Microsoft Cloud Workshop

Big data and visualization

Whiteboard design session student guide

January 2018

Information in this document, including URL and other Internet Web site references, is subject to change without notice. Unless otherwise noted, the example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious, and no association with any real company, organization, product, domain name, e-mail address, logo, person, place or event is intended or should be inferred. Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

The names of manufacturers, products, or URLs are provided for informational purposes only and Microsoft makes no representations and warranties, either expressed, implied, or statutory, regarding these manufacturers or the use of the products with any Microsoft technologies. The inclusion of a manufacturer or product does not imply endorsement of Microsoft of the manufacturer or product. Links may be provided to third party sites. Such sites are not under the control of Microsoft and Microsoft is not responsible for the contents of any linked site or any link contained in a linked site, or any changes or updates to such sites. Microsoft is not responsible for webcasting or any other form of transmission received from any linked site. Microsoft is providing these links to you only as a convenience, and the inclusion of any link does not imply endorsement of Microsoft of the site or the products contained therein.

© 2018 Microsoft Corporation. All rights reserved.

Microsoft and the trademarks listed at https://www.microsoft.com/en-us/legal/intellectualproperty/Trademarks/Usage/General.aspx are trademarks of the Microsoft group of companies. All other trademarks are property of their respective owners.

Contents

Big data and visualization whiteboard design session student guide

Abstract and learning objectives

In this workshop, you will complete a web app using Machine Learning to predict travel delays given flight delay data and weather conditions, plan the bulk data import operation, followed by preparation tasks, such as cleaning and manipulating the data for testing, and training your Machine Learning model.

By attending this workshop, you will be better able to build a complete Azure Machine Learning (ML) model for predicting if an upcoming flight will experience delays. In addition, you will learn to:

- Integrate the Azure ML web service in a Web App for both one at a time and batch predictions
- Analyze batch data with SQL Data Warehouse
- Visualize batch predictions on a map using Power BI

This whiteboard design session is designed to provide exposure to many of Microsoft's transformative line of business applications built using Microsoft big data and advanced analytics. The goal is to show an end-to-end solution, leveraging many of these technologies, but not necessarily doing work in every component possible. The architecture includes:

- Azure Machine Learning (Azure ML)
- Azure Data Factory (ADF)
- Azure Storage
- HDInsight Spark
- Power BI Desktop
- Azure App Service

©2018 Microsoft Corporation

Step 1: Review the customer case study

Outcome

Analyze your customer's needs.

Facilitator/subject matter expert (SME) presentation of customer case study

Timeframe: 15 minutes

Directions: With all participants in the session, the facilitator/SME presents an overview of the customer case study along with technical tips.

1. Meet your table participants and trainer.
2. Read all of the directions for Steps 1–3 in the Student guide.
3. As a table team, review the following customer case study.

©2018 Microsoft Corporation

Customer situation

AdventureWorks Travel (AWT) 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.

AWT is investigating ways that they can 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 7 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), AWT believes an innovative solution might come in the form of giving the customer an assessment of the risk of encountering flight delays. For low risk flights, the customer may choose to book with a narrower travel window, giving them more precious time at home and less on the road spent arriving too early to a destination. AWT is interested in applying data science to the problem to discover if the weather forecast coupled with their historical flight delay data could be used to provide a meaningful input into the customer's decision-making process.

AWT plans to pilot this solution internally, whereby the small population of customer support who service AWT'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.

AWT 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 up to 10 days. They have a history of the historical weather condition for each flight 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",
```

 ©2018 Microsoft Corporation

```
      "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 AWT, 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 datacenter 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.

©2018 Microsoft Corporation

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 from which 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 machine learning model that takes as input their historical data on flight delays and weather conditions in order to identify whether a flight is likely to be delayed or not.
6. Need web-based visualizations of the flight delay predictions.

