Big Data and Visualization

Customer Scenario:

Margie's Travel (MT) provides concierge services for business travelers. In an increasingly crowded market, they are always looking for ways to differentiate themselves and provide added value to their corporate customers.

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?"

While there are many factors that customer service uses to tailor their guidance to the customer (such as cost and travel duration), MT believes an innovative solution might come in the form of giving the customer an assessment of the risk of encountering flight delays. The customer may choose to book with a narrower travel window for low-risk flights, giving them more precious time at home and less on the road spent arriving too early to a destination. 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.

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 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)

They receive updates to this data monthly, where the flight data and other related files total about 1 GB. In total, their solution currently manages about 2 TB worth of data.

Additionally, they receive current and forecasted weather data from a third-party service. This service gives them the ability to receive weather forecasts around any airport and provides forecasts for up to 10 days. They have a history of each flight's historical weather condition as CSV files, but acquiring the weather forecasts requires a call to a REST API that returns a JSON (JavaScript Object Notation) structure. Each airport of interest needs to be queried individually. An excerpt of the weather forecast for a single day at the Seattle-Tacoma International airport is as follows:


```
{
 "DATE": {
  "EPOCH": "1444701600",
  "PRETTY": "7:00 PM PDT ON OCTOBER 12, 2015",
  "DAY": 12,
  "MONTH": 10,
  "YEAR": 2015,
  "YDAY": 284,
  "HOUR": 19,
  "MIN": "00",
  "SEC": 0,
  "AMPM": "PM",
  "TZ_SHORT": "PDT",
  "TZ_LONG": "AMERICA/LOS_ANGELES"
 },
 "HIGH": {
  "FAHRENHEIT": "64",
  "CELSIUS": "18"
 },
 "LOW": {
  "FAHRENHEIT": "54",
  "CELSIUS": "12"
 },
 "CONDITIONS": "OVERCAST",
 "MAXWIND": {
  "MPH": 15,
  "KPH": 24,
  "DIR": "SSW",
  "DEGREES": 209
 },
 "AVEWIND": {
  "MPH": 10,
  "KPH": 16,
  "DIR": "SSW",
  "DEGREES": 209
 },
 "AVEHUMIDITY": 70,
 "MAXHUMIDITY": 0,
 "MINHUMIDITY": 0
}
```


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 long-standing relationship with Microsoft, he would like to see if Azure can meet his needs.

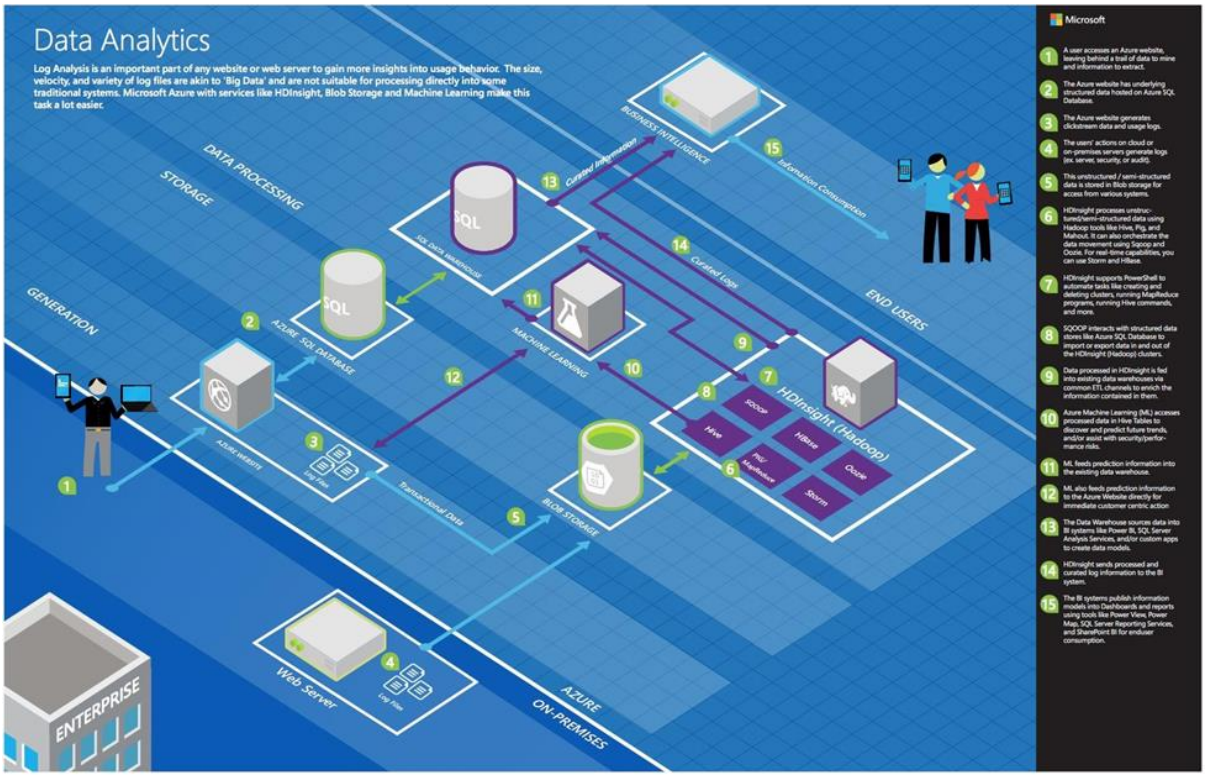
Customer Needs:

