High Level Document

Amazon Sales Data Analysis

Revision Number: 1.0

Last date of revision: 16/01/2022

Date Issued	Version	Description	Author
16th Jan 2022	1.0	First Version of Complete HLD	Dipankar Modak, Akash Sahu

Contents

Sales and revenue are an important reflection of the company's progress. These indicators of a company's progress help the company to get a fair idea of their target customers, most profitable products, and most revenue-generating strategies. The aim of the project is to visualize the sales, profit, revenue trends and other important KPIs

Abstract

The Amazon Food Sales report describes the sales of the various range of products in Domestic as well as in international countries. This report may help to take the necessary steps after seeing the key insights and analysis. The Detailed analysis of the Sales of Amazon for the year 2017 to the Year 2019 by using the tableau tool which can showcase key insights of the sales from the Given data.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all the design aspects and define them in detail
- Describe the user interface is implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
  - Security
  - Reliability
  - Maintainability
  - Portability
  - Reusability
  - Application compatibility
  - Resource utilization
  - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

2 General Description

2.1 Product Perspective & Problem Statement

Sales, Revenue, Profit are important indicators of a company's growth. In this project, we display the sales trend along with location wise trends of sales quantity, revenue, profits of various products, we also display the customer type and their highest profit incurring products.

The objective of the project is to perform data visualization techniques to understand the

insight of the data. This project aims to apply various Business Intelligence tools such as Tableau online to get a visual understanding of the data.
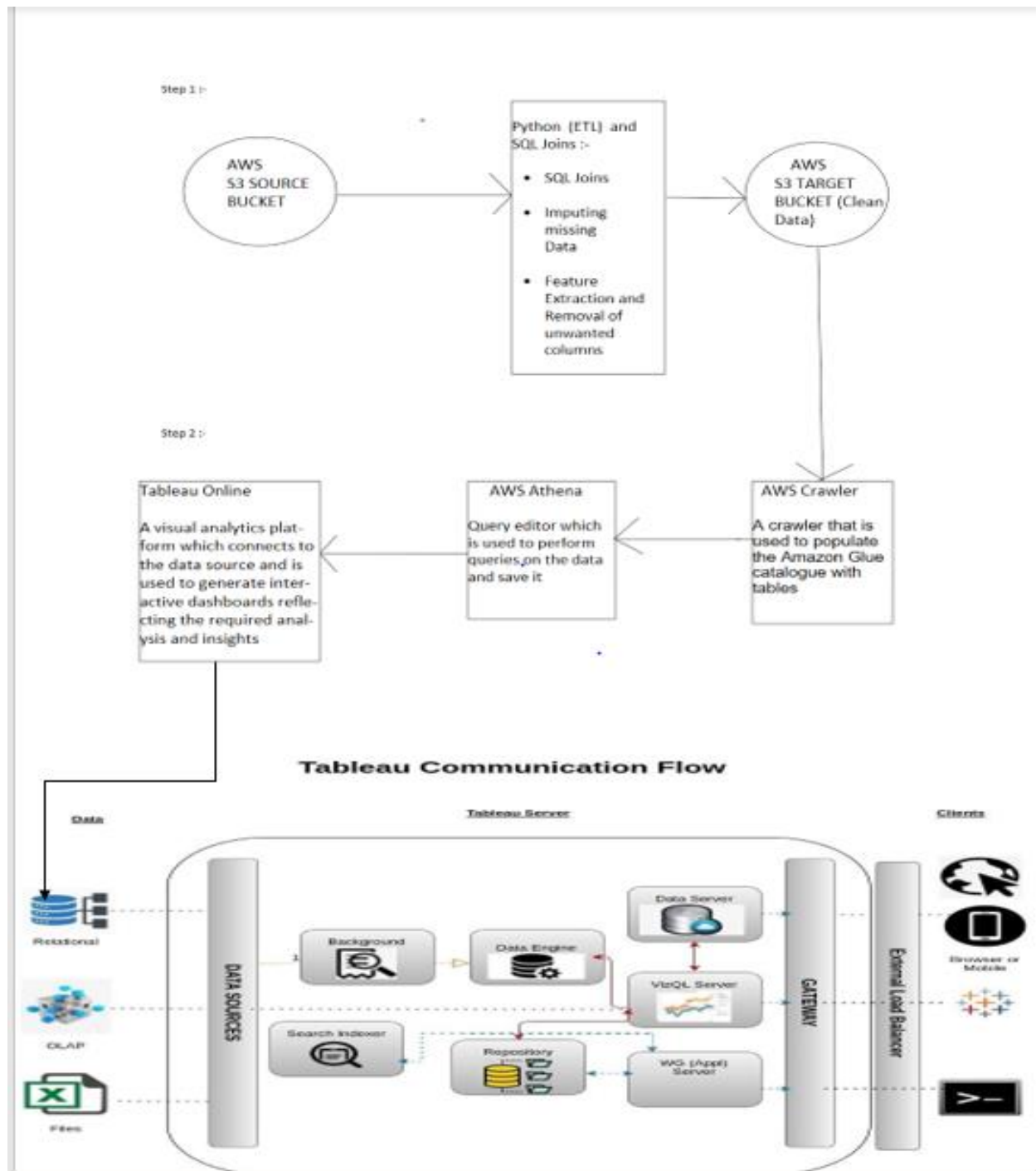
2.2 Tools used

Business Intelligence tools and libraries work such as NumPy, Pandas, Python, Tableau online, AWS glue and AWS Athena are used to build the whole framework.

