**BIG DATA AND VISUALIZATION**

**TEAM – 4**

RISHABH VEDANT

ROHIT RAJSHEKAR CHIMKODE

RONALI SAHU

SAI CHANDANA GUMMADI

SHIVA PRASAD KARTAN

## Table of Contents

[Executive Summary](#_Toc3498)

[Introduction](#_Toc3499)

[Analysis](#_Toc3500)

[Requirements](#_Toc3501)

[Implementation Plan](#_Toc3502)

[References](#_Toc3505)

**Executive Summary**

Purpose:  
 The purpose of this documentation is to provide an consolidated workflow of the job executions for the project – Big Data and Visualization.

This project encapsulates concepts related to Big data analysis using Azure Cloud platform.

Problem addressed:

This project is meant to assist the customers to get an idea about the tentative time delay while they book their tours. We have used Machine learning predictive analysis to get the delay, which will help the customers to select the best travels based on the likely hood of delay.



**Introduction**

Margie's Travel (MT) provides concierge services for business travelers. MT is investigating ways to capitalize on their existing data assets to provide new insights that provide them a strategic advantage against their competition.

In planning their product, they heard much fanfare about machine learning and came up with the idea of using predictive analytics to help customers best select their travels based on the likelihood of a delay. When reviewing their customer transaction histories, they discovered that their most premium customers often book their travel within seven days of departure. In speaking with customer service, they learned that these customers often ask questions like, "I don't have to be there until Tuesday, so is it better for me to fly out on Sunday or Monday?"

They have historical flight data provided to them by the United States Department of Transportation (USDOT).

In order to enhance the customer experience, they are in search of a solution which can improve the process of flight selection based on the flight delay prediction model curated using machine learning process.

MT plans to pilot this solution internally, whereby the small population of customer support who service MT's premium tier of business travelers would begin using the solution and offering it as an additional data point for travel optimization. They would like to provide their customer support agents a web-based solution that enables them to map the predicted delays for a particular customer's departure airport(s) of choice.

MT is interested in applying data science to the problem to discover if the weather forecast and their historical flight delay data could provide meaningful input into the customer's decision-making process.

**Analysis**

MT has over 30 years of historical flight data along with each flight's historical weather condition as CSV files, apart from these, they are fetching weather forecast via third party application.

Jack Tradewinds, the CIO of MT, is looking to modernize their data story. He has heard a great deal of positive news about Spark SQL on HDInsight and its ability to query exactly the type of files he has in a performant way, but also in a way that is more familiar to his analysts and developers because they are all familiar with the SQL syntax that it supports.

He would love to understand if they can move this data away from their on-premises data center into the cloud and enhance their ability to load, process, and analyze it going forward. Given his longstanding relationship with Microsoft, he would like to see if Azure can meet his needs.

Keeping in mind the customer preference, we have come up with an architecture which leverages Azure Databricks for Data preparation and cleaning along with development of machine learning model for flight delay prediction.

MT has over 30 years of historical flight data provided to them by the United States Department of Transportation (USDOT), which among other data points, includes flight delay information for every flight. The data arrives in flat, comma-separated value (CSV) files with a schema of the following:

(Year, Month, DayOfMonth, Airline, TailNum, FlightNum, OriginAirport, DestinationAirport, ScheduledDepartureTime, ActualDepartureTime, ScheduledArrivalTime, DepartureDelay, AirTime, Distance, Cancelled, CancellationCode)

In addition, for all data since 2003, each row includes new fields describing the type of delay experienced, where the value for each type is the number of minutes the delay was experienced for that source of delay:

(CarrierDelay, WeatherDelay, NationalAirSystemDelay, SecurityDelay, LateAircraftDelay)

**Requirements**

**Customer Needs**

1. Want to modernize their analytics platform without sacrificing the ability to query their data using SQL.

2. Need an approach that can store all of their data, including the unmodified source data and the cleansed data they query for production purposes.

3. Want to understand how they will load their large quantity of historical data into Azure.

4. Need to be able to query the weather forecast and use it as input to their flight delay predictions.

5. Desire a proof of concept (PoC) machine learning model that takes as input their historical data on flight delays and weather conditions to identify whether a flight is likely to be delayed or not.

