

Contents

Team Data Science Process Documentation

Overview

Lifecycle

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer acceptance

Roles and tasks

Group manager

Team lead

Project lead

Individual contributor

Project planning

Development

Agile development

Collaborative coding with Git

Execute data science tasks

Code testing

Track progress

Operationalization

DevOps - CI/CD

Worked-out examples

Spark with PySpark and Scala

Explore and model data

Advanced data exploration and modeling

Score models

Hive with HDInsight Hadoop

U-SQL with Azure Data Lake

R, Python and T-SQL with SQL Server

T-SQL and Python with SQL DW

Utilities & tools

Data exploration & modeling utils (GitHub)

Training

For data scientists

For DevOps

How To

Set up data science environments

Azure storage accounts

Platforms and tools

R and Python on HDInsight clusters

Introduction to Spark on HDInsight

Create an Apache Spark cluster in Azure HDInsight

PySpark kernels for Jupyter Notebook

R Server on HDInsight

Get started using R Server on HDInsight

Azure Machine Learning workspace

Analyze business needs

Identify your scenario

Acquire and understand data

Ingest data

Overview

Move to/from Blob storage

Overview

Use Storage Explorer

Use AzCopy

Use Python

Use SSIS

Move to SQL on a VM

Move to Azure SQL Database

Move to Hive tables

[Move to SQL partitioned tables](#)

[Move from on-prem SQL](#)

[Explore and visualize data](#)

[Prepare data](#)

[Explore data](#)

[Overview](#)

[Explore Azure Blob storage](#)

[Explore SQL on a VM](#)

[Explore Hive tables](#)

[Sample data](#)

[Overview](#)

[Use blob storage](#)

[Use SQL Server](#)

[Use Hive tables](#)

[Process data](#)

[Access with Python](#)

[Process blob data](#)

[Use Azure Data Lake](#)

[Use SQL VM](#)

[Use data pipeline](#)

[Use Spark](#)

[Use Scala and Spark](#)

[Develop models](#)

[Engineer features](#)

[Overview](#)

[Use SQL+Python](#)

[Use Hive queries](#)

[Select features](#)

[Create and train models](#)

[Choose algorithms](#)

[Algorithm cheat sheet](#)

[Deploy models in production](#)

Related

[Azure Machine Learning](#)

[Microsoft Cognitive Toolkit - CNTK](#)

[Azure AI Gallery](#)

[Cortana Analytics](#)

[Anomaly detection](#)

[Cognitive Services](#)

[Predictive maintenance](#)

[Overview](#)

[Architecture](#)

[Technical guide](#)

Resources

[Blogs](#)

[Recent Updates for the TDSP](#)

[Using the TDSP in Azure Machine Learning](#)

[Data Science Utilities v.0.11 for the TDSP](#)

[Latest utilities for the TDSP](#)

[Data Science Project Planning Template](#)

[Cortana Intelligence and Machine Learning Blog](#)

[Organizations using TDSP](#)

[New Signature](#)

[Blue Granite](#)

[Related Microsoft resources](#)

[MSDN forum](#)

[Stack Overflow](#)

[Videos](#)

What is the Team Data Science Process?

2/26/2019 • 4 minutes to read

The Team Data Science Process (TDSP) is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently. TDSP helps improve team collaboration and learning. It contains a distillation of the best practices and structures from Microsoft and others in the industry that facilitate the successful implementation of data science initiatives. The goal is to help companies fully realize the benefits of their analytics program.

This article provides an overview of TDSP and its main components. We provide a generic description of the process here that can be implemented with a variety of tools. A more detailed description of the project tasks and roles involved in the lifecycle of the process is provided in additional linked topics. Guidance on how to implement the TDSP using a specific set of Microsoft tools and infrastructure that we use to implement the TDSP in our teams is also provided.

Key components of the TDSP

TDSP comprises of the following key components:

- A **data science lifecycle** definition
- A **standardized project structure**
- **Infrastructure and resources** for data science projects
- **Tools and utilities** for project execution

Data science lifecycle

The Team Data Science Process (TDSP) provides a lifecycle to structure the development of your data science projects. The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

If you are using another data science lifecycle, such as [CRISP-DM](#), [KDD](#) or your organization's own custom process, you can still use the task-based TDSP in the context of those development lifecycles. At a high level, these different methodologies have much in common.

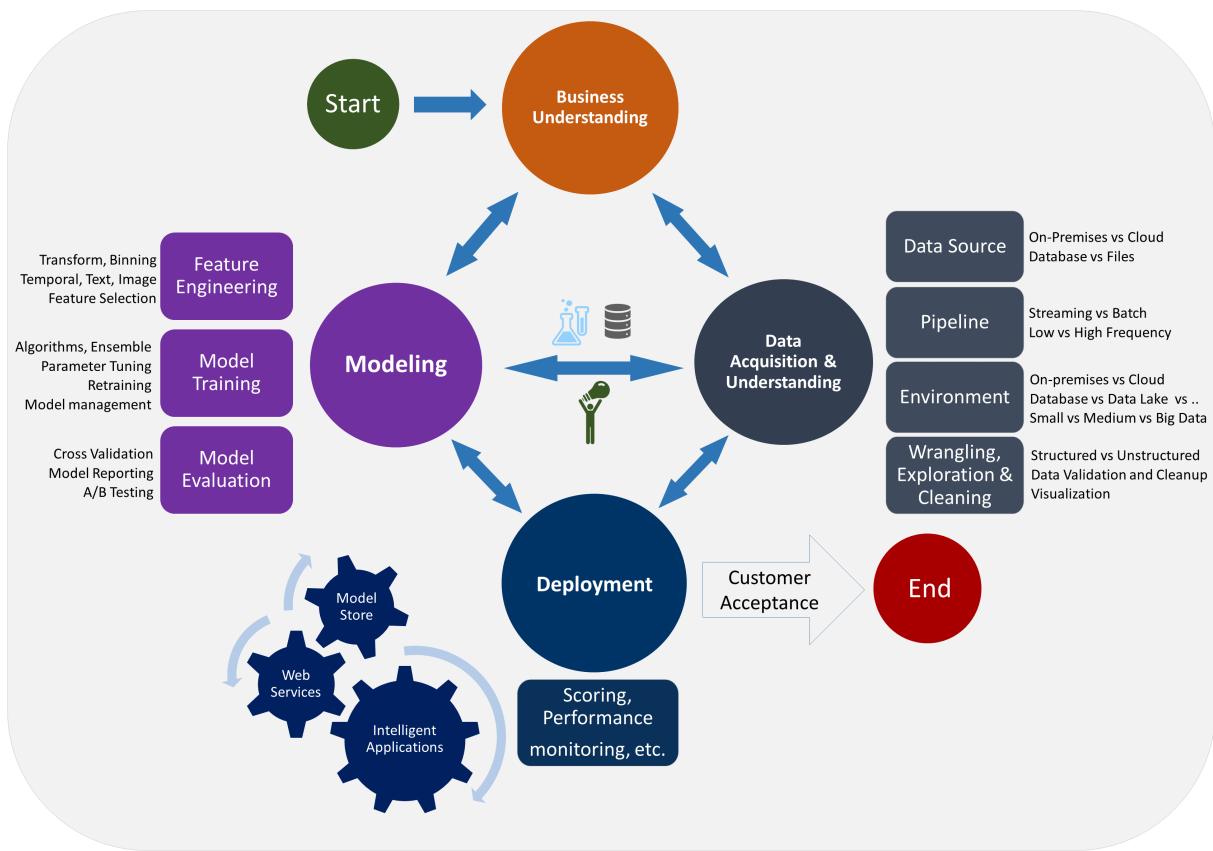
This lifecycle has been designed for data science projects that ship as part of intelligent applications. These applications deploy machine learning or artificial intelligence models for predictive analytics. Exploratory data science projects or ad hoc analytics projects can also benefit from using this process. But in such cases some of the steps described may not be needed.

The lifecycle outlines the major stages that projects typically execute, often iteratively:

- **Business Understanding**
- **Data Acquisition and Understanding**
- **Modeling**
- **Deployment**
- **Customer Acceptance**

Here is a visual representation of the **Team Data Science Process lifecycle**.

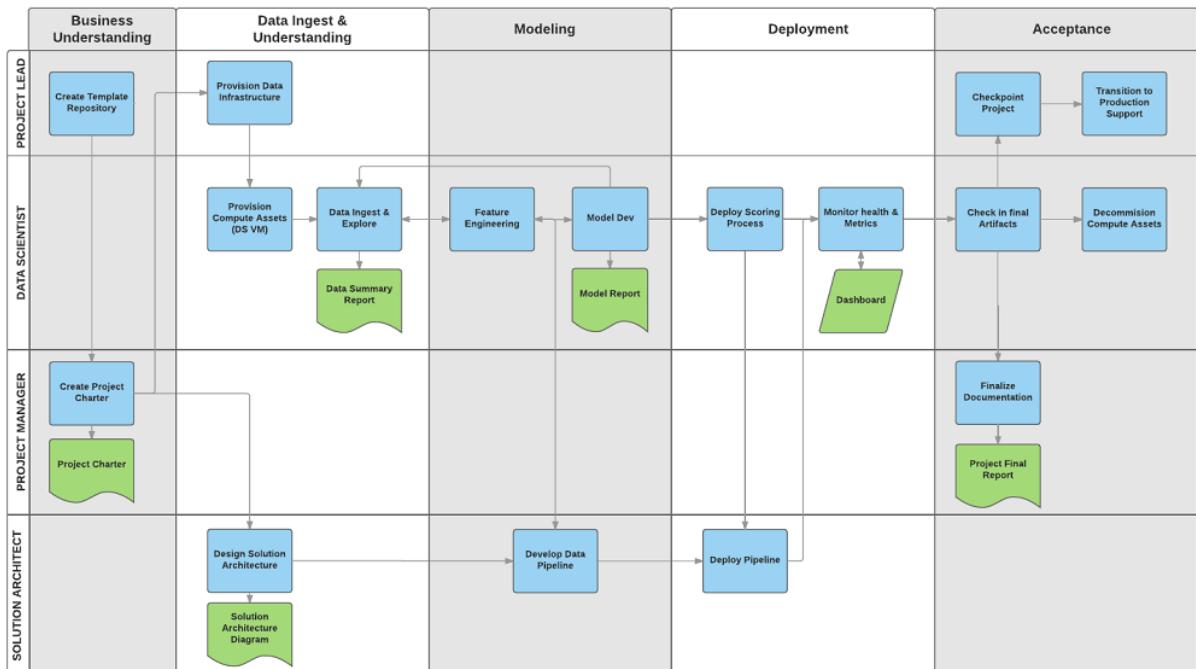
Data Science Lifecycle



The goals, tasks, and documentation artifacts for each stage of the lifecycle in TDSP are described in the [Team Data Science Process lifecycle](#) topic. These tasks and artifacts are associated with project roles:

- Solution architect
- Project manager
- Data scientist
- Project lead

The following diagram provides a grid view of the tasks (in blue) and artifacts (in green) associated with each stage of the lifecycle (on the horizontal axis) for these roles (on the vertical axis).

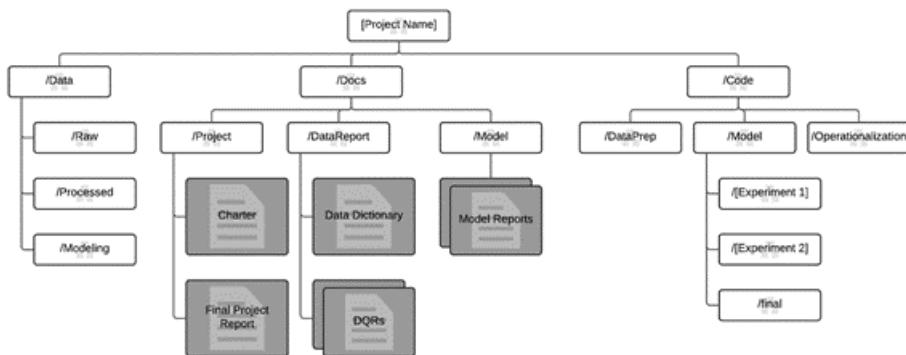


Standardized project structure

Having all projects share a directory structure and use templates for project documents makes it easy for the team members to find information about their projects. All code and documents are stored in a version control system (VCS) like Git, TFS, or Subversion to enable team collaboration. Tracking tasks and features in an agile project tracking system like Jira, Rally, and Azure DevOps allows closer tracking of the code for individual features. Such tracking also enables teams to obtain better cost estimates. TDSP recommends creating a separate repository for each project on the VCS for versioning, information security, and collaboration. The standardized structure for all projects helps build institutional knowledge across the organization.

We provide templates for the folder structure and required documents in standard locations. This folder structure organizes the files that contain code for data exploration and feature extraction, and that record model iterations. These templates make it easier for team members to understand work done by others and to add new members to teams. It is easy to view and update document templates in markdown format. Use templates to provide checklists with key questions for each project to insure that the problem is well-defined and that deliverables meet the quality expected. Examples include:

- a project charter to document the business problem and scope of the project
- data reports to document the structure and statistics of the raw data
- model reports to document the derived features
- model performance metrics such as ROC curves or MSE



The directory structure can be cloned from [GitHub](#).

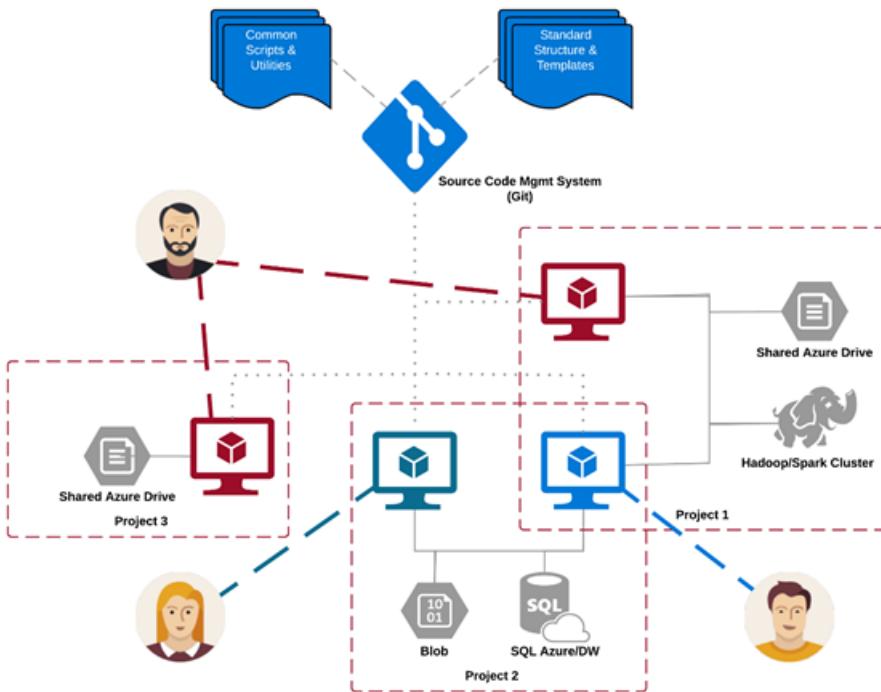
Infrastructure and resources for data science projects

TDSP provides recommendations for managing shared analytics and storage infrastructure such as:

- cloud file systems for storing datasets
- databases
- big data (Hadoop or Spark) clusters
- machine learning service

The analytics and storage infrastructure can be in the cloud or on-premises. This is where raw and processed datasets are stored. This infrastructure enables reproducible analysis. It also avoids duplication, which can lead to inconsistencies and unnecessary infrastructure costs. Tools are provided to provision the shared resources, track them, and allow each team member to connect to those resources securely. It is also a good practice to have project members create a consistent compute environment. Different team members can then replicate and validate experiments.

Here is an example of a team working on multiple projects and sharing various cloud analytics infrastructure components.



Tools and utilities for project execution

Introducing processes in most organizations is challenging. Tools provided to implement the data science process and lifecycle help lower the barriers to and increase the consistency of their adoption. TDSP provides an initial set of tools and scripts to jump-start adoption of TDSP within a team. It also helps automate some of the common tasks in the data science lifecycle such as data exploration and baseline modeling. There is a well-defined structure provided for individuals to contribute shared tools and utilities into their team's shared code repository. These resources can then be leveraged by other projects within the team or the organization. TDSP also plans to enable the contributions of tools and utilities to the whole community. The TDSP utilities can be cloned from [GitHub](#).

Next steps

[Team Data Science Process: Roles and tasks](#) Outlines the key personnel roles and their associated tasks for a data science team that standardizes on this process.

The Team Data Science Process lifecycle

1/30/2019 • 2 minutes to read

The Team Data Science Process (TDSP) provides a recommended lifecycle that you can use to structure your data-science projects. The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed. If you use another data-science lifecycle, such as the Cross Industry Standard Process for Data Mining ([CRISP-DM](#)), Knowledge Discovery in Databases ([KDD](#)), or your organization's own custom process, you can still use the task-based TDSP.

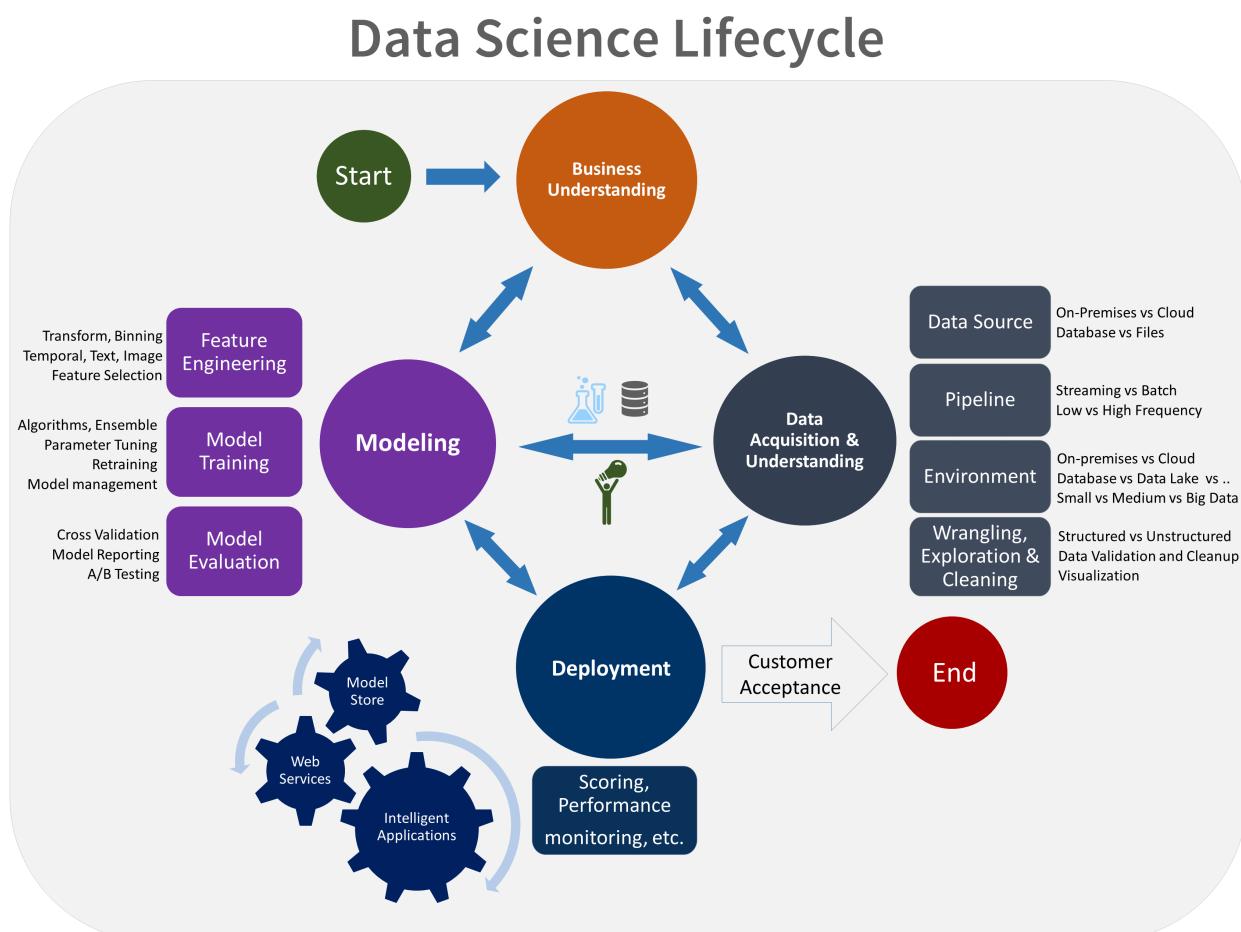
This lifecycle is designed for data-science projects that are intended to ship as part of intelligent applications. These applications deploy machine learning or artificial intelligence models for predictive analytics. Exploratory data-science projects and ad hoc analytics projects can also benefit from the use of this process. But for those projects, some of the steps described here might not be needed.

Five lifecycle stages

The TDSP lifecycle is composed of five major stages that are executed iteratively. These stages include:

1. [Business understanding](#)
2. [Data acquisition and understanding](#)
3. [Modeling](#)
4. [Deployment](#)
5. [Customer acceptance](#)

Here is a visual representation of the TDSP lifecycle:



The TDSP lifecycle is modeled as a sequence of iterated steps that provide guidance on the tasks needed to use predictive models. You deploy the predictive models in the production environment that you plan to use to build the intelligent applications. The goal of this process lifecycle is to continue to move a data-science project toward a clear engagement end point. Data science is an exercise in research and discovery. The ability to communicate tasks to your team and your customers by using a well-defined set of artifacts that employ standardized templates helps to avoid misunderstandings. Using these templates also increases the chance of the successful completion of a complex data-science project.

For each stage, we provide the following information:

- **Goals:** The specific objectives.
- **How to do it:** An outline of the specific tasks and guidance on how to complete them.
- **Artifacts:** The deliverables and the support to produce them.

