Architecture Design

Airbnb Data Analysis

Written By	Kumar Shubham
Document Version	
Last Revised Date	

Document Version Control

Date Issued	Version	Description	Author
12th October 2022	1.0	First Version of Complete Architecture Design	Kumar Shubham

Contents

Introduction

What is Architecture design document?

Any software needs the architectural design to represents the design of software. IEEE defines architectural design as “the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system.” The software that is built for computer-based systems can exhibit one of these many architectures.

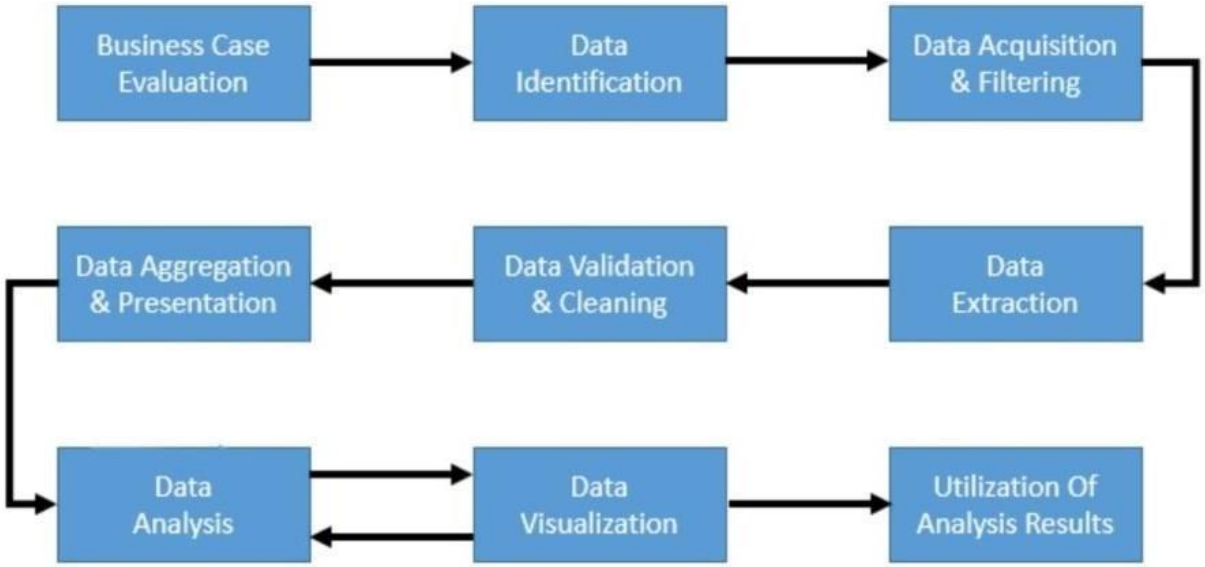
Each style will describe a system category that consists of :

- A set of components (eg: a database, computational modules) that will perform a function required by the system.
- The set of connectors will help in coordination, communication, and cooperation between the components.
- Conditions that how components can be integrated to form the system.
- Semantic models that help the designer to understand the overall properties of the system.

Scope

Architecture Design Document (ADD) is an architecture design process that follows a step-by-step refinement process. The process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the design principles may be defined during requirement analysis and then refined during architectural design work.

Architecture



Business Case Evaluation

Offers a clear understanding of the support, motivation, and goals of carrying out the analysis. Claims that a business case is created, assessed and approved previous to proceeding with the actual hands-on analysis tasks. This evaluation of the business case benefits understands the business resources that will require to be utilized and which business challenges the analysis will stop.

Data Identification

Knowing the datasets needed for the analysis project and their sources. Recognizing a wider kind of data source may enhance the possibility of finding hidden patterns and correlations. Depending on the business range of the analysis project and the nature of the business problems being addressed, the needed datasets and their sources can be internal and/or external to the company.

Data Acquisition And Filtering

Data is collected from all of the data sources that were recognized during the previous stage. The collected data is then subjected to automated filtering for the elimination of corrupt data or data that has been considered to have no value to the analysis objectives. Depending on the type of data source may need API integration, such as with Twitter.

Data Extraction

Some of the data recognized may arrive in a format unsuitable with the Big Data solution. Different types of data are more possible with data from external sources. Extract diverse data and transform it into a form that can be used for the purpose of the data analysis. The amount of extraction and transformation needed depends on the types of analytics and capacities of the Big Data solution.