6. Need web-based visualizations of the flight delay predictions.

**Customer Objections:**

1. We have heard that creating a machine learning model takes a month to build and another 2-3 months to operationalize so that it is useable from our production systems. Is this true?

2. Once our model is operationalized, how do we retrain and redeploy it? Will this process break clients currently accessing the deployed model?

3. Can we query flat files in the file system using SQL?

4. Does Azure provide anything that would speed up querying (and exploration) files in Hadoop Distributed File Systems (HDFS)?

5. Does Azure provide any tools for visualizing our data? Ideally, access to these could be managed with Active Directory.

6. Can we use Azure Active Directory accounts for our users? If so, can we restrict who can access Azure Databricks when they can access it, require two-factor authentication, and restrict access if there is suspicious activity on their account?

7. Is Azure Databricks our only option for running SQL on Hadoop solutions in Azure?

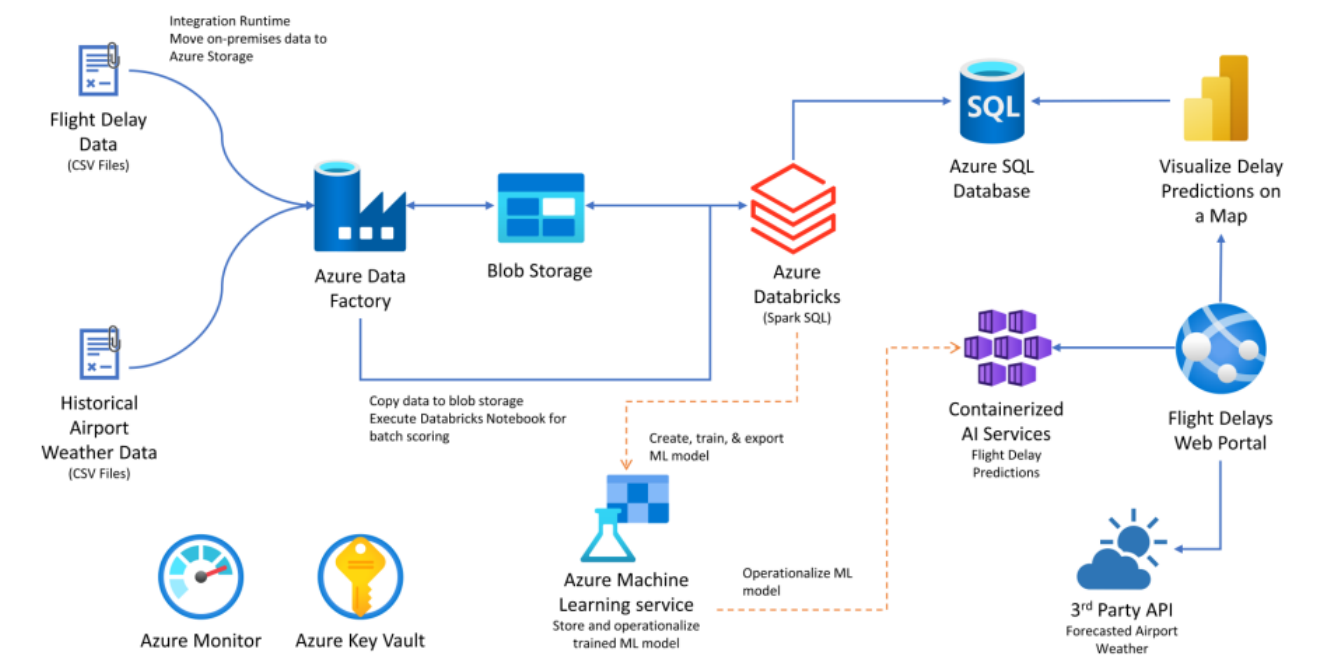
8. We have heard of Azure Data Lake, but we are not clear about whether this is currently a good fit for our PoC solution or whether we should be using it for interactive analysis of our data.