Next steps

We provide full end-to-end walkthroughs that demonstrate all the steps in the process for specific scenarios. The [Example walkthroughs](#) article provides a list of the scenarios with links and thumbnail descriptions. The walkthroughs illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

For examples of how to execute steps in TDSPs that use Azure Machine Learning Studio, see [Use the TDSP with Azure Machine Learning](#).

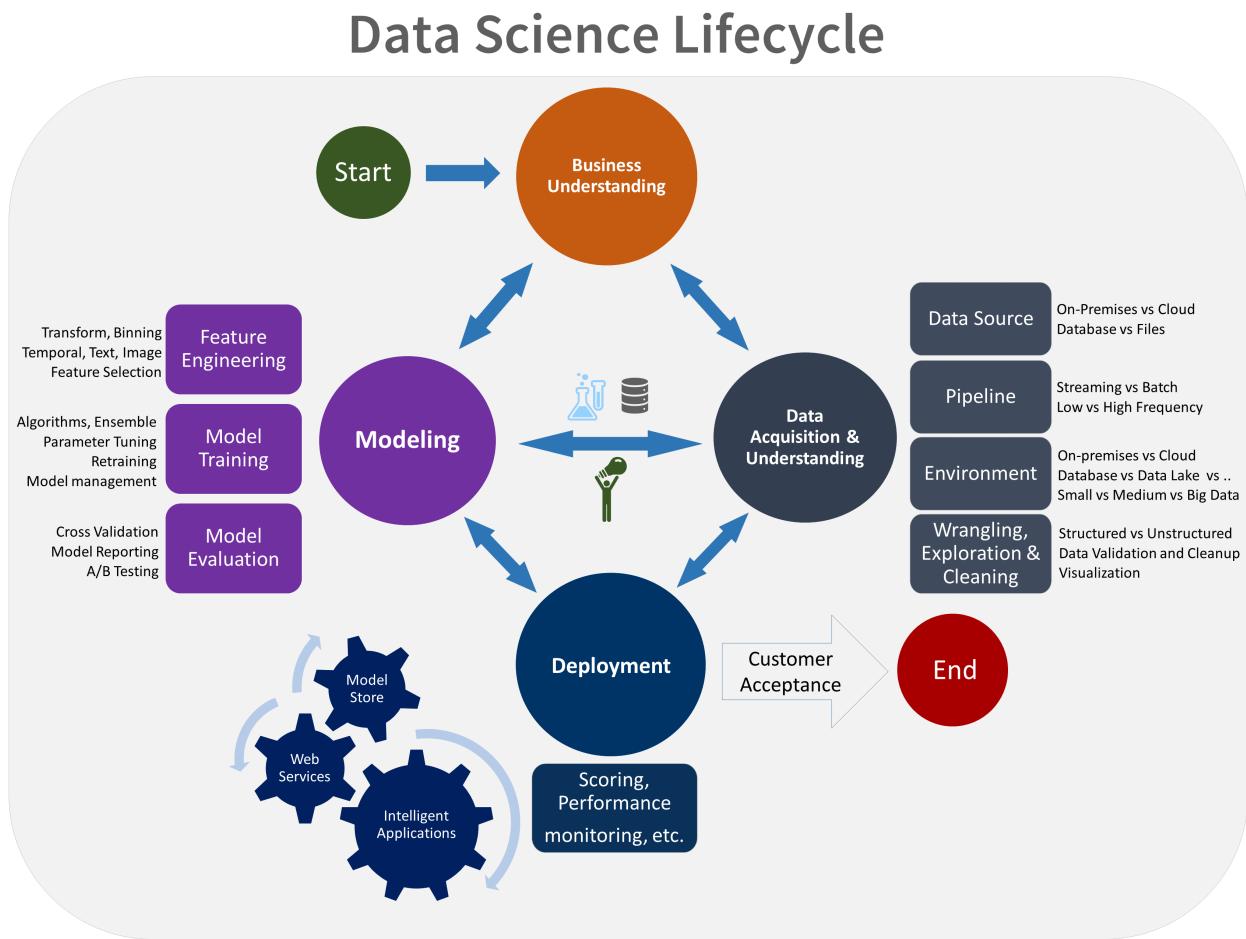
The business understanding stage of the Team Data Science Process lifecycle

1/30/2019 • 3 minutes to read

This article outlines the goals, tasks, and deliverables associated with the business understanding stage of the Team Data Science Process (TDSP). This process provides a recommended lifecycle that you can use to structure your data-science projects. The lifecycle outlines the major stages that projects typically execute, often iteratively:

1. **Business understanding**
2. **Data acquisition and understanding**
3. **Modeling**
4. **Deployment**
5. **Customer acceptance**

Here is a visual representation of the TDSP lifecycle:



Goals

- Specify the key variables that are to serve as the model targets and whose related metrics are used determine the success of the project.
- Identify the relevant data sources that the business has access to or needs to obtain.

How to do it

There are two main tasks addressed in this stage:

- **Define objectives:** Work with your customer and other stakeholders to understand and identify the business problems. Formulate questions that define the business goals that the data science techniques can target.
- **Identify data sources:** Find the relevant data that helps you answer the questions that define the objectives of the project.

Define objectives

1. A central objective of this step is to identify the key business variables that the analysis needs to predict. We refer to these variables as the *model targets*, and we use the metrics associated with them to determine the success of the project. Two examples of such targets are sales forecasts or the probability of an order being fraudulent.
2. Define the project goals by asking and refining "sharp" questions that are relevant, specific, and unambiguous. Data science is a process that uses names and numbers to answer such questions. For more information on asking sharp questions, see the [How to do data science](#) blog. You typically use data science or machine learning to answer five types of questions:
 - How much or how many? (regression)
 - Which category? (classification)
 - Which group? (clustering)
 - Is this weird? (anomaly detection)
 - Which option should be taken? (recommendation)

Determine which of these questions you're asking and how answering it achieves your business goals.

3. Define the project team by specifying the roles and responsibilities of its members. Develop a high-level milestone plan that you iterate on as you discover more information.
4. Define the success metrics. For example, you might want to achieve a customer churn prediction. You need an accuracy rate of "x" percent by the end of this three-month project. With this data, you can offer customer promotions to reduce churn. The metrics must be **SMART**:
 - **S**pecific
 - **M**easurable
 - **A**chievable
 - **R**elevant
 - **T**ime-bound

Identify data sources

Identify data sources that contain known examples of answers to your sharp questions. Look for the following data:

- Data that's relevant to the question. Do you have measures of the target and features that are related to the target?
- Data that's an accurate measure of your model target and the features of interest.

For example, you might find that the existing systems need to collect and log additional kinds of data to address the problem and achieve the project goals. In this situation, you might want to look for external data sources or update your systems to collect new data.

Artifacts

Here are the deliverables in this stage:

- [Charter document](#): A standard template is provided in the TDSP project structure definition. The charter

document is a living document. You update the template throughout the project as you make new discoveries and as business requirements change. The key is to iterate upon this document, adding more detail, as you progress through the discovery process. Keep the customer and other stakeholders involved in making the changes and clearly communicate the reasons for the changes to them.

- **Data sources:** The **Raw data sources** section of the **Data definitions** report that's found in the TDSP project **Data report** folder contains the data sources. This section specifies the original and destination locations for the raw data. In later stages, you fill in additional details like the scripts to move the data to your analytic environment.
- **Data dictionaries:** This document provides descriptions of the data that's provided by the client. These descriptions include information about the schema (the data types and information on the validation rules, if any) and the entity-relation diagrams, if available.

Next steps

Here are links to each step in the lifecycle of the TDSP:

1. [Business understanding](#)
2. [Data acquisition and understanding](#)
3. [Modeling](#)
4. [Deployment](#)
5. [Customer acceptance](#)

We provide full end-to-end walkthroughs that demonstrate all the steps in the process for specific scenarios. The [Example walkthroughs](#) article provides a list of the scenarios with links and thumbnail descriptions. The walkthroughs illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

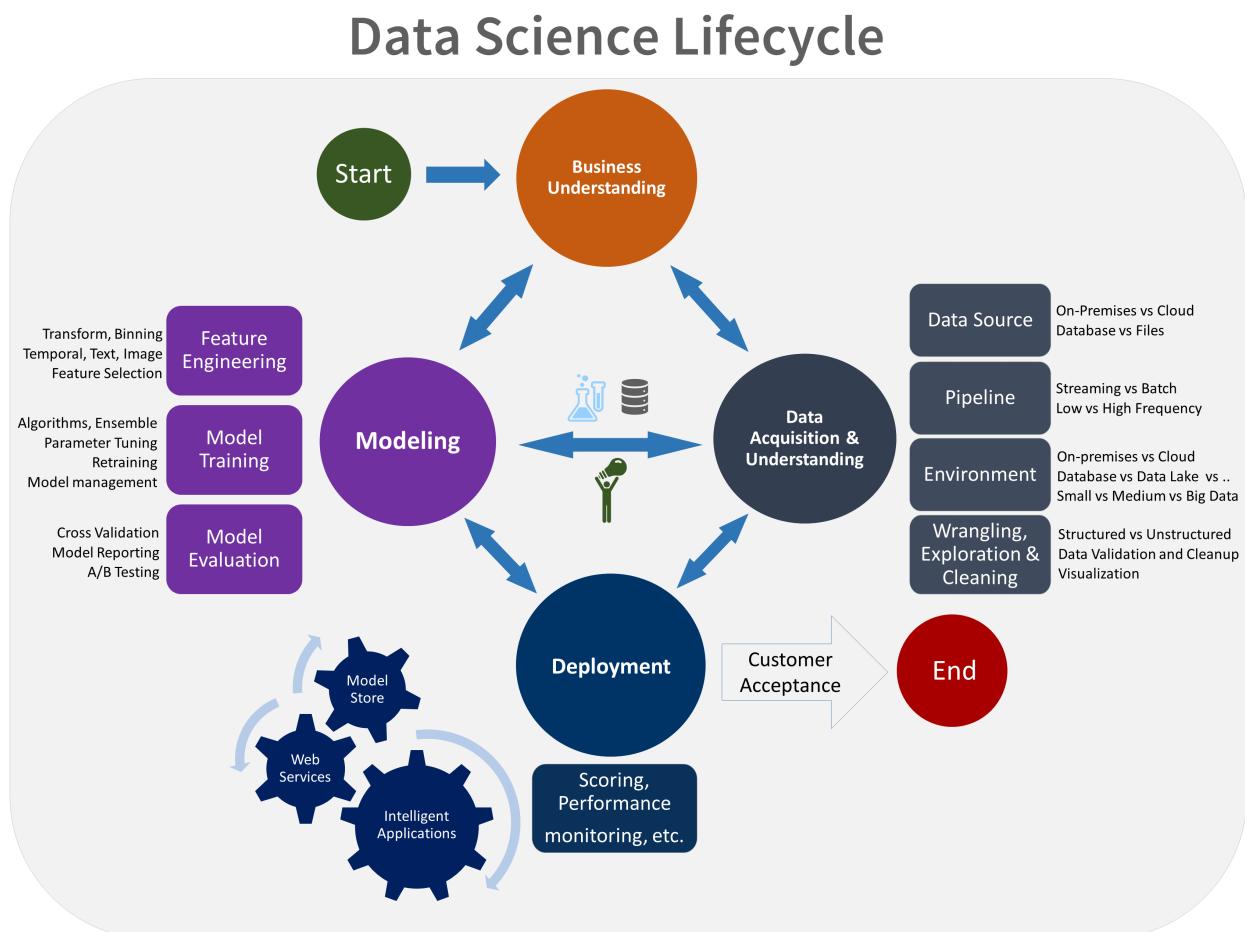
Data acquisition and understanding stage of the Team Data Science Process

3/5/2019 • 4 minutes to read

This article outlines the goals, tasks, and deliverables associated with the data acquisition and understanding stage of the Team Data Science Process (TDSP). This process provides a recommended lifecycle that you can use to structure your data-science projects. The lifecycle outlines the major stages that projects typically execute, often iteratively:

1. **Business understanding**
2. **Data acquisition and understanding**
3. **Modeling**
4. **Deployment**
5. **Customer acceptance**

Here is a visual representation of the TDSP lifecycle:



Goals

- Produce a clean, high-quality data set whose relationship to the target variables is understood. Locate the data set in the appropriate analytics environment so you are ready to model.
- Develop a solution architecture of the data pipeline that refreshes and scores the data regularly.

How to do it

There are three main tasks addressed in this stage:

- **Ingest the data** into the target analytic environment.
- **Explore the data** to determine if the data quality is adequate to answer the question.
- **Set up a data pipeline** to score new or regularly refreshed data.

Ingest the data

Set up the process to move the data from the source locations to the target locations where you run analytics operations, like training and predictions. For technical details and options on how to move the data with various Azure data services, see [Load data into storage environments for analytics](#).

Explore the data

Before you train your models, you need to develop a sound understanding of the data. Real-world data sets are often noisy, are missing values, or have a host of other discrepancies. You can use data summarization and visualization to audit the quality of your data and provide the information you need to process the data before it's ready for modeling. This process is often iterative.

TDSP provides an automated utility, called [IDEAR](#), to help visualize the data and prepare data summary reports. We recommend that you start with IDEAR first to explore the data to help develop initial data understanding interactively with no coding. Then you can write custom code for data exploration and visualization. For guidance on cleaning the data, see [Tasks to prepare data for enhanced machine learning](#).

After you're satisfied with the quality of the cleansed data, the next step is to better understand the patterns that are inherent in the data. This helps you choose and develop an appropriate predictive model for your target. Look for evidence for how well connected the data is to the target. Then determine whether there is sufficient data to move forward with the next modeling steps. Again, this process is often iterative. You might need to find new data sources with more accurate or more relevant data to augment the data set initially identified in the previous stage.

Set up a data pipeline

In addition to the initial ingestion and cleaning of the data, you typically need to set up a process to score new data or refresh the data regularly as part of an ongoing learning process. You do this by setting up a data pipeline or workflow. The [Move data from an on-premises SQL Server instance to Azure SQL Database with Azure Data Factory](#) article gives an example of how to set up a pipeline with [Azure Data Factory](#).

In this stage, you develop a solution architecture of the data pipeline. You develop the pipeline in parallel with the next stage of the data science project. Depending on your business needs and the constraints of your existing systems into which this solution is being integrated, the pipeline can be one of the following:

- Batch-based
- Streaming or real time
- A hybrid

Artifacts

The following are the deliverables in this stage:

- **Data quality report:** This report includes data summaries, the relationships between each attribute and target, variable ranking, and more. The [IDEAR](#) tool provided as part of TDSP can quickly generate this report on any tabular data set, such as a CSV file or a relational table.
- **Solution architecture:** The solution architecture can be a diagram or description of your data pipeline that you use to run scoring or predictions on new data after you have built a model. It also contains the pipeline to retrain your model based on new data. Store the document in the [Project](#) directory when you use the TDSP directory structure template.

- **Checkpoint decision:** Before you begin full-feature engineering and model building, you can reevaluate the project to determine whether the value expected is sufficient to continue pursuing it. You might, for example, be ready to proceed, need to collect more data, or abandon the project as the data does not exist to answer the question.

Next steps

Here are links to each step in the lifecycle of the TDSP:

1. [Business understanding](#)
2. [Data acquisition and understanding](#)
3. [Modeling](#)
4. [Deployment](#)
5. [Customer acceptance](#)

We provide full end-to-end walkthroughs that demonstrate all the steps in the process for specific scenarios. The [Example walkthroughs](#) article provides a list of the scenarios with links and thumbnail descriptions. The walkthroughs illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

For examples of how to execute steps in TDSPs that use Azure Machine Learning Studio, see [Use the TDSP with Azure Machine Learning](#).

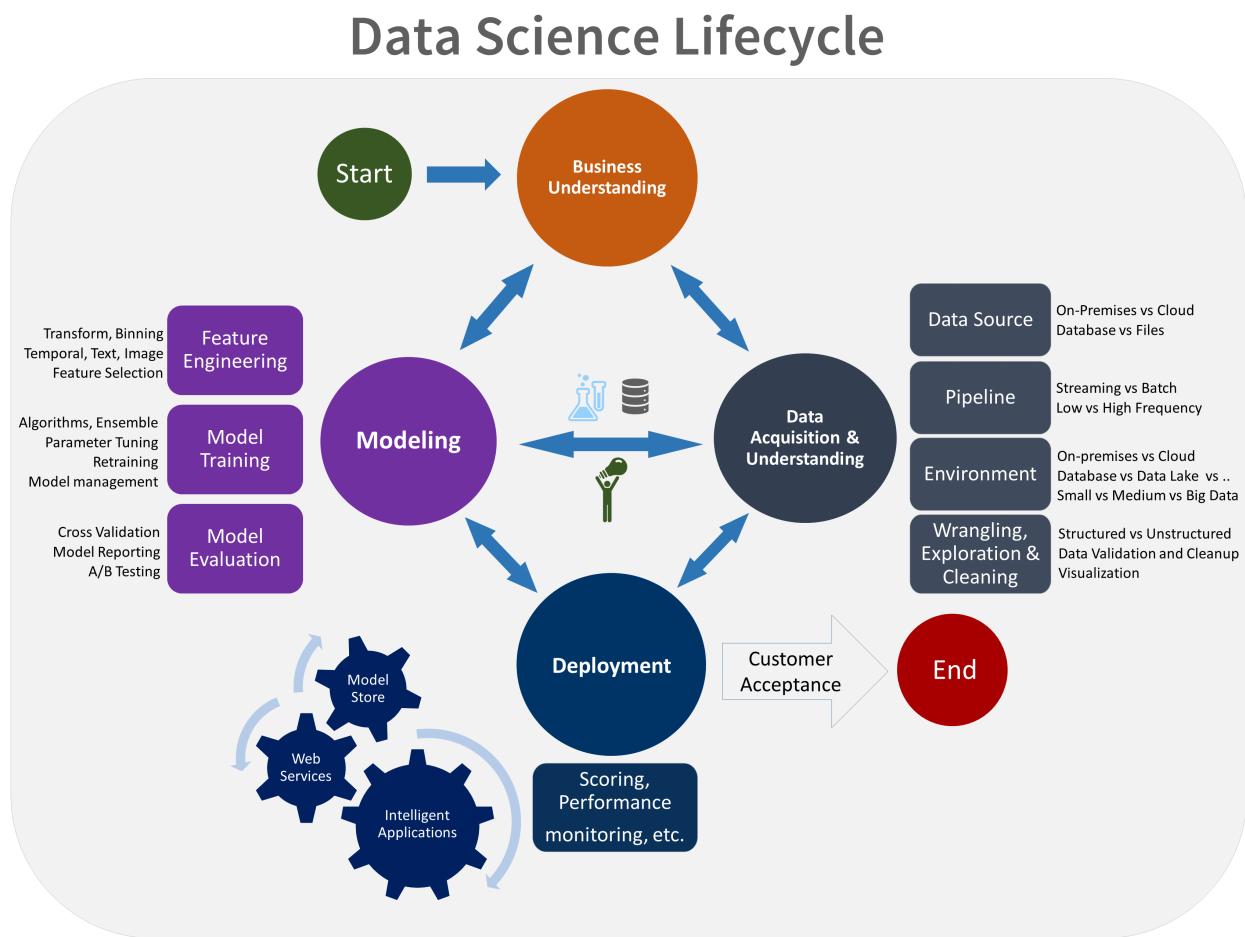
Modeling stage of the Team Data Science Process lifecycle

1/30/2019 • 4 minutes to read

This article outlines the goals, tasks, and deliverables associated with the modeling stage of the Team Data Science Process (TDSP). This process provides a recommended lifecycle that you can use to structure your data-science projects. The lifecycle outlines the major stages that projects typically execute, often iteratively:

1. **Business understanding**
2. **Data acquisition and understanding**
3. **Modeling**
4. **Deployment**
5. **Customer acceptance**

Here is a visual representation of the TDSP lifecycle:



Goals

- Determine the optimal data features for the machine-learning model.
- Create an informative machine-learning model that predicts the target most accurately.
- Create a machine-learning model that's suitable for production.

How to do it

There are three main tasks addressed in this stage:

- **Feature engineering:** Create data features from the raw data to facilitate model training.
- **Model training:** Find the model that answers the question most accurately by comparing their success metrics.
- Determine if your model is **suitable for production**.

Feature engineering

Feature engineering involves the inclusion, aggregation, and transformation of raw variables to create the features used in the analysis. If you want insight into what is driving a model, then you need to understand how the features relate to each other and how the machine-learning algorithms are to use those features.

This step requires a creative combination of domain expertise and the insights obtained from the data exploration step. Feature engineering is a balancing act of finding and including informative variables, but at the same time trying to avoid too many unrelated variables. Informative variables improve your result; unrelated variables introduce unnecessary noise into the model. You also need to generate these features for any new data obtained during scoring. As a result, the generation of these features can only depend on data that's available at the time of scoring.

For technical guidance on feature engineering when make use of various Azure data technologies, see [Feature engineering in the data science process](#).

Model training

Depending on the type of question that you're trying to answer, there are many modeling algorithms available. For guidance on choosing the algorithms, see [How to choose algorithms for Microsoft Azure Machine Learning](#). Although this article uses Azure Machine Learning, the guidance it provides is useful for any machine-learning projects.

The process for model training includes the following steps:

- **Split the input data** randomly for modeling into a training data set and a test data set.
- **Build the models** by using the training data set.
- **Evaluate** the training and the test data set. Use a series of competing machine-learning algorithms along with the various associated tuning parameters (known as a *parameter sweep*) that are geared toward answering the question of interest with the current data.
- **Determine the “best” solution** to answer the question by comparing the success metrics between alternative methods.

