High Level Design (HLD) Black Friday Sales Prediction

Revision Number: 1.0

Last date of revision: 25/09/2021

Soumit Kar

# Contents

[Document Version Control 2](#_TOC_250010)

[Abstract 3](#_TOC_250009)

1. Introduction 5
   1. [Why this High-Level Design Document? 5](#_TOC_250008)
   2. [Scope 5](#_TOC_250007)
2. General Description 6
   1. [Product Perspective & Problem Statement](#_TOC_250006)
   2. [Tools used 6](#_TOC_250005)
3. [Design Details 7](#_TOC_250004)
   1. [Functional Architecture and Model Building 7](#_TOC_250003)
   2. [Optimization 8](#_TOC_250002)
4. [KPIs 9](#_TOC_250001)
   1. [KPIs (Key Performance Indicators) 9](#_TOC_250000)
5. Deployment

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| **26th SEP 2021** | 1.0 | First Version of Complete HLD | Soumit Kar |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Abstract

Black Friday marks the beginning of the Christmas shopping festival across the US. On Black Friday big shopping giants like Amazon, Flipkart, etc. lure customers by offering discounts and deals on different product categories. The product categories range from electronic items, Clothing, kitchen appliances. Research has been carried out to predict sales by various researchers. The analysis of this data serves as a basis to provide discounts on various product items. With the purpose of analyzing and predicting the sales, we have used three models. The dataset Black Friday Sales Dataset available on Kaggle has been used for analysis and prediction purposes. The models used for prediction are linear regression, lasso regression, ridge regression, Decision Tree Regressor, and Random Forest Regressor. Mean Squared Error (MSE) is used as a performance evaluation measure. Random Forest Regressor outperforms the other models with the least MSE score.

## 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

## 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.

## Product Perspective & Problem Statement

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month.

The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

.

## Tools used

Business Intelligence tools and libraries such as **ggplot2, dplyr, corrplot, caTools, rpart, Excel, Power BI** are used to build the whole framework.