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?

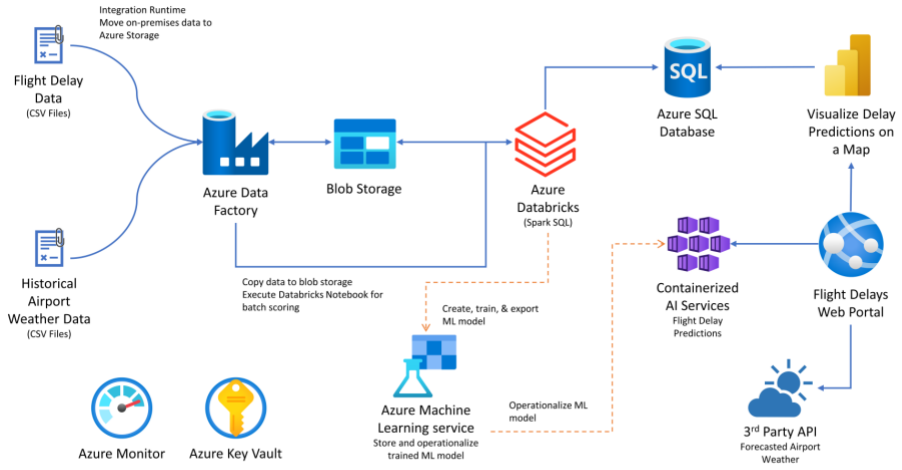
Common Scenarios:



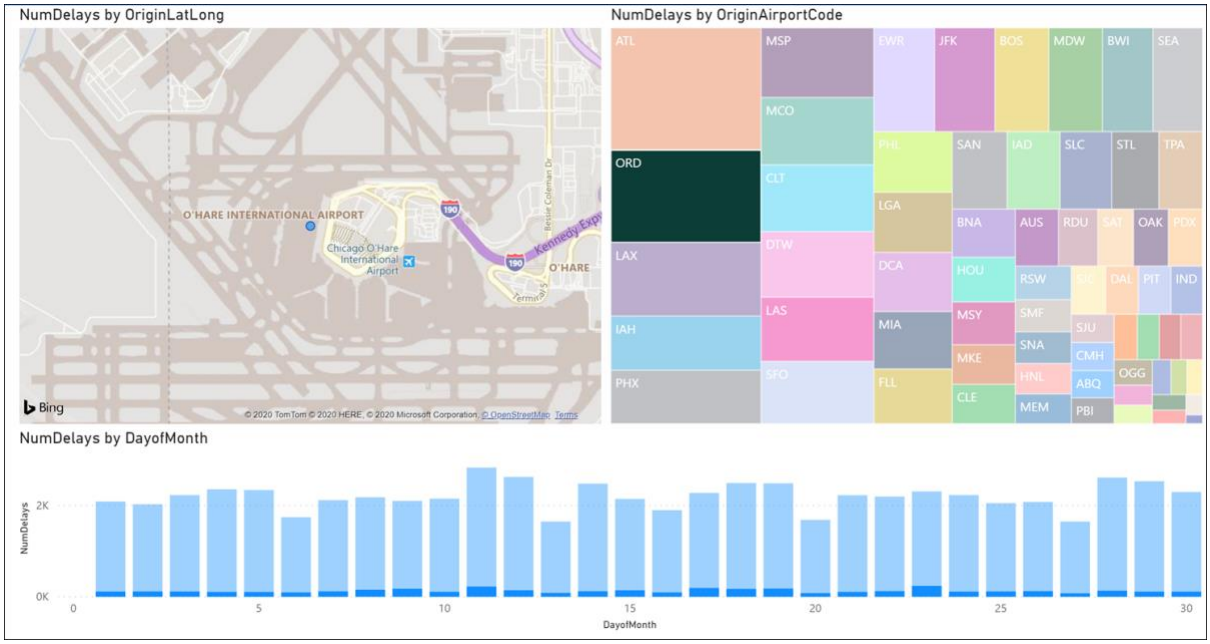
Your Challenge:

