Big data and visualization

Whiteboard design session trainer guide

April 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

[Trainer information 1](#_Toc510622896)

[Big data and visualization whiteboard design session student guide 4](#_Toc510622897)

[Abstract and learning objectives 4](#_Toc510622898)

[Step 1: Review the customer case study 5](#_Toc510622899)

[Step 2: Design a proof of concept solution 12](#_Toc510622900)

[Step 3: Present the solution 14](#_Toc510622901)

[Wrap-up 15](#_Toc510622902)

[Additional references 16](#_Toc510622903)

[Big data and visualization whiteboard design session trainer guide 17](#_Toc510622904)

[Step 1: Review the customer case study 17](#_Toc510622905)

[Step 2: Design a proof of concept solution 17](#_Toc510622906)

[Step 3: Present the solution 17](#_Toc510622907)

[Wrap-up 17](#_Toc510622908)

[Preferred target audience 18](#_Toc510622909)

[Preferred solution 19](#_Toc510622910)

[Checklist of preferred objection handling 27](#_Toc510622911)

[Customer quote (to be read back to the attendees at the end) 29](#_Toc510622912)

# Trainer information

Thank you for taking time to support the whiteboard design sessions as a trainer!

#### Role of the trainer

An amazing trainer:

* Creates a safe environment in which learning can take place.
* Stimulates the participant’s thinking.
* Involves the participant in the learning process.
* Manages the learning process (on time, on topic, and adjusting to benefit participants).
* Ensures individual participant accountability.
* Ties it all together for the participant.
* Provides insight and experience to the learning process.
* Effectively leads the whiteboard design session discussion.
* Monitors quality and appropriateness of participant deliverables.
* Effectively leads the feedback process.

#### Whiteboard design session flow

Each whiteboard design session uses the following flow:

**Step 1: Review the customer case study (15 minutes)**

Outcome: Analyze your customer’s needs

* Customer’s background, situation, needs and technical requirements
* Current customer infrastructure and architecture
* Potential issues, objectives and blockers

**Step 2: Design a proof of concept solution (60 minutes)**

Outcome: Prepare to present a solution for your target customer audience

* Determine your target customer audience
* Determine customer’s business needs to address your solution
* Design and diagram your solution
* Prepare to present your solution

**Step 3: Present the solution (30 minutes)**

Outcome: Present solution to your customer

* Present solution
* Respond to customer objections
* Receive feedback

**Wrap-up (15 minutes)**

* Review preferred solution

#### Before the whiteboard design session: How to prepare

Before conducting your first whiteboard design session:

* Read the Student guide (including the case study) and Trainer guide.
* Become familiar with all key points and activities.
* Plan the point you want to stress, which questions you want to drive, transitions, and be ready to answer questions.
* Prior to the whiteboard design session, discuss the case study to pick up more ideas.
* Make notes for later.

#### During the whiteboard design session: Tips for an effective whiteboard design session

**Refer to the Trainer guide** to stay on track and observe the timings.

**Do not expect to memorize every detail** of the whiteboard design session.

When participants are doing activities, you can **look ahead to refresh your memory**.

* **Adjust activity and whiteboard design session pace** as needed to allow time for presenting, feedback, and sharing.
* **Add examples, points, and stories** from your own experience. Think about stories you can share that help you make your points clearly and effectively.
* **Consider creating a “parking lot”** to record issues or questions raised that are outside the scope of the whiteboard design session or can be answered later. Decide how you will address these issues, so you can acknowledge them without being derailed by them.

***Have fun****! Encourage participants to have fun and share!*

**Involve your participants.** Talk and share your knowledge but always involve your participants, even while you are the one speaking.

**Ask questions** and get them to share to fully involve your group in the learning process.

**Ask first**, whenever possible. Before launching into a topic, learn your audience’s opinions about it and experiences with it. Asking first enables you to assess their level of knowledge and experience, and leaves them more open to what you are presenting.

**Wait for responses**. If you ask a question such as, “What’s your experience with (fill in the blank)?” then wait. Do not be afraid of a little silence. If you leap into the silence, your participants will feel you are not serious about involving them and will become passive. Give participants a chance to think, and if no one answers, patiently ask again. You will usually get a response.