Data Validation And Cleaning

Wrong data can skew and distort analysis results. Data input into Big Data analyses can be disorganized without any evidence of validity. Its complexity can more make it difficult to come at a set of proper validation constraints. This phase sets often complex validation rules and eliminates any known wrong data.

Data Aggregation and Representation

Integrating various datasets together to arrive at a combined view. Data may be dispersed across various datasets, claiming that datasets be joined together via common fields. Example: date or ID. Related data fields may appear in multiple datasets.

Data Analysis

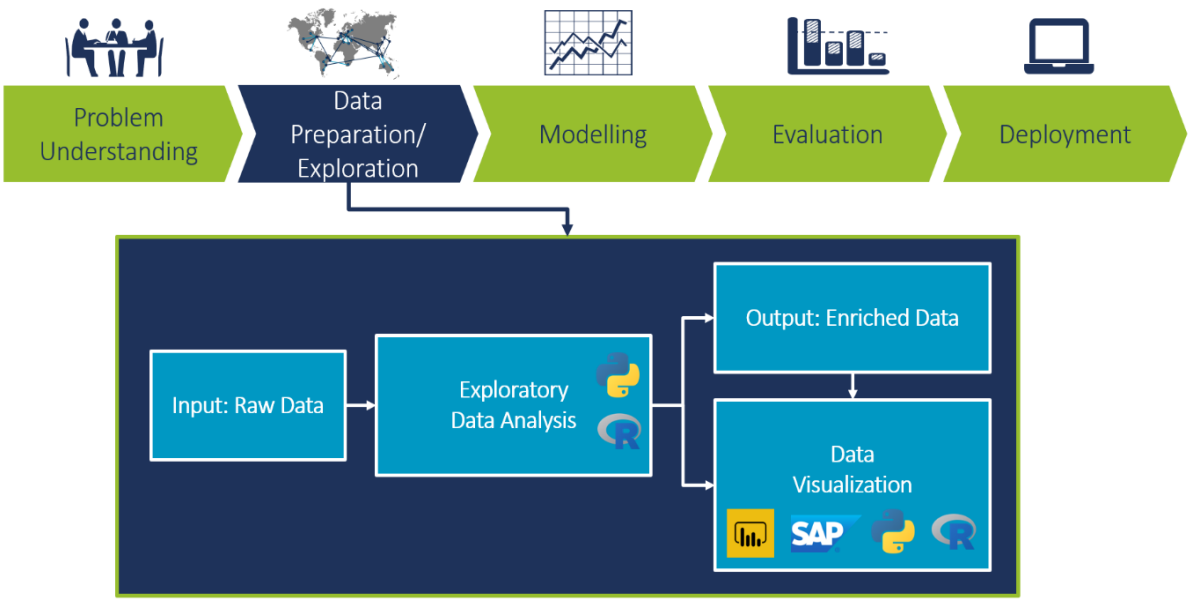
Typically includes one or more types of analytics. Iterative in nature, particularly if the data analysis is exploratory; The analysis is reproduced till the proper pattern or correlation is uncovered. It can be as easy as querying a dataset to compute an aggregate for comparison. Or as challenging as joining data mining and complex statistical analysis techniques.

Data Visualization

Used to graphically demonstrate the analysis results for effective analysis by company users. Present users with the ability to make visual analysis, allotting for the discovery of answers to questions that users have not yet even formed. The same results may be performed in a number of various ways, which can change the presentation of the results. Use the most proper visualization technique by keeping the business domain in context.

Utilization Of Analysis Result

Committed to determining how and where prepared analysis data can be further leveraged.
Typical areas that are investigated during this stage include:
Input for Enterprise Systems
Business Process Optimization



- Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is an approach to extract the information enfolded in the data and summarize the main characteristics of the data. It is considered to be a crucial step in any data science project. Most people underestimate the importance of data preparation and data exploration. EDA is essential for a well-defined and structured data science project and it should be performed before any statistical or machine learning modeling phase.

- Data Preparation:

Data preparation is cleaning and organizing the real-world data, which is known to take up more than 80% of the time of a data scientist's work. Real-world data or raw data is dirty, full of missing values, duplicates and in some cases wrong information. Most machine learning algorithms cannot deal with missing values; hence, data needs to be converted and cleaned. Common solutions of handling missing values would be dropping rows, linear interpolation, using mean values etc. Depending on the importance of the feature and amount of the missing values one of these solutions can be employed.

we mainly use Python and R (Programming languages commonly used by data scientists on their daily work) for data preparation and data pre-processing.

