuring AWS infrastructure (S3, Redshift, EMR)

03

Writing and scheduling ETL jobs in Apache Airflow.

04

Ensuring Data Quality and Integrity through validation checks.

05

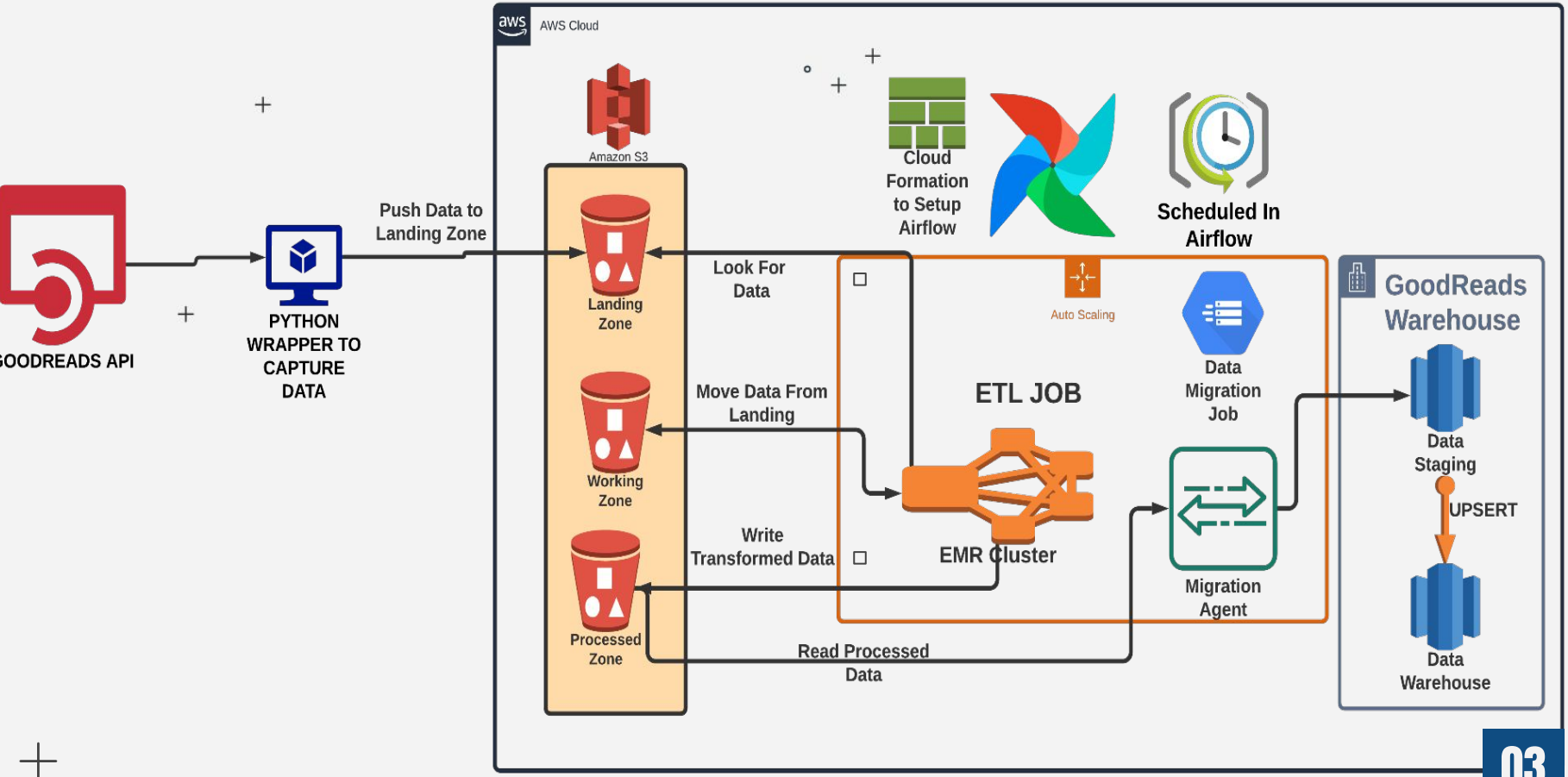
Performing performance tuning and scalability testing.

06

Identify and resolving technical issues within the ETL pipeline.

About the Project Concept

Concept of Solution



Technical Challenges

- Real Time Data Capture from the Goodreads API.
- Efficiently Managing Large Volumes of data in S3.
- Optimizing Sparks Jobs For performance
- Ensuring smooth data transfer between S3, EMR and Redshift.
- Implementing robust data quality checks and error handling.

Solution Approach

- Developed Modular ETL Jobs to handle each stage of data processing.
- Automated data transfer and transformation processes.
- Implemented data quality checks at multiple stages.
- Conducted Scalability testing using synthetic data.

