Using Microsoft Azure for predicting flight delays

Prepared by: Muhammad Anas Mansoori

Introduction

Background

The case study pertains to a challenge faced by an airline company. The company wants to improve its flight delay predictions (flights delayed by 15 minutes or longer) so that its customers are preemptively informed about any challenges they might face in their travel. The company wants to to provide this service to their agents who can access this feature via web service. To overcome this challenge, the company wants to leverage its historical data on flight delays, weather conditions and other factors that may contribute to flight delays. Furthermore the airline wants the solution to use cloud platforms using machine learning techniques.

Solution

Microsoft Azure Services are used to provide a solution to the problem and to modernise the system. Some of the key tools and techniques used include:

- Azure Data Factory: Allows the users to create data driven workflows. It organises the flow of data between storage spaces and the computation environments. The data factory service allows to create pipelines on a specified schedule. It connects and collects data from storage space such as SasS, FTP etc. and moves them to a on premise location or cloud. Next the data is transformed using compute services such as Hadoop and Spark. Then the transformed data is moved to SQL server for consumption by Power BI and analytics tools and other applications [1].

- Azure Databricks: Cloud-based Data Engineering tool that processes and transforms large quantities of data. This is used to process and transform extensive amounts of data and explore it through Machine Learning models. This is used as an alternative to MapReduce and provides just-in-time cloud-based platform [2],[3].

- Azure Machine Learning service: It is collecting of tools and techniques that builds and deploys machine learning models.

- Power BI: A Business Intelligence platform that has visualisation and analytical tools to help develop and gather insights

- Azure SQL Database: A relational database with cloud based service It is built on the SQL Server database engine and provides a set of features that are compatible with SQL Server.

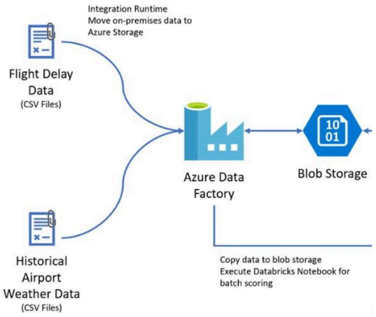
Architecture - Refer Figure 1 in Appendix

1) Data Ingestion: Azure Data Factory is used to ingest data from on-premises sources into the cloud enabled Azure Blob storage - shown in figure 2.

2) Data Storage: Azure Blob Storage stores the relational data utilising the Azure SQL Database

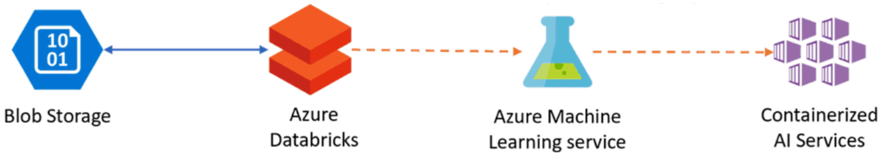
3) Data Processing and Analytics: Azure Databricks processes and analyses the big data relying on Apache Spark based analytics platform for building data pipelines, machine learning models and data analytics solutions. The data is stored back to blob storage. Refer figure 2.

Figure 2 - Data ingestion and storage


Integration Runtime
Move on-premises data to Azure Storage
Flight Delay Data
(CSV Files)
Azure Data Factory
10
01
Blob Storage
Historical Airport Weather Data
(CSV Files)
Copy data to blob storage
Execute Databricks Notebook for batch scoring

4) Machine Learning: Azure Machine Learning service builds and deploys machine learning models. The model is built and trained in Azure Databricks notebook using Python and Spark SQL. The trained model is used to fit the data using Machine Learning libraries - azureml-sdk[databricks] and mlflow. After the model is trained and tested, it is deployed to a web service and for batch scoring - refer figure 3

Figure 3 - Machine learning with Azure Databricks



5) Reporting and Visualisation: Power BI creates dashboards and report to visualise the data. It reads data from Azure SQL Database.

Solution Implementation

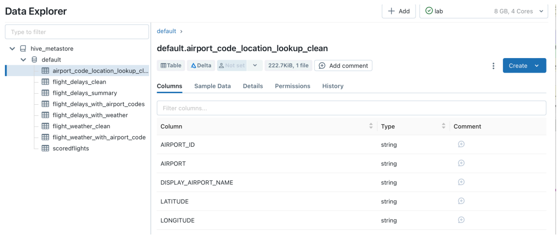
Creating a Databricks Cluster

A databricks cluster was created to run computations on data stored in blob storage, machine learning, data analytics works load. The clusters used a single mode with version 7.3 of databricks runtime. The machine learning libraries installed were azureml-sdk[databricks] and mlflow.

