=========================================

System Design Specification (SDS)

For

Northwoods Airlines

=========================================

**Publish Date:** 4th Mar 2021

**Prepared By:** Jitendra Nikhare

**Document Version:** V1.0

**Contents:**

Introduction …………………………………………………………………………………………………………………………………… 3

Ingestion ………………………………………………………………………………………………………………………………………… 3

Extraction & Annotation .………………………………………………………………………………………………………………… 3

Knowledge Store for Data Insights .……………………………….………………………………………………………………… 3

System Flow Diagram .…………………………………………………………………………………………………………………… 4

Code Inventory………………………………………………………………………………………………………………………………… 4

**Introduction:**

Northwoods Airlines headquarter in Northwoods city US, is working in airline business and deal in passenger travelling and courier and freight services. It is working in domestic and international flight services and very famous for prompt services, cheap flights rates from Northwoods to all destinations.

Northwoods Airlines want to collect the data into system for data analytics to understand business and can take decision for their business improvement.

**Data Ingestion:**

phData forms the data lake system for Northwoods Airlines, wherein data from varied source system is brought into .csv format. The data is available in landing zone for each source under separate directory like Airlines, Flights, Airport etc.

Ingestion process takes the data from each sources and load into respective Snowflake temporary tables with the same table name, column name and structure.

Each ingestion looks a delta data for daily transactions throughout the world. Ingestion process have established through Databricks pipeline process, it read the data from .csv files and load into SnowSQL temporary tables. Every execution will confirm the data loading process with overwrite option so that it rid of data duplication during data loading into Snowflake tables. During data ingestion process, it will make sure for quality checks as per business requirements. After loading the data into respective temporary tables, ingestion process has stored the data into Azure blob storage to keep it for always-on data availability for auditing.

**Data Extraction & Annotation:**

Once data is available into SnowSQL tables, we apply business logics, conditions and transformation rules for data validation, enrichment and annotation process. We apply the logic, rules to applicable columns as per ETL sheet and load the transformed data into final snapshot table. The Snapshot table is full data table (ie. first business day to today’s date) gets the validate data as per business requirement. After loading the data into respective Snowflake tables, process has stored the data into Azure blob storage to keep it for always-on data availability for their downstream users.

Data extraction and transformation process is done through Databricks pipelines which take SnowSQL temporary table as sources and load into final SnowSQL snapshot tables for their respective sources.

**Knowledge Store for Data Insights:**

Knowledge Store process is middle layer between Extraction and data insight process. It is use to create the SnowSQL views or materialized Views as per business requirement to collect aggregate data into summary form which will use to show into data analytics process. Once the data has been loaded into the respective platform(s), Clients will request various insights or KPI reports derived from the provided data.

The summary views are connecting into report tool and showing live data to determine how their competition is performing.

● Total number of flights by airline and airport on a monthly basis

● On-time percentage of each airline for the year 2015

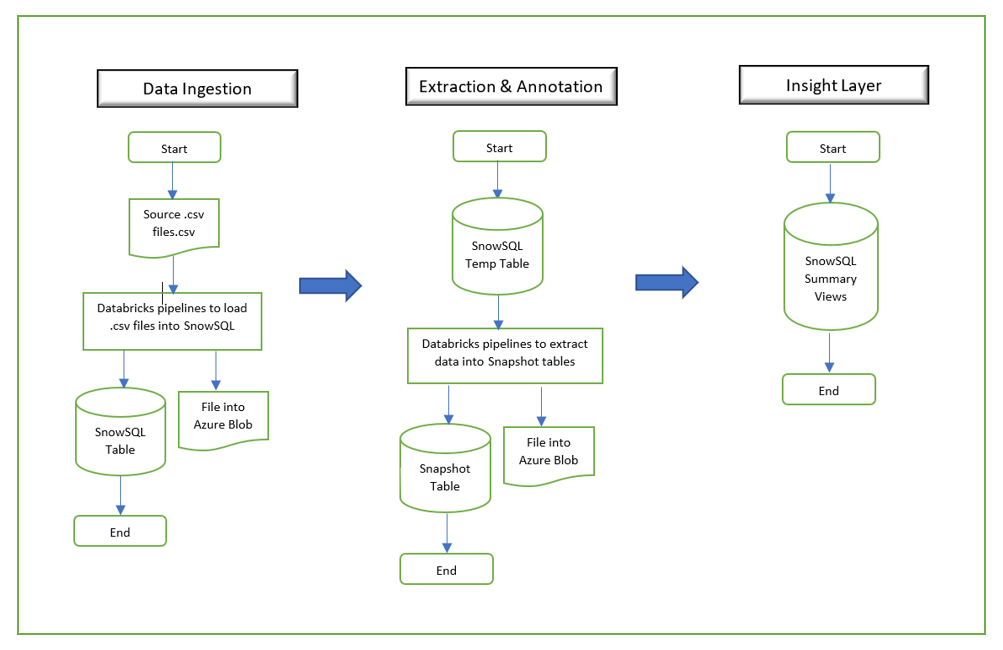
● Airlines with the largest number of delays

● Cancellation reasons by airport

● Delay reasons by airport

● Airline with the most unique routes

**System Flow Diagram:**



**Technologies and Tools:**

**Snowflake:** Cloud based database use to store data into structured format.

* SnowSQL: Temporary Tables, External Tables, Regular Views

**Databricks:** Developed program to extract, enrich and process data.

* PythonSpark

**Azure Cloud:** Use services to storage data.

* Microsoft Azure Blob Storage

**Code Inventory:**

<https://github.com/jsnikhare/phData_Case-Study>