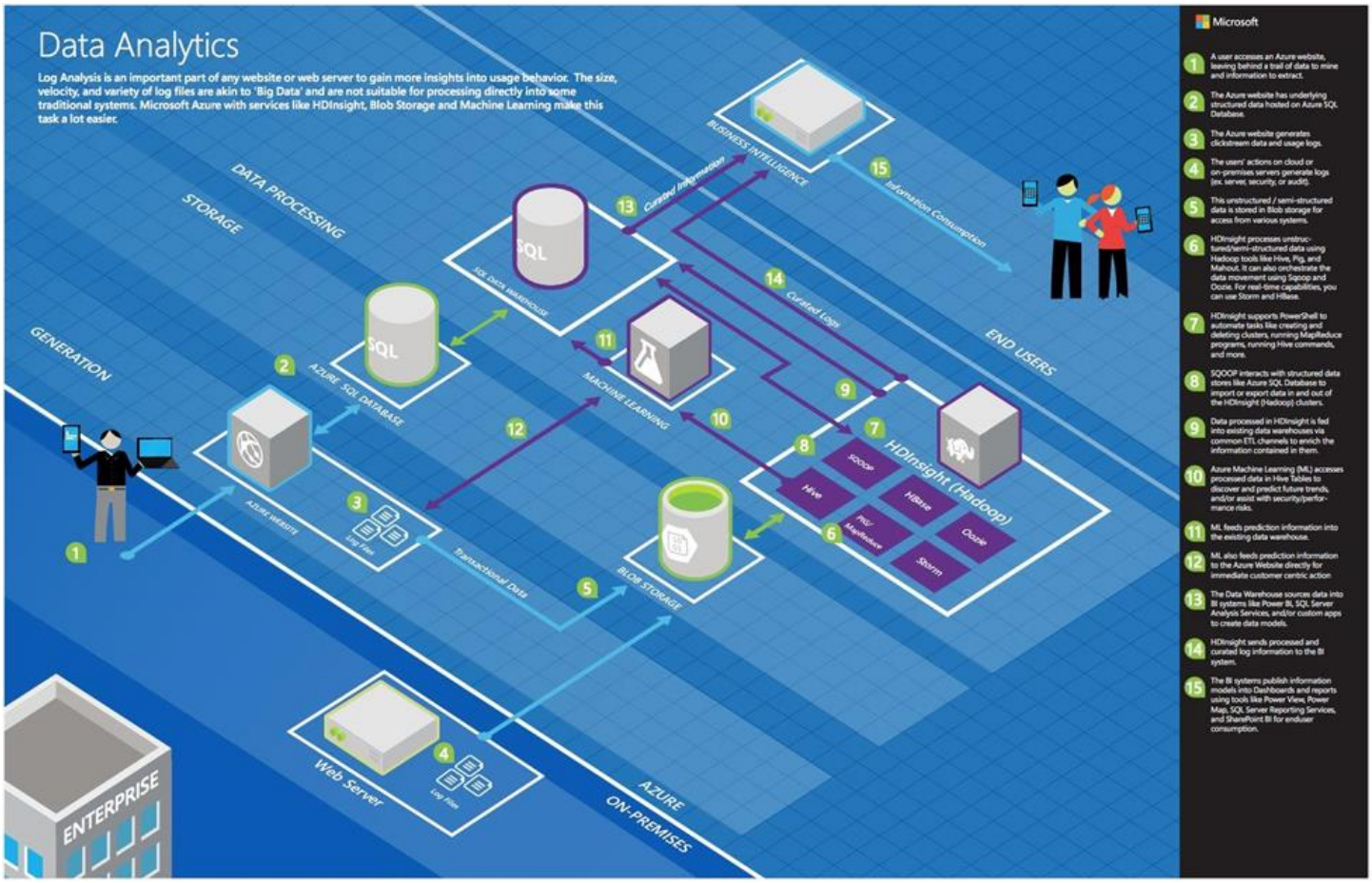
©2018 Microsoft Corporation

Customer objections

1. Does Azure offer a machine learning solution that does not require a PhD in statistics?
2. 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?
3. Can we query flat files in the file system using SQL?
4. Does Azure provide anything that would speed up querying (and exploration) of files in HDFS?
5. Does Azure provide any tools for visualizing our data? Ideally access to these could be managed with Active Directory.
6. While our Proof of Concept (PoC) does not have any sensitive data, if it is successful we would like to include customer data that contains personally identifiable information (PII) and transaction history so we could achieve new insights combining our flight delay predictions with our customers' profiles. Are there any additional services in the Azure Marketplace we could use to identify data loaded that contains PII, monitor access to sensitive data, and protect the data at rest (via encryption or masking)?
7. Is HDInsight 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'd like our operationalized models to be flexible in the inputs they support. In some cases, we want to provide both the flight and weather data to get a prediction. In others we just want to provide flight data and have the weather looked up. Is this possible?

©2018 Microsoft Corporation

Infographic for common scenarios



Step 2: Design a proof of concept solution

Outcome

Prepare to present a solution to the target customer audience in a 15-minute chalk-talk format.

Timeframe: 60 minutes

Business needs

Directions: With all participants at your table, answer the following questions and list the answers on a flip chart.

1. Who should you present this solution to? Who is your target customer audience? Who are the decision makers?
2. What customer business needs do you need to address with your solution?

Design

Directions: With all participants at your table, respond to the following questions on a flip chart.