Load Sample data

Data was uploaded to Databricks workspace in the hive metastore. Once the file was uploaded, a table was created which would be used for computational purposes - refer figure 4. At the same step the notebooks with Python scripts were also uploaded in the Databricks workspace. These scripts were used for data preparation, exploring, training the models and deployment.

Figure 4 - Load sample data



default.airport_code_location_lookup_clean

Column	Type	Comment
AIRPORT_ID	string	
AIRPORT	string	
DISPLAY_AIRPORT_NAME	string	
LATITUDE	string	
LONGITUDE	string	

Setup the Data Factory

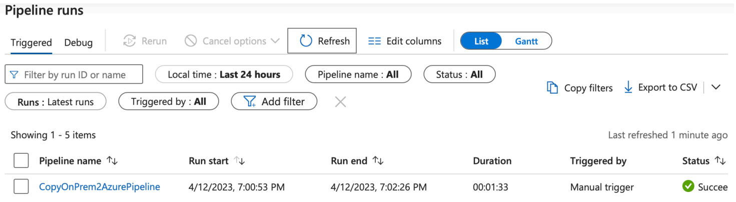
Microsoft Integration Runtime service was used to provide compute power and provide appropriate access to data sources with the cloud service. The unique key produced by the Azure Data Factory was used to register the integration runtime. The self hosted service allows the on-premises data to be used on cloud.

Data Factory Pipeline

The Azure data factory pipeline was created so that data can be moved from on-premises to Azure Storage using the Integration Runtime that was installed in the step earlier. A monthly schedule was established that fetches the data from on-premises location. The data uploaded was stored in Azure Blob Storage.

Operationalise Machine Learning model

Earlier the notebooks were executed for data preparation and training of the Machine Learning model. The next step was to deploy the machine learning model. The pipeline created is connected to the notebook and connections are established with Azure Databricks. Here it is ensured that the copy activity must complete processing and store files in the storage account before notebook activity is execute. The work flow is triggered and pipeline activity is monitored for execution as show in figure 5.



Local time : Last 24 hours
Pipeline name : All
Status : All
Runs : Latest runs
Triggered by : All
Add filter

Showing 1 - 5 items

Last refreshed 1 minute ago

☐	Pipeline name	Run start	Run end	Duration	Triggered by	Status
☐	CopyOnPrem2AzurePipeline	4/12/2023, 7:00:53 PM	4/12/2023, 7:02:26 PM	00:01:33	Manual trigger	✓ Succee

Figure 5 - Pipeline is successfully executed

Summarize data

The notebook is executed where data is fetched using Spark SQL service and summary table is created and saved that will be used by Power BI for visualisation.

Visualize data

Power BI application is used to create visual representation of the selected data. Once Power BI is installed, it is connected to the cluster using JDBC server and Spark cluster connection in Power BI. After connection is established, the data is automatically fetched into Power BI application and it can be visualised from tables using drag/drop techniques utilising various data representation platforms. Figure 6 shows map visualisation of airport locations and number of delays. On the same window a stacked column chart is also used that shows days of month vs number of delays on a particular day. And the third visualisation uses a Tree map that shows Airport code vs number of delays. These three visuals are interactive and can respond in real time - Refer Figure 6.

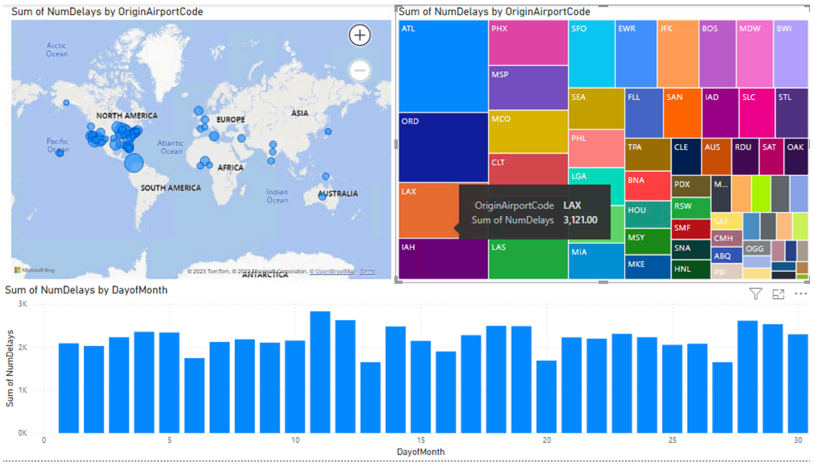


Figure 6 - Data visualisation in Power BI

Evaluation

The solution enables the airline company to scale their data processing and analysis capabilities as their needs grow. This flexibility ensures that the airline can adapt to changing business needs and requirements.

