Project Report

LFPG | Copyright by Holger Zengler | 2011-06-19 | Airport-Data.com

Group 3

Chetan M Jadhav
Soumavo Guria
Sumeet Gaglani
Vinisha Garg

Table of Contents

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.

- Data Acquisition
- Data Processing
- 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.

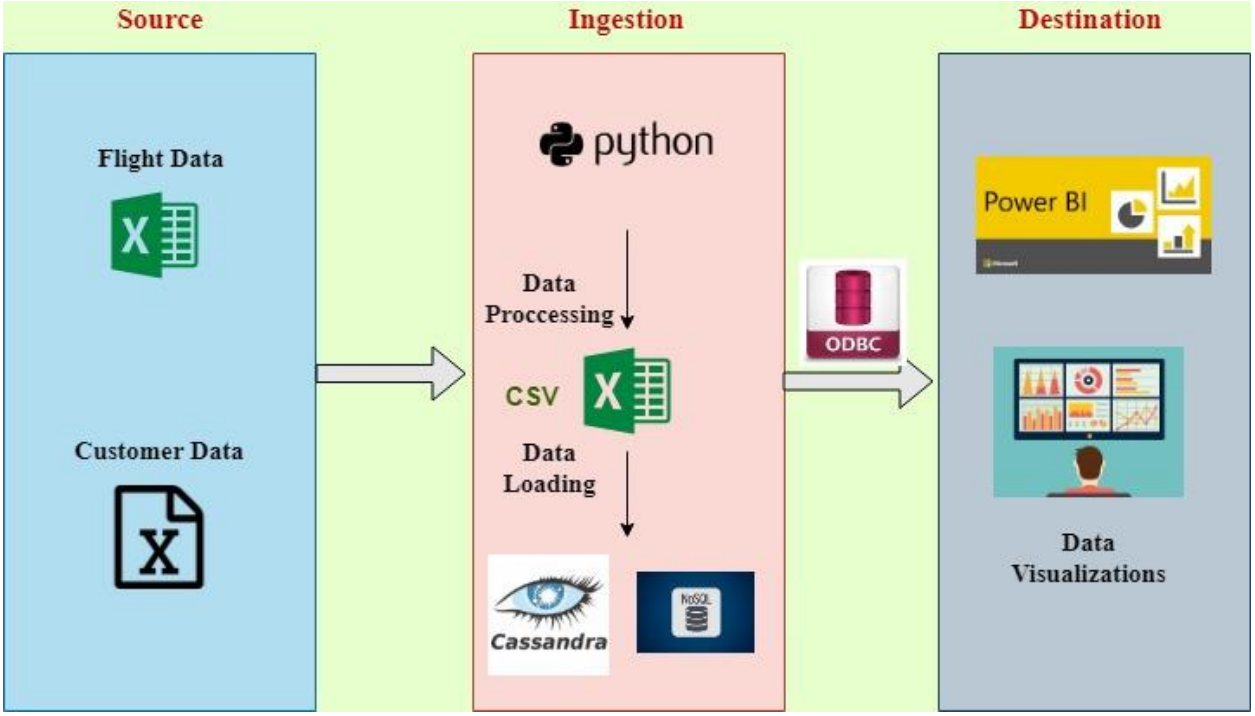
Integrity:

The system's process and output should have the same integrity as of the source.

Flow control:

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.

Vision Diagram:



According to the above vision diagram, below mentioned are the tools used for this project to complete.

Tools Used:

Language & Tools:

- CQL (Cassandra Query Language):

Cassandra Query Language (CQL) is a query language for the Cassandra database. This release of CQL works with Cassandra 3.0 for Linux. The Cassandra Query Language (CQL) is the primary language for communicating with the Cassandra database.

- Python:

Python is used to Clean the data and to process the data. Python is an interpreted, high-level, general-purpose programming language

Database:

- Cassandra NoSQL Database:

Apache Cassandra is a free and open-source, distributed, wide column store, NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple datacenters, with asynchronous master less replication allowing low latency operations everyone who uses it.

Visualization:

- Power BI:

Power BI is a business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards.