# Big data and visualization whiteboard design session student guide

## Abstract and learning objectives

In this workshop, you will deploy a web app using Machine Learning (ML) to predict travel delays given flight delay data and weather conditions. Plan a bulk data import operation, followed by preparation, 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
* Use Azure Data Factory (ADF) for data movement and operationalizing ML scoring
* Summarize data with HDInsight and Spark SQL
* 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

## 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.

#### 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",

"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.

#### 

#### 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 (PoC) 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.

#### 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 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. 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?

#### Infographic for common scenarios

The Data Analytics diagram is broken into three sections: On-Premises, Azure, and End Users. On-Premises includes a Web Server with log files, and an end user with a computer and a portable device. The Azure section includes three parts: Generation (Azure website and log files), Storage (Azure SQL Database and Blob Storage), and Data Processing (SQL Data Warehouse, Machine Learning, and HDInsight (Hadoop). The End Users section has Business Intelligence, and End Users with portable devices.

At this time, we are unable to capture all of the steps in the diagram. Future versions of this course should address this.

## Step 2: Design a proof of concept solution

**Outcome**

Design a solution and 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:
   1. What is the high-level flow of your machine learning model? Diagram this.
   2. 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.
   3. 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?
   4. 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:
   1. 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?
   2. 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?
   3. 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.

## Step 3: Present the solution

**Outcome**

Prepare to 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.

## Wrap-up

Timeframe: 15 minutes

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

## Additional references

|  |  |  |
| --- | --- | --- |
| **Item** | **Description** | **Links** |
| Infographic | Hi-resolution version of data analytics blueprint | [https://msdn.microsoft.com/dn630664 - fbid=rVymR\_3WSRo](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> |

# Big data and visualization whiteboard design session trainer guide

## Step 1: Review the customer case study

* Check in with your table participants to introduce yourself as the trainer.
* Ask, “What questions do you have about the customer case study?”
* Briefly review the steps and timeframes of the whiteboard design session.
* Ready, set, go! Let the table participants begin.

## Step 2: Design a proof of concept solution

* Check in with your tables to ensure that they are transitioning from step to step on time.
* Provide some feedback on their responses to the business needs and design.
  + Try asking questions first that will lead the participants to discover the answers on their own.
* Provide feedback for their responses to the customer’s objections.
  + Try asking questions first that will lead the participants to discover the answers on their own.

## Step 3: Present the solution

* Determine which table will be paired with your table before Step 3 begins.
* For the first round, assign one table as the Microsoft team and the other table as the customer.
* Have the Microsoft team present their solution to the customer team.
  + Have the customer team provide one objection for the Microsoft team to respond to.
  + The presentation and objections should be no longer than 10 minutes.
* Have participants on the customer team give feedback to the Microsoft team.
  + The feedback should be no longer than 5 minutes.
  + If needed, the trainer may also provide feedback.

## Wrap-up

* Have the table participants reconvene with the larger session group to hear a SME share the following preferred solution.

## Preferred target audience

Jack Tradewinds, CIO of AdventureWorks Travel (AWT)

