High Level Design (HLD)

Analyzing world's best wine reviews dataset

Revision Number: 1.0
Last date of revision: 10/11/2022

Harshvardhan
Kushwaha

Document Version Control

Date Issued	Version	Description	Author
10th November 2022	1.0	First Version of Complete HLD	Harshvardhan Kushwaha

Contents

Document Version Control 1
Abstract 4
1 Introduction 5
1.1 Why this High-Level Design Document? 5
1.2 Scope 5
2 General Description 5
2.1 Product Perspective & Problem Statement 5
2.2 Tools used 6
3 Design Details 6
3.1 Functional Architecture 6
3.2 Optimization 7
4 KPIs 8
4.1 KPIs (Key Performance Indicators) 9

Abstract

While the wine industry generally has been very proactive about dealing with climate change, from capturing fermentation carbon to trialling new varieties, this is ultimately a political problem.

In the world of rising new technology and innovation, the Business is also advancing with the role of Data Science and Analytics. Data analysis can help them to understand their business in a quite different manner and helps to improve the quality of the service by identifying the weak areas of the business. This study demonstrates how different analysis help to make better business decisions and help analyse wine reviews, which can lead to new and better products and services. Different analyses were performed on a variety of use cases to get the key insights from this data based on which business decisions will be taken.

The massive dataset of wines reviews with information as the professional description, price, points and so on. The aim of the project is to give an overview of a worldwide wine clues, for example, the relation between price and quality(points), the countries where the wines are made or perform an analysis of the most used words to describe wines. To find an effective solution for this problem you were asked to help in ETL process.

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 of the design aspects and define them in detail
- Describe the user interface being 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

While the wine industry generally has been very proactive about dealing with climate change, from capturing fermentation carbon to trialling new varieties, this is ultimately a political problem. To find an effective solution for this problem you were asked to help in ETL process. The objective of the project is to perform data visualization techniques to understand the insight of the data.

2.2 Tools used

Business Intelligence tools like Power BI and cloud platform Snowflake are used to build the whole framework.

