University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)

TECHNICAL REPORT TEMPLATE

Spring 2023

CONTENTS

Optimizing Blood donation logistics

Executive Summary

In this comprehensive ETL pipeline, data flows seamlessly from the transactional database to the data warehouse database." Our goal is to automate this data movement process and overcome the problems that our customer faces while transferring data from the transactional database to the data warehouse. We hope to expedite operations, improve data integrity, and optimize analytical processes to make more informed decisions by deploying this automated system.

Team Members:

Aryanadh Kommineni
Vivekananda Arekatla
Venkata Raja Varaprasad Sanayila
Rahul Kumar Mavuri

Questions?
Contact : akomm5@unh.newhaven.edu

Technical Report

Optimizing Blood donation logistics

Highlights of Project

Utilizing AWS capabilities to automate data transfer from Data Lake to Data Warehouse.

Submitted on: 04-23-2024

Abstract

This abstract summarizes our project's ETL pipeline design, with a focus on the efficient transfer of data from transactional databases to data warehouse databases. We hope that by automating this process, we can help our client overcome the obstacles of seamlessly migrating data between these different database types. We secure the reliability and correctness of data transfers by implementing modern ETL strategies and technologies such as data validation, transformation algorithms, and scheduling mechanisms, while also improving efficiency and lowering operational overhead. Our solution not only addresses present data transmission challenges, but also creates a scalable architecture that can accommodate future data expansion and changing business needs.

Introductory Section

We are primarily focusing on developing an ETL pipeline using AWS services to make the client more comfortable in their business, hence reducing human-caused errors.

Review of available research

- The major resources that we are using are AWS Lambda, AWS S3, AWS Redshift and PowerBI.
- **Aws S3 -** Amazon S3, also known as Amazon Simple Storage Service, is a service provided by Amazon Web Services that stores objects via a web interface.
- **AWS Lambda** - AWS Lambda is an event-driven, serverless computing technology offered by Amazon as part of its Amazon Web Services.
- **AWS Redshift -** Amazon Redshift is a data warehousing product that is part of the wider cloud computing platform Amazon Web Services.
- **Python 3.10**
- **Pandas -** Pandas is a Python-based software toolkit for data manipulation and analysis.
- **PowerBI-** Microsoft developed Power BI, an interactive data visualization software application aimed primarily at business intelligence.

Methodology

Data Understanding

- Dataset contains the following features
    1. ID
    2. Name
    3. Age
    4. Gender
    5. Race
    6. Date
    7. City
    8. State
    9. Blood Group

Data Preparation

- Remove or redirect records with null values in the age and blood group columns.

The source of this Dataset is Kaggle.

Business Understanding

- The client requires blood donation data and donor information stored in the Data Warehouse.
- Reasons for storing data in a Data Warehouse:
    1. Plan drives based on company requirements.
    2. Analyze the commercial factors of the business.

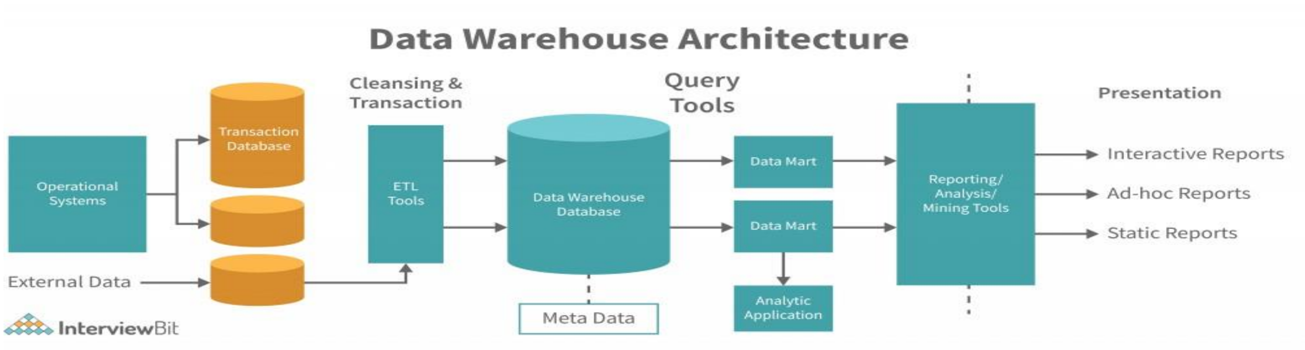
Problem

- Source data is received in the form of csv files, which are then cleansed and stored in the Data Warehouse using ETL tools.
- Disadvantages of this process:

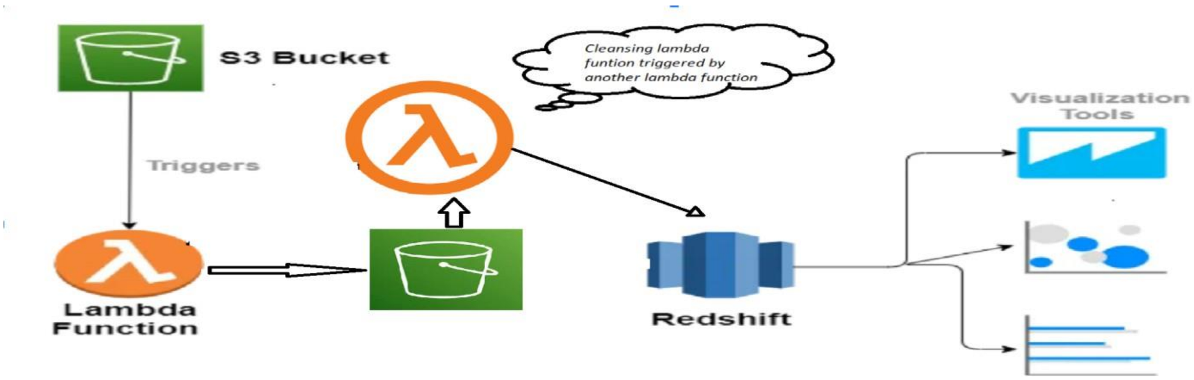
1. Time consuming
2. Lack of resources
3. Beyond the budget estimations

Solution

- Automating the movement of received files from Data Lake to Data Warehouse.

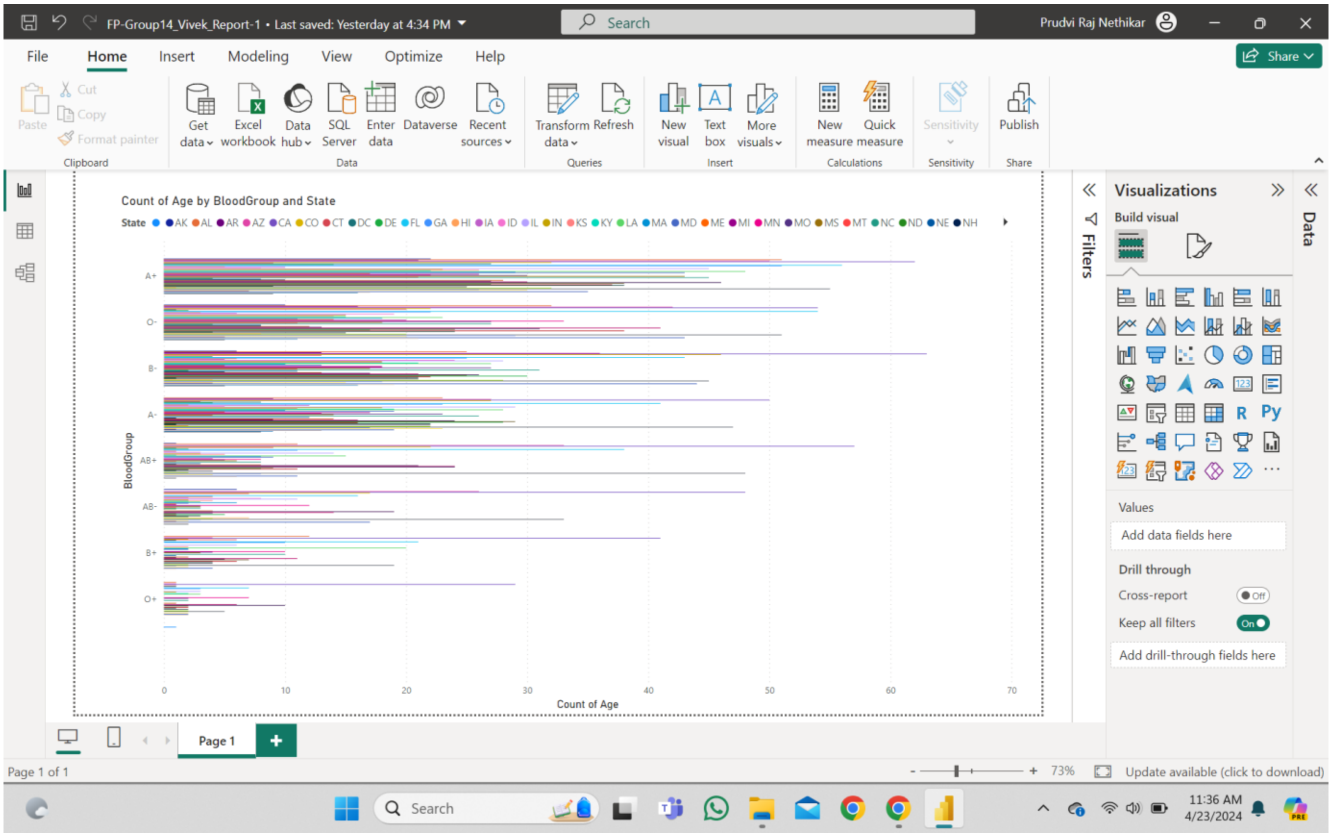


Modelling of Pipeline


S3 Bucket
Triggers
Lambda Function
Cleansing lambda funtion triggered by another lambda function
Redshift
Visualization Tools

Results Section

This is the final example report based on the data that was imported into Redshift.



Conclusion

- We provide data to business teams for report building, reducing the following issues.
  1. Time-consuming.
  2. Limited resources
  3. Excessive budget estimates