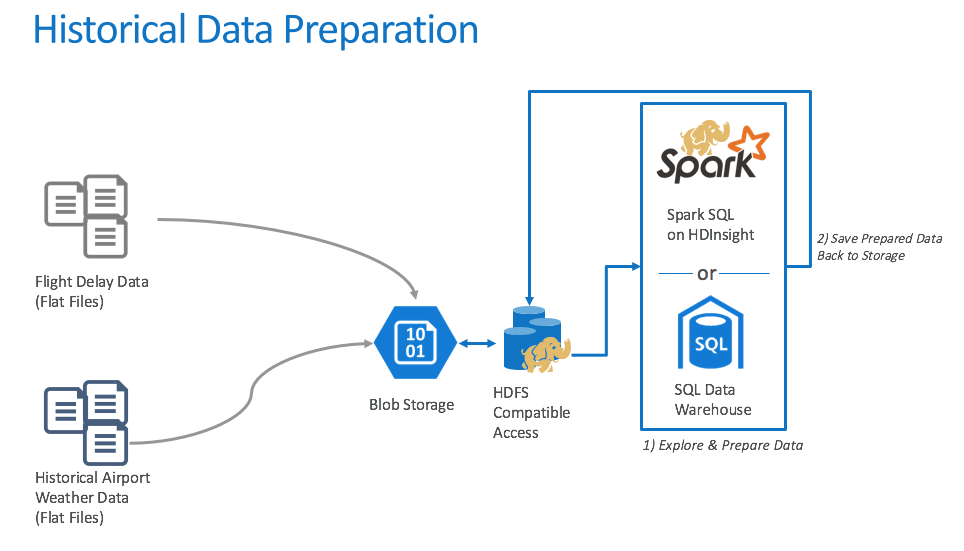
The primary audience is the business decision makers and technology decision makers. From the case study scenario, this would include the Director of Analytics. Usually we talk to the infrastructure managers who report to the chief information officers (CIOs), or to application sponsors (like a vice president [VP] line of business [LOB], or chief marketing officer [CMO]), or to those that represent the business unit IT or developers that report to application sponsors.

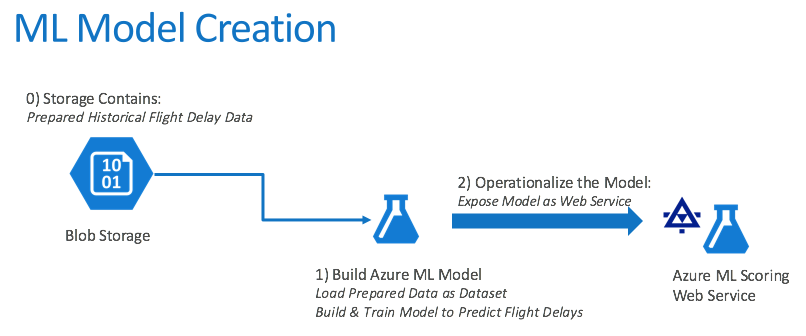
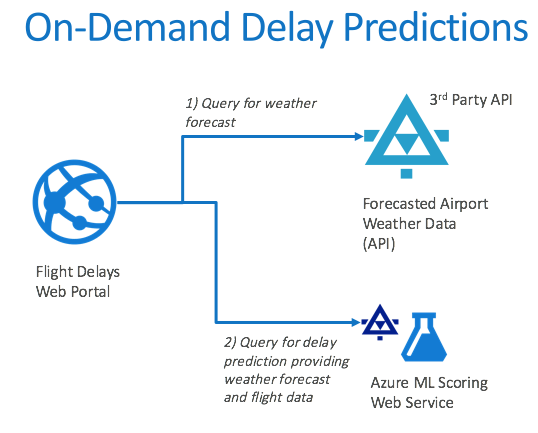
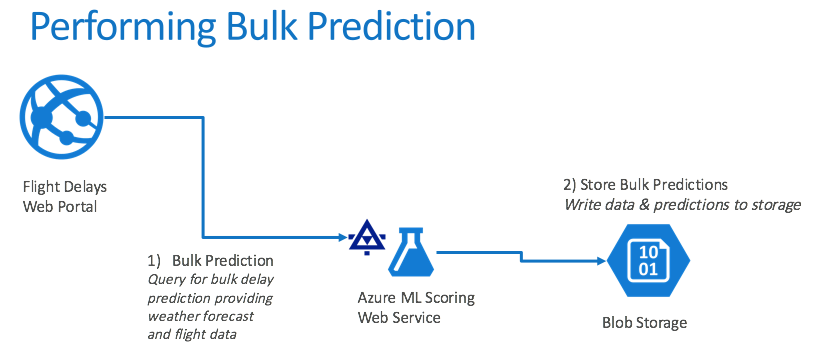
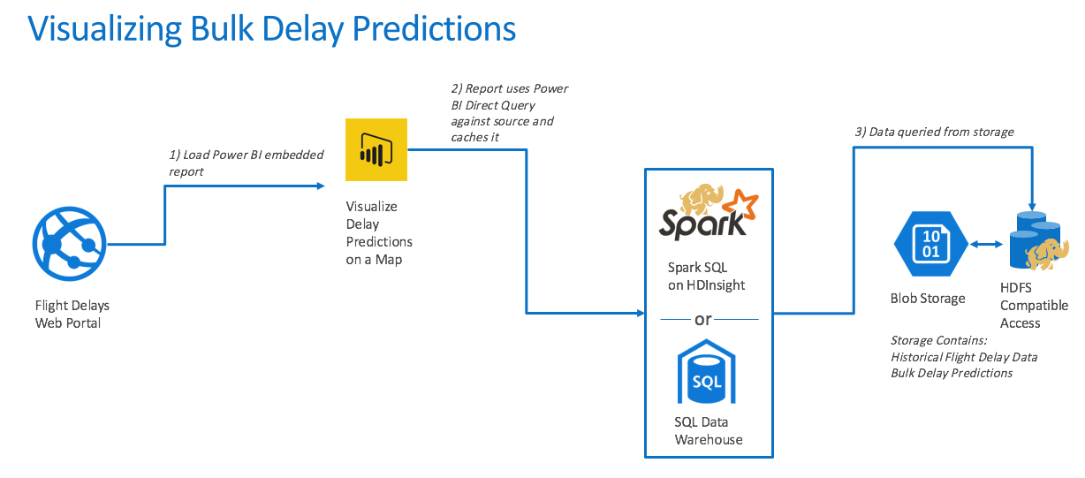
## Preferred solution