# Design Details

## Functional Architecture and Model Building

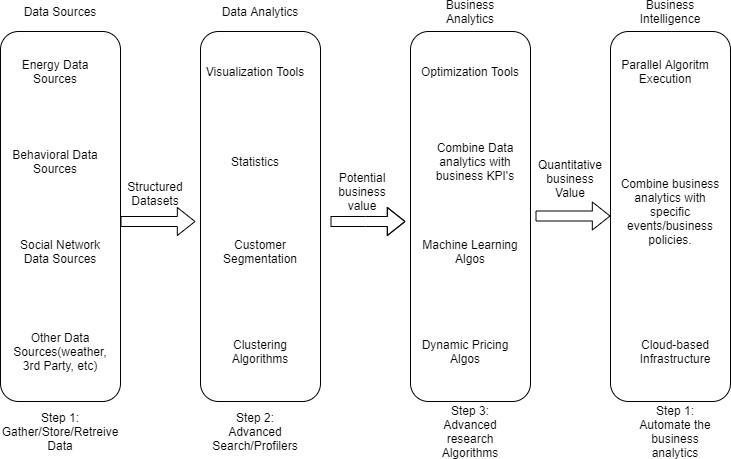
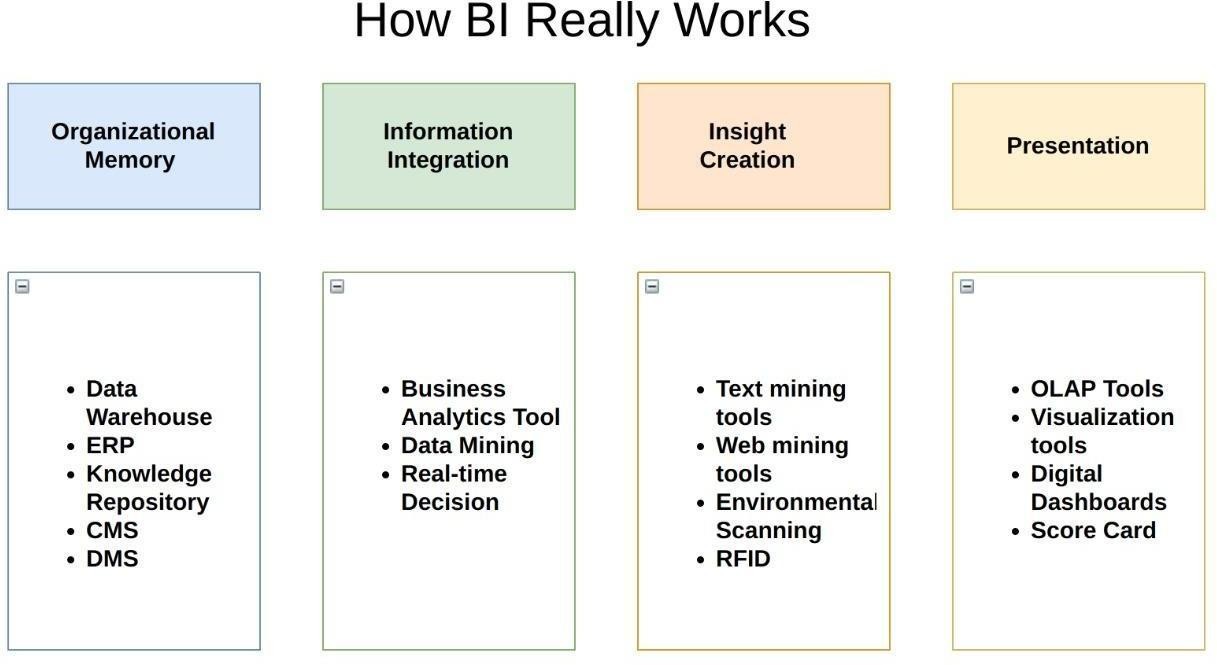
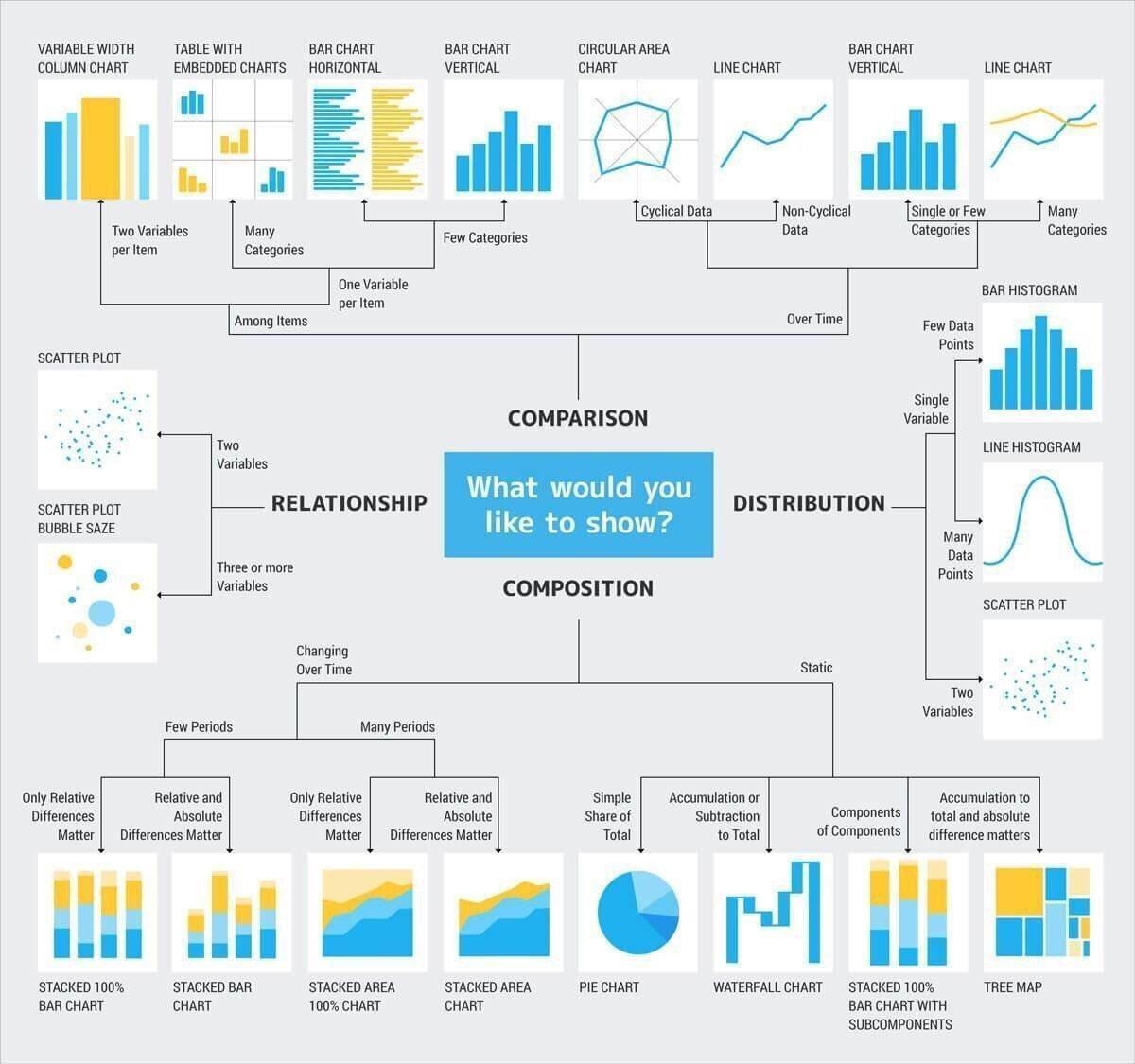
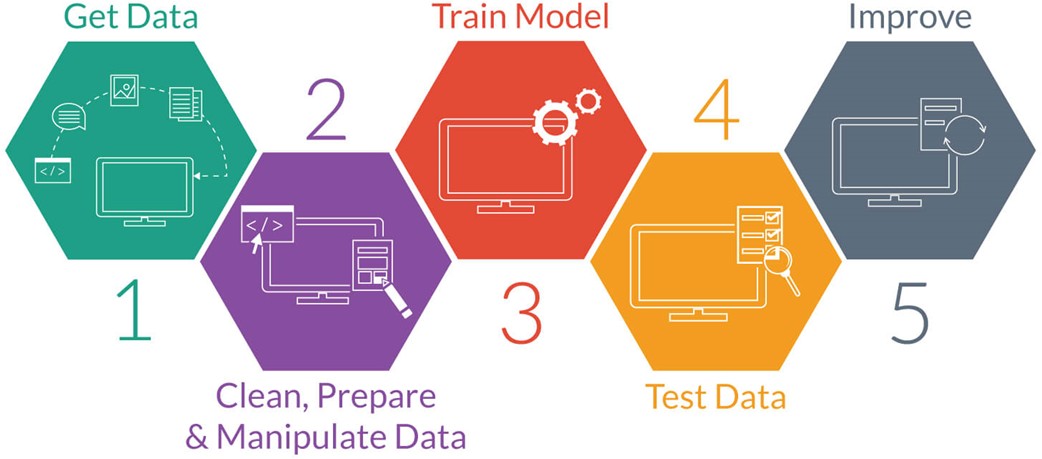