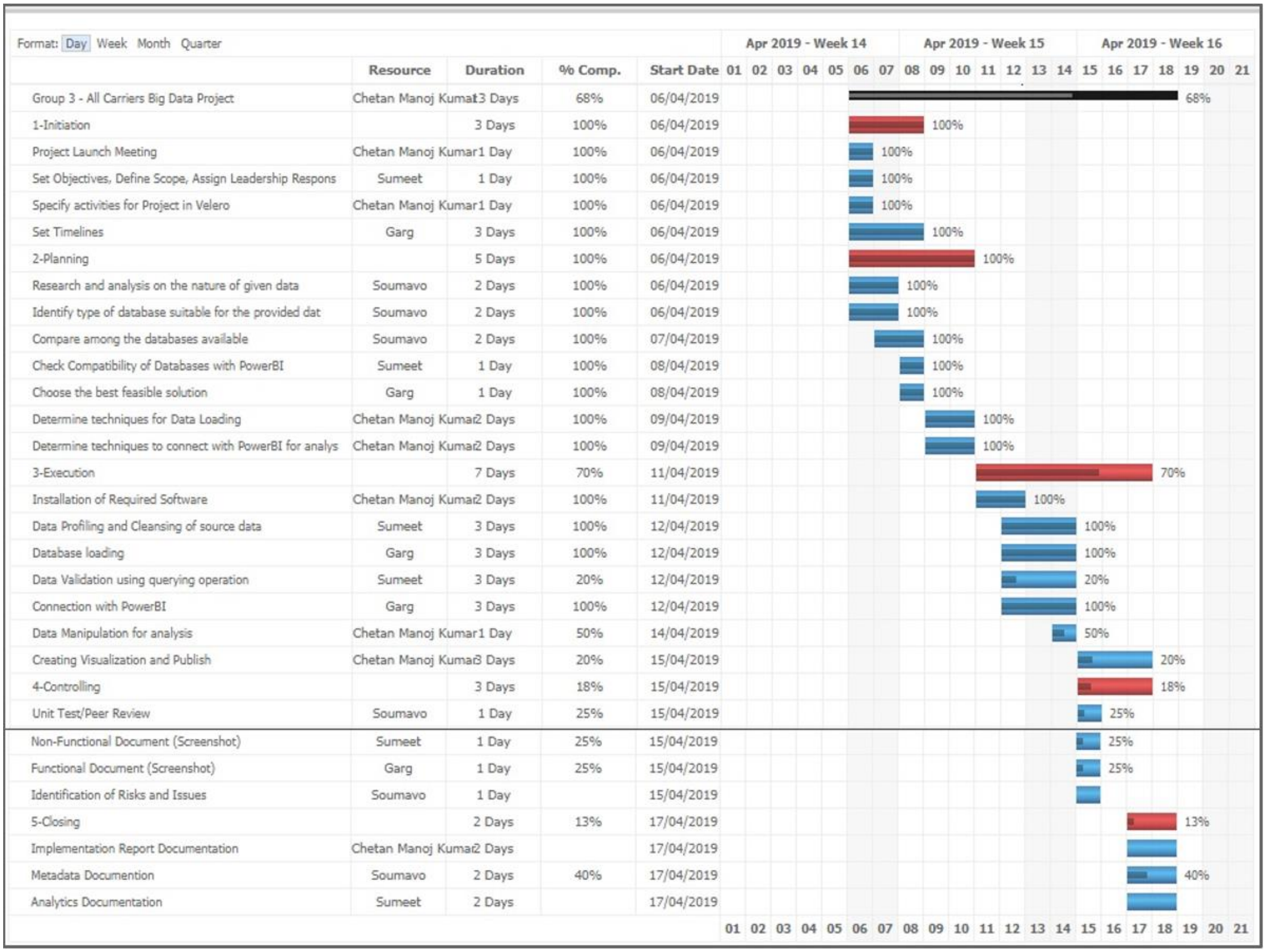
ODBC Connector:

- Simba ODBC Cassandra Connector:

The Simba Technologies Cassandra ODBC and JDBC Drivers enable direct SQL query translation to the Cassandra Query Language (CQL), offering users unparalleled performance at scale. The built-in Collaborative Query Execution (CQE) passes down filters and aggregations to provide high-performance access to Cassandra.

Gantt Chart:

A Gantt chart is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph shows the duration of each activity.