NOTE

Avoid leakage: You can cause data leakage if you include data from outside the training data set that allows a model or machine-learning algorithm to make unrealistically good predictions. Leakage is a common reason why data scientists get nervous when they get predictive results that seem too good to be true. These dependencies can be hard to detect. To avoid leakage often requires iterating between building an analysis data set, creating a model, and evaluating the accuracy of the results.

We provide an [automated modeling and reporting tool](#) with TDSP that's able to run through multiple algorithms and parameter sweeps to produce a baseline model. It also produces a baseline modeling report that summarizes the performance of each model and parameter combination including variable importance. This process is also iterative as it can drive further feature engineering.

Artifacts

The artifacts produced in this stage include:

- **Feature sets:** The features developed for the modeling are described in the **Feature sets** section of the **Data definition** report. It contains pointers to the code to generate the features and a description of how the feature was generated.
- **Model report:** For each model that's tried, a standard, template-based report that provides details on each experiment is produced.
- **Checkpoint decision:** Evaluate whether the model performs well enough to deploy it to a production system. Some key questions to ask are:
 - Does the model answer the question with sufficient confidence given the test data?
 - Should you try any alternative approaches? Should you collect additional data, do more feature engineering, or experiment with other algorithms?

Next steps

Here are links to each step in the lifecycle of the TDSP:

1. [Business understanding](#)
2. [Data acquisition and understanding](#)
3. [Modeling](#)
4. [Deployment](#)
5. [Customer acceptance](#)

We provide full end-to-end walkthroughs that demonstrate all the steps in the process for specific scenarios. The [Example walkthroughs](#) article provides a list of the scenarios with links and thumbnail descriptions. The walkthroughs illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

For examples of how to execute steps in TDSPs that use Azure Machine Learning Studio, see [Use the TDSP with Azure Machine Learning](#).

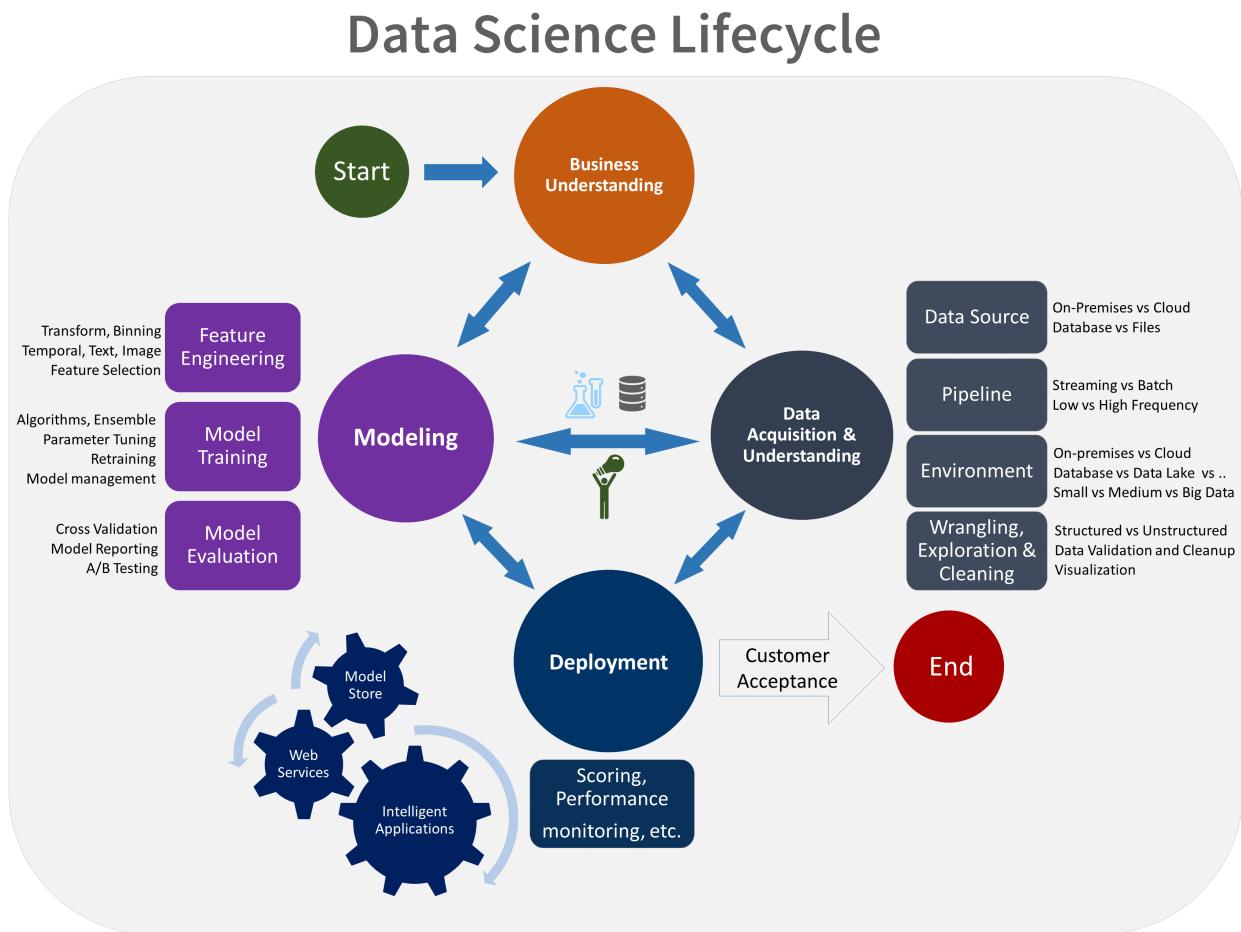
Deployment stage of the Team Data Science Process lifecycle

1/30/2019 • 2 minutes to read

This article outlines the goals, tasks, and deliverables associated with the deployment of the Team Data Science Process (TDSP). This process provides a recommended lifecycle that you can use to structure your data-science projects. The lifecycle outlines the major stages that projects typically execute, often iteratively:

1. **Business understanding**
2. **Data acquisition and understanding**
3. **Modeling**
4. **Deployment**
5. **Customer acceptance**

Here is a visual representation of the TDSP lifecycle:



Goal

Deploy models with a data pipeline to a production or production-like environment for final user acceptance.

How to do it

The main task addressed in this stage:

Operationalize the model: Deploy the model and pipeline to a production or production-like environment for application consumption.

Operationalize a model

After you have a set of models that perform well, you can operationalize them for other applications to consume. Depending on the business requirements, predictions are made either in real time or on a batch basis. To deploy models, you expose them with an open API interface. The interface enables the model to be easily consumed from various applications, such as:

- Online websites
- Spreadsheets
- Dashboards
- Line-of-business applications
- Back-end applications

For examples of model operationalization with an Azure Machine Learning web service, see [Deploy an Azure Machine Learning web service](#). It is a best practice to build telemetry and monitoring into the production model and the data pipeline that you deploy. This practice helps with subsequent system status reporting and troubleshooting.

Artifacts

- A status dashboard that displays the system health and key metrics
- A final modeling report with deployment details
- A final solution architecture document

Next steps

Here are links to each step in the lifecycle of the TDSP:

1. [Business understanding](#)
2. [Data Acquisition and understanding](#)
3. [Modeling](#)
4. [Deployment](#)
5. [Customer acceptance](#)

We provide full end-to-end walkthroughs that demonstrate all the steps in the process for specific scenarios. The [Example walkthroughs](#) article provides a list of the scenarios with links and thumbnail descriptions. The walkthroughs illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

For examples of how to execute steps in TDSPs that use Azure Machine Learning Studio, see [Use the TDSP with Azure Machine Learning](#).

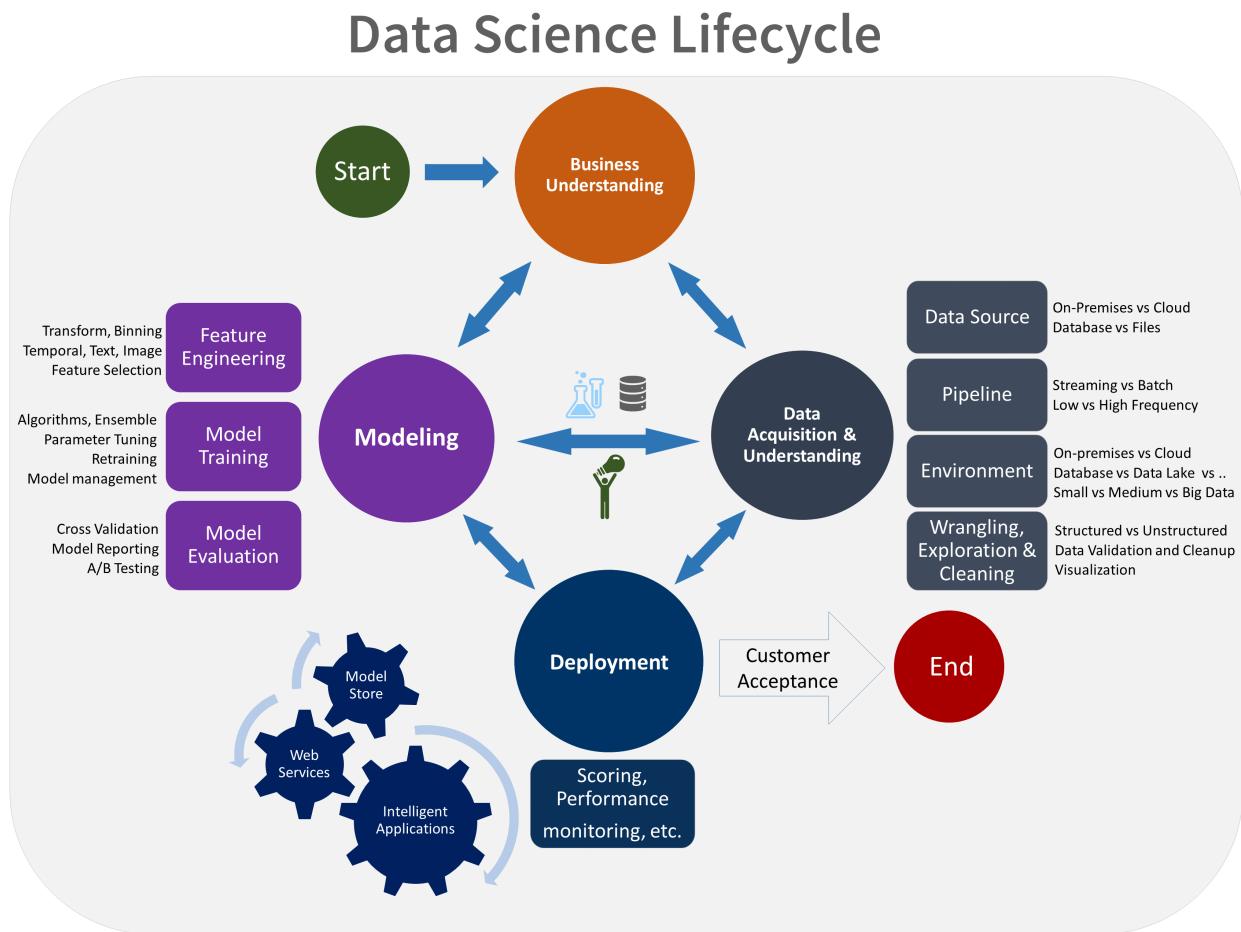
Customer acceptance stage of the Team Data Science Process lifecycle

1/30/2019 • 2 minutes to read

This article outlines the goals, tasks, and deliverables associated with the customer acceptance stage of the Team Data Science Process (TDSP). This process provides a recommended lifecycle that you can use to structure your data-science projects. The lifecycle outlines the major stages that projects typically execute, often iteratively:

1. **Business understanding**
2. **Data acquisition and understanding**
3. **Modeling**
4. **Deployment**
5. **Customer acceptance**

Here is a visual representation of the TDSP lifecycle:



Goal

Finalize the project deliverables: Confirm that the pipeline, the model, and their deployment in a production environment satisfy the customer's objectives.

How to do it

There are two main tasks addressed in this stage:

- **System validation:** Confirm that the deployed model and pipeline meet the customer's needs.
- **Project hand-off:** Hand the project off to the entity that's going to run the system in production.

The customer should validate that the system meets their business needs and that it answers the questions with acceptable accuracy to deploy the system to production for use by their client's application. All the documentation is finalized and reviewed. The project is handed-off to the entity responsible for operations. This entity might be, for example, an IT or customer data-science team or an agent of the customer that's responsible for running the system in production.

Artifacts

The main artifact produced in this final stage is the **Exit report of the project for the customer**. This technical report contains all the details of the project that are useful for learning about how to operate the system. TDSP provides an [Exit report](#) template. You can use the template as is, or you can customize it for specific client needs.

Next steps

Here are links to each step in the lifecycle of the TDSP:

1. [Business understanding](#)
2. [Data acquisition and understanding](#)
3. [Modeling](#)
4. [Deployment](#)
5. [Customer acceptance](#)

We provide full end-to-end walkthroughs that demonstrate all the steps in the process for specific scenarios. The [Example walkthroughs](#) article provides a list of the scenarios with links and thumbnail descriptions. The walkthroughs illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

For examples of how to execute steps in TDSPs that use Azure Machine Learning Studio, see [Use the TDSP with Azure Machine Learning](#).

Team Data Science Process roles and tasks

1/30/2019 • 7 minutes to read

The Team Data Science Process is a framework developed by Microsoft that provides a structured methodology to build predictive analytics solutions and intelligent applications efficiently. This article outlines the key personnel roles, and their associated tasks that are handled by a data science team standardizing on this process.

This introduction links to tutorials that provide instructions on how to set up the TDSP environment for the entire data science group, data science teams, and projects. It provides detailed guidance using Azure DevOps in the tutorials. Azure DevOps provides a code-hosting platform and agile planning tool to manage team tasks, control access, and manage the repositories.

You can use this information to implement TDSP on your own code-hosting and agile planning tool.

Structures of data science groups and teams

Data science functions in enterprises may often be organized in the following hierarchy:

1. **Data science group/s**
2. **Data science team/s within group/s**

In such a structure, there are group and team leads. Typically, a data science project is done by a data science team, which may be composed of project leads (for project management and governance tasks) and data scientists or engineers (individual contributors / technical personnel) who will execute the data science and data engineering parts of the project. Prior to execution, the setup and governance is done by the group, team, or project leads.

Definition of four TDSP roles

With the above assumption, there are four distinct roles for the team personnel:

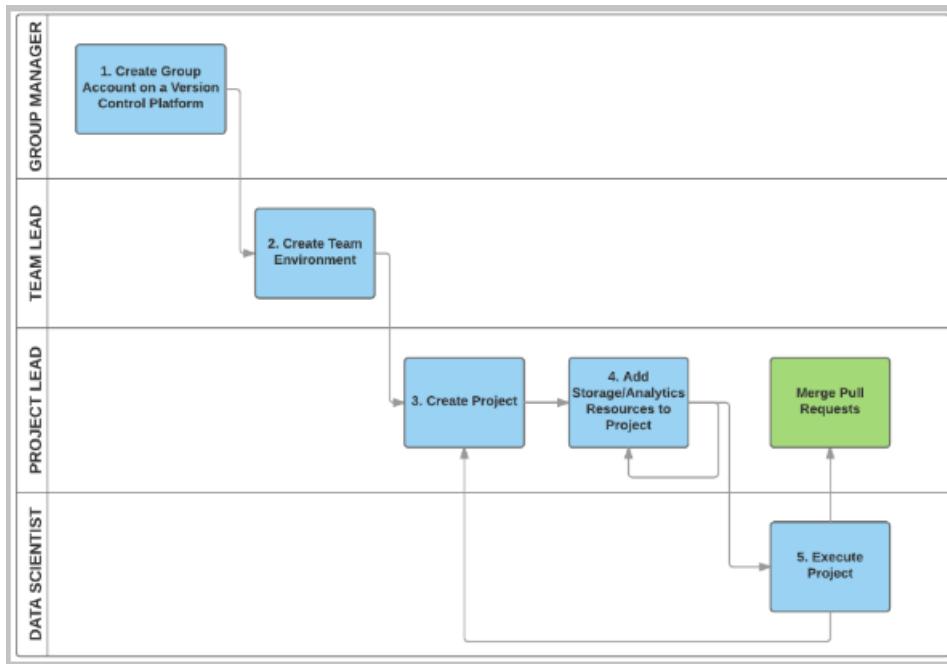
1. **Group Manager.** Group Manager is the manager of the entire data science unit in an enterprise. A data science unit might have multiple teams, each of which is working on multiple data science projects in distinct business verticals. A Group Manager might delegate their tasks to a surrogate, but the tasks associated with the role do not change.
2. **Team Lead.** A team lead is managing a team in the data science unit of an enterprise. A team consists of multiple data scientists. For data science unit with only a small number of data scientists, the Group Manager and the Team Lead might be the same person.
3. **Project Lead.** A project lead manages the daily activities of individual data scientists on a specific data science project.
4. **Project Individual Contributor.** Data Scientist, Business Analyst, Data Engineer, Architect, etc. A project individual contributor executes a data science project.

NOTE

Depending on the structure in an enterprise, a single person may play more than one role OR there may be more than one person working on a role. This may frequently be the case in small enterprises or enterprises with a small number of personnel in their data science organization.

Tasks to be completed by four personnel

The following picture depicts the top-level tasks for personnel by role in adopting and implementing the Team Data Science Process as conceptualized by Microsoft.



This schema and the following, more detailed outline of tasks that are assigned to each role in the TDSP should help you choose the appropriate tutorial based on your responsibilities in the organization.

NOTE

The following instructions show steps of how to set up a TDSP environment and complete other data science tasks in Azure DevOps. We specify how to accomplish these tasks with Azure DevOps because that is what we are using to implement TDSP at Microsoft. Azure DevOps facilitates collaboration by integrating the management of work items that track tasks and a code hosting service used to share utilities, organize versions, and provide role-based security. You are able to choose other platforms, if you prefer, to implement the tasks outlined by the TDSP. But depending on your platform, some features leveraged from Azure DevOps may not be available.

Instructions here also use the [Data Science Virtual Machine \(DSVM\)](#) on the Azure cloud as the analytics desktop with several popular data science tools pre-configured and integrated with various Microsoft software and Azure services. You can use the DSVM or any other development environment to implement TDSP.

Group Manager tasks

The following tasks are completed by the Group Manager (or a designated TDSP system administrator) to adopt the TDSP:

- Create a **group account** on a code hosting platform (like GitHub, Git, Azure DevOps, or others)
- Create a **project template repository** on the group account, and seed it from the project template repository developed by Microsoft TDSP team. The TDSP project template repository from Microsoft
 - provides a **standardized directory structure** including directories for data, code, and documents,
 - provides a set of **standardized document templates** to guide an efficient data science process.
- Create a **utility repository**, and seed it from the utility repository developed by Microsoft TDSP team. The TDSP utility repository from Microsoft provides
 - a set of useful utilities to make the work of a data scientist more efficient, including utilities for interactive data exploration, analysis, and reporting, and for baseline modeling and reporting.
- Set up the **security control policy** of these two repositories on your group account.

For detailed step-by-step instructions, see [Group Manager tasks for a data science team](#).

Team Lead tasks

The following tasks are completed by the Team Lead (or a designated project administrator) to adopt the TDSP:

- If Azure DevOps is selected to be the code hosting platform for versioning and collaboration, create a **project** on the group's Azure DevOps Services. Otherwise, this task can be skipped.
- Create the **project template repository** under the project, and seed it from the group project template repository set up by your group manager or the delegate of the manager.
- Create the **team utility repository**, and add the team-specific utilities to the repository.
- (Optional) Create **Azure file storage** to be used to store data assets that can be useful for the entire team. Other team members can mount this shared cloud file store on their analytics desktops.
- (Optional) Mount the Azure file storage to the **Data Science Virtual Machine** (DSVM) of the team lead and add data assets on it.
- Set up the **security control** by adding team members and configure their privileges.

For detailed step-by-step instructions, see [Team Lead tasks for a data science team](#).

Project Lead tasks

The following tasks are completed by the Project Lead to adopt the TDSP:

- Create a **project repository** under the project, and seed it from the project template repository.
- (Optional) Create **Azure file storage** to be used to store data assets of the project.
- (Optional) Mount the Azure file storage to the **Data Science Virtual Machine** (DSVM) of the Project Lead and add project data assets on it.
- Set up the **security control** by adding project members and configure their privileges.

For detailed step-by-step instructions, see [Project Lead tasks for a data science team](#).

Project Individual Contributor tasks

The following tasks are completed by a Project Individual Contributor (usually a Data Scientist) to conduct the data science project using the TDSP:

- Clone the **project repository** set up by the project lead.
- (Optional) Mount the shared **Azure file storage** of the team and project on their **Data Science Virtual Machine** (DSVM).
- Execute the project.

For detailed step-by-step instructions for on-boarding onto a project, see [Project Individual Contributors for a data science team](#).

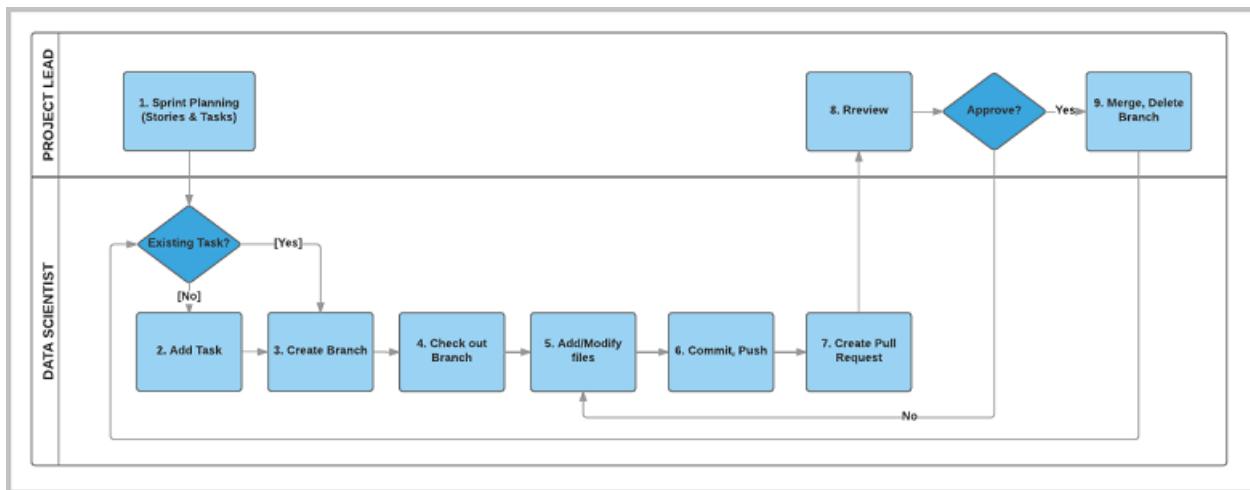
Data science project execution

By following the relevant set of instructions, data scientists, project lead, and team leads can create work items to track all tasks and stages that a project needs from its beginning to its end. Using git also promotes collaboration among data scientists and ensures that the artifacts generated during project execution are version controlled and shared by all project members.

