Data Warehousing With AWS: Extraction, Transformation, Loading

1. Introduction

This report offers an insight into a data warehousing project undertaken by PayMinute, a prominent FinTech company with operations in Nigeria and Kenya. The primary focus of this project is the Extraction, Transformation, and Loading (ETL) process, utilizing Amazon Web Services (AWS) to optimize data management. Key concepts and pertinent questions pertaining to the project will be addressed in the following sections.

2. Data Integration Process

The core of the project revolves around the Data Integration process, a fundamental part of ETL:

Extraction: Data is extracted from the source system, predominantly a PostgreSQL database in this case.

Transformation: Extracted data undergoes essential transformations to align with the desired data warehousing schema. These transformations may involve data cleansing, aggregation, or restructuring.

Loading: Transformed data finds its new home in the target data storage system, an AWS S3 Data Lake in this project.

3. Project Overview

Business Case: PayMinute FinTech

PayMinute is a thriving FinTech company boasting more than 5000 active users across Nigeria and Kenya. However, its data analysts have encountered issues stemming from growing transaction volumes, including delays and concerns about

data accuracy. In response to these challenges, PayMinute has enlisted the services of a skilled Big Data Engineer.

4. Process Flow

The project follows a meticulously crafted process flow encompassing the following stages:

Fetching Data from PostgreSQL to AWS S3: Data extraction is accomplished from a PostgreSQL database utilizing Pandas and Boto3. Extracted data is then securely stored in an AWS S3 bucket.

Copying Tables from S3 to Redshift (Raw): Tables from the S3 bucket are copied into Amazon Redshift, signifying the "Raw Data" schema.

Executing Create Statements in the Warehouse's Staging Environment: SQL create statements are meticulously executed in the Redshift staging environment, establishing the foundational schema structure.

Transforming the Dataset: Data transformation becomes a pivotal phase, where data is reshaped to align with the predefined Star Schema or other chosen schemas.

Loading Data into the Staging Environment: The transformed data is skillfully loaded into the Redshift staging environment using SQL scripts.

5. Business Case

PayMinute's primary objective is to elevate data retrieval efficiency and enhance data accuracy for its data analysts. Consequently, the organization has decided to

embrace a data warehousing solution. By harnessing AWS services, the project aims to deliver a robust ETL process, thereby optimizing data management.

6. Guidelines

The project adheres to a set of guiding principles:

Data extraction from PostgreSQL to AWS S3 is achieved using Pandas and Boto3.

Tables are systematically copied from the S3 bucket to Amazon Redshift, residing in the "Raw Data" schema.

SQL create statements are meticulously executed within the Redshift staging environment.

Data undergoes transformation to fit the predefined Star Schema or other chosen schemas.

The transformed data is then expertly loaded into the staging environment using SQL scripts.

7. Code Implementation

The provided code snippet serves as a demonstration of the data extraction and loading process, leveraging AWS services and PostgreSQL. This code adheres to the ETL paradigm, ensuring that data is appropriately transformed and loaded into the desired schema, ready for further analysis.

It's essential to note that the code presented here is a partial representation, and a comprehensive project implementation would entail additional components for data transformation, loading into a star schema, and scheduling ETL jobs as needed.

This will be updated soon.