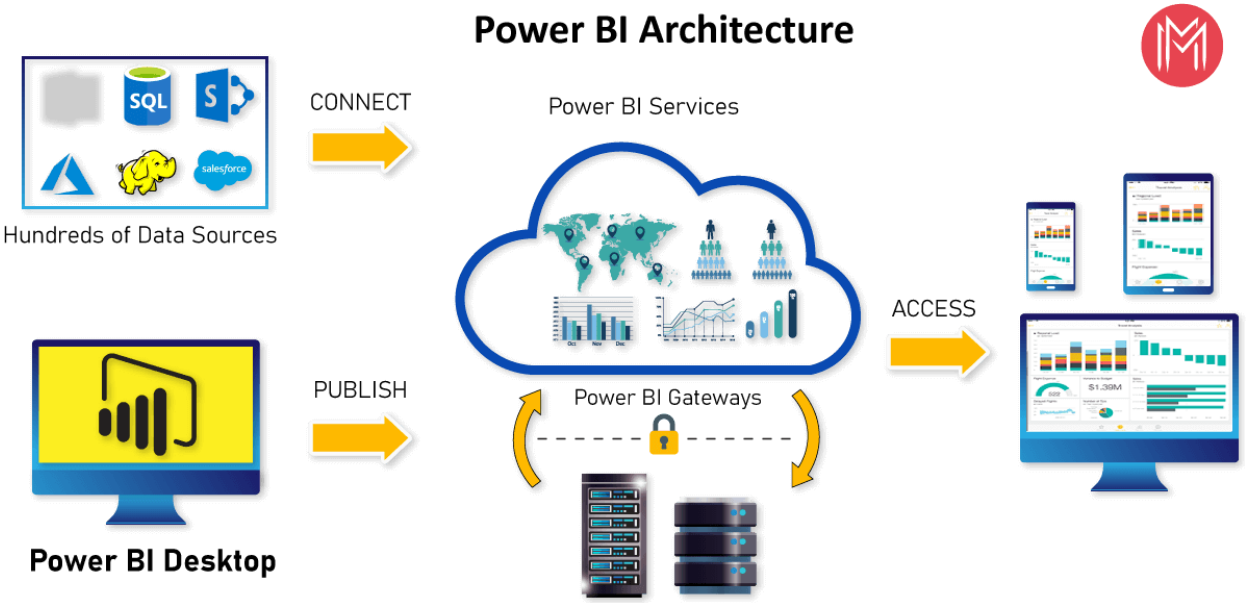
- Data Exploration

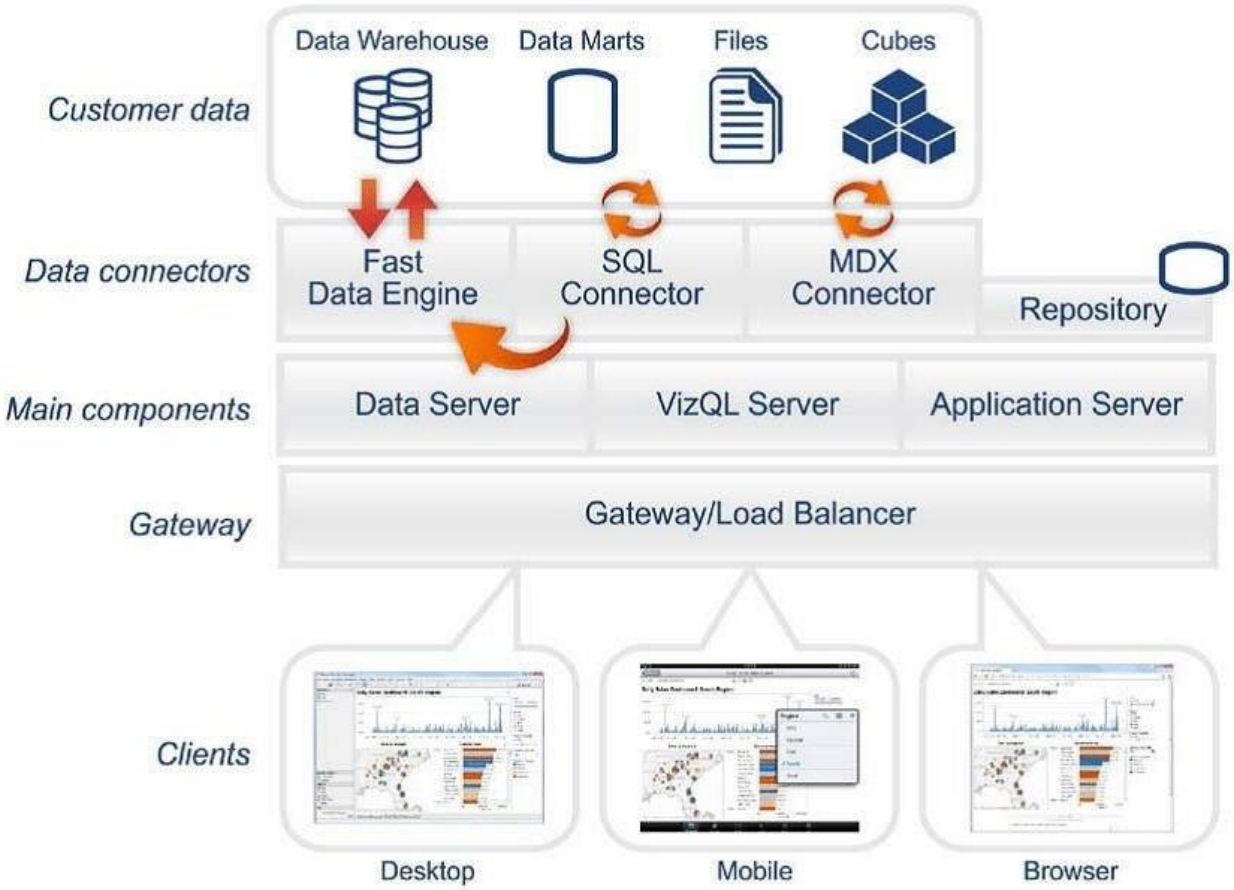
Having a clean dataset in hand, we need to understand the data, summarize its characteristics, and visualize it.
Understanding the data is an iterative process between the data science team and the experts from the business side. It can help both sides to identify and construct important features, and later to build suitable machine learning models.
An essential part of data exploration is data transformation. Let's clarify it by an example. Assume a forecasting problem in the field of logistics for a number of deliveries in different locations by different suppliers. One data transformation option is "filtering". It is possible to filter one specific location, or one specific group of suppliers, and make the forecast on the filtered data to get a fast insight. Another tactic is called "aggregation". If we have daily data, by making weekly or even monthly aggregation we can have a new data set which brings us insights to the existing seasonality and trends.

Some commonly used plots for EDA are:

- Histograms: to check the distribution of a specific variable
- Scatter plots: to check the dependency between two variables
- Maps: to show distribution of a variable on a regional or world map
- Feature correlation plot (heatmap): to understand the dependencies between multiple variables
- Time series plots: to identify trends and seasonality in time dependent data

Power BI Architecture





Power BI is a business platform that includes several technologies to work together. It delivers outstanding business intelligence solutions. Power BI Architecture contains four steps.

these four steps giving insightful information about each one of them.