The instructions provided for project execution have been developed based on the assumption that both work items and project git repositories are on Azure DevOps. Using Azure DevOps for both allows you to link your work items with the Git branches of your project repositories. In this way, you can easily track what has been done

for a work item.

The following figure outlines this workflow for project execution using the TDSP.



The workflow includes steps that can be grouped into three activities:

- Sprint planning (Project Lead)
- Developing artifacts on git branches to address work items (Data Scientist)
- Code review and merging branches with master branches (Project Lead or other team members)

For detailed step-by-step instructions on project execution workflow, see [Execution of data science projects](#).

Project structure

Use this [project template repository](#) to support efficient project execution and collaboration. This repository gives you a standardized directory structure and document templates you can use for your own TDSP project.

Next steps

Explore more detailed descriptions of the roles and tasks defined by the Team Data Science Process:

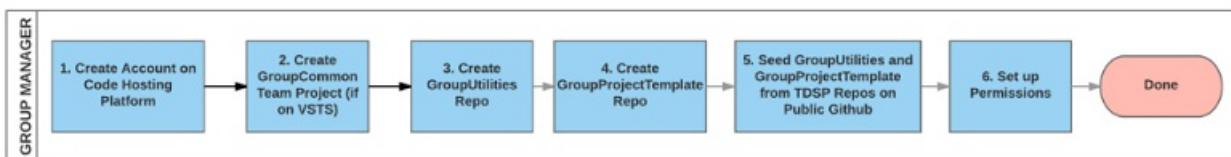
- [Group Manager tasks for a data science team](#)
- [Team Lead tasks for a data science team](#)
- [Project Lead tasks for a data science team](#)
- [Project Individual Contributors for a data science team](#)

Tasks for a group manager on a data science team project

1/30/2019 • 11 minutes to read

This topic outlines the tasks that a Group Manager is expected to complete for his/her data science organization. The objective is to establish collaborative group environment that standardizes on the [Team Data Science Process \(TDSP\)](#). For an outline of the personnel roles and their associated tasks that are handled by a data science team standardizing on this process, see [Team Data Science Process roles and tasks](#).

The **Group Manager** is the manager of the entire data science unit in an enterprise. A data science unit may have multiple teams, each of which is working on multiple data science projects in distinct business verticals. A Group Manager may delegate their tasks to a surrogate, but the tasks associated with the role are the same. There are six main tasks as shown in the following diagram:



NOTE

We outline the steps needed to set up a TDSP group environment using Azure DevOps Services in the instructions that follow. We specify how to accomplish these tasks with Azure DevOps Services because that is how we implement TDSP at Microsoft. If another code hosting platform is used for your group, the tasks that need to be completed by the group manager generally do not change. But the way to complete these tasks is going to be different.

1. Set up **Azure DevOps Services** for the group.
2. Create a **group project** on Azure DevOps Services (for Azure DevOps Services users)
3. Create the **GroupProjectTemplate** repository
4. Create the **GroupUtilities** repository
5. Seed the **GroupProjectTemplate** and **GroupUtilities** repositories for the Azure DevOps Services with content from the TDSP repositories.
6. Set up the **security controls** for team members to access to the GroupProjectTemplate and GroupUtilities repositories.

Each of the preceding steps is described in detail. But first, we familiarize you with the abbreviations and discuss the pre-requisites for working with repositories.

Abbreviations for repositories and directories

This tutorial uses abbreviated names for repositories and directories. These definitions make it easier to follow the operations between the repositories and directories. This notation is used in the following sections:

- **G1:** The project template repository developed and managed by TDSP team of Microsoft.
- **G2:** The utilities repository developed and managed by TDSP team of Microsoft.
- **R1:** The GroupProjectTemplate repository on Git you set up on your Azure DevOps group server.
- **R2:** The GroupUtilities repository on Git you set up on your Azure DevOps group server.
- **LG1** and **LG2:** The local directories on your machine that you clone G1 and G2 to, respectively.
- **LR1** and **LR2:** The local directories on your machine that you clone R1 and R2 to, respectively.

Pre-requisites for cloning repositories and checking code in and out

- Git must be installed on your machine. If you are using a Data Science Virtual Machine (DSVM), Git has been pre-installed and you are good to go. Otherwise, see the [Platforms and tools appendix](#).
- If you are using a **Windows DSVM**, you need to have [Git Credential Manager \(GCM\)](#) installed on your machine. In the README.md file, scroll down to the **Download and Install** section and click the *latest installer*. This step takes you to the latest installer page. Download the .exe installer from here and run it.
- If you are using **Linux DSVM**, create an SSH public key on your DSVM and add it to your group Azure DevOps Services. For more information about SSH, see the **Create SSH public key** section in the [Platforms and tools appendix](#).

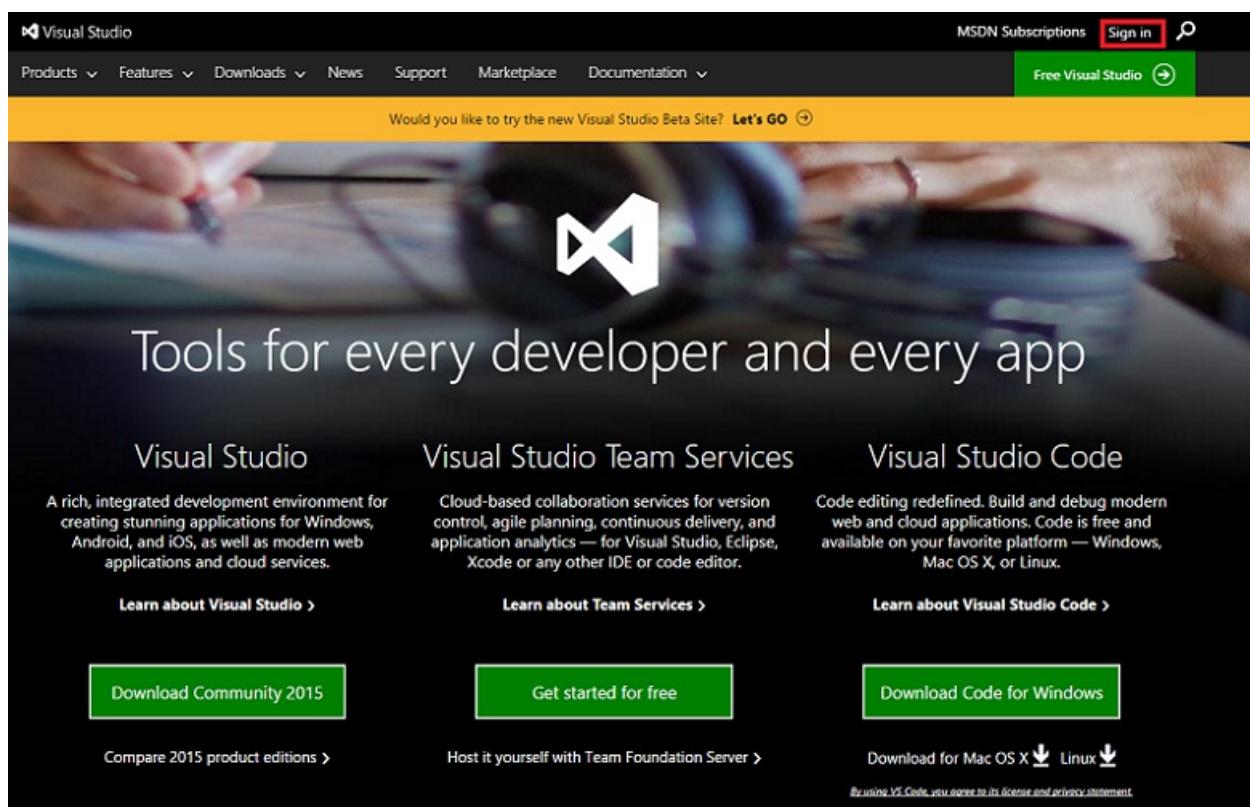
1. Create Account on Azure DevOps Services

The Azure DevOps Services hosts the following repositories:

- **group common repositories:** General-purpose repositories that can be adopted by multiple teams within a group for multiple data science projects. For example, the *GroupProjectTemplate* and *GroupUtilities* repositories.
- **team repositories:** Repositories for specific teams within a group. These repositories are specific for a team's need, and can be adopted by multiple projects executed by that team, but not general enough to be useful to multiple teams within a data science group.
- **project repositories:** Repositories available for specific projects. Such repositories may not be general enough to be useful to multiple projects performed by a team, and to multiple teams in a data science group.

Setting up the Azure DevOps Services Sign into your Microsoft account

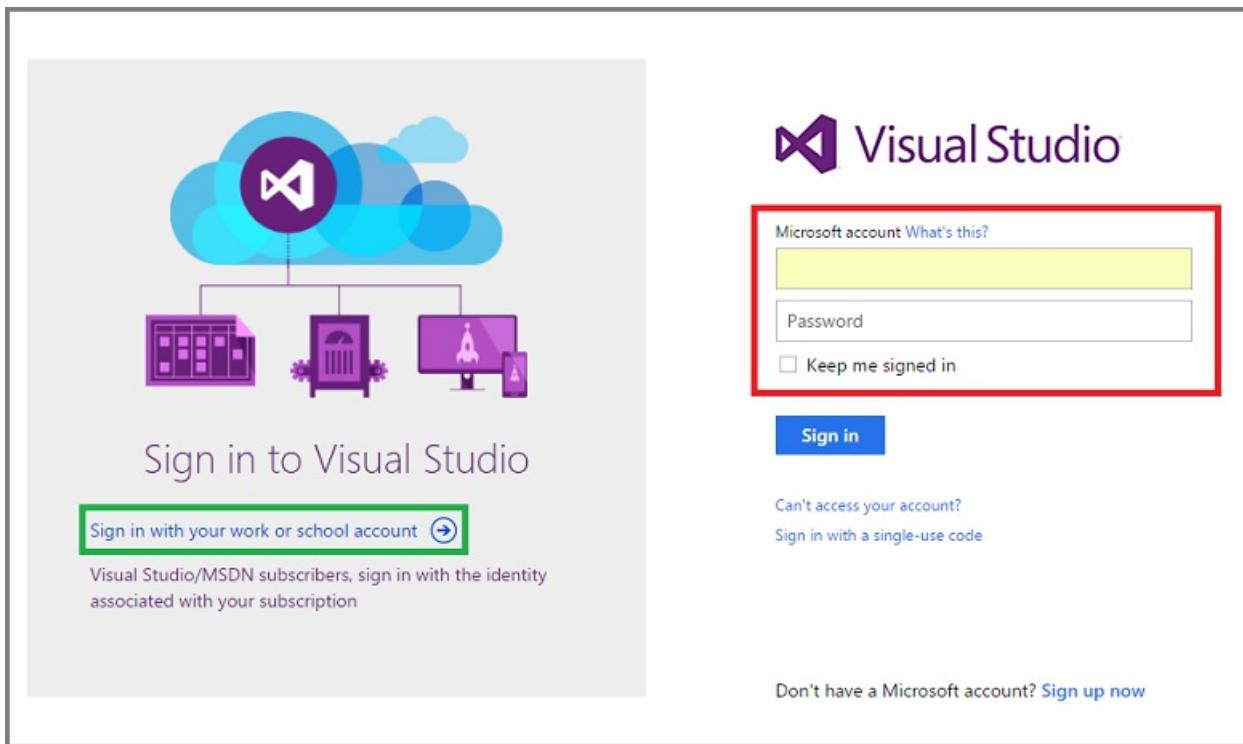
Go to [Visual Studio online](#), click **Sign in** in the upper right corner and sign into your Microsoft account.



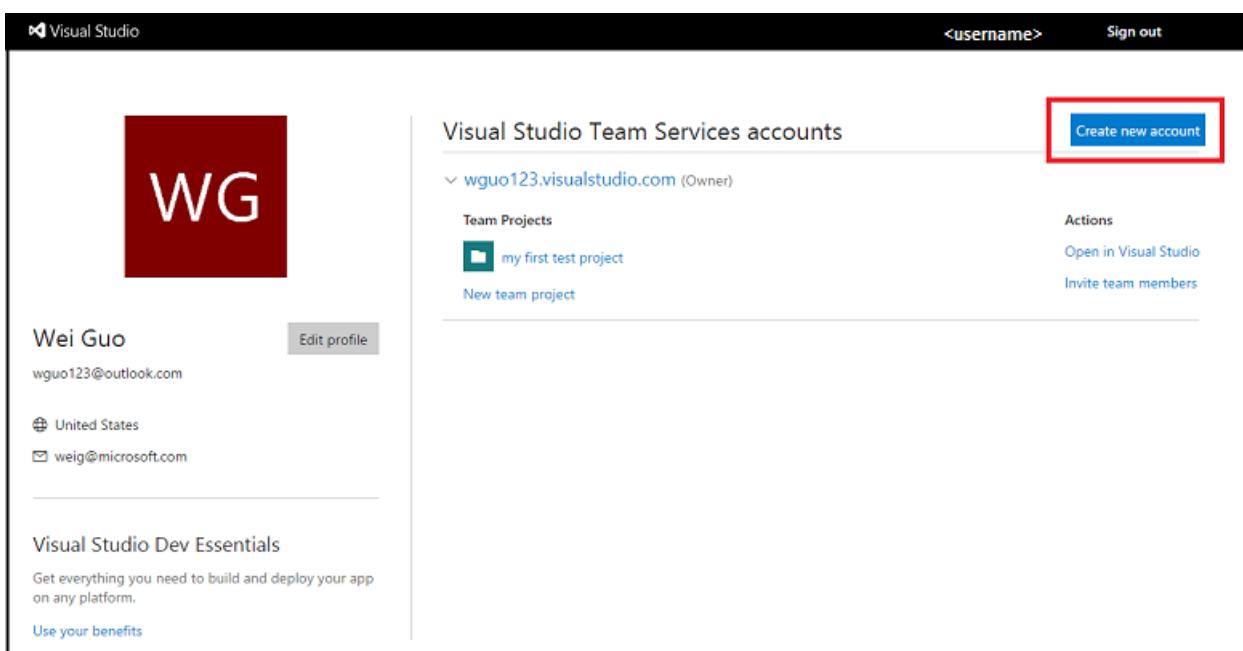
The screenshot shows the Visual Studio website homepage. At the top, there is a navigation bar with links for 'Products', 'Features', 'Downloads', 'News', 'Support', 'Marketplace', 'Documentation', 'MSDN Subscriptions', 'Sign in' (which is highlighted with a red box), and a search icon. Below the navigation bar, a yellow banner asks if you want to try the new Visual Studio Beta Site, with a 'Let's GO' button. The main content area features a large image of hands writing on a tablet with a stylus. Overlaid on the image is the Visual Studio logo. Below the image, the text 'Tools for every developer and every app' is displayed. There are three main sections: 'Visual Studio', 'Visual Studio Team Services', and 'Visual Studio Code'. Each section includes a brief description, a 'Learn about' link, and a green 'Download' or 'Get started for free' button. At the bottom, there are links for 'Compare 2015 product editions', 'Host it yourself with Team Foundation Server', and download links for 'Download for Mac OS X' and 'Linux'.

If you do not have a Microsoft account, click **Sign up now** to create a Microsoft account, and then sign in using this account.

If your organization has a Visual Studio/MSDN subscription, click the green **Sign in with your work or school account** box and sign in with the credentials associated with this subscription.



After you sign in, click **Create New Account** in the upper right corner as shown in the following image:



Fill in the information for the Azure DevOps Services that you want to create in the **Create your account** wizard with the following values:

- **Server URL:** Replace *mysamplegroup* with your own *server name*. The URL of your server is going to be: <https://<servername>.visualstudio.com>.
- **Manage code using:** Select **Git**.
- **Project name:** Enter *GroupCommon*.
- **Organize work using:** Choose *Agile*.
- **Host your projects in:** Choose a geo location. In this example, we choose *South Central US*.

You're on your way to managing projects
with Visual Studio Team Services

Create your account

mysamplegroup .visualstudio.com

Manage code using:

Git
 Team Foundation Version Control

Project name:

GroupCommon

Organize work using:

Agile

Host your projects in:

South Central US

You can share the work with other users of:

Continue

NOTE

If you see the following pop-up window after you click **Create new account**, then you need to click **Change details** to display all the fields itemized.

You're on your way to managing projects
with Visual Studio Team Services

Create your account

Pick a memorable name .visualstudio.com

Manage code using:

Git
 Team Foundation Version Control

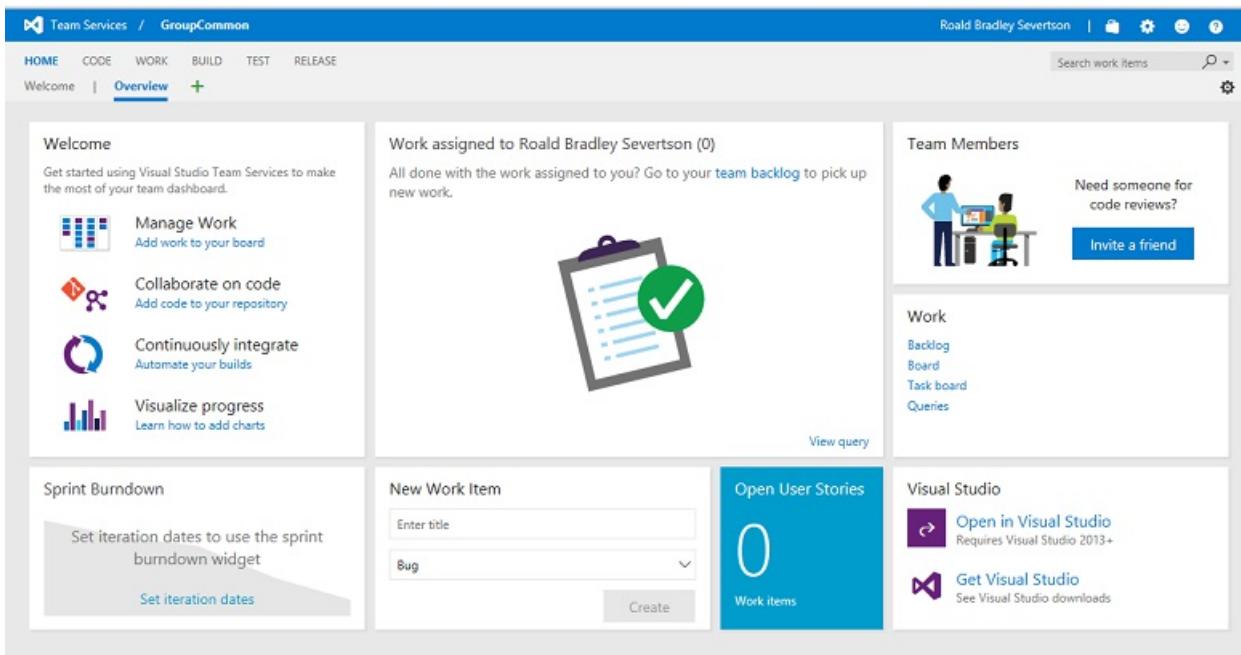
We will host your projects in South Central US region.
[Change details](#)

Continue

Click **Continue**.

2. GroupCommon Project

The **GroupCommon** page (<https://<servername>.visualstudio.com/GroupCommon>) opens after your Azure DevOps Services is created.