*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.

After speaking with its supportive team at Microsoft, AWT decided that Azure would in fact be the right choice for their platform. They decided to load data into blob storage; explore and prepare it using Spark SQL on HDInsight; train a model using Machine Learning; operationalize the model as a web service; and visualize the result using a map visualization in Power BI.

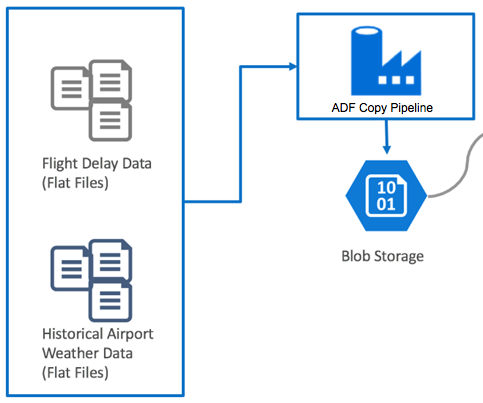


  
  
  
  
*NOTE: The preferred solution is only one of many possible, viable approaches.*

*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?

AWT should consider using Azure Data Factory (ADF) for copying their historical data into Azure. By setting up a continuous pipeline containing a copy activity configured to copy time partitioned source data, they could pull all their historical information, as well as ingest any future data, into Azure blob storage through a scheduled, and continuously running pipeline. Because their historical data is stored on-premises, AWT would need to install and configure an Azure Data Factory Integration Runtime (formerly known as a Data Management Gateway). Once in place, this would allow ADF to copy data from their local data store to a container is blob storage. Their pipeline would be configured to run monthly, as that is the frequency at which new data is received, and this would still allow for all their historical data to be copied without delay.

1. 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 United States Department of Transportation (USDOT) using SQL?

They could use either HDInsight or SQL Data Warehouse (SQL DW), as both provide support for SQL over HDFS (Hadoop Distributed File System).

1. 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?

They should use Spark SQL. They would need to create the appropriate external tables atop the flight delay files available in HDFS, which reads from the Azure Storage account attached to their HDInsight instance.

1. If you suggested SQL Data Warehouse, explain how you would configure the SQL DW instance.

They could query the files in HDFS using PolyBase. They would need to create the appropriate external tables atop the flight delay files available in HDFS, which live in the Azure Storage account that is registered with their SQL Data Warehouse instance.

1. Why did you recommend HDInsight over SQL Data Warehouse or vice versa?

All other items being equal, HDInsight and Spark SQL would be the preferred solution because the customer expressed interest in it. HDInsight allows developers to use either Scala or Python to easily work with files stored in Azure storage, while Spark SQL allows them to leverage the existing SQL skills. Using Spark SQL, they can write data to Hive tables in storage. These Hive tables are accessible via Power BI, allowing easy visualization and access to reporting on data scored through their Machine Learning models.

