Accelerating Analytics with Data bricks and AWS S3
Solution Design Document

Created by: Anisha Kale
Date: 19-04-2023

Project Overview:

Our project aims to build a data warehouse in Databricks by importing data from AWS S3 and extracting insights. This will enable you to make data-driven business decisions and gain important insight into sales data. The Purpose of the Document is to outlines the proposed solution for the project based on the user stories identified.

User stories

S.no	User Stories
1.	Data collection and Normalization
2.	Connecting AWS S3 with data bricks.

Solution:

The solution of above two will be achieved following AWS services and workforce.

1. Excel: Data engineer will use excel for cleaning the data.
2. Normalization: Manually normalization will be performed.
3. Lucid charts: Used for creation of schema after normalizing.
4. AWS S3: For storing the tables created after normalization.
5. Data bricks: For creation of data warehouse and mounting s3 over it.

Design:

1. Excel: After collecting data from client, data will be cleaned using excel by data engineers. Following things kept in mind while cleaning the data,

    a. No repeated rows in the data
    b. Removal of null values.
    c. Checking for unique value for all the entities.

2. Normalization :- While normalizing the data following things are kept in mind and schema will be created

    a. Avoiding data redundancy: Ensuring that each piece of data is stored only once in the database.
    b. 1 NF: It states that an attribute of table cannot hold multiple values. It must hold only single-valued attribute.
    c. 2 NF: In the second normal form, all non-key attributes are fully functional dependent on the primary key. Identifying the dependencies between the attributes of the tables and ensuring that they are in their simplest form.

d. Maintain consistency: Ensure that the same data is represented consistently throughout the database.
e. Maintain data integrity: Ensure that data constraints are enforced so that the database maintains data integrity.
f. Understand the business requirements: Understand the business requirements and ensure that the normalized database structure meets those requirements.

3. Lucid chart: Above this chart schema will be designed, the type of schema fit will be snowflake type. Below is complete description of fact table and dimension table design on the data.
   a. Fact table : This table include column such as
      1. Sales_id: Sales id given to sales data.
      2. Product_id: Product id given to each product
      3. Year_id: Year id of the years
      4. City_id: City id for 10 given cities.
      5. Month_id: Months of year
      6. Quantity: Quantity of product.
      7. Cost_price: Includes the cost price of data.
      8. Selling_price: Updated selling price based on demands
      9. MRP:- Maximum retail price

   b. Dimension tables : Following will be the dimension tables
      1. Products table:- Include product name and product ID
      2. Year table: Include years for sales data.
      3. Month table: Include month names and month ID.
      4. Cities table: Include city name and IDs.
      5. Subs table: Include sub name of product and there id.
      6. Categories: Include categories of each product and ID.
      7. Supplier table: Include supplier name and supplier ID.
      8. Sub_pro table: Include supplier ID, product ID, ID. This table further have dimension as supplier table which include supplier name and supplier ID.
      9. Pro_sub table: Include product ID, subname ID, and ID.

4. AWS S3: Configuration in s3 includes following things
   a. From AWS console accessing S3 and creating fresh s3 bucket
   b. Configuring s3 and creating its access keys and secret keys.

5. Data Bricks: Task in data bricks would be to mount the s3 bucket over data bricks in a secure way using the access keys and secret keys created. The link to code is attached in document for your reference.
   Repository:
   https://github.com/Aniishak/AAwDS3

Work Flow:

The workflow for above user stories will look like,

1. Data collection and cleaning is performed on the data which will ease the further steps and remove complexities.
2. Schemas are created for the data and normalization.

S3 bucket is mounted over data bricks using access and secret keys so that data analyst can excess the data without any obstacles related to permission or policies.

Security

Security is provided to s3 bucket, Person with access keys and secret keys only can access it. In the code as well we followed best coding practices and keys are not freely visible, they are stored in CSV file type.