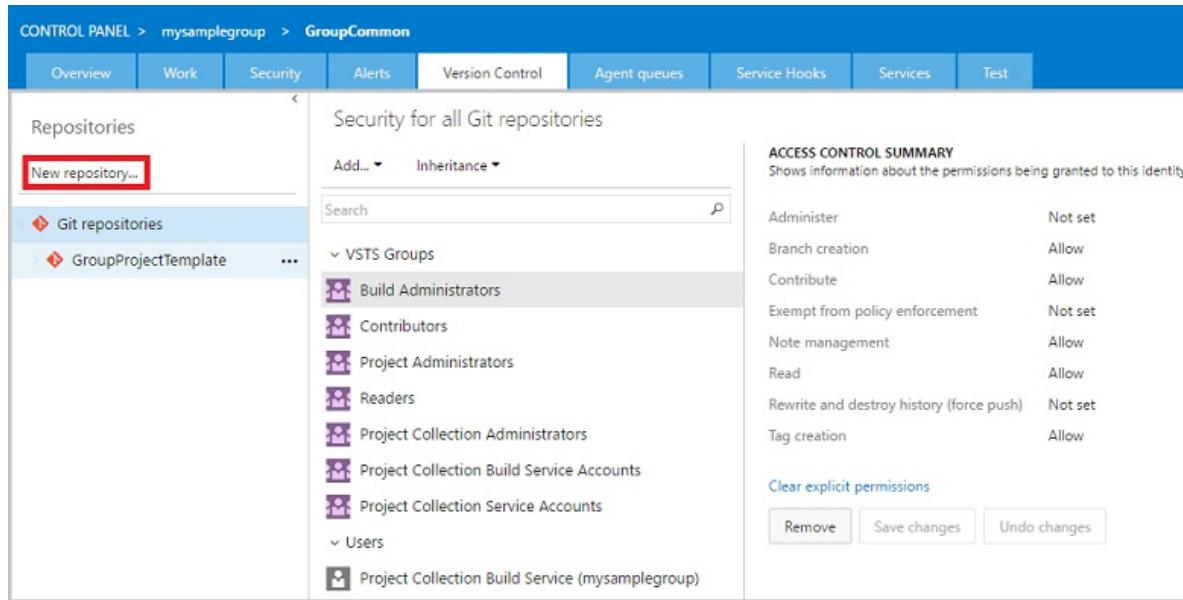


The screenshot shows the Azure DevOps Services GroupCommon dashboard. At the top, there's a navigation bar with links for HOME, CODE, WORK, BUILD, TEST, and RELEASE. The 'Overview' tab is selected. On the left, there's a 'Welcome' section with four cards: 'Manage Work' (Add work to your board), 'Collaborate on code' (Add code to your repository), 'Continuously integrate' (Automate your builds), and 'Visualize progress' (Learn how to add charts). In the center, there's a large 'Work assigned to Roald Bradley Severtson (0)' card with a clipboard icon and a green checkmark. To the right, there's a 'Team Members' section with a 'Need someone for code reviews?' link and an 'Invite a friend' button. Below that is a 'Work' section with links for Backlog, Board, Task board, and Queries. Further down, there's a 'Sprint Burndown' section with a 'Set iteration dates' button, a 'New Work Item' form, and an 'Open User Stories' section showing 0 work items. On the far right, there's a 'Visual Studio' section with links to 'Open in Visual Studio' (Requires Visual Studio 2013+) and 'Get Visual Studio'.

3. Create the GroupUtilities (R2) repository

To create the **GroupUtilities** (R2) repository under Azure DevOps Services:

- To open the **Create a new repository** wizard, click **New repository** on the **Version Control** tab of your project.

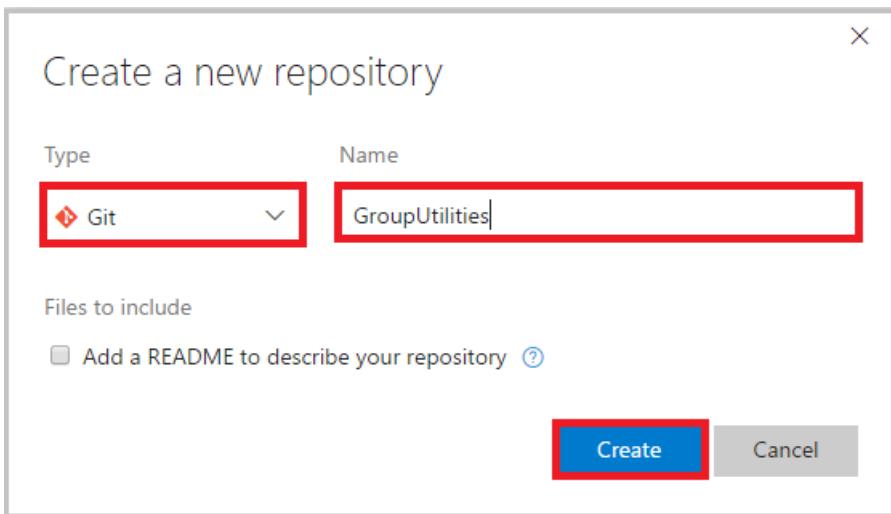


The screenshot shows the 'Version Control' screen in the Azure DevOps Services control panel. The 'Version Control' tab is selected. On the left, there's a 'Repositories' section with a 'New repository...' button highlighted with a red box. Below it are sections for 'Git repositories' and 'GroupProjectTemplate'. On the right, there's a 'Security for all Git repositories' section. It shows a search bar and a list of groups: VSTS Groups (Build Administrators, Contributors, Project Administrators, Readers, Project Collection Administrators, Project Collection Build Service Accounts, Project Collection Service Accounts), Users, and Project Collection Build Service (mysamplegroup). To the right of the group list is an 'ACCESS CONTROL SUMMARY' table:

ACCESS CONTROL SUMMARY	
Shows information about the permissions being granted to this identity	
Administer	Not set
Branch creation	Allow
Contribute	Allow
Exempt from policy enforcement	Not set
Note management	Allow
Read	Allow
Rewrite and destroy history (force push)	Not set
Tag creation	Allow

At the bottom of the security section are buttons for 'Remove', 'Save changes', and 'Undo changes'.

- Select **Git** as the **Type**, and enter *GroupUtilities* as the **Name**, and then click **Create**.



Now you should see two Git repositories **GroupProjectTemplate** and **GroupUtilities** in the left column of the **Version Control** page:

Access Control Summary	Description
Administrator	Not set
Branch creation	Inherited allow
Contribute	Inherited allow
Exempt from policy enforcement	Not set
Note management	Inherited allow
Read	Inherited allow
Rewrite and destroy history (force push)	Not set
Tag creation	Inherited allow

4. Create the GroupProjectTemplate (R1) repository

The setup of the repositories for the Azure DevOps group server consists of two tasks:

- Rename the default **GroupCommon** repository **GroupProjectTemplate**.
- Create the **GroupUtilities** repository on the Azure DevOps Services under project **GroupCommon**.

Instructions for the first task are contained in this section after remarks on naming conventions or our repositories and directories. The instructions for the second task are contained in the following section for step 4.

Rename the default GroupCommon repository

To rename the default **GroupCommon** repository as *GroupProjectTemplate* (referred as **R1** in this tutorial):

- Click **Collaborate on code** on the **GroupCommon** project page. This takes you to the default Git repository page of the project **GroupCommon**. Currently, this Git repository is empty.

The screenshot shows the 'Overview' page of a team project named 'GroupCommon'. The top navigation bar includes links for HOME, CODE, WORK, BUILD, TEST, and RELEASE. The 'Overview' tab is selected. The main content area is divided into several sections:

- Welcome:** A general introduction to the service.
- Manage Work:** Options to add work to the board and collaborate on code (highlighted with a red box).
- Collaborate on code:** Option to add code to your repository.
- Continuously integrate:** Option to automate your builds.
- Visualize progress:** Option to learn how to add charts.
- Open User Stories:** A section showing a query result with 0 items.
- Team Members:** Shows a single member and a button to invite friends.
- Work:** Links to Backlog, Board, Task board, and Queries.
- Sprint Burndown:** A section for setting iteration dates.
- New Work Item:** A form to enter a title (e.g., 'Bug') and a dropdown menu, with a 'Create' button.
- Open User Stories:** A summary card showing 0 work items.
- Visual Studio:** Options to open in Visual Studio (requires 2013+) and get Visual Studio.

- Click **GroupCommon** on the top left corner (highlighted with a red box in the following figure) on the Git repository page of **GroupCommon** and select **Manage repositories** (highlighted with a green box in the following figure). This procedure brings up the **CONTROL PANEL**.
- Select the **Version Control** tab of your project.

Team Services / GroupCommon

HOME CODE WORK BUILD TEST RELEASE

GroupCommon Explorer History Branches Pull Requests

Filter repositories

GroupCommon

+ New repository...

Manage repositories...

Empty. Add some code!

Command line or another Git client

Clone URL

HTTPS | SSH https://mysamplegroup.visualstudio.com/_git/GroupCommon

Generate Git credentials

Download Git for Windows

Plug-ins and credential managers

These provide the best experience with single sign in, multi-factor auth, and integration with p

IntelliJ IDEA Android Studio Eclipse Windows command line

Command line instructions

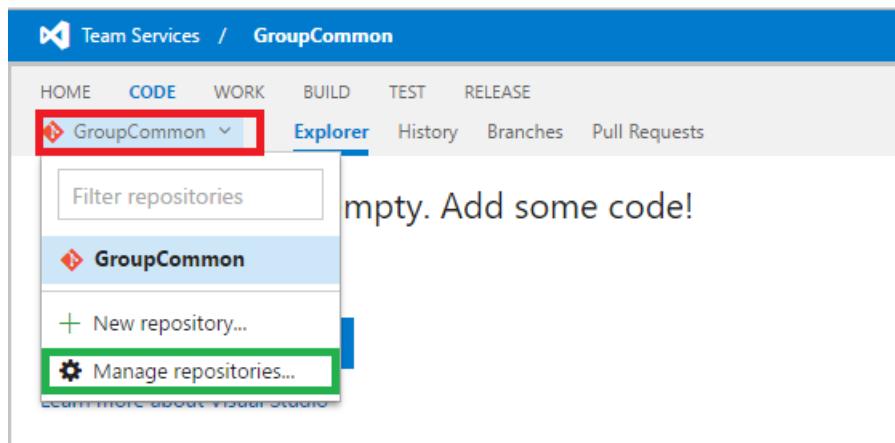
Clone this repository

```
git clone https://mysamplegroup.visualstudio.com/_git/GroupCommon
```

Push an existing repository

```
git remote add origin https://mysamplegroup.visualstudio.com/_gi  
t/GroupCommon
```

```
git push -u origin --all
```



Empty. Add some code!

Command line or another Git client

Clone URL

HTTPS | SSH https://mysamplegroup.visualstudio.com/_git/GroupCommon

Generate Git credentials

Download Git for Windows

Plug-ins and credential managers

These provide the best experience with single sign in, multi-factor auth, and integration with p

IntelliJ IDEA Android Studio Eclipse Windows command line

Command line instructions

Clone this repository

```
git clone https://mysamplegroup.visualstudio.com/_git/GroupCommon
```

Push an existing repository

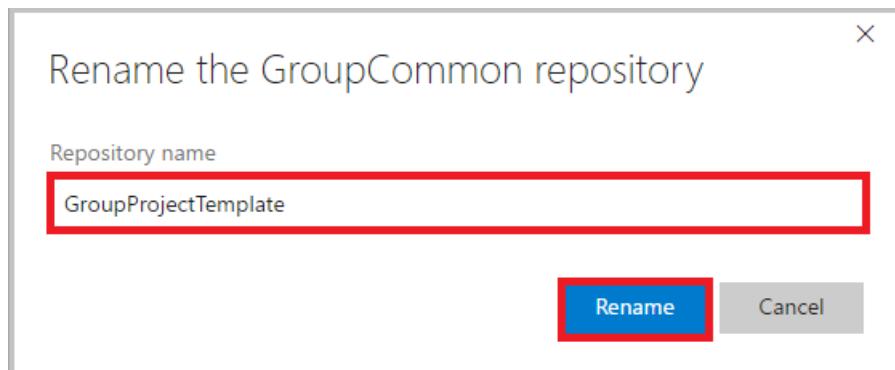
```
git remote add origin https://mysamplegroup.visualstudio.com/_gi  
t/GroupCommon
```

```
git push -u origin --all
```

- Click the ... to the right of the **GroupCommon** repository on the left panel, and select **Rename repository**.

Permission	Value
Administer	Not set
Branch creation	Allow
Contribute	Allow
Exempt from policy enforcement	Not set
Note management	Allow
Read	Allow
Rewrite and destroy history (force push)	Not set
Tag creation	Allow

- In the **Rename the GroupCommon repository** wizard that pops up, enter *GroupProjectTemplate* in the **Repository name** box, and then click **Rename**.



5. Seed the R1 & R2 repositories on the Azure DevOps Services

In this stage of the procedure, you seed the *GroupProjectTemplate* (R1) and *GroupUtilities* (R2) repositories that you set up in the previous section. These repositories are seeded with the **ProjectTemplate (G1)** and **Utilities (G2)** repositories that are managed by Microsoft for the Team Data Science Process. When this seeding is completed:

- your R1 repository is going to have the same set of directories and document templates that the G1 does
- your R2 repository is going to contain the set of data science utilities developed by Microsoft.

The seeding procedure uses the directories on your local DSVM as intermediate staging sites. Here are the steps followed in this section:

- G1 & G2 - cloned to -> LG1 & LG2
- R1 & R2 - cloned to -> LR1 & LR2
- LG1 & LG2 - files copied into -> LR1 & LR2
- (Optional) customization of LR1 & LR2
- LR1 & LR2 - contents add to -> R1 & R2

Clone G1 & G2 repositories to your local DSVM

In this step, you clone the Team Data Science Process (TDSP) ProjectTemplate repository (G1) and Utilities (G2) from the TDSP GitHub repositories to folders in your local DSVM as LG1 and LG2:

- Create a directory to serve as the root directory to host all your clones of the repositories.

- In the Windows DSVM, create a directory `C:\GitRepos\TDSPCommon`.
- In the Linux DSVM, create a directory `GitRepos\TDSPCommon` in your home directory.
- Run the following set of commands from the `GitRepos\TDSPCommon` directory.

```
git clone https://github.com/Azure/Azure-TDSP-ProjectTemplate
git clone https://github.com/Azure/Azure-TDSP-Utilities
```

```
PS C:\GitRepos\TDSPCommon> git clone https://github.com/Azure/Azure-TDSP-ProjectTemplate
Cloning into 'Azure-TDSP-ProjectTemplate'...
remote: Counting objects: 41, done.
remote: Total 41 (delta 0), reused 0 (delta 0), pack-reused 41
Unpacking objects: 100% (41/41), done.
PS C:\GitRepos\TDSPCommon> git clone https://github.com/Azure/Azure-TDSP-Utilities
Cloning into 'Azure-TDSP-Utilities'...
remote: Counting objects: 96, done.
remote: Compressing objects: 100% (89/89), done.
remote: Total 96 (delta 12), reused 0 (delta 0), pack-reused 6
Unpacking objects: 100% (96/96), done.
```

- Using our abbreviated repository names, this is what these scripts have achieved:
 - G1 - cloned into -> LG1
 - G2 - cloned into -> LG2
- After the cloning is completed, you should be able to see two directories, `ProjectTemplate` and `Utilities`, under `GitRepos\TDSPCommon` directory.

Clone R1 & R2 repositories to your local DSVM

In this step, you clone the GroupProjectTemplate repository (R1) and GroupUtilities repository (R2) on local directories (referred as LR1 and LR2, respectively) under `GitRepos\GroupCommon` on your DSVM.

- To get the URLs of the R1 and R2 repositories, go to your **GroupCommon** home page on Azure DevOps Services. This usually has the URL `https://<Your Azure DevOps Services Name>.visualstudio.com/GroupCommon`.
- Click **CODE**.
- Choose the **GroupProjectTemplate** and **GroupUtilities** repositories. Copy and save each of the URLs (HTTPS for Windows; SSH for Linux) from the **Clone URL** element, in turn, for use in the following scripts:

The screenshot shows the 'GroupCommon' repository page on Azure DevOps Services. The 'GroupProjectTemplate' repository is selected. The 'Clone in your favorite IDE' section has a 'Clone in Visual Studio' button. The 'Command line or another Git client' section has a 'Clone URL' input field containing 'HTTPS | SSH https://weig-ds.visualstudio.com/GroupCommon/_git/Group..'. A red box highlights the 'Clone URL' field.

- Change into the `GitRepos\GroupCommon` directory on your Windows or Linux DSVM and run one of

the following sets of commands to clone R1 and R2 into that directory.

Here are the Windows and Linux scripts:

```
# Windows DSVM

git clone <the HTTPS URL of the GroupProjectTemplate repository>
git clone <the HTTPS URL of the GroupUtilities repository>
```

```
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2> git clone https://weig-ds.visualstudio.com/GroupCommon/_git/GroupProjectTemplate
Cloning into 'GroupProjectTemplate'...
warning: You appear to have cloned an empty repository.
Checking connectivity... done.
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2> git clone https://weig-ds.visualstudio.com/GroupCommon/_git/GroupUtilities
Cloning into 'GroupUtilities'...
warning: You appear to have cloned an empty repository.
Checking connectivity... done.
```

```
# Linux DSVM

git clone <the SSH URL of the GroupProjectTemplate repository>
git clone <the SSH URL of the GroupUtilities repository>
```

```
[dsl@weiglinuxdsvm1 test2]$ git clone ssh://weig-ds@weig-ds.visualstudio.com:22/GroupCommon/_git/GroupProjectTemplate
Cloning into 'GroupProjectTemplate'...
warning: You appear to have cloned an empty repository.
[dsl@weiglinuxdsvm1 test2]$ git clone ssh://weig-ds@weig-ds.visualstudio.com:22/GroupCommon/_git/GroupUtilities
Cloning into 'GroupUtilities'...
warning: You appear to have cloned an empty repository.
```

NOTE

Expect to receive warning messages that LR1 and LR2 are empty.

- Using our abbreviated repository names, this is what these scripts have achieved:
 - R1 - cloned into -> LR1
 - R2 - cloned into -> LR2

Seed your GroupProjectTemplate (LR1) and GroupUtilities (LR2)

Next, in your local machine, copy the content of ProjectTemplate and Utilities directories (except the metadata in the .git directories) under GitRepos\TDSPCommon to your GroupProjectTemplate and GroupUtilities directories under **GitRepos\GroupCommon**. Here are the two tasks to complete in this step:

- Copy the files in GitRepos\TDSPCommon\ProjectTemplate (**LG1**) to
GitRepos\GroupCommon\GroupProjectTemplate (**LR1**)
- Copy the files in GitRepos\TDSPCommon\Utilities (**LG2**) to GitRepos\GroupCommon\Utilities (**LR2**).

To achieve these two tasks, run the following scripts in PowerShell console (Windows) or Shell script console (Linux). You are prompted to input the complete paths to LG1, LR1, LG2, and LR2. The paths that you input are validated. If you input a directory that does not exist, you are asked to input it again.

```
# Windows DSVM

wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-
DataScience/master/Misc/TDSP/tdsp_local_copy_win.ps1" -outfile "tdsp_local_copy_win.ps1"
.\tdsp_local_copy_win.ps1 1
```

```

PS D:\AML_Projects\TDSP-Linux\test0912> wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/TDSP/tdsp_local_copy_win.ps1" -outfile "tdsp_local_copy_win.ps1"
PS D:\AML_Projects\TDSP-Linux\test0912> .\tdsp_local_copy_win.ps1
Please input the full path to ProjectTemplate repository from Microsoft TDSP team (source directory): D:\AML_Projects\TDSP-Linux\test0912\ProjectTemplate
Please input the full path to Your GroupProjectTemplate repository (destination directory): D:\AML_Projects\TDSP-Linux\test0912\GroupProjectTemplate
Start copying files (except files in .git directory) from D:\AML_Projects\TDSP-Linux\test0912\ProjectTemplate to D:\AML_Projects\TDSP-Linux\test0912\GroupProjectTemplate...
Please input the full path to Utilities repository from Microsoft TDSP team (source directory): D:\AML_Projects\TDSP-Linux\test0912\Utilities
Please input the full path to Your GroupUtilities repository (destination directory): D:\AML_Projects\TDSP-Linux\test0912\GroupUtilities
Start copying files (except files in .git directory) from D:\AML_Projects\TDSP-Linux\test0912\Utilities to D:\AML_Projects\TDSP-Linux\test0912\GroupUtilities...
PS D:\AML_Projects\TDSP-Linux\test0912>

```

Now you can see that files in directories LG1 and LG1 (except files in the .git directory) have been copied to LR1 and LR2, respectively.

```

PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $src = "ProjectTemplate"
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $dest = "GroupProjectTemplate"
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $SourceDirectory = $PWD.Path+"\\"+$src
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $DestinationDirectory = $PWD.Path+"\\"+$dest
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $ExcludeSubDirectory = $SourceDirectory+'\.git'
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $files = Get-ChildItem $SourceDirectory -Recurse | Where-Object { $ExcludeSubDirectory -notcontains $_.DirectoryName }
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing>
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> foreach ($file in $files)
>> {
>>     $CopyPath = Join-Path $DestinationDirectory $file.FullName.Substring($SourceDirectory.length)
>>     Copy-Item $file.FullName -Destination $CopyPath
>> }
>>
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing>
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing>
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $src = "Utilities"
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $dest = "GroupUtilities"
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $SourceDirectory = $PWD.Path+"\\"+$src
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $DestinationDirectory = $PWD.Path+"\\"+$dest
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $ExcludeSubDirectory = $SourceDirectory+'\.git'
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> $files = Get-ChildItem $SourceDirectory -Recurse | Where-Object { $ExcludeSubDirectory -notcontains $_.DirectoryName }
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing>
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing> foreach ($file in $files)
>> {
>>     $CopyPath = Join-Path $DestinationDirectory $file.FullName.Substring($SourceDirectory.length)
>>     Copy-Item $file.FullName -Destination $CopyPath
>> }
>>
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing>

```

```

# Linux DSVM

wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/TDSP/tdsp_local_copy_linux.sh"
bash tdsp_local_copy_linux.sh

```

```

[dsl@linuxdsvm test0912]$ wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/TDSP/tdsp_local_copy_linux.sh"
--2016-09-12 16:22:43 -- https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/TDSP/tdsp_local_copy_linux.sh
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.48.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.48.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2057 (2.0K) [text/plain]
Saving to: 'tdsp_local_copy_linux.sh'

100%[=====] 2,057      ...-k/s   in 0s

2016-09-12 16:22:43 (34.6 MB/s) - 'tdsp_local_copy_linux.sh' saved [2057/2057]

[dsl@linuxdsvm test0912]$ hash tdsp_local_copy_linux.sh
Please input the full path to ProjectTemplate repository from Microsoft TDSP team (source directory): /home/dsl/test0912/ProjectTemplate
Please input the full path to Your GroupProjectTemplate repository (destination directory): /home/dsl/test0912/GroupProjectTemplate
Code/
Code/DataPrep/
Code/DataPrep/dataPrep.py
Code/Model/
Code/Model/model.R
Code/Operationalization/
Code/Operationalization/Operationalization.py
Data/

```

Now you see that the files in the two folders (except files in the .git directory) are copied to GroupProjectTemplate and GroupUtilities respectively.

```
[dsl@weiglinuxdsvm ~]$ src="ProjectTemplate"
[dsl@weiglinuxdsvm ~]$ dest="GroupProjectTemplate"
[dsl@weiglinuxdsvm ~]$ SourceDirectory=$PWD/$src
[dsl@weiglinuxdsvm ~]$ DestinationDirectory=$PWD/$dest
[dsl@weiglinuxdsvm ~]$ cd $SourceDirectory
[dsl@weiglinuxdsvm ProjectTemplate]$ git archive HEAD --format=tar | (cd $DestinationDirectory; tar xvf -)
Code/
Code/DataPrep/
Code/DataPrep/dataPrep.py
Code/Model/
Code/Model/model.R
Code/Operationalization/
Code/Operationalization/operationalization.py
Data/
Data/Modeling/
Data/Modeling/modelling.md
Data/Processed/
Data/Processed/processed.md
Data/Raw/
Data/Raw/rawData.md
Docs/
Docs/DataDictionaries/
Docs/DataDictionaries/ReadMe.md
Docs/DataDictionaries/dict1.csv
Docs/DataReport/
Docs/DataReport/Data Defintion.md
Docs/DataReport/DataPipeline.txt
Docs/DataReport/DataQualityReport.md
Docs/DataReport/ReadMe.md
Docs/Model/
Docs/Model/Baseline/
Docs/Model/Baseline/Baseline Models.md
Docs/Model/FinalReport.md
Docs/Model/Model 1/
Docs/Model/Model 1/Model Report.md
Docs/Project/
Docs/Project/Charter.md
Docs/Project/Exit Report.md
Docs/Project/Glossary.md
Docs/Project/System Architecture.docx
Docs/Project/UseCases.md
README.md
```

- Using our abbreviated repository names, this is what these scripts have achieved:
 - LG1 - files copied into -> LR1
 - LG2 - files copied into -> LR2

Option to customize the contents of LR1 & LR2

If you want to customize the contents of LR1 and LR2 to meet the specific needs of your group, this is the stage of the procedure where that is appropriate. You can modify the template documents, change the directory structure, and add existing utilities that your group has developed or that are helpful for your entire group.