1. How would you suggest AWT integrate weather forecast data?

They could retrieve the weather forecast data from a third-party service that provides a REST API. An example of such a service is weatherunderground.com.

*Machine learning modeling*

1. What technology would you recommend that AWT use for implementing their machine learning model?

Azure Machine Learning (Azure ML).

1. How would you guide AWT to load data, so it can be processed by the machine learning model?

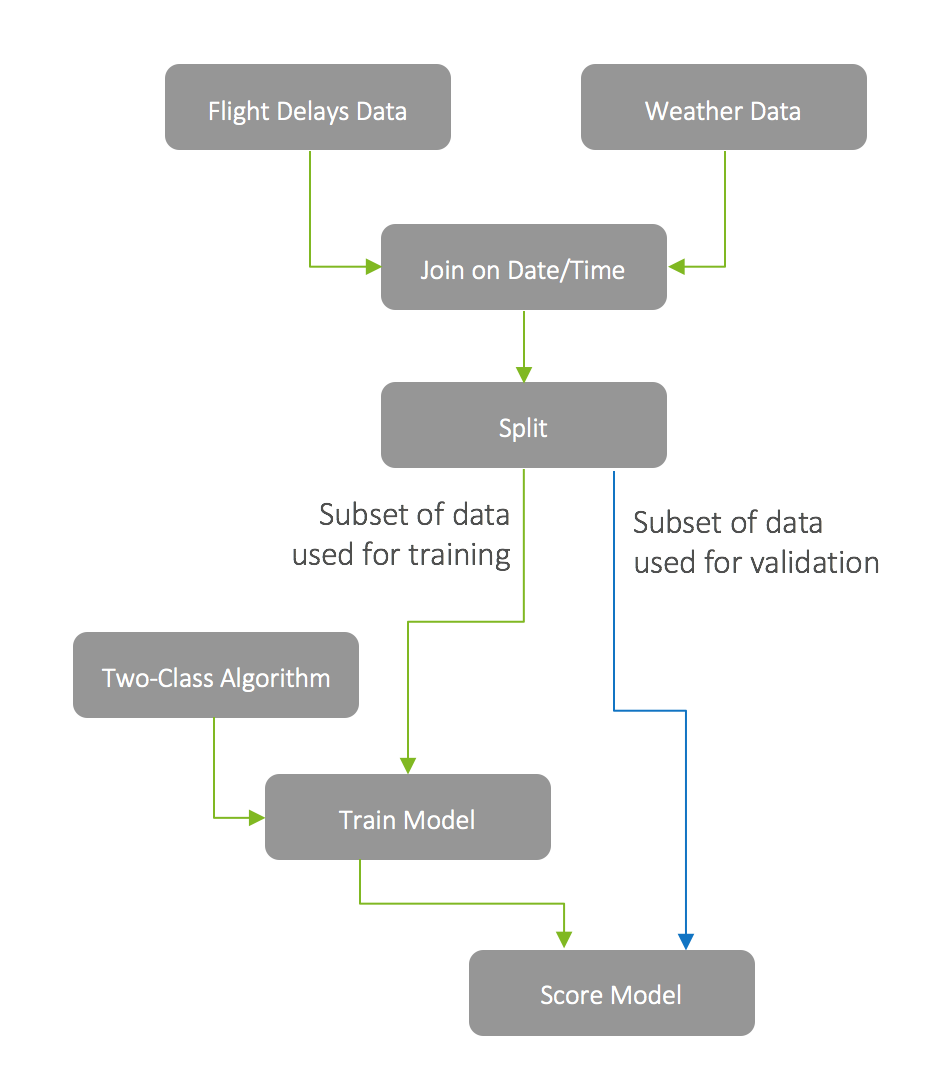
The data used for training would need to be uploaded into a dataset or accessed using the Reader module from Azure Storage blobs. Some amount of data preparation work would need to be done first, possibly using Spark to create a subset of the data that is used for training and validation.