9. We are hiring a data scientist who prefers to use MLflow to track model training run metrics and artifacts. Can the proposed Azure-based solution support this library?

**Implementation Plan**

The implementation model which we have come up with includes Azure cloud services such as Azure Data Factory, Azure Blob Storage, Azure Databricks, Power BI.

The Workflow commences with loading of Data sets from the local system into Azure Databricks. The Data sets such as the historical flight delay data and the historical airport weather data loaded into the azure databricks file system’s default database.

****

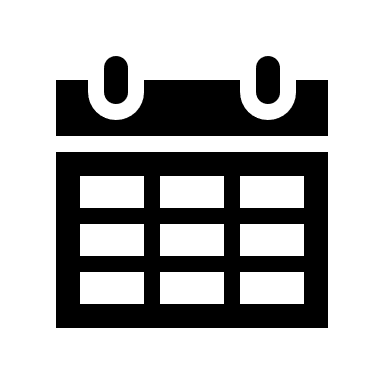
**Storage solutions used are :**

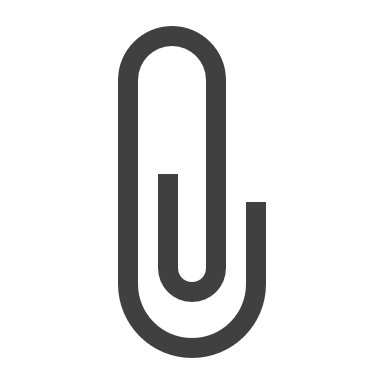
**1.Azure SQL Database** : Azure SQL Database is a fully managed platform as a service (PaaS) database engine that handles most of the database management functions such as upgrading, patching, backups, and monitoring without user involvement.

**2. Azure Blob Storage :** Azure Blob storage is Microsoft's object storage solution for the cloud. Blob storage is optimized for storing massive amounts of unstructured data, such as text or binary data. Blob storage is ideal for: Serving images or documents directly to a browser.

**Data Loading**

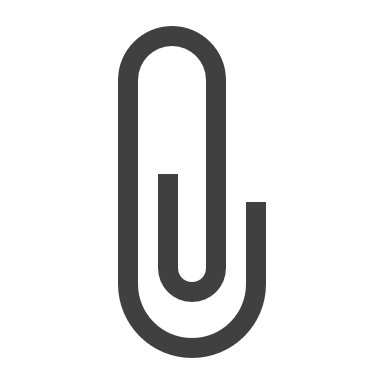
The historical and weather data can be loaded into Azure Data Factory as it will help us to create ELT pipeline which will help in orchestrating data movement and transforming data at scale.

****



Flight Delay Data

(CSV Files)



Historical Airport Weather Data

(CSV Files)

Blob Storage

Azure Data Factory

Install Integration Runtime  
to move on-premises data to Azure Storage

Monthly copy activity

**Data Preparation**

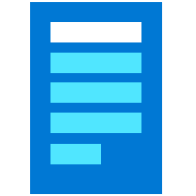
The large amount of weather and flight data than need to be prepared for creating a model can be done using Azure Databricks Workspace . With the help of this environment we can access all the Databricks assets such as Notebook and clusters.

By creating New clusters we load the data into Notebook and then perform SQL queries for selecting the data required and then we perform clean up on data and remove all null values.