Add the contents in LR1 & LR2 to R1 & R2 on group server

Now, you need to add the contents in LR1 and LR2 to repositories R1 and R2. Here are the git bash commands you can run in either Windows PowerShell or Linux.

Run the following commands from the GitRepos\GroupCommon\GroupProjectTemplate directory:

```
git status
git add .
git commit -m"push from DSVM"
git push
```

The -m option lets you set a message for your git commit.

```

PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2\GroupProjectTemplate> git status
On branch master

Initial commit

Untracked files:
  (use "git add <file>..." to include in what will be committed)

    Code/
    Data/
    Docs/
    README.md

nothing added to commit but untracked files present (use "git add" to track)
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2\GroupProjectTemplate> git add .
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2\GroupProjectTemplate> git commit -m"push from local Win DSVM"
[master (root-commit) 4170390] push from local Win DSVM
 21 files changed, 233 insertions(+)
 create mode 100644 Code/DataPrep/dataPrep.py
 create mode 100644 Code/Model/model.R
 create mode 100644 Code/Operationalization/operationalization.py
 create mode 100644 Data/Modeling/modelling.md
 create mode 100644 Data/Processed/processed.md
 create mode 100644 Data/Raw/rawData.md
 create mode 100644 Docs/DataDictionaries/ReadMe.md

```

You can see that in your group's Azure DevOps Services, in the GroupProjectTemplate repository, the files are synced instantly.

Name	Last Change	Comments
Code	a minute ago	push from local Win DSVM - Wei Guo
Data	a minute ago	push from local Win DSVM - Wei Guo
Docs	a minute ago	push from local Win DSVM - Wei Guo
README.md	a minute ago	push from local Win DSVM - Wei Guo

Finally, change to the **GitRepos\GroupCommon\GroupUtilities** directory and run the same set of git bash commands:

```

git status
git add .
git commit -m"push from DSVM"
git push

```

NOTE

If this is the first time you commit to a Git repository, you need to configure global parameters `user.name` and `user.email` before you run the `git commit` command. Run the following two commands:

```

git config --global user.name <your name>
git config --global user.email <your email address>

```

If you are committing to multiple Git repositories, use the same name and email address when you commit to each of them. Using the same name and email address proves convenient later on when you build PowerBI dashboards to track your Git activities on multiple repositories.

- Using our abbreviated repository names, this is what these scripts have achieved:
 - LR1 - contents add to -> R1
 - LR2 - contents add to -> R2

6. Add group members to the group server

From your group Azure DevOps Services's homepage, click the **gear icon** next to your user name in the upper right corner, then select the **Security** tab. You can add members to your group here with various permissions.

The screenshot shows the Azure DevOps Control Panel with the Security tab selected. On the left, there is a search bar for 'Create VSTS group' and a list of users. A user named 'yuso' is selected, showing their details: 'yuso' and 'yuso@microsoft.com'. Below this, it says 'Showing 1 result' and lists several groups: 'Project Collection Build Administrators', 'Project Collection Build Service Accounts', 'Project Collection Proxy Service Accounts', 'Project Collection Service Accounts', 'Project Collection Test Service Accounts', 'Project Collection Valid Users', and 'Security Service Group'. On the right, the 'weig-ds > Project Collection Administrators' page is displayed. It has tabs for 'Permissions', 'Members', and 'Member of'. Under 'Permissions', it says 'Members of this application group can perform all privileged operations or Administrator group can not be modified.' Below this, a long list of permissions is shown, each with an 'Allow' status: Administer build resource permissions, Administer process permissions, Administer Project Server integration, Administer shelved changes, Administer workspaces, Alter trace settings, Create a workspace, Create new projects, Create process, Delete process, Delete team project, Edit collection-level information, Edit process, Make requests on behalf of others, Manage build resources, Manage test controllers, Trigger events, Use build resources, View build resources, View collection-level information, and View system synchronization information. At the bottom, there is a 'Clear explicit permissions' link and two buttons: 'Save changes' and 'Undo changes'.

Next steps

Here are links to the more detailed descriptions of the roles and tasks defined by the Team Data Science Process:

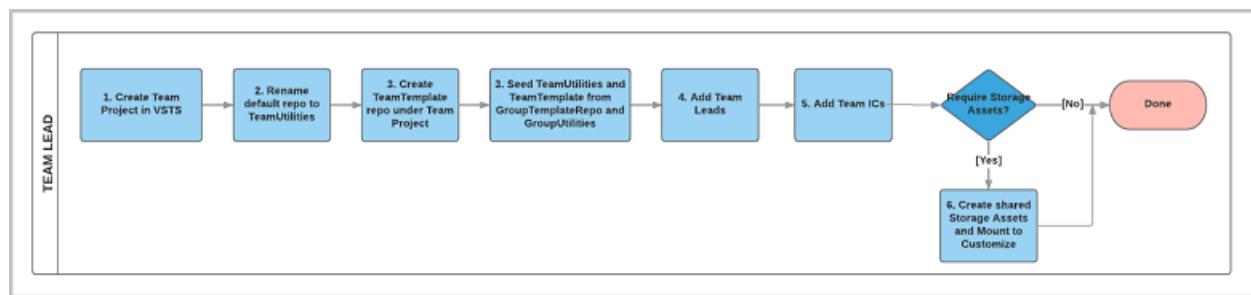
- [Group Manager tasks for a data science team](#)
- [Team Lead tasks for a data science team](#)
- [Project Lead tasks for a data science team](#)
- [Project Individual Contributors for a data science team](#)

Tasks for the team lead in the Team Data Science Process Team

1/30/2019 • 16 minutes to read

This topic outlines the tasks that a team lead is expected to complete for their data science team. The objective is to establish collaborative team environment that standardizes on the [Team Data Science Process](#) (TDSP). TDSP is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently. It is designed to help improve collaboration and team learning. The process is a distillation of the best practices and structures from both Microsoft as well as from the industry, needed for successful implementation of data science initiatives to help companies fully realize the benefits of their analytics programs. For an outline of the personnel roles and their associated tasks that are handled by a data science team standardizing on this process, see [Team Data Science Process roles and tasks](#).

A **Team Lead** manages a team in the data science unit of an enterprise. A team consists of multiple data scientists. For data science unit with only a small number of data scientists, the **Group Manager** and the **Team Lead** might be the same person or they could delegate their task to a surrogate. But the tasks themselves do not change. The workflow for the tasks to be completed by team leads to set up this environment are depicted in the following figure:



[AZURE.NOTE] The tasks in blocks 1 and 2 of the figure are needed if you are using Azure DevOps as the code hosting platform and you want to have a separate Azure DevOps project for your own team. Once these tasks are completed, all repositories of your team can be created under this project.

After several prerequisites tasks specified in a following section are satisfied by the group manager, there are the five principal tasks (some optional) that you complete in this tutorial. These tasks correspond the main numbered sections of this topic:

1. Create a **project** on the group's Azure DevOps Services of the group and two team repositories in the project:
 - **ProjectTemplate repository**
 - **TeamUtilities repository**
2. Seed the team **ProjectTemplate** repository from the **GroupProjectTemplate** repository which has been set up by your group manager.
3. Create team data and analytics resources:
 - Add the team-specific utilities to the **TeamUtilities** repository.
 - (Optional) Create an **Azure file storage** to be used to store data assets that can be useful for the entire team.
4. (Optional) Mount the Azure file storage to the **Data Science Virtual Machine** (DSVM) of the team lead and add data assets on it.
5. Set up the **security control** by adding team members and configure their privileges.

[AZURE.NOTE] We outline the steps needed to set up a TDSP team environment using Azure DevOps in the following instructions. We specify how to accomplish these tasks with Azure DevOps because that is how we implement TDSP at Microsoft. If another code hosting platform is used for your group, the tasks that need to be completed by the team lead generally do not change. But the way to complete these tasks is going to be different.

Repositories and directories

This topic uses abbreviated names for repositories and directories. These names make it easier to follow the operations between the repositories and directories. This notation (**R** for Git repositories and **D** for local directories on your DSVM) is used in the following sections:

- **R1:** The **GroupProjectTemplate** repository on Git that your group manager set up on your Azure DevOps group server.
- **R3:** The team **ProjectTemplate** repository on Git you set up.
- **R4:** The **TeamUtilities** repository on Git you set up.
- **D1:** The local directory cloned from R1 and copied to D3.
- **D3:** The local directory cloned from R3, customize, and copied back to R3.
- **D4:** The local directory cloned from R4, customize, and copied back to R4.

The names specified for the repositories and directories in this tutorial have been provided on the assumption that your objective is to establish a separate project for your own team within a larger data science group. But there are other options open to you as team lead:

- The entire group can choose to create a single project. Then all projects from all data science teams would be under this single project. To achieve this, you can designate a git administrator to follow these instructions to create a single project. This scenario might be valid, for example, for:
 - a small data science group that does not have multiple data science teams
 - a larger data science group with multiple data science teams that nevertheless wants to optimize inter-team collaboration with activities such as group-level sprint planning.
- Teams can choose to have team-specific project templates or team-specific utilities under the single project for the entire group. In this case, the team leads should create project template repositories and/or team utilities repositories under the same project. Name these repositories <*TeamName*>*ProjectTemplate* and <*TeamName*>*Utilities*, for instance, *TeamJohnProjectTemplate* and *TeamJohnUtilities*.

In any case, team leads need to let their team members know which template and utilities repositories to adopt when they are setting up and cloning the project and utilities repositories. Project leads should follow the [Project Lead tasks for a data science team](#) to create project repositories, whether under separate projects or under a single project.

0. Prerequisites

The prerequisites are satisfied by completing the tasks assigned to your group manager outlined in [Group Manager tasks for a data science team](#). To summarize here, the following requirements need to meet before you begin the team lead tasks:

- Your **group Azure DevOps Services** (or group account on some other code hosting platform) has been set up by your group manager.
- Your **GroupProjectTemplate repository** (R1) has been set up on your group account by your group manager on the code hosting platform you plan to use.
- You have been **authorized** on your group account to create repositories for your team.
- Git must be installed on your machine. If you are using a Data Science Virtual Machine (DSVM), Git has been pre-installed and you are good to go. Otherwise, see the [Platforms and tools appendix](#).

- If you are using a **Windows DSVM**, you need to have [Git Credential Manager \(GCM\)](#) installed on your machine. In the README.md file, scroll down to the **Download and Install** section and click the *latest installer*. This takes you to the latest installer page. Download the .exe installer from here and run it.
- If you are using **Linux DSVM**, create an SSH public key on your DSVM and add it to your group Azure DevOps Services. For more information about SSH, see the **Create SSH public key** section in the [Platforms and tools appendix](#).

1. Create a project and repositories

Complete this step if you are using Azure DevOps as your code hosting platform for versioning and collaboration. This section has you create three artifacts in the Azure DevOps Services of your group:

- **MyTeam** project in Azure DevOps
- **MyProjectTemplate** repository (**R3**) on Git
- **MyTeamUtilities** repository (**R4**) on Git

Create the MyTeam project

- Go to your group's Azure DevOps Services homepage at URL
<https://<Azure DevOps Services Name>.visualstudio.com>.
- Click **New** to create a project.

The screenshot shows the Azure DevOps Services homepage with the 'Overview' tab selected. The main content area includes sections for 'Recent projects & teams' (with a 'New' button), 'Recent team rooms' (with a note to 'Browse to the Rooms hub'), and 'Enjoy free benefits' (listing 'Download software, use Azure services, view online training and more' with icons for GitHub, Microsoft Azure, and Pluralsight). A large callout on the right side promotes 'Visualize data across your enterprise' using Power BI, showing a monitor displaying charts and graphs.

- A Create project window asks you to input the Project name (**MyTeam** in this example). Make sure that you select **Agile** as the **Process template** and **Git** as the **Version control**.

Create team project

Project name
MyTeam

Description

Process template
Agile

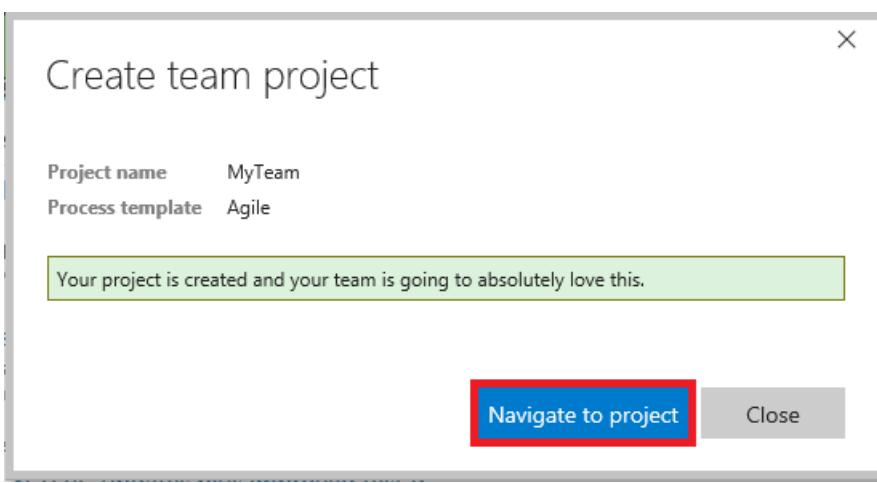
This template is flexible and will work great for most teams using Agile planning methods, including those practicing Scrum.

Version control
Git

Git is a Distributed Version Control System (DVCS) that uses a local repository to track and version files. Changes are shared with other developers by pushing and pulling changes through a remote, shared repository.

Create project **Cancel**

- Click **Create project**. Your project **MyTeam** is created in less than 1 minute.
- After the project **MyTeam** is created, click **Navigate to project** button, to be directed to the home page of your project.



- If you see a **Congratulations!** popup window, click the **Add code** (button in red box). Otherwise, click **Code** (in yellow box). This directs you to the Git repository page of your project.

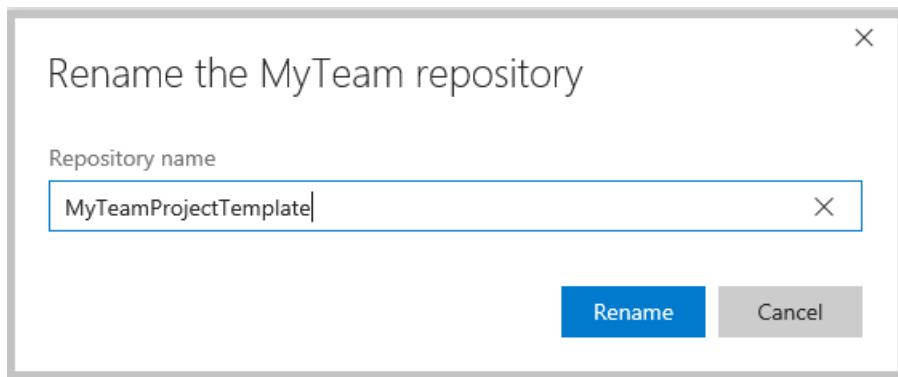
Create the MyProjectTemplate repository (R3) on Git

- On the Git repository page of your project, click the downward arrow beside repository name **MyTeam**, and select **Manage repositories....**

- On the **Version control** tab of the control panel of your project, click **MyTeam**, then select **Rename repository....**

Role	Permission
Administer	Inherited allow
Branch creation	Inherited allow
Contribute	Inherited allow
Exempt from policy enforcement	Not set
Note management	Inherited allow
Read	Inherited allow
Rewrite and destroy history (force push)	Inherited allow
Tag creation	Inherited allow

- Input a new name to the repository in the **Rename the MyTeam repository** window. In this example, *MyTeamProjectTemplate*. You can choose something like <*Your team name*>*ProjectTemplate*. Click **Rename** to continue.

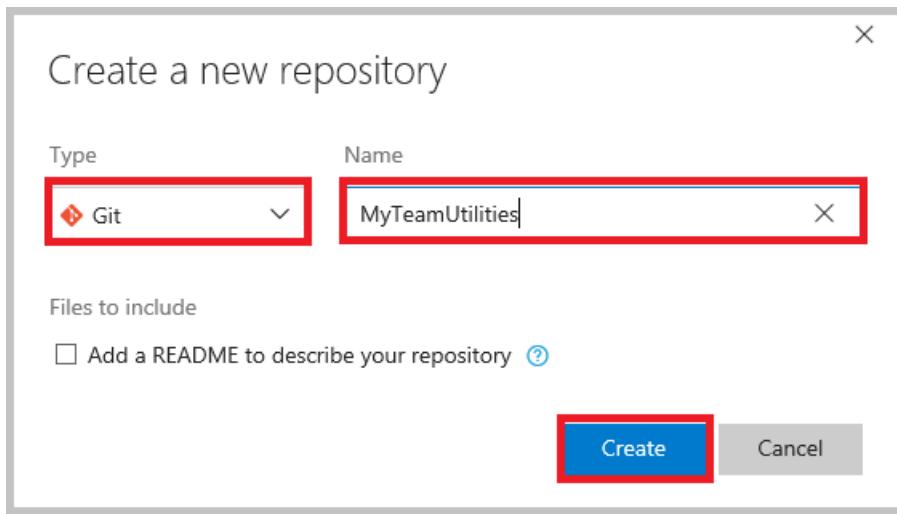


Create the MyTeamUtilities repository (R4) on Git

- To create a new repository <*your team name*>*Utilities* under your project, click **New repository...** on the **Version control** tab of your project's control panel.

Role	Permission
Administer	Inherited allow
Branch creation	Inherited allow
Contribute	Inherited allow
Exempt from policy enforcement	Not set
Note management	Inherited allow
Read	Inherited allow
Rewrite and destroy history (force push)	Inherited allow
Tag creation	Inherited allow

- In the **Create a new repository** window that pops up, provide a name for this repository. In this example, we name it as *MyTeamUtilities*, which is **R4** in our notation. Choose something like <*your team name*>*Utilities*. Make sure that you select **Git** for **Type**. Then, click **Create** to continue.



- Confirm that you see the two new Git repositories created under your project **MyTeam**. In this example:
- **MyTeamProjectTemplate** (R3)
- **MyTeamUtilities** (R4).

Access Control Summary	Description
Administer	Inherited allow
Branch creation	Inherited allow
Contribute	Inherited allow
Exempt from policy enforcement	Not set
Note management	Inherited allow
Read	Inherited allow
Rewrite and destroy history (force push)	Inherited allow
Tag creation	Inherited allow

2. Seed your ProjectTemplate and TeamUtilities repositories

The seeding procedure uses the directories on your local DSVM as intermediate staging sites. If you need to customize your **ProjectTemplate** and **TeamUtilities** repositories to meet some specific team needs, you do so in the penultimate step of following procedure. Here is a summary of the steps used to seed the content of the **MyTeamProjectTemplate** and **MyTeamUtilities** repositories for a data science team. The individual steps correspond to the subsections in the seeding procedure:

- Clone group repository into local directory: team R1 - cloned to -> local D1
- Clone your team repositories into local directories: team R3 & R4 - cloned to -> local D3 & D4
- Copy the group project template content to the local team folder: D1 - contents copied to -> D3
- (Optional) customization of local D3 & D4
- Push local directory content to team repositories: D3 & D4 - contents add to -> team R3 & R4

Initialize the team repositories

In this step, you initialize your project template repository from the group project template repository:

- **MyTeamProjectTemplate** repository (R3) from your **GroupProjectTemplate** (R1) repository

Clone group repositories into local directories

To begin this procedure:

- Create directories on your local machine:
 - For **Windows**: **C:\GitRepos\GroupCommon** and **C:\GitRepos\MyTeam**
 - For **Linux**: **GitRepos\GroupCommon** and **GitRepos\MyTeam** on your home directory
- Change to directory **GitRepos\GroupCommon**.
- Run the following command, as appropriate, on the operating system of your local machine.