1. What category of machine learning algorithm would you recommend to AWT for use in constructing their model? For this scenario, your options are clustering, regression or two-class classification. Why?

Given that AWT only wants a binary prediction of flight delayed or flight not delayed, they should proceed with a two-class classification algorithm.

1. Assuming you selected an algorithm that requires training, address the following model design questions:
   1. What is the high-level flow of your machine learning model? Diagram this.

The diagram could look like the following:



* 1. 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 modules or libraries they could use and how they would apply them against the data.

There are multiple approaches AWT could use to perform feature selection and to identify the data attributes that are the most helpful in accurately predicting a delay. They should start with any domain knowledge they have—this would likely point in the direction of the flight attribute’s airline, departure airport, destination airport, and time of day as well as the weather attribute’s wind speed, temperature, and precipitation conditions. Additionally, they should use the Evaluate Statistics module to identify and remove fields that do not add value (i.e., because they are mostly empty or only have a constant value). From there, they could construct a preliminary model and validate how it performs against the training data. In additional passes, they might choose to use the Feature Selection modules or the Principal Component Analysis module that statistically rank the fields that provide the greatest predictive value against the training data.

* 1. 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?

Data munging can be best accomplished using R or Python, languages familiar to data scientists and developers. These languages provide powerful data transformation capabilities, and allow for flexibility in how data cleanup occurs, while reducing the overall complexity of ML models. These can be expressed using Execute R and Execute Python Script modules.

* 1. How would you recommend AWT measure the success of their model?

For a classification model, they should measure the success of their model against a data set whereby a large portion of the data is used for training purposes, but a smaller set is reserved as examples that will be used to “test” the model to see how it performs against a known outcome. The output of the prediction as well as the “correct” output is passed to the Evaluate Model module which provides scores such as those from the Confusion Matrix that give a perspective on accuracy and in what situations the models makes mistakes. A more sophisticated validation model could involve using an ensemble approach, such as that provided by the Cross Validate Model module which takes multiple passes through the input data set to ascertain the performance.

*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.

They should operationalize their Azure ML Model by using ML Studio to create a Predictive Web Service around it. This will package the scoring function of the model so that it can be invoked via a REST call that takes the weather conditions and flight information as input and returns a response with the classification. The service is hosted by Azure ML.

1. 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?

After operationalizing the model, they would need to use Azure ML’s Batch Execution API. This can be implemented via Azure Data Factory by adding a Linked Service to the Azure ML batch request URL, and including an AzureMLBatchExecution activity in a continuous pipeline. By using ADF, the can apply the operational model to data as it is moved to the proper location in Azure storage. The scored results will be available in Blob storage after processing, and with the scheduled pipeline new data will be processed automatically, on the schedule indicated by the pipeline.

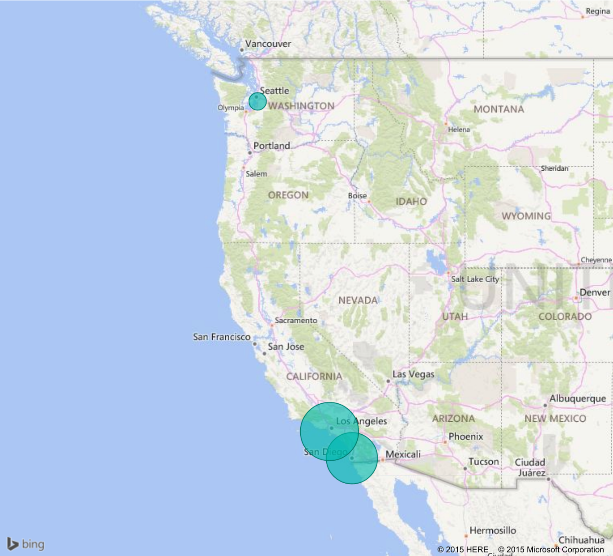
*Visualization and reporting*

1. Is Power BI an option for AWT to use in visualizing the flight delays?

Yes, Power BI is a good option for AWT. Power BI can perform what is called a Direct Query against HDInsight Spark hive data sources as well as an Import query that copies the data into Power BI managed datasets from Spark.