High-level architecture

1. Without getting into the details (the following sections will address the details), diagram your initial vision for handling the top-level requirements for data loading, data preparation, storage, machine learning modeling, and reporting. You will refine this diagram as you proceed.

Data loading

1. How would you recommend that AWT get their historical data into Azure? What services would you suggest and what are the specific steps they would need to take to prepare the data, to transfer the data, and where would the loaded data land?
2. Update your diagram with the data loading process with the steps you identified.

Data preparation

1. What service would you recommend AWT capitalize on to explore the flat files they get from the USDOT using SQL?
2. If you suggested HDInsight, what specific configuration would you use? What components of the Hadoop stack would you use to allow AWT analysts to query and prep the data? How would they author and execute these data prep tasks?
3. If you suggested SQL Data Warehouse (DW), explain how you would configure the SQL DW instance.
4. Why did you recommend HDInsight over SQL Data Warehouse or vice versa?
5. How would you suggest AWT integrate weather forecast data?

Machine learning modeling

1. What technology would you recommend that AWT use for implementing their machine learning model?
2. How would you guide AWT to load data, so it can be processed by the machine learning model?
3. What category of machine learning algorithm would you recommend to AWT for use in constructing their model? For this scenario your option is clustering, regression or two-class classification. Why?
4. Assuming you selected an algorithm that requires training, address the following model design questions:
    a. What is the high-level flow of your machine learning model? Diagram this.

 ©2018 Microsoft Corporation

   b. What attributes of the flight and weather data do you think AWT should use in predicting flight delays? How would you recommend that AWT identify the columns that provide the most predictive value in determining if a flight will be delayed? Be specific on the particular modules or libraries they could use and how they would apply them against the data.
   c. Some of the data may need a little touching up: columns need to be removed, data types need to be changed. How would these steps be applied in your model?
   d. How would you recommend AWT measure the success of their model?

Operationalizing machine learning

1. How can AWT release their model for production use and avoid their concerns about extremely long delays operationalizing the model? Be specific on how your model is packaged, hosted, and invoked.
2. AWT has shown interest in not only scoring a flight at a time (based on a customer's request), but also doing scoring in large chunks so that they could show summaries of predicted flight delays across the United States. What changes would you need to make to your ML model to support this?

Visualization and reporting

1. Is Power BI an option for AWT to use in visualizing the flight delays?
2. If so, explain:
   a. How would AWT load the data and plot it on a map? What specific components would you use and how would you configure them to display the data?
   b. If they need to make minor changes, such as a change to the data types of a column in the model, how would they perform this in Power BI?
   c. How could they secure access to these reports to only their internal customer service agents?

Prepare

Directions: With all participants at your table:

1. Identify any customer needs that are not addressed with the proposed solution.
2. Identify the benefits of your solution.
3. Determine how you will respond to the customer's objections.

Prepare a 15-minute chalk-talk style presentation to the customer.

©2018 Microsoft Corporation

Step 3: Present the solution

Outcome

Present a solution to the target customer audience in a 15-minute chalk-talk format.

Presentation

Timeframe: 30 minutes

Directions

1. Pair with another table.
2. One table is the Microsoft team and the other table is the customer.
3. The Microsoft team presents their proposed solution to the customer.
4. The customer makes one of the objections from the list of objections.
5. The Microsoft team responds to the objection.
6. The customer team gives feedback to the Microsoft team.
7. Tables switch roles and repeat Steps 2–6.

©2018 Microsoft Corporation

Wrap-up

Timeframe: 15 minutes

- Tables reconvene with the larger group to hear a SME share the preferred solution for the case study.

©2018 Microsoft Corporation

Additional references

Item	Description	Links
Infographic	Hi-resolution version of data analytics blueprint	https://msdn.microsoft.com/dn630664 -fbid=rVymR_3WSRo
Machine Learning	Azure ML algorithm cheat sheet	https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/
Azure Data Factory	What is Azure Data Factory?	https://docs.microsoft.com/azure/data-factory/introduction
HDInsight Spark	Overview: Apache Spark on HDInsight	https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-overview/
Power BI	Power BI overview	https://support.powerbi.com/knowledgebase/articles/430814-get-started-with-power-bi
Travel data	Sample data source Bureau of Transportation Statistics, United States Department of Transportation Database: Airline On-Time Performance Data Table: On-Time Performance Table	http://www.transtats.bts.gov/Tables.asp?DB_ID=120
Weather data	Sample REST API for weather forecasts	http://www.wunderground.com/weather/api/d/docs
ARM Templates	Understand the structure and syntax of ARM templates	https://docs.microsoft.com/azure/azure-resource-manager/resource-group-authoring-templates

©2018 Microsoft Corporation