Windows

```
git clone https://<Your Azure DevOps Services name>.visualstudio.com/GroupCommon/_git/GroupProjectTemplate
```

```
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test_team> git clone https://weig-ds.visualstudio.com/GroupCommon/_git/GroupProjectTemplate
Cloning into 'GroupProjectTemplate'...
remote:
remote:          vSTs
remote:          vSTSVSTSv
remote:          vSTSVSTSvST
remote: VSTS      vSTSVSTSvSTSV
remote: VSTSvS    vSTSVSTSv STSVs
remote: VSTSvSTSvTSVSTSvS  TSVST
remote: VS tSVSTSvSTSVs  STSVs
remote: VS lSVSTSvST    SVSTS
remote: VS tSVSTSvSTSVsts  VSTSV
remote: VSTSvST  SVSTSvSTSs VSTSV
remote: VSTSv     STSVSTSvSTSv
remote:          VSTSvSTSvST
remote:          VSTSvSTSs
remote:          VSTSs   (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Unpacking objects: 100% (34/34), done.
Checking connectivity... done.
```

Linux

```
git clone ssh://<Your Azure DevOps Services name>@<Your Azure DevOps Services
name>.visualstudio.com:22/GroupCommon/_git/GroupProjectTemplate
```

```
[dsl@weiglinuxdsvml test_team]$ git clone ssh://weig-ds@weig-ds.visualstudio.com:22/GroupCommon/_git/GroupProjectTemplate
Cloning into 'GroupProjectTemplate'...
remote:
remote:          vSTs
remote:          vSTSVSTSv
remote:          vSTSVSTSvST
remote: VSTS      vSTSVSTSvSTSV
remote: VSTSvS    vSTSVSTSv STSVs
remote: VSTSvSTSvTSVSTSvS  TSVST
remote: VS tSVSTSvSTSVs  STSVs
remote: VS tSVSTSvST    SVSTS
remote: VS tSVSTSvSTSVsts  VSTSV
remote: VSTSvST  SVSTSvSTSs VSTSV
remote: VSTSv     STSVSTSvSTSv
remote:          VSTSvSTSvST
remote:          VSTSvSTSs
remote:          VSTSs   (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Receiving objects: 100% (34/34), 14.65 KiB | 0 bytes/s, done.
Resolving deltas: 100% (2/2), done.
```

These commands clone your **GroupProjectTemplate** (R1) repository on your group Azure DevOps Services to local directory in **GitRepos\GroupCommon** on your local machine. After cloning, directory

GroupProjectTemplate (D1) is created in directory **GitRepos\GroupCommon**. Here, we assume that your group manager created a project **GroupCommon**, and the **GroupProjectTemplate** repository is under this project.

Clone your team repositories into local directories

These commands clone your **MyTeamProjectTemplate** (R3) and **MyTeamUtilities** (R4) repositories under your project **MyTeam** on your group Azure DevOps Services to the **MyTeamProjectTemplate** (D3) and

MyTeamUtilities (D4) directories in **GitRepos\MyTeam** on your local machine.

- Change to directory **GitRepos\MyTeam**
- Run the following commands, as appropriate, on the operating system of your local machine.

Windows

```
git clone https://<Your Azure DevOps Services name>.visualstudio.com/<Your Team Name>/_git/MyTeamProjectTemplate  
git clone https://<Your Azure DevOps Services name>.visualstudio.com/<Your Team Name>/_git/MyTeamUtilities
```

```
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test_team> git clone https://weig-ds.visualstudio.com/DS_Team_1/_git/DS_Team_1_ProjectTemplate  
Cloning into 'DS_Team_1_ProjectTemplate'...  
warning: You appear to have cloned an empty repository.  
Checking connectivity... done.  
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test_team> git clone https://weig-ds.visualstudio.com/DS_Team_1/_git/DS_Team_1_Utils  
Cloning into 'DS_Team_1_Utils'...  
warning: You appear to have cloned an empty repository.  
Checking connectivity... done.
```

Linux

```
git clone ssh://<Your Azure DevOps Services name>@<Your Azure DevOps Services name>.visualstudio.com:22/<Your Team Name>/_git/MyTeamProjectTemplate  
git clone ssh://<Your Azure DevOps Services name>@<Your Azure DevOps Services name>.visualstudio.com:22/<Your Team Name>/_git/MyTeamUtilities
```

```
[dsl@weiglinuxdsvml test_team]$ git clone ssh://weig-ds@weig-ds.visualstudio.com:22/DS_Team_1/_git/DS_Team_1_Utils  
Cloning into 'DS_Team_1_Utils'...  
warning: You appear to have cloned an empty repository.  
[dsl@weiglinuxdsvml test_team]$ git clone ssh://weig-ds@weig-ds.visualstudio.com:22/DS_Team_1/_git/DS_Team_1_ProjectTemplate  
Cloning into 'DS_Team_1_ProjectTemplate'...  
warning: You appear to have cloned an empty repository.
```

After cloning, two directories **MyTeamProjectTemplate** (D3) and **MyTeamUtilities** (D4) are created in directory **GitRepos\MyTeam**. We have assumed here that you named your project template and utilities repositories **MyTeamProjectTemplate** and **MyTeamUtilities**.

Copy the group project template content to the local project template directory

To copy the content of the local **GroupProjectTemplate** (D1) folder to the local **MyTeamProjectTemplate** (D3), run one of the following shell scripts:

From the PowerShell command-line for Windows

```
wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-  
DataScience/master/Misc/TDSP/tdsp_local_copy_win.ps1" -outfile "tdsp_local_copy_win.ps1"  
.\\tdsp_local_copy_win.ps1 2
```

```
PS D:\AML_Projects\TDSP-Linux\test0912> .\tdsp_local_copy_win.ps1 2  
Please input the full path to Your GroupProjectTemplate repository (source directory): D:\AML_Projects\TDSP-Linux\test0912\GroupProjectTemplate  
Please input the full path to Your TeamProjectTemplate repository (destination directory): D:\AML_Projects\TDSP-Linux\test0912\DS_Team_1_ProjectTemplate  
Start copying files (except files in .git directory) from D:\AML_Projects\TDSP-Linux\test0912\GroupProjectTemplate to D:\AML_Projects\TDSP-Linux\test0912\DS_Team_1_ProjectTemplate...  
PS D:\AML_Projects\TDSP-Linux\test0912>
```

From the Linux shell for the Linux DSVM

```
wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-  
DataScience/master/Misc/TDSP/tdsp_local_copy_linux.sh"  
bash tdsp_local_copy_linux.sh 2
```

```
[dsl@linuxdsvm6 test0912]$ bash tdsp_local_copy_linux.sh 2
Please input the full path to Your GroupProjectTemplate repository (source directory): /home/dsl/test0912/GroupProjectTemplate
Please input the full path to Your TeamProjectTemplate repository (destination directory): /home/dsl/test0912/DS_Team_1_ProjectTemplate
Code/
Code/DataPrep/
Code/DataPrep/dataPrep.py
Code/Model/
Code/Model/model.R
Code/Operationalization/
Code/Operationalization/operationalization.py
Data/
Data/Modeling/
Data/Modeling/modelling.md
```

The scripts exclude the contents of the .git directory. The scripts prompt you to provide the **complete paths** to the source directory D1 and to the destination directory D3.

Customize your project template or team utilities (optional)

Customize your **MyTeamProjectTemplate** (D3) and **MyTeamUtilities** (D4), if needed, at this stage of the setup process.

- If you want to customize the contents of D3 to meet the specific needs of your team, you can modify the template documents or change the directory structure.
- If your team has developed some utilities that you want to share with your entire team, copy and paste these utilities into directory D4.

Push local directory content to team repositories

To add the contents in the (optionally customized) local directories D3 and D4 to the team repositories R3 and R4, run the following git bash commands either from a Windows PowerShell console or from the Linux shell. Run the commands from the **GitRepos\MyTeam\MyTeamProjectTemplate** directory.

```
git status
git add .
git commit -m"push from DSVM"
git push
```

```
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2\GroupProjectTemplate> git status
On branch master

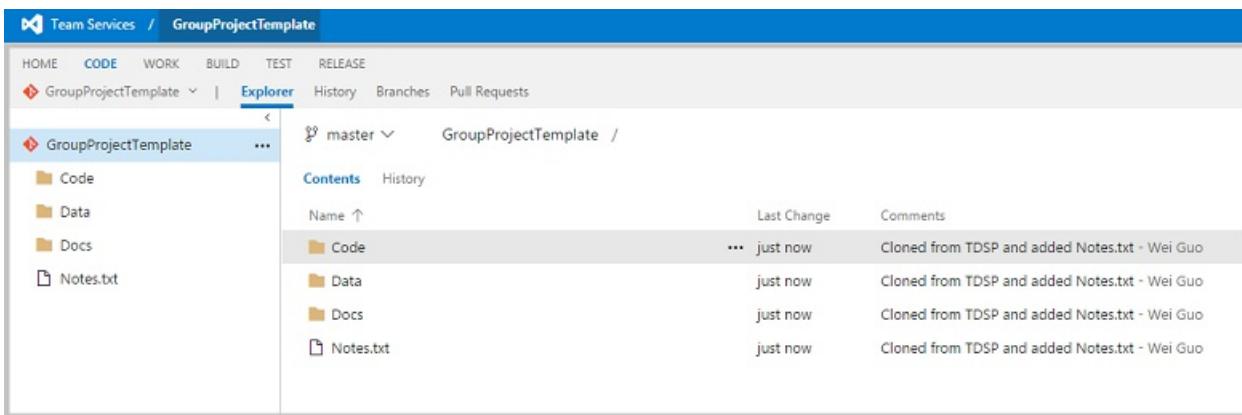
Initial commit

Untracked files:
  (use "git add <file>..." to include in what will be committed)

    Code/
    Data/
    Docs/
    README.md

nothing added to commit but untracked files present (use "git add" to track)
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2\GroupProjectTemplate> git add .
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test2\GroupProjectTemplate> git commit -m"push from local Win DSVM"
[master (root-commit) 4170390] push from local Win DSVM
 21 files changed, 233 insertions(+)
 create mode 100644 Code/DataPrep/dataPrep.py
 create mode 100644 Code/Model/model.R
 create mode 100644 Code/Operationalization/operationalization.py
 create mode 100644 Data/Modeling/modelling.md
 create mode 100644 Data/Processed/processed.md
 create mode 100644 Data/Raw/rawData.md
 create mode 100644 Docs/DataDictionary/ReadMe.md
```

The files in the MyTeamProjectTemplate repository of your group's Azure DevOps Services are synced nearly instantly when this script is run.



Now run the same set of four git commands from the **GitRepos\MyTeam\MyTeamUtilities** directory.

[AZURE.NOTE] If this is the first time you commit to a Git repository, you need to configure global parameters *user.name* and *user.email* before you run the `git commit` command. Run the following two commands:

```
git config --global user.name <your name>
git config --global user.email <your email address>
```

If you are committing to multiple Git repositories, use the same name and email address when you commit to each of them. Using the same name and email address proves convenient later on when you build PowerBI dashboards to track your Git activities on multiple repositories.

```
[ds1@linuxdsvm6 ~]$ [ds1@linuxdsvm6 ~]$ git config --global user.name "weig"
[ds1@linuxdsvm6 ~]$ git config --global user.email "weig@microsoft.com"
[ds1@linuxdsvm6 ~]$
```

3. Create team data and analytics resources (Optional)

Sharing data and analytics resources with your entire team has performance and cost benefits: team members can execute their projects on the shared resources, save on budgets, and collaborate more efficiently. In this section, we provide instructions on how to create Azure file storage. In the next section, we provide instruction on how to mount Azure file storage to your local machine. For additional information on sharing other resources, such as Azure Data Science Virtual Machines, Azure HDInsight Spark Clusters, see [Platforms and tools](#). This topic provides you guidance from a data science perspective on selecting resources that are appropriate for your needs, and links to product pages and other relevant and useful tutorials that we have published.

[AZURE.NOTE] To avoid data transmitting cross data centers, which might be slow and costly, make sure that the resource group, storage account, and the Azure VM (e.g., DSVM) are in the same Azure data center.

Run the following scripts to create Azure file storage for your team. Azure file storage for your team can be used to store data assets that are useful for your entire team. The scripts prompt you for your Azure account and subscription information, so have these credentials ready to enter.

Create Azure file storage with PowerShell from Windows

Run this script from the PowerShell command-line:

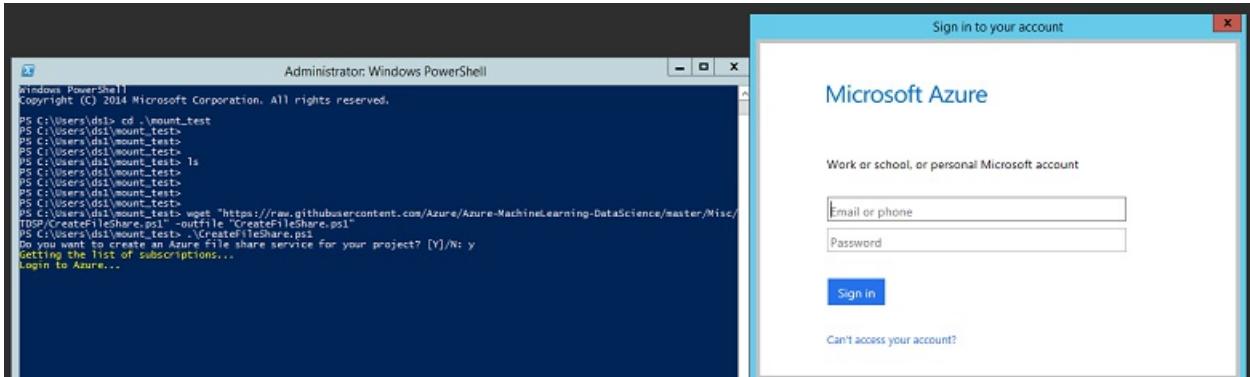
```
wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/TDSP/CreateFileShare.ps1" -outfile "CreateFileShare.ps1"
.\CreateFileShare.ps1
```

```

PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test_scripts> wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/TDSP/CreateFileShare.ps1" -outfile "CreateFileShare.ps1"
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\test_scripts> .\CreateFileShare.ps1
Do you want to create an Azure file share service for your team? [Y]/N: y
Getting the list of subscriptions...
[1]: Microsoft Azure Internal Consumption
[2]: Microsoft Azure Internal - Wei
[3]: Microsoft Azure Internal - Demos
[4]: ACE-Test Subscription
[5]: BDHadoopTeamPMTTestDemo
[6]: ADLTrainingMS
[7]: Office CLE - External
[8]: BigDataDemosExternal
Enter the index of the subscription name where resources will be created(1-8): 1
You selected subscription [1]:Microsoft Azure Internal Consumption. [Y]-yes to continue/N-no to reselect:

```

Log in to your Microsoft Azure account when prompted:



Select the Azure subscription you want to use:

```

[1]: Microsoft Azure Internal Consumption
[2]: Microsoft Azure Internal - Wei
[3]: Microsoft Azure Internal - Demos
[4]: ACE-Test Subscription
[5]: BDHadoopTeamPMTTestDemo
[6]: ADLTrainingMS
[7]: Office CLE - External
[8]: BigDataDemosExternal
Enter the index of the subscription name where resources will be created(1-8): 2

```

Select which storage account to use or create a new one under your selected subscription:

```

Here are the storage account names under your subscription Microsoft Azure Internal - Wei
[1]: weigdsvm2rsc458
[2]: weiglinuxdsvm5664
[3]: weiglinuxresource39870
[4]: weiglinuxtest8
[5]: weigresourcegroup1825
[6]: weigstorage08092016
[7]: weigstorageforsvm
Do you want to create a new storage account for your Azure file share?[Y]/N: n
Enter the index of the storage account to use(1-7): 6

```

Enter the name of the Azure file storage to create. Only lower case characters, numbers and - are accepted:

```

Enter the name of the file share service to create (lower case characters, numbers, and - are accepted): fs0823
ServiceClient : Microsoft.WindowsAzure.Storage.File.CloudFileClient
Uri          : https://weigstorage08092016.file.core.windows.net/fs0823/data
StorageUri   : Primary = 'https://weigstorage08092016.file.core.windows.net/fs0823/data'; Secondary = ''
Properties    : Microsoft.WindowsAzure.Storage.File.FileDirectoryProperties
Metadata     : {}
Share        : Microsoft.WindowsAzure.Storage.File.CloudFileShare
Parent       : Microsoft.WindowsAzure.Storage.File.CloudFileDirectory
Name         : data

ServiceClient : Microsoft.WindowsAzure.Storage.File.CloudFileClient
Uri          : https://weigstorage08092016.file.core.windows.net/fs0823/data
StorageUri   : Primary = 'https://weigstorage08092016.file.core.windows.net/fs0823/data'; Secondary = ''
Properties    : Microsoft.WindowsAzure.Storage.File.FileDirectoryProperties
Metadata     : {}
Share        : Microsoft.WindowsAzure.Storage.File.CloudFileShare
Parent       : Microsoft.WindowsAzure.Storage.File.CloudFileDirectory
Name         : data

An Azure file share service created. It can be later mounted to the Azure virtual machines created for your team project $.
Please keep a note for the information of the Azure file share service. It will be needed in the future when mounting it to Azure virtual machines
Do you want to output the file share information to a file in your current directory?[Y]/N: 

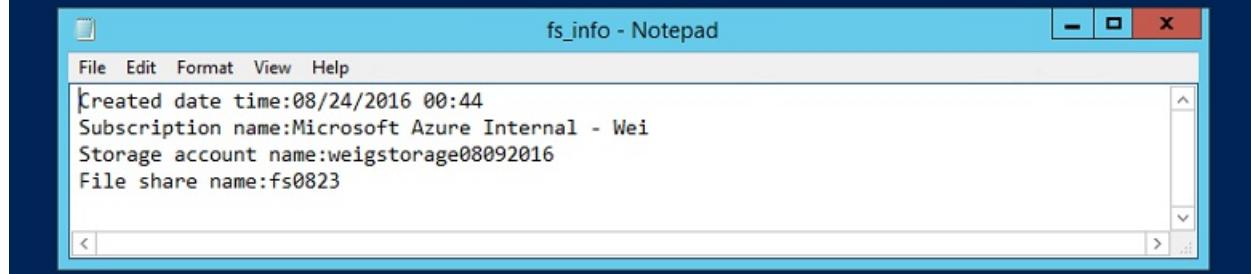
```

To facilitate mounting and sharing this storage after it is created, save the Azure file storage information into a text file and make a note of the path to its location. In particular, you need this file to mount your Azure file storage to your Azure virtual machines in the next section.

It is a good practice to check in this text file into your ProjectTemplate repository. We recommend to put in the directory **Docs\Dictionary**s. Therefore, this data asset can be accessed by all projects in your team.

```
Do you want to output the file share information to a file in your current directory? [Y]/N: y
Please provide the file name. This file under the current working directory will be created to store the file share information.: fs_info.txt
File share information output to C:\Users\ds1\mount_test\fs_info.txt. Share it with your team members who want to mount it to their virtual machines.
```

```
PS C:\Users\ds1\mount_test> notepad.exe .\fs_info.txt
PS C:\Users\ds1\mount_test>
```



Create Azure file storage with a Linux script

Run this script from the Linux shell:

```
wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/TDSP/CreateFileShare.sh"
bash CreateFileShare.sh
```

Log in to your Microsoft Azure account following the instructions on this screen:

```
Do you want to create an Azure file share service for your team? [Y]/N y
info: Executing command config mode
info: New mode is arm
info: config mode command OK
Follow directions on screen to login to your Azure account
info: Executing command login
Authenticating...info: To sign in, use a web browser to open the page https://aka.ms/devicelogin. Enter the code FDUJKQJGQ to authenticate.
/
```

Select the Azure subscription that you want to use:

```
Here are your subscriptions:
1 Microsoft Azure Internal Consumption
2 Microsoft Azure Internal - Wei
3 Microsoft Azure Internal - Demos
4 ACE-Test Subscription
5 BDHadoopTeamPMTestDemo
6 ADLTrainingMS
7 Office CLE - External
8 BigDataDemosExternal

Enter the index of the subscription name where resources will be created (1-8): 2
You selected 2: Microsoft Azure Internal - Wei. [Y]-yes to continue/N-no to reselect: y
info: Executing command account set
info: Setting subscription to "Microsoft Azure Internal - Wei" with id "49bb74df-a9b8-4275-9439-198b33ae0f5f".
```

Select which storage account to use or create a new one under your selected subscription:

```
Here are the storage account names under your subscription Microsoft Azure Internal - Wei:
```

```
1 weigdsvm2rsc458
2 weiglinuxdsvm5664
3 weiglinuxresource39870
4 weiglinuxtest8
5 weigresourcegroup1825
6 weigstorage08092016
7 weigstoragefordsvm
```

```
Do you want to create a new storage account for your Azure file share? [Y]/N: n
Enter the index of the storage account to use (1 - 7 ): 6
You selected storage account weigstorage08092016. [Y]-continue/N-reselect: y
```

Enter the name of the Azure file storage to create, only lower case characters, numbers and - are accepted:

```

Enter the name of the file share service to create (lower case characters, numbers, and - are accepted): linuxfs0823
info: Executing command storage share create
+ Creating storage file share linuxfs0823
+ Getting Storage share information
data:  {
data:    name: 'linuxfs0823',
data:    metadata: {},
data:    etag: '"0x803CBBB44FCB459"',
data:    lastModified: 'Wed, 24 Aug 2016 01:09:16 GMT',
data:    requestId: '22e551c5-001a-012f-52a4-fd5137000000',
data:    quota: '5120',
data:    shareUsage: '0'
data:  }
info: storage share create command OK
An Azure file share service created. It can be later mounted to the Azure virtual machine created for your team projects.
Please keep a note for the information of the Azure file share service. It will be needed in the future when mounting it to Azure virtual machines.
Do you want to output the file share information to a file in your current directory? [Y]/N: ■

```

To facilitate accessing this storage after it is created, save the Azure file storage information into a text file and make a note of the path to its location. In particular, you need this file to mount your Azure file storage to your Azure virtual machines in the next section.