	Resource	Duration	% Comp.	Start Date
Group 3 - All Carriers Big Data Project	Chetan Manoj Kuma	13 Days	68%	06/04/2019
1-Initiation		3 Days	100%	06/04/2019
Project Launch Meeting	Chetan Manoj Kumar	1 Day	100%	06/04/2019
Set Objectives, Define Scope, Assign Leadership Respons	Sumeet	1 Day	100%	06/04/2019
Specify activities for Project in Velero	Chetan Manoj Kumar	1 Day	100%	06/04/2019
Set Timelines	Garg	3 Days	100%	06/04/2019
2-Planning		5 Days	100%	06/04/2019
Research and analysis on the nature of given data	Soumavo	2 Days	100%	06/04/2019
Identify type of database suitable for the provided dat	Soumavo	2 Days	100%	06/04/2019
Compare among the databases available	Soumavo	2 Days	100%	07/04/2019
Check Compatibility of Databases with PowerBI	Sumeet	1 Day	100%	08/04/2019
Choose the best feasible solution	Garg	1 Day	100%	08/04/2019
Determine techniques for Data Loading	Chetan Manoj Kuma	2 Days	100%	09/04/2019
Determine techniques to connect with PowerBI for analys	Chetan Manoj Kuma	2 Days	100%	09/04/2019
3-Execution		7 Days	70%	11/04/2019
Installation of Required Software	Chetan Manoj Kuma	2 Days	100%	11/04/2019
Data Profiling and Cleansing of source data	Sumeet	3 Days	100%	12/04/2019
Database loading	Garg	3 Days	100%	12/04/2019
Data Validation using querying operation	Sumeet	3 Days	20%	12/04/2019
Connection with PowerBI	Garg	3 Days	100%	12/04/2019
Data Manipulation for analysis	Chetan Manoj Kumar	1 Day	50%	14/04/2019
Creating Visualization and Publish	Chetan Manoj Kuma	3 Days	20%	15/04/2019
4-Controlling		3 Days	18%	15/04/2019
Unit Test/Peer Review	Soumavo	1 Day	25%	15/04/2019
Non-Functional Document (Screenshot)	Sumeet	1 Day	25%	15/04/2019
Functional Document (Screenshot)	Garg	1 Day	25%	15/04/2019
Identification of Risks and Issues	Soumavo	1 Day		15/04/2019
5-Closing		2 Days	13%	17/04/2019
Implementation Report Documentation	Chetan Manoj Kuma	2 Days		17/04/2019
Metadata Documention	Soumavo	2 Days	40%	17/04/2019
Analytics Documentation	Sumeet	2 Days		17/04/2019

Issues & Risks:

Risk is the possibility of losing something of value. Values can be gained or lost when taking risk resulting from a given action or inaction, foreseen or unforeseen (planned or not planned). Risk can also be defined as the intentional interaction with uncertainty. Uncertainty is a potential, unpredictable, and uncontrollable outcome; risk is a consequence of action taken despite uncertainty.

Issues are present focused, always negative and need to be fixed to be on the track of the project's goal. Issues is when an even that has in fact has occurred and the team now must work on resolution.

Data Dictionary:

The information about the columns of the air carrier csv weren't informative which created an issue in understanding the data as in what it represents and what the data means.

To resolve this, we had to come up with our own metadata document which we created by reading the data dictionary which was given by the site and with our understanding of the data.

Choosing NoSQL database:

After we identified that we need to use the "Key Value NoSQL database", we now had a pool of key valued databases to choose from and as our team who has never worked on any of the NoSQL databases. It posed as an issue for us to move forward and a road block in our way to reach goal.

To resolve this, we divided the databases; did the research on which is better amongst other key valued NoSQL databases and which one has a better connection with our PowerBI data visualization tool.

ODBC Connector:

This was the main issue and roadblock which we faced while doing this project. Choosing the ODBC connectors which connects Cassandra database with the PowerBI and the information must be input to make the connection was a task, must check quite a few connectors before we chose to go with SIMBA ODBC CONNECTOR.

Primary Key:

For a key valued type NoSQL database, identifying primary key is mandatory and as the primary key column wasn't known for us. We must find a column which can be identifies the unique row.

As there were many columns, it posed as an issue to identify which column can be primary key and upon various permutation and combination to find out 8 column composite primary key.

Data Loading:

Loading from the CSV was a quite a task where we came with so many errors while loading. The error log had little or no information at all which had no help in debugging.

We resolved this by deleting all the empty rows and columns and then tried to load.

Visualization:

The data when loaded to the power BI data visualization tool, there were couple of columns where we need to change the data type to get the better visualization for example: data type of geographical columns so that the visualization of the same on the map can be done.

The data had to be aggregated and calculated manually on all most all the columns in the power BI tool to get the analysis done.

Missing Revenue/Sales column:

There was a missing revenue/sales column which interprets fact of the database, on which the analysis can be based on.

As the column was missing, we had to do our analysis on given set of the columns.

***** The End *****