Figure 1: Functional Architecture of Business Intelligence





## Model Building



## 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](http://onlinehelp.tableau.com/current/pro/online/mac/en-us/help.htm#filtering_add_dragfields_dates.html). 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 [parameters](http://onlinehelp.tableau.com/current/pro/online/en-us/help.htm#parameters.html) and [action filters](http://onlinehelp.tableau.com/current/pro/online/en-us/help.htm#actions.html). 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 table calculations in the view. The more granular the calculation, the longer it takes.

o Calculations - the more marks in the view, the longer it will take to calculate.

[Where possible, use MIN or MAX instead of AVG](http://onlinehelp.tableau.com/current/pro/online/windows/en-us/help.htm#calculations_aggregation.html). AVG requires more processing than MIN or MAX. Often rows will be duplicated and display the same result with MIN, MAX, or AVG.

[Use Booleans or numeric calculations instead of string calculations](http://onlinehelp.tableau.com/current/pro/online/mac/en-us/help.htm#functions_functions_string.html). Computers can process integers and Booleans (t/f) much faster than strings.

Boolean>Int>Float>Date>DateTime>String

# KPIs

Dashboards will be implemented to display and indicate certain KPIs and relevant indicators for the customer lifetime value.

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

## KPIs (Key Performance Indicators)

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

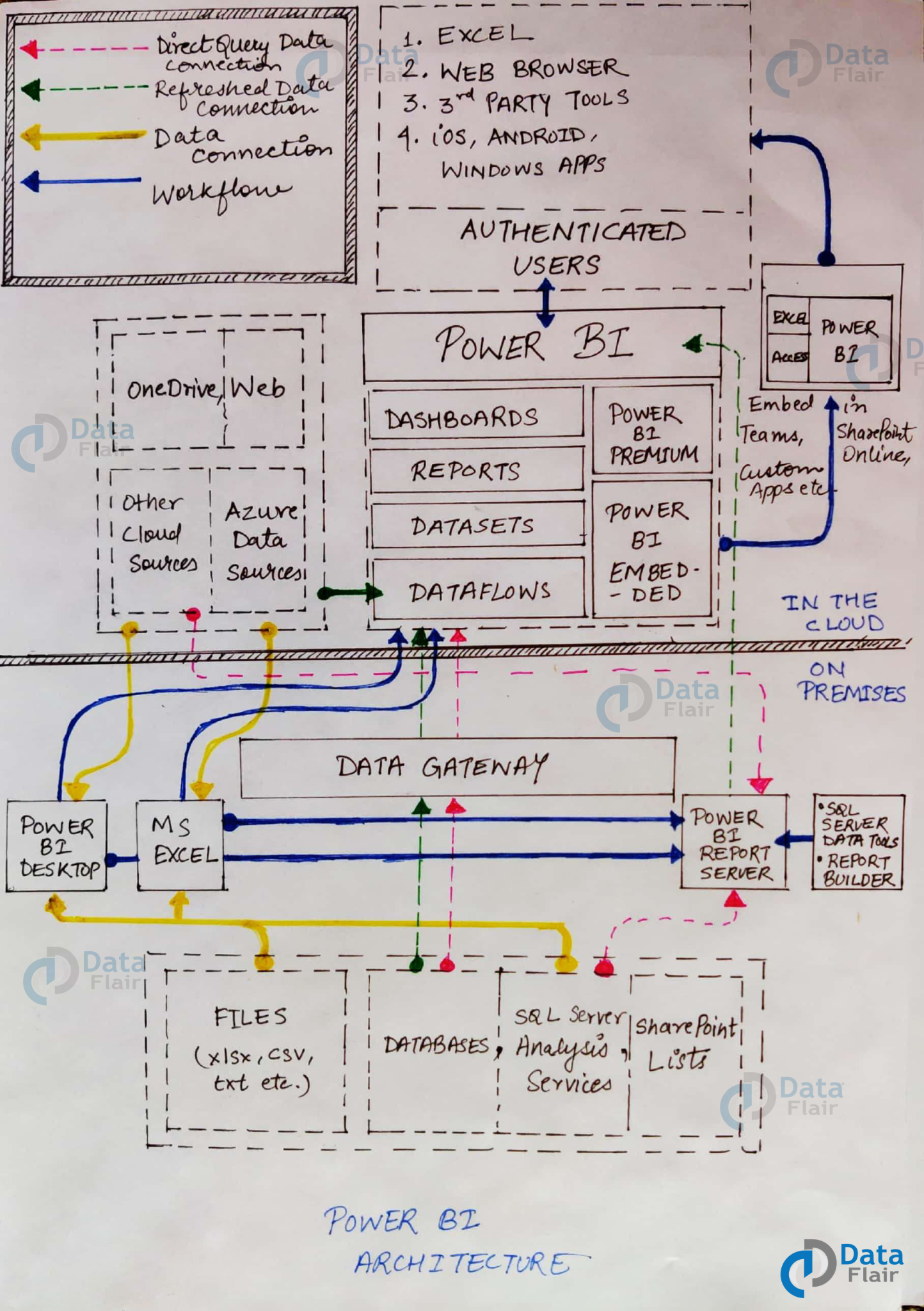
Score card for revenue and unit sold product

Slicer for product id, user id, marital status, city type and gender.

# Deployment

Prioritizing data and analytics couldn’t come at a better time. Companies, no matter what size, are 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 Power BI at scale, as well as organizing, orchestrating, and unifying disparate sources of data for business users and experts alike to author and consume content.

**Deployment Process**



**1. Power BI Desktop**

This desktop-based software is loaded with tools and functionalities to connect to data sources, transform data, data modeling and create reports.

**2. PowerBI Service**

Power BI Service is a web-based platform from where you can share reports made on Power BI Desktop, collaborate with other users, and create dashboards.

**3. PowerBI Report Server**

The Power BI Report Server is similar to the Power BI Service. The only difference between these two is that Power BI Report Server is an on-premise platform. It is used by organizations who do not want to publish their reports on the cloud and are concerned about the security of their data.

**4. PowerBI Gateway**

This component is used to connect and access on-premise data in secured networks. Power BI Gateways are generally used in organizations where data is kept in security and watch. Gateways help to extract out such data through secure channels to Power BI platforms for analysis and reporting.