Secondly, the use of machine learning techniques allows the airline company to analyze historical data and identify patterns that contribute to flight delays. This information can be used to develop predictive models that provide insights into potential delays. This enables the airline company to preemptively inform their customers of potential delays, which can improve customer satisfaction.

Thirdly, the use of Power BI for data visualization and reporting provides the airline company with real-time insights into their operations. This enables them to make informed decisions about their business and improve their overall performance.

Suggestions

The Decision Tree Machine Learning Model used in the task resulted in only 62% accuracy which is quite low and not recommended for use in business decision making. Instead of Decision Tree, the preferred model would be Random Forest as it is an ensemble learning technique that combines multiple decision trees to improve accuracy of predictions. Support Vector Machines can also be used as it can handle high dimensionality data and works well with pre-processed data.

Further, more data cleansing techniques can be used for e.g. removing outliers. This will reduce noise in the data and the currently used decision tree algorithm could perform better.

References

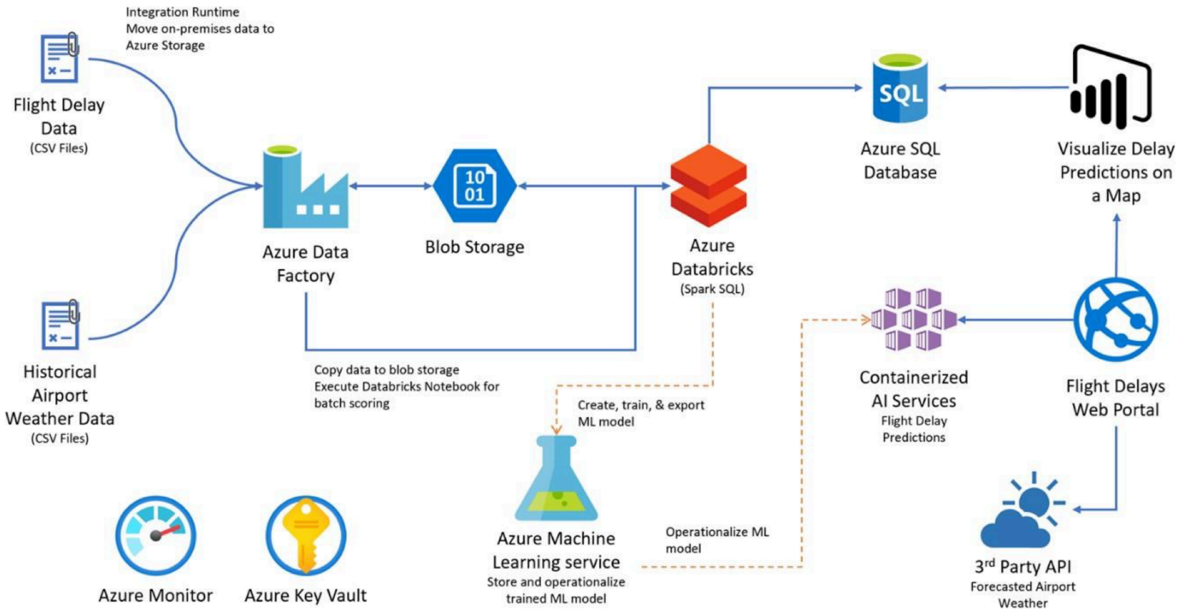
[1]C. Marketing, “What is Azure Data Factory?,” Convergytics, Mar. 05, 2020. https://convergytics.net/what-is-azure-data-factory/#:~:text=AzureDataFactoryisacloud-baseddataintegrationservicethatallowsyoutocreatedata-drivenworkflowsinthecloudfororchestratingandautomatingdatamovementanddatatransformation. (accessed Apr. 16, 2023).

[2]“What is DataBricks? | Definition from TechTarget,” WhatIs.com. https://www.techtarget.com/whatis/definition/DataBricks

[3]“What is Databricks: The Best Guide for Beginners 101,” Dec. 01, 2021. https://hevodata.com/learn/what-is-databricks/

Appendix

Figure 1 - Solution Architecture


Integration Runtime
Move on-premises data to Azure Storage
Flight Delay Data
(CSV Files)
Historical Airport Weather Data
(CSV Files)
Azure Data Factory
Blob Storage
10
01
Azure Databricks
(Spark SQL)
SQL
Azure SQL Database
Visualize Delay Predictions on a Map
Copy data to blob storage
Execute Databricks Notebook for batch scoring
Create, train, & export ML model
Containerized AI Services
Flight Delay Predictions
Flight Delays Web Portal
Azure Monitor
Azure Key Vault
Azure Machine Learning service
Store and operationalize trained ML model
Operationalize ML model
3rd Party API
Forecasted Airport Weather