Design an deploy a solution to meet the customer needs. Be prepared to present your design documentation to a customer panel (15 mins), anytime during the assessment phase, the solution documents must contain:

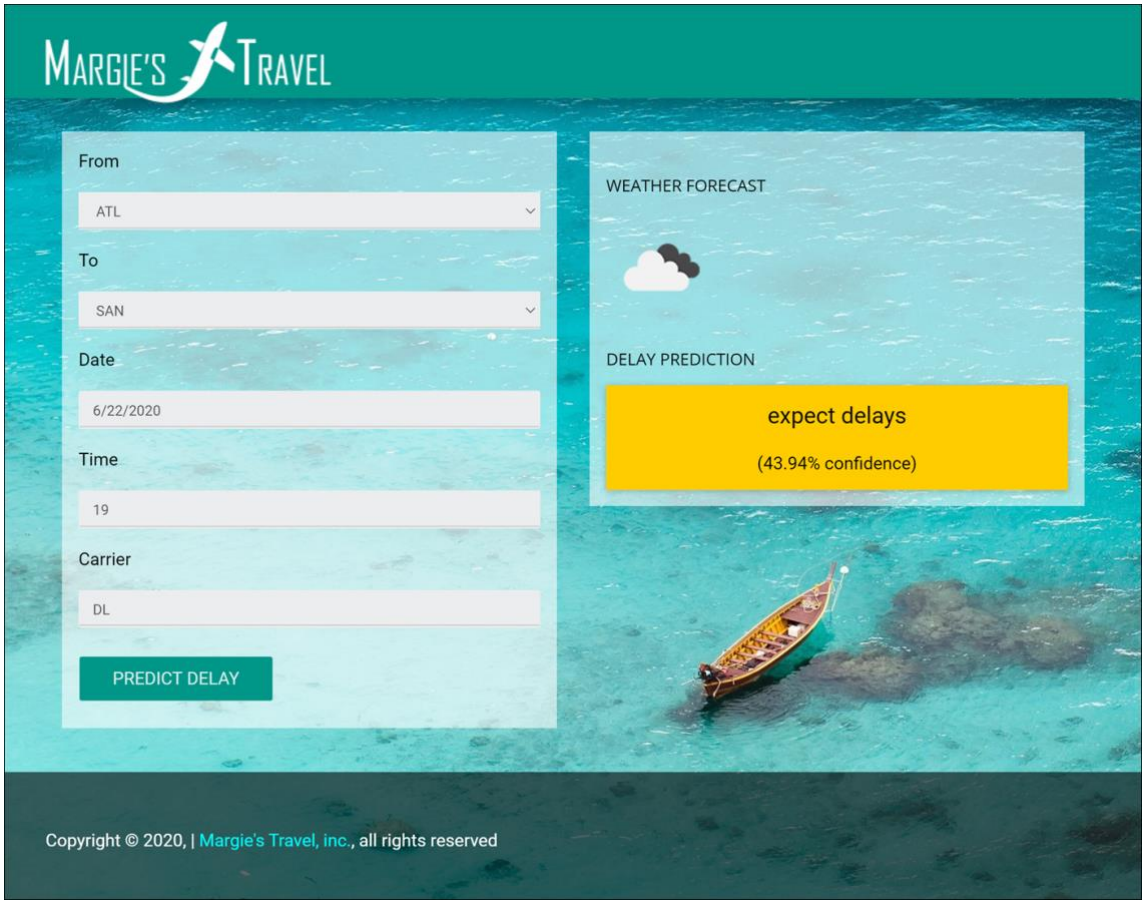
- Deploy the solution as per the below Architecture



- Data loading and Preparation
- Machine Learning Modelling
- Operationalizing Machine Learning
- Develop a Data Factory pipeline to for data movement
- Visualization and Reporting – Flight Delay Data over Map



- Deploy Customer application in Azure Application Service



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