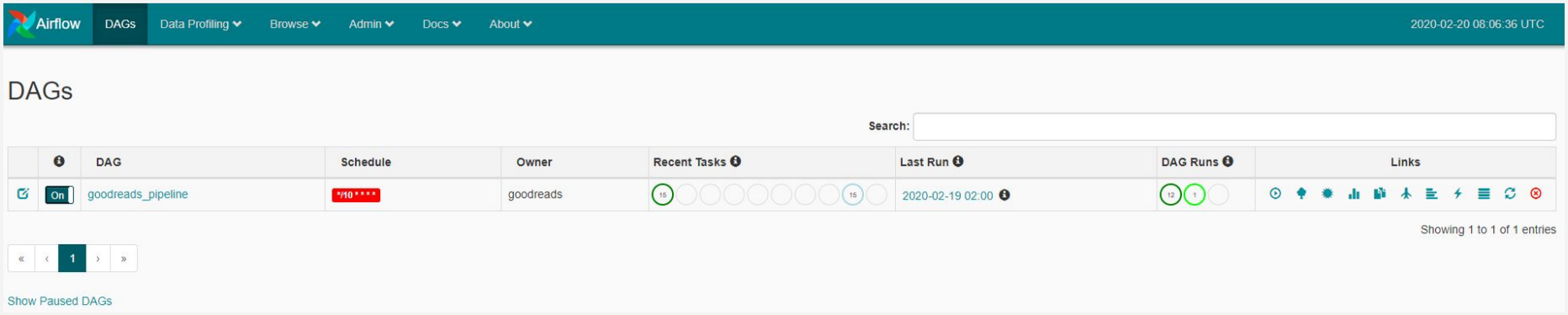
Results and Impacts

1. Successful implementation of real-time ETL pipeline.

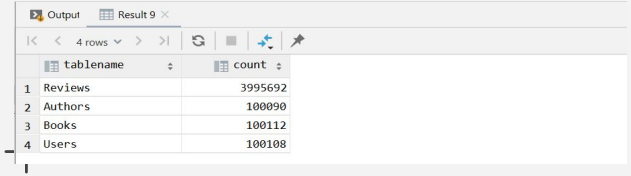
2. Efficient handling of large data volumes (up to 1.6 TB/day).

3. Enhanced data analysis with fast query performance in Redshift.

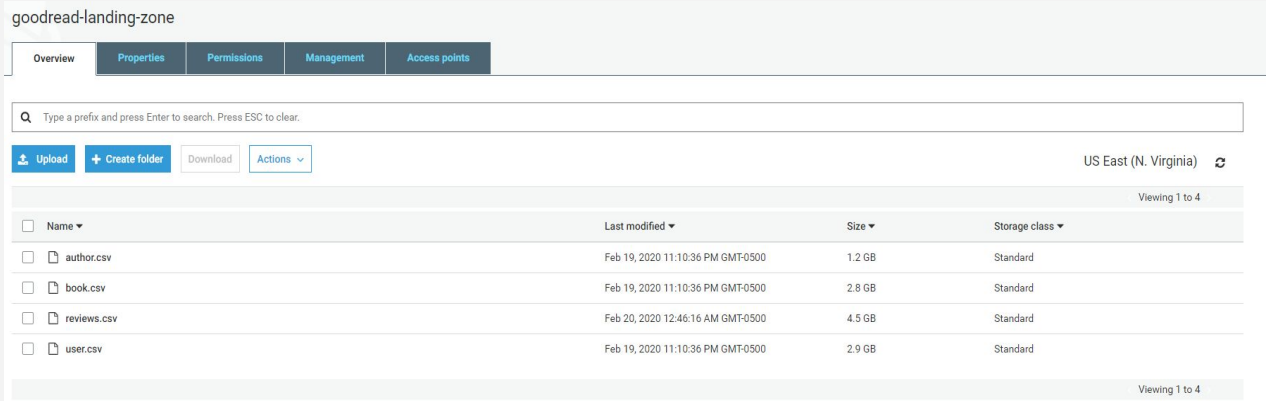
Results



```
select 'Authors' as TableName, count(*) from goodreads_warehouse.authors
union
select 'Reviews' as TableName,count(*) from goodreads_warehouse.reviews
union
select 'Books' as TableName,count(*) from goodreads_warehouse.books
union
select 'Users' as TableName,count(*) from goodreads_warehouse.users;
```



	tablename	count
1	Reviews	3995692
2	Authors	100090
3	Books	100112
4	Users	100108



Lesson Learned

- Importance of thorough planning for the ETL processes.
- Need for robust error handling and data quality checks.
- Value of automation in managing large-scale operations.
- Continuous performance tuning and monitoring are essential for maintaining efficiency.

Technical Tools And Requirements

Goodreads API: Real time Data capture

AWS S3: Scalable and data storage

Apache Spark: Data Transformation and processing.

Apache Airflow: ETL job scheduling and Orchestration.

AWS Redshift: Data Warehousing and Analytics.

Python: Scripting and automation

Boto3: AWS SDK for python used for S3 operations.

Psycopg2: PostgreSQL adapter for python used for Redshift.

Closing Statement

- Demonstrated ability to handle complex data engineering challenges.
- Proven expertise in using a variety of tools and technologies to build scalable, efficient ETL pipelines.
- Effective collaboration and adaptability in evolving project requirements.
- Commitment to ensuring data quality and integrity throughout the process.

Sponsors and Acknowledgements

Implemented by:

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

KFW

Special Thanks To Facilitators and Organizers: Derek Degbedzui ,Wonders Aggor and All other Staffs

THANKS!

Do you have any questions?
Follow the project updates

Joamofa@st.ug.edu.gh

https://github.com/iamamofa/Final_project_Justice_goodreads_etl_pipeline