1. If so, explain:
   1. 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?

The data would be set up as a query against Spark that loads the predicted flight delay for each airport location. They would use the map visualization to display a delay indicator for a particular flight at a particular airport.



* 1. 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?

They could do this using the Query Editor component of the Power BI Desktop application. They could then upload this file to the Power BI service.

* 1. How could they secure access to these reports to only their internal customer service agents?

By utilizing the Power BI service, they can create a Content Pack that contains only the desired Dashboards, Reports, and Datasets and restrict access to those Groups in Azure Active Directory to which the customer service agents belong.

## Checklist of preferred objection handling

1. Does Azure offer a machine learning solution that does not require a PhD in statistics?

Yes. Azure Machine Learning provides tools that are powerful enough for the data scientist, but are approachable by the developer.

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?

This is true in the traditional process for creating machine learning models, whereby the data scientist creates a model (e.g., in R) and then hands it over to developers who translate it into Java or C#—which can take months to get the translation correct and performant. With Azure ML, operationalizing a model is done by converting a scoring experiment into a Web service, which takes only a few clicks.

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

Yes. There are many options for using a SQL syntax to query files in Blob storage such as SQL Data Warehouse and HDInsight with Spark SQL or Hive.

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

Yes. Hive on HDInsight provides the ability to create indices on flat file content, as does SQL Data Warehouse. Spark SQL on HDInsight provides a distributed in-memory cache that can aid exploratory queries that often repeat queries against the same subsets of data.

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

Power BI is available as a service and provides tools for creating both dashboards and reports whose access can be restricted by group membership in Azure Active Directory.

1. While our 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 customer’s profile. Are there any additional services in the Azure Marketplace we could use to apply data-centric security—that is to identify data loaded that contains PII, monitor access to sensitive data, and protect the data at rest (via encryption or masking)?

DataGuise provides the DGSecure product line that provides support for the automated detection of sensitive information at the time of data ingest by means of its DGSecure Ingest Agents (which include agents for relational databases, FTP, and Apache Flume). In addition, sensitive information can also be identified after it is stored in HDFS or Blob by using a Hadoop or HDInsight cluster to run periodic map/reduce scans for sensitive data on disk. The actions taken after discovery of sensitive data can range from flagging the data as sensitive (for monitoring and notification purposes), to modifying the data where the sensitive data is masked or encrypted and then updated in the data store. Finally, DGMonitor provides the ability to centralize monitoring of access to sensitive data, as well as raising alerts when suspicious activity is detected.  
Details of the DgSecure Architecture diagram are described in the previous text.

1. Is HDInsight our only option for running SQL on Hadoop solutions in Azure?

No. Hortonworks, Cloudera, and MapR provide fully supported distributions that are available in the Azure Marketplace and provide environments for running HiveQL Spark SQL in Azure.

1. 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.

There are two offerings related to Azure Data Lake that should be called out separately. Azure Data Lake Analytics provides a distributed querying engine whose costs are based solely on the number of resources used in the query, the amount of time those resources were made available to the query, and a nominal job completion charge. In its current form, Azure Data Lake Analytics supports only Batch workloads, so it is not the right tool for interactive analytic workloads.

Azure Data Lake Store provides the ability to store an unlimited number of items, each of unlimited size, having no upper limit on the total storage capacity. This makes it more capable than Azure Blob Storage since it removes the 500TB account size limit, the 1TB page blob size limit, and the ~4.75TB block blob limit imposed by Blob storage. Today, it is not possible to query data stored in Azure Data Lake Store from Spark running on HDInsight, although support for this is on the roadmap. In the long run, Azure Data Lake Store is where the customer should consider landing all of their data, but for the purposes of the PoC, using Azure Blob Storage is easier to implement because of its built-in support from the portal.

1. 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?

Yes. It is possible to add multiple web service input modules to the experiment, allowing you to add multiple sources, joining them at runtime or processing them separately.

## Customer quote (to be read back to the attendees at the end)

*“We are flying into the future with Azure, helping our customers more aggressively schedule their travel and optimize their non-travel time.”*

Jack Tradewinds, CIO of AdventureWorks Travel