Feature Engine

3 Design Details

3.1 System Architecture

SYSTEM ARCHITECTURE

3.2 Tools used description:

Python: - A scripting language used for data cleaning, feature extraction and other ETL operations

S3 bucket: - A public cloud storage resource available in the Amazon web services.

AWS crawler: - A crawler that is used to populate the Amazon Glue catalogue with tables. It can crawl multiple data stores in a single run.

AWS Athena: - It is a service that enables data analysts to perform interactive queries in the web-based cloud storage service. It is used with large scale datasets.

Tableau: - Tableau is a visual analytics platform used in the business intelligence industry.   Data analytics and creating interactive dashboards become faster-using tableau. Tableau connects to all kinds of resources like excel, JSON, databases. Tableau services can be accessed through tableau online, tableau server, tableau public as per the difficulty of the task.

3.3 Optimization

Your data strategy drives performance

- Minimize the number of fields
- Minimize the number of records
- Optimize extracts to speed up future queries by materializing calculations, removing columns and the use of accelerated views

Reduce the marks (data points) in your view

- Practice guided analytics. There's no need to fit everything you plan to show in a single view. Compile related views and connect them with action filters to travel from overview to highly granular views at the speed of thought.
- Remove unneeded dimensions from the detail shelf.
- Explore. Try displaying your data in different types of views.

Limit your filters by number and type

- Reduce the number of filters in use. Excessive filters on a view will create a more complex query, which takes longer to return results. Double-check your filters and remove any that aren't necessary.

- Use an include filter. Exclude filters load the entire domain of a dimension, while include filters do not. An include filter runs much faster than an exclude filter, especially for dimensions with many members.

- Use a continuous date filter. Continuous date filters (relative and range-of-date filters) can take advantage of the indexing properties in your database and are faster than discrete date filters.

- Use Boolean or numeric filters. Computers process integers and Booleans (t/f)

much faster than strings.

Use [parameters](#) and [action filters](#). These reduce the query load (and work across data sources).

Optimize and materialize your calculations

- ⬚ Perform calculations in the database
- ⬚ Reduce the number of nested calculations.

- ⬚ Reduce the granularity of LOD or table calculations in the view. The more granular the calculation, the longer it takes.
  - o LODs - Look at the number of unique dimension members in the calculation.
  - o Table Calculations - the more marks in the view, the longer it will take to calculate.

- ⬚ Where possible, use MIN or MAX instead of AVG. AVG requires more processing than MIN or MAX. Often rows will be duplicated and display the same result with MIN, MAX, or AVG.

- ⬚ Make groups with calculations. Like include filters, calculated groups load only named members of the domain, whereas Tableau's group function loads the entire domain.

- ⬚ Use Booleans or numeric calculations instead of string calculations. Computers can process integers and Booleans (t/f) much faster than strings. Boolean>Int>Float>Date>Date Time>String

4 KPIs and Charts

Dashboards will be implemented to display and indicate certain KPIs and relevant indicators for the sales trends, location wise sales quantity, revenue, profit, most revenue, and profit incurring products per customer type.

4.1 KPIs (Key Performance Indicators)

Key indicators displaying a summary of the sales Price and revenue its relationship with different metrics

1. Sales quantity by month and year
2. Revenue trend by month and year
3. Profit margin by month and year
4. Cost of goods sold by the month and year
5. Top 10 products by profit margin, revenue for each customer segment

4.3 Charts

Charts displaying an understanding of Amazon Food Sales Data

1. Revenue by month, year, year-month (Line Chart)
2. Profit Margin % by month, year, year-month (Line Chart)
3. Sales Quantity by month, year, year-month (Line Chart)
4. Cost of Goods Sold by month, year, year-month (Line Chart)
5. Revenue by Location (Map Chart)
6. Profit Margin % by Location (Map Chart)
7. Sales Quantity by Location (Map Chart)
8. Cost of Goods Sold by Location (Map Chart)
9. Customer Type vs Top 10 items based on revenue (Bar chart)
10. Customer Type vs Top 10 items based on Profit Margin % (Bar chart)

5 Deployment

Prioritizing data and analytics couldn't come at a better time. Your company, no matter what size, is already collecting data and most likely analyzing just a portion of it to solve business problems, gain competitive advantages, and drive enterprise transformation. With the explosive growth of enterprise data, database technologies, and the high demand for analytical skills, today's most effective IT organizations have shifted their focus to enabling self-service by deploying and operating Tableau at scale, as well as organizing, orchestrating, and unifying disparate sources of data for business users and experts alike to author and consume content.

Amazon Food Sales data has been cleaned using ETL Pipeline created using Python and analyzed on Tableau which will provide a better key insight for the data and tell you a better

story of the raw data. The Tableau report is published on workspace where you can play with the data for necessary insights.