1. Data Integration
2. Data Transforming
3. Report & Publish
4. Creating and Dashboard

Data Integration:

Data is extracted from different sources which can be different servers or databases. The data from various sources can be in different types and formats. If you import the file into the Power BI, it compresses the data sets up to 1GB, and it uses a direct query if the compressed data sets exceed more than 1GB. Then the data is integrated into a standard format and stored at a place called a staging area. There are two choices for big data sets. They are as follows.
Azure Analytics Services
Power BI premium

Data Transforming:

Integrated data is not ready to visualize data because the data should be transformed. To transform the data, it should be cleaned or pre-processed. For example, redundant or missing values are removed from the data sets. After data is pre-processed or cleaned, business rules are applied to transform the data. After processing the data, it is loaded into the data warehouse.

Report & Publish:

After sourcing and cleaning the data, you can create the reports. Reports are the visualization of the data in the form of slicers, graphs, and charts. Power BI offers a lot of custom visualization to create the reports. After creating reports, you can publish them to power bi services and also publish them to an on-premise power bi server.

Creating Dashboards:

You can create dashboards after publishing reports to Power BI services, by holding the individual elements. The visual retains the filter when the report is holding the individual elements to save the report. Pinning the live report page allows the dashboard users to interact with the visual by selecting slicers and filters.