3 Design Details

3.1 Functional Architecture

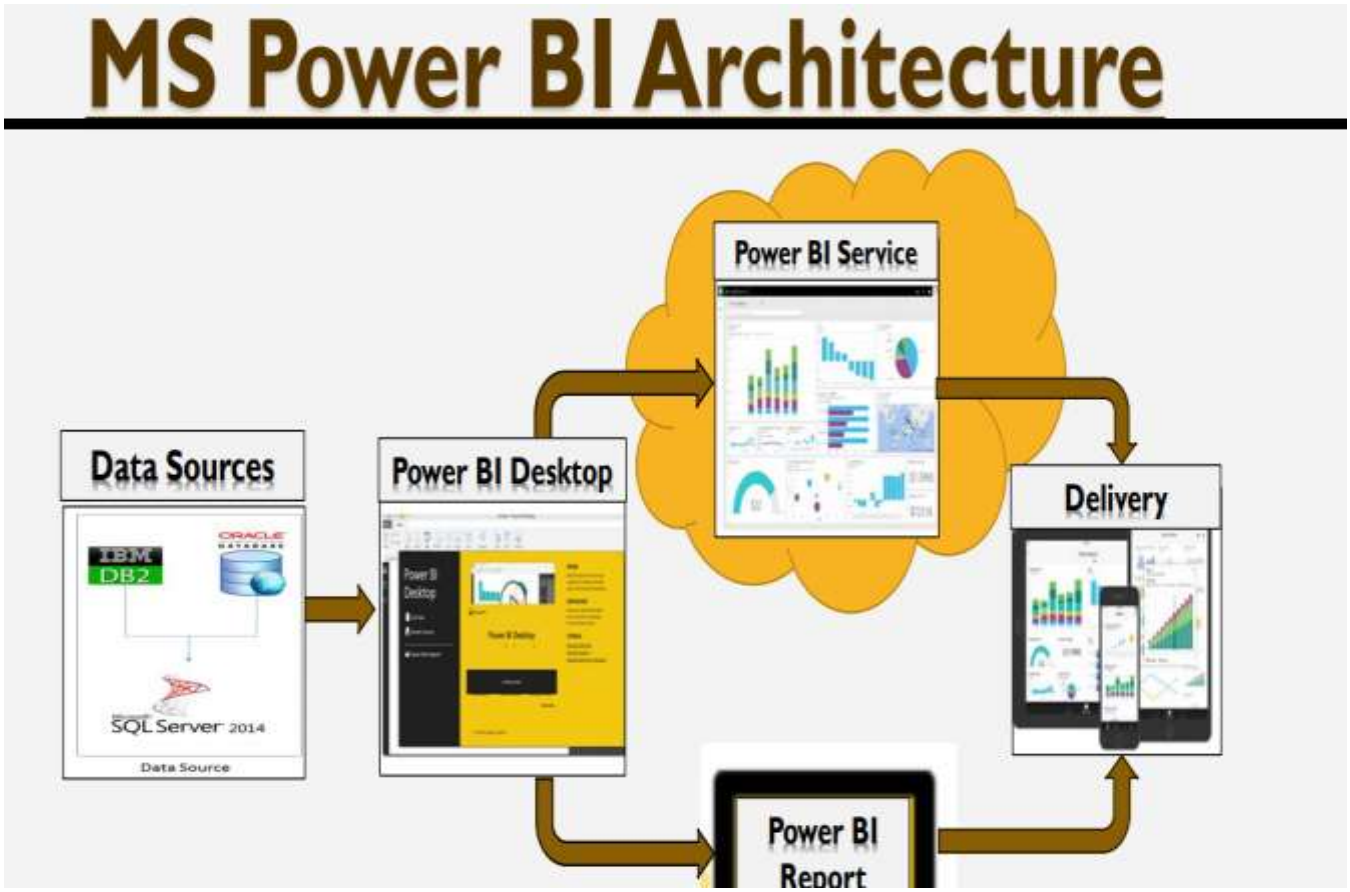
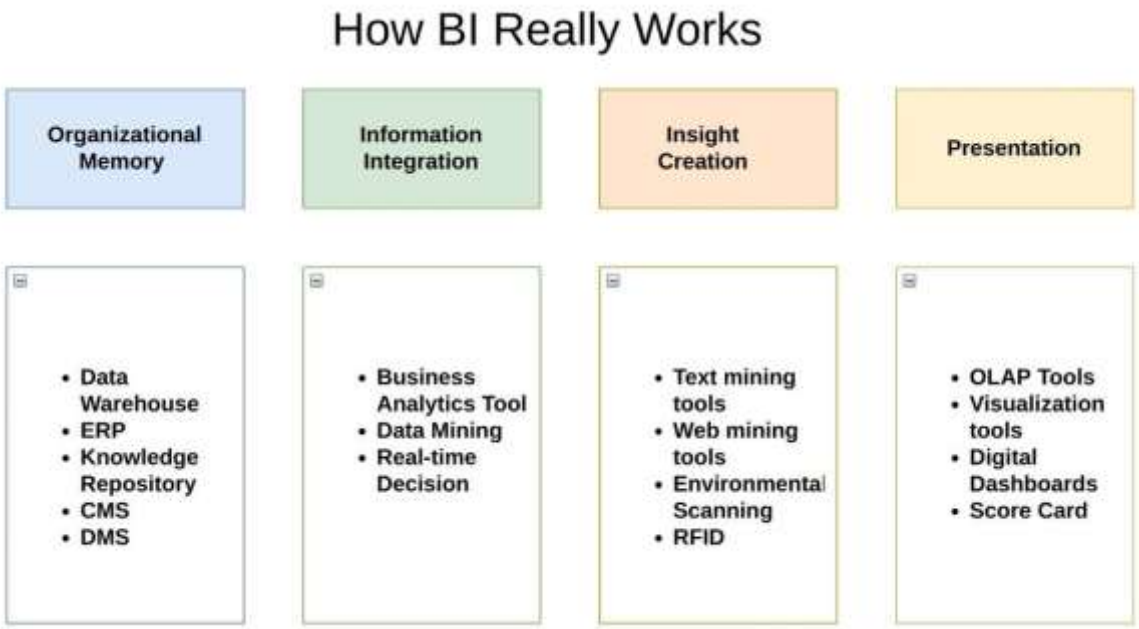


Figure 1: Functional Architecture of Business Intelligence



3.2 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.
  - LODs - Look at the number of unique dimension members in the calculation.
  - 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

Dashboards will be implemented to display and indicate certain KPIs and relevant indicators for the dataset.

As and when, the system starts to capture the historical/periodic data for a user, the dashboards will be included to display charts over time with progress on various indicators or factors

4.1 KPIs (Key Performance Indicators)

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

1. Total no. of Countries.
2. Total no. of Provinces.
3. Total no. of Wineries.
4. Total no. of Varieties of Wines.
5. Average Point and Price of various wines.
6. Top Wines based on price, points and review.