Azure Databricks Workspace

Spark SQL



Explore & Prepare Data



Databricks notebook used

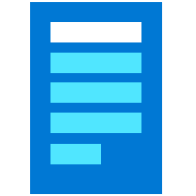
By AWT analysts

**Machine Learning Model**



Azure Databricks Workspace

Spark SQL

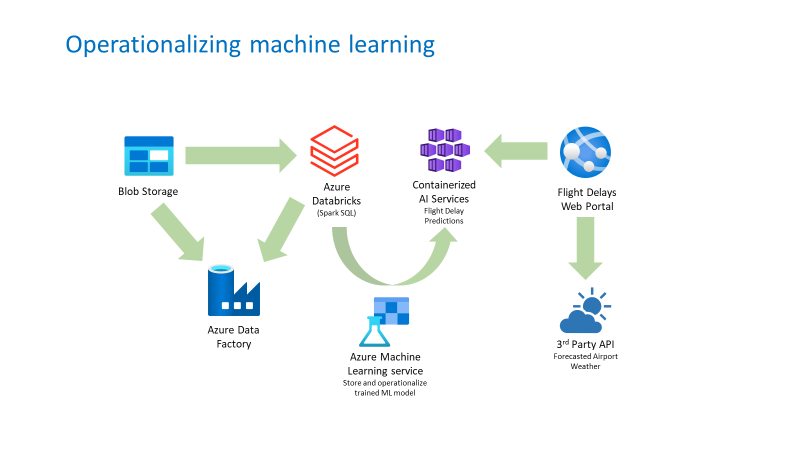


Prepare Data



Spark MLlib two-class logistic

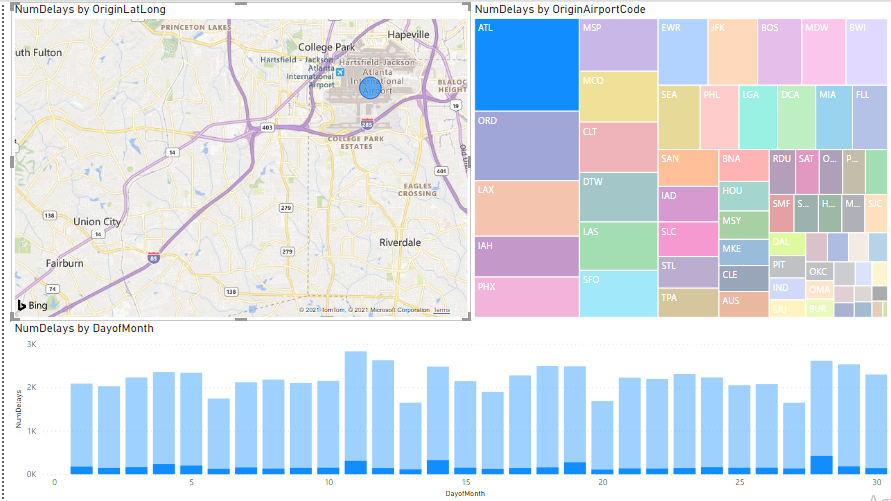
regression model created and   
trained within Databricks notebook

****

**Visualization and Reporting**

Visualizing the Data in Power BI provides deeper data insights .  
In the Visualization, Map visualization represents distinct airports with number of delays.

Likewise, the Airport code gets selected in the TreeMap, along with this, the stacked column chart represents the number delays occurred on a daily basis for that particular airport

****

|  |  |
| --- | --- |
| **Additional references Description** | **Links** |
| Azure solution architectures | Https://azure.microsoft.com/en-us/solutions/architecture/ |
| Azure machine learning services | Https://docs.microsoft.com/en-us/azure/machine-learning/service/ |
|  |  |
| Machine learning algorithms | Https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/ |
|  |  |
| Azure data factory | <Https://docs.microsoft.com/azure/data-factory/introduction/> |
| Azure databricks | <Https://docs.microsoft.com/en-us/azure/azure-databricks//> |
| Power bi | <Https://support.powerbi.com/knowledgebase/articles/430814-get-started-with-power-bi/> |
| Travel data | <Https://www.transtats.bts.gov/homepage.asp/> |
| Weather data | <Https://openweathermap.org/api/one-call-api> |
| Arm templates | <Https://docs.microsoft.com/azure/azure-resource-manager/resource-group-authoring-templates/> |
| Azure ad conditional access | Https://docs.microsoft.com/azure/active-directory/conditional-access/ |
|  |  |
| Azure sql database | <Https://docs.microsoft.com/azure/azure-sql/database/sql-database-paas-overview> |
| Write to azure sql database from a dataframe | <Https://docs.microsoft.com/azure/databricks/data/data-sources/sql-databases#write-data-to-jdbc> |
| Azure key vault | <Https://docs.microsoft.com/azure/key-vault/key-vault-overview> |
| Azure monitor | Https://docs.microsoft.com/azure/azure-monitor/overview |