**Functional Requirements**

**DATA COLLECTION**

* For any task the collecting data is exceptionally critical and important. Anticipating the data and how to store it will have vital job in the venture's result
* The data here will stream in from different sources, for example, the different airlines, customer review and customers purchase
* As the data is colossal and as the data is streaming in different from and different sources the information will be unstructured. Hence, we will be utilizing Big Data architecture.

**DATA CLEANSING**

* Processed and relevant data will be accessible for analysts and different recipients. These data will be introduced in a clean structure using **python** that is effectively justifiable over recipients.
* As the data sets are large, often a big data solution must process data files using running batch jobs to filter, aggregate, and prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to new NoSQL database (**Cassandra**)

**DATA VISUALIZATION**

* We will need Simba Cassandra ODBC driver to connect Cassandra to PowerBI.
* The cleaned and appropriately sorted out information will be utilized to create dashboard in **Power BI.**
* The reports and dashboards need to be created so that the data analyst and business executives will be able to track KPIs, metrics, and other key data points relevant to a business.

**ROBUST FRAMEWORK & ORCHESTRATION**

* Any sort of Big Data frameworks is intended to adapt to the issues of Variety, Volume and Velocity. Key highlights to take in account are iterative preparing, storage and information ingestion
* The framework should allow repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and load the processed data into an analytical data store, or push the results straight to a report or dashboard generating tool.

**Non-functional requirements**

* Performance**:** The performance of the data collection, data integration and data visualization like the speed of operation with integrity in the data. Following are the Performance consideration for any operation.

1. Data Acquisition

2. Data Processing

3. Storage

* Reliability**:** The system should be continuing to work the way it is supposed to be without any failure. The system’s solution should be reliable to use and make decisions.
* Usability**:** The usability should be straight forward. We will be creating a metadata documentation on how to use and troubleshoot the system step by step.
* Confidentiality**:** As we handle customer related data, we need to ensure that the sensitive data is protected, and only authorized users can only view the highly sensitive data. We will create user-based dashboard so that the user is restricted to only specific data only.
* Efficiency**:** The system should handle the glitches and the unexpected inputs and the volume of records, and yet the changes or the glitches shouldn’t affect the efficiency.
* Reusability**:** The system should be used multiple times to use, the process should be iterative.
* Integrity**:**  The system’s process and output should have the same integrity as of the source.
* Flowcontrol**:** The system should handle the data input flux and the flow should be well defined, change in the input flux shouldn’t affect the flow of the project.