It is a good practice to check in this text file into your ProjectTemplate repository. We recommend to put in the directory **Docs\DataDictionaries**. Therefore, this data asset can be accessed by all projects in your team.

```

Do you want to output the file share information to a file in your current directory? [Y]/N: y
Please provide the file name. This file under the current working directory will be created to store the file share information: linuxfs0823.txt
File share information output to linuxfs0823.txt. Share it with your team members who want to mount it to their virtual machines.
[ds1@weiqlinuxdsvm5 ~]$ vim linuxfs0823.txt
[ds1@weiqlinuxdsvm5 ~]$
Created date time:Wed Aug 24 01:10:23 UTC 2016
Subscription name:Microsoft Azure Internal - Wei
Storage account name:weiqistorage08092016
File share name:linuxfs0823

```

4. Mount Azure file storage (Optional)

After Azure file storage is created successfully, it can be mounted to your local machine using one of the following PowerShell or Linux scripts.

Mount Azure file storage with PowerShell from Windows

```

wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-
DataScience/master/Misc/TDSP/AttachFileShare.ps1" -outfile "AttachFileShare.ps1"
.\AttachFileShare.ps1

```

You are asked to log in first, if you have not logged in.

Click **Enter** or **y** to continue when you are asked if you have an Azure file storage information file, and then input the ****complete path and name*** of the file you create in previous step. The information to mount an Azure file storage is read directly from that file and you are ready to go to the next step.

```

PS C:\Users\ds1\mount_test> wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-Datascience/master/Misc/
TDSP/AttachFileShare.ps1" -outfile "AttachFileShare.ps1"
PS C:\Users\ds1\mount_test> .\AttachFileShare.ps1
Do you want to mount an Azure file share service to your Azure virtual machine? [Y]-yes/N-no to quit.: y
Start getting the list of subscriptions under your Azure account...
Here are the subscription names under your Azure account:
[1]: Microsoft Azure Internal Consumption
[2]: Microsoft Azure Internal - Wei
[3]: Microsoft Azure Internal - Demos
[4]: ACE-Test Subscription
[5]: BDHadoopTeamPMTestDemo
[6]: ADLTrainingMS
[7]: Office CLE - External
[8]: BigDataDemosExternal
Do you have a file with the information of the file share you want to mount?[Y]-yes/N-no: y
Please provide the name of the file with the information of the file share you want to mount: fs_info.txt

```

[AZURE.NOTE] If you do not have a file containing the Azure file storage information, the steps to input the information from keyboard are provided at the end of this section.

Then you are asked to enter the name of the drive to be added to your virtual machine. A list of existing drive names is printed on the screen. You should provide a drive name that does not already exist in the list.

```

Environment      : AzureCloud
Account         : weig@microsoft.com
TenantId        : 72f988bf-86f1-41af-91ab-2d7cd011db47
SubscriptionId  : 49bb74df-a9b8-4275-9439-198b33ae0f5f
SubscriptionName : Microsoft Azure Internal - Wei
CurrentStorageAccount : weigstorage08092016

Start getting the list of storage accounts under your subscription Microsoft Azure Internal - Wei
Here are the storage account names under subscription Microsoft Azure Internal - Wei
[1]: weigdsvm2rsc458
[2]: weiglinuxdsvm5664
[3]: weiglinuxresource39870
[4]: weiglinuxtest8
[5]: weigresourcegroup1825
[6]: weigstorage08092016
[7]: weigstoragefordsvm
Existing disk names are:
[1]: A:\ 
[2]: C:\ 
[3]: D:\ 
[4]: E:\ 
Enter the name of the drive to be added to your virtual machine. This name should be different from the disk names your
virtual machine has.: F
File share fs0823 will be mounted to your virtual machine as drive F:
CMDKEY: Credential added successfully.
The command completed successfully.

Do you want to mount other Azure file share services? [Y]-yes/N-no to quit.: n

```

Confirm that a new F drive has been successfully mounted to your machine.

```

PS C:\Users\ds1\mount_test> gdr -PSProvider 'FileSystem'

Name      Used (GB)    Free (GB) Provider   Root           CurrentLocation
----      -----    -----   -----   -----
A          40.04       86.95   FileSystem  A:\ 
C          1.26        283.74  FileSystem  C:\ 
D          5120.00     5120.00 FileSystem  D:\ 
E          5120.00     5120.00 FileSystem  E:\ 
F          5120.00     5120.00 FileSystem  F:\ 


```

How to enter the Azure file storage information manually: If you do not have your Azure file storage information on a text file, you can follow the instructions on the following screen to type in the required subscription, storage account, and Azure file storage information:

```

PS C:\Users\ds1\mount_test> .\AttachFileShare.ps1
Do you want to mount an Azure file share service to your Azure virtual machine? [Y]-yes/N-no to quit.: y
Start getting the list of subscriptions under your Azure account...
Here are the subscription names under your Azure account:
[1]: Microsoft Azure Internal Consumption
[2]: Microsoft Azure Internal - Wei
[3]: Microsoft Azure Internal - Demos
[4]: ACE-Test Subscription
[5]: BDHadoopTeamPMTTestDemo
[6]: ADLTrainingMS
[7]: Office CLE - External
[8]: BigDataDemosExternal
Do you have a file with the information of the file share you want to mount?[Y]-yes/N-no: n

```

Type in your Azure subscription name, select the storage account where the Azure file storage is created, and type in the Azure file storage name:

```

Enter the subscription name where the Azure file share service has been created: Microsoft Azure Internal - Wei

Environment      : AzureCloud
Account         : weig@microsoft.com
TenantId        : 72f988bf-86f1-41af-91ab-2d7cd011db47
SubscriptionId  : 49bb74df-a9b8-4275-9439-198b33ae0f5f
SubscriptionName : Microsoft Azure Internal - Wei
CurrentStorageAccount : weigstorage08092016

Start getting the list of storage accounts under your subscription Microsoft Azure Internal - Wei
Here are the storage account names under subscription Microsoft Azure Internal - Wei
[1]: weigdsvm2rsc458
[2]: weiglinuxdsvm5664
[3]: weiglinuxresource39870
[4]: weiglinuxtest8
[5]: weigresourcegroup1825
[6]: weigstorage08092016
[7]: weigstoragefordsvm
Enter the index of the storage account name where your Azure file share you want to mount is created (1-7): 6
Enter the name of the file share to mount (lower case only): fs0823

```

Mount Azure file storage with a Linux script

```

wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-
DataScience/master/Misc/TDSP/AttachFileShare.sh"
bash AttachFileShare.sh

```

```

Do you want to mount an Azure File share service to your Azure Virtual Machine? [Y]-yes/N -no to quit: y
Start getting the list of subscriptions under your Azure account...info: Executing command config mode
info: New mode is arm
info: config mode command OK
Here are the subscriptions :

1 Microsoft Azure Internal Consumption
2 Microsoft Azure Internal - Wei
3 Microsoft Azure Internal - Demos
4 ACE-Test Subscription
5 BDHadoopTeamPMTTestDemo
6 ADLTrainingMS
7 Office CLE - External
8 BigDataDemosExternal

```

You are asked to log in first, if you have not logged in.

Click **Enter** or **y** to continue when you are asked if you have an Azure file storage information file, and then input the ****complete path and name*** of the file you create in previous step. The information to mount an Azure file storage is read directly from that file and you are ready to go to the next step.

```

Do you have a file with the information of the file share you want to mount? Y/N y
Please provide the name of the file with the information of the file share you want to mount: linuxfs0823.txt
info: Executing command account set
info: Setting subscription to "Microsoft Azure Internal - Wei" with id "49bb74df-a9b8-4275-9439-190b33ae0f5f".
info: Changed saved
info: account set command OK
Start getting the list of storage accounts under your subscription Microsoft Azure Internal - Wei Here are the storage account names under subscription Microsoft Azure Internal - Wei
1 weigdsvm2rsc458
2 weiglinuxdsvm5664
3 weiglinuxresource39870
4 weiglinuxtest8
5 weigresourcegroup1825
6 weigstorage08092016
7 weigstoragefordsvm

```

Then you are asked to enter the name of the drive to be added to your virtual machine. A list of existing drive names is printed on the screen. You should provide a drive name that does not already exist in the list.

```

Existing disk names are:
Filesystem      Size  Used Avail Use% Mounted on
/dev/sdal       30G   14G   17G  47% /
devtmpfs        3.4G   0     3.4G  0% /dev
tmpfs          3.5G   0     3.5G  0% /dev/shm
tmpfs          3.5G  361M  3.1G  11% /run
tmpfs          3.5G   0     3.5G  0% /sys/fs/cgroup
/dev/sdb1       281G  65M   267G  1% /mnt/resource
tmpfs          697M   0    697M  0% /run/user/1005
Enter the name of the drive to be added to your virtual machine. This name should be different from the disk names your virtual machine has: fs0823
File share linuxfs0823 will be mounted to your virtual machine as drive fs0823 [sudo] password for dsl:
Do you want to mount other Azure File share services? [Y]-yes/N -no to quit: n

```

Confirm that a new F drive has been successfully mounted to your machine.

```

[dsl@weiglinuxdsvm5 ~]$ df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sdal       30G   14G   17G  47% /
devtmpfs        3.4G   0     3.4G  0% /dev
tmpfs          3.5G   0     3.5G  0% /dev/shm
tmpfs          3.5G  361M  3.1G  11% /run
tmpfs          3.5G   0     3.5G  0% /sys/fs/cgroup
/dev/sdb1       281G  65M   267G  1% /mnt/resource
tmpfs          697M   0    697M  0% /run/user/1005
//weigstorage08092016,file.core.windows.net/linuxfs0823  5.0T   0  5.0T  0% /fs0823

```

How to enter the Azure file storage information manually: If you do not have your Azure file storage information on a text file, you can follow the instructions on the following screen to type in the required subscription, storage account, and Azure file storage information:

- Input **n**.
- Select the index of the subscription name where the Azure file storage was created in the previous step:

```

Do you want to mount an Azure File share service to your Azure Virtual Machine? [Y]-yes/N -no to quit: y
Start getting the list of subscriptions under your Azure account...info: Executing command config mode
info: New mode is arm
info: config mode command OK
Here are the subscriptions :

1 Microsoft Azure Internal Consumption
2 Microsoft Azure Internal - Wei
3 Microsoft Azure Internal - Demos
4 ACE-Test Subscription
5 BDHadoopTeamPMTTestDemo
6 ADLTrainingMS
7 Office CLE - External
8 BigDataDemosExternal

Do you have a file with the information of the file share you want to mount? Y/N n
Enter the subscription name where the Azure file share service has been created: Microsoft Azure Internal - Wei

```

- Select the storage account under your subscription and type in the Azure file storage name:

```
info: Executing command account set
info: Setting subscription to "Microsoft Azure Internal - Wei" with id "49bb74df-a9b8-4275-9439-198b33ae0f5f".
info: Changes saved
info: account set command OK
Start getting the list of storage accounts under your subscription Microsoft Azure Internal - Wei Here are the storage account names:
Internal - Wei
1 weigdsvm2rsc458
2 weiglinuxdsvm5664
3 weiglinuxresource39870
4 weiglinuxtest8
5 weigresourcegroup1825
6 weigstorage08092016
7 weigstoragefordsvm
Enter the index of the storage account name where your Azure file share you want to mount is created: 6
Enter the name for the file share to mount (lower case only): fs0823
```

- Enter the name of drive to be added to your machine, which should be distinct from any existing ones:

```
Existing disk names are:
Filesystem      Size  Used  Avail Use% Mounted on
/dev/sdal       30G   14G   17G  47% /
/dev/tmpfs      3.4G   0     3.4G  0% /dev
tmpfs          3.5G   0     3.5G  0% /dev/shm
tmpfs          3.5G  361M  3.1G  1% /run
tmpfs          3.5G   0     3.5G  0% /sys/fs/cgroup
/dev/sdb1       281G  65M   287G  1% /mnt/resource
tmpfs          697M   0     697M  0% /run/user/1005
Enter the name of the drive to be added to your virtual machine. This name should be different from the disk names your virtual machine has: fileshare0823
File share fs0823 will be mounted to your virtual machine as drive fileshare0823 Do you want to mount other Azure File share services? [Y]-yes/N -no to quit: n
```

5. Set up security control policy

From your group Azure DevOps Services's homepage, click the **gear icon** next to your user name in the upper right corner, then select the **Security** tab. You can add members to your team here with various permissions.

The screenshot shows the Azure DevOps Security interface for the 'DS_Team_1' team. The left sidebar lists 'Teams' (selected) and 'VSTS Groups'. Under 'Teams', 'DS_Team_1 Team' is selected. Under 'VSTS Groups', there are five groups: 'Build Administrators', 'Contributors', 'Project Administrators', 'Project Valid Users', and 'Readers'. The main area displays the 'Members' tab for the 'DS_Team_1 Team'. It shows one member: 'Wei Guo' (wguo123@outlook.com). There are buttons for '+ Add...', 'Edit...', and 'Search'.

Display Name	Username Or Scope	Action
Wei Guo	wguo123@outlook.com	Remove

Next steps

Here are links to the more detailed descriptions of the roles and tasks defined by the Team Data Science Process:

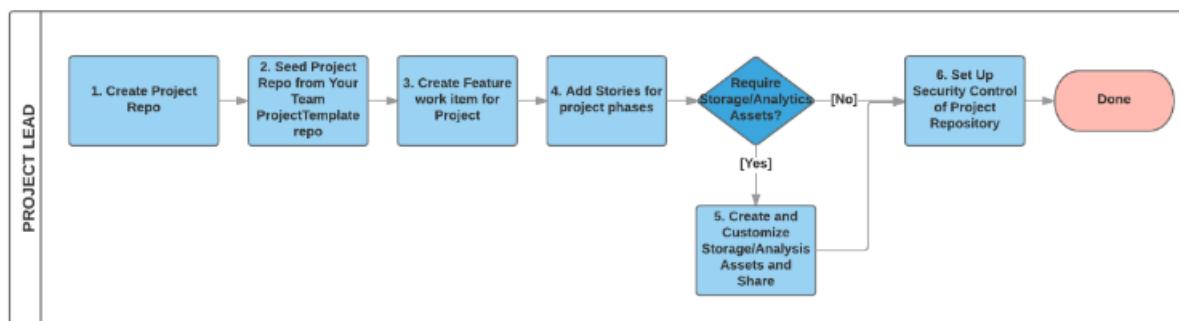
- [Group Manager tasks for a data science team](#)
- [Team Lead tasks for a data science team](#)
- [Project Lead tasks for a data science team](#)
- [Project Individual Contributors for a data science team](#)

Tasks for the project lead in the Team Data Science Process

3/14/2019 • 7 minutes to read

This tutorial outlines the tasks that a project lead is expected to complete for his/her project team. The objective is to establish collaborative team environment that standardizes on the [Team Data Science Process \(TDSP\)](#). The TDSP is a framework developed by Microsoft that provides a structured sequence of activities to execute cloud-based, predictive analytics solutions efficiently. For an outline of the personnel roles and their associated tasks that are handled by a data science team standardizing on this process, see [Team Data Science Process roles and tasks](#).

A **Project Lead** manages the daily activities of individual data scientists on a specific data science project. The workflow for the tasks to be completed by project leads to set up this environment are depicted in the following figure:



This topic currently covers tasks 1,2 and 6 of this workflow for project leads.

NOTE

We outline the steps needed to set up a TDSP team environment for a project using Azure DevOps in the following instructions. We specify how to accomplish these tasks with Azure DevOps because that is how we implement TDSP at Microsoft. If another code-hosting platform is used for your group, the tasks that need to be completed by the team lead generally do not change. But the way to complete these tasks is going to be different.

Repositories and directories

This tutorial uses abbreviated names for repositories and directories. These names make it easier to follow the operations between the repositories and directories. This notation (R for Git repositories and D for local directories on your DSVM) is used in the following sections:

- **R3:** The team **ProjectTemplate** repository on Git your team lead has set up.
- **R5:** The project repository on Git you setup for your project.
- **D3:** The local directory cloned from R3.
- **D5:** The local directory cloned from R5.

0. Prerequisites

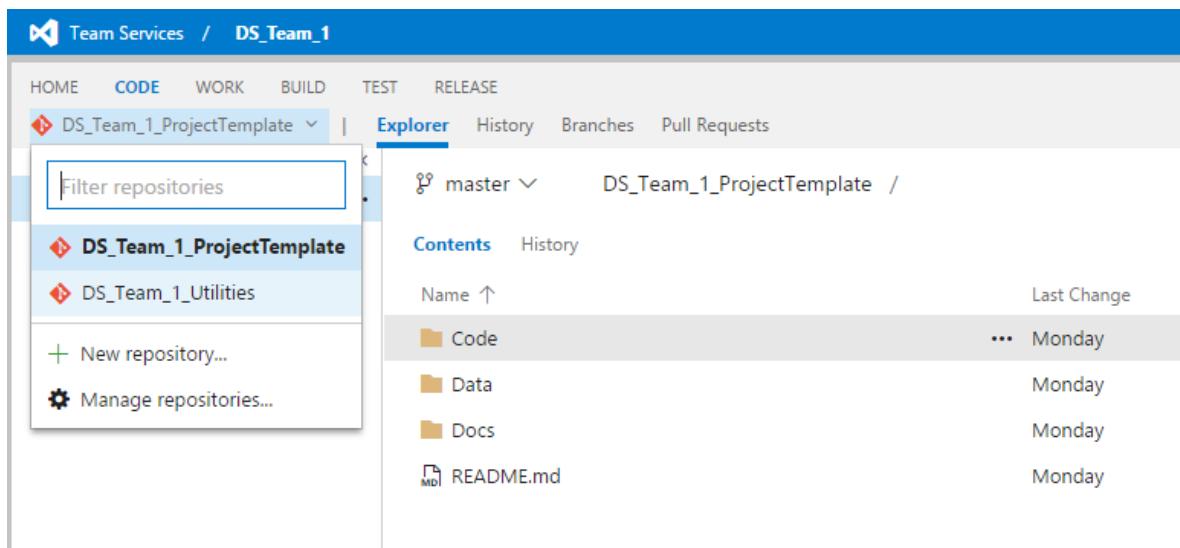
The prerequisites are satisfied by completing the tasks assigned to your group manager outlined in [Group Manager tasks for a data science team](#) and to your team lead outlined in [Team lead tasks for a data science team](#).

To summarize here, the following requirements need to meet before you begin the team lead tasks:

- Your **group Azure DevOps Services** (or group account on some other code-hosting platform) has been set up by your group manager.
- Your **TeamProjectTemplate repository** (R3) has been set up under your group account by your team lead on the code-hosting platform you plan to use.
- You have been **authorized** by your team lead to create repositories on your group account for your team.
- Git must be installed on your machine. If you are using a Data Science Virtual Machine (DSVM), Git has been pre-installed and you are good to go. Otherwise, see the [Platforms and tools appendix](#).
- If you are using a **Windows DSVM**, you need to have [Git Credential Manager \(GCM\)](#) installed on your machine. In the README.md file, scroll down to the **Download and Install** section and click the *latest installer*. This takes you to the latest installer page. Download the .exe installer from here and run it.
- If you are using **Linux DSVM**, create an SSH public key on your DSVM and add it to your group Azure DevOps Services. For more information about SSH, see the **Create SSH public key** section in the [Platforms and tools appendix](#).

1. Create a project repository (R5)

- Log in to your group Azure DevOps Services at <https://<Azure DevOps Services Name>.visualstudio.com>.
- Under **Recent projects & teams**, click **Browse**. A window that pops up lists all projects on the Azure DevOps Services.



The screenshot shows the Azure DevOps Services interface. The top navigation bar includes HOME, CODE, WORK, BUILD, TEST, and RELEASE. The CODE tab is selected. Below the navigation is a breadcrumb trail: Team Services / DS_Team_1. The main area is titled 'DS_Team_1_ProjectTemplate' with a dropdown menu. The 'Explorer' tab is active. On the left, there's a sidebar with a 'Filter repositories' input field and options for 'DS_Team_1_ProjectTemplate', 'DS_Team_1_Utils', 'New repository...', and 'Manage repositories...'. The main content area shows a 'master' branch for 'DS_Team_1_ProjectTemplate'. It displays a 'Contents' table with columns for 'Name' (sorted by Last Change) and 'Last Change'. The contents include 'Code' (Monday), 'Data' (Monday), 'Docs' (Monday), and 'README.md' (Monday). There are also 'History' and 'Pull Requests' tabs at the top of the main content area.

- Click the project name in which you are going to create your project repository. In this example, click **MyTeam**.
- Then, click **Navigate** to be directed to the home page of the project **MyTeam**:

Search...

- mysamplegroup
 - GroupCommon
 - MyTeam**

Team name **MyTeam Team**

Project MyTeam

Members 1

Description The default project team.

Navigate **Close**

- Click **Collaborate on code** to be directed to the git home page of your project.

Welcome

Get started using Visual Studio Team Services to make the most of your team dashboard.

Collaborate on code
Add code to your repository

Continuously integrate
Automate your builds

Visualize progress
Learn how to add charts

Open User Stories

Query returned no results. Create new work items or edit query to see results.

Team Members

It's lonely in here...
Invite a friend

Work

Backlog
Board
Task board
Queries

Sprint Burndown

New Work Item

Enter title
Bug
Create

Open User Stories

0 Work items

Visual Studio

Open in Visual Studio
Requires Visual Studio 2013+
Get Visual Studio
See Visual Studio downloads

- Click the downward arrow at the top left corner, and select **+ New repository**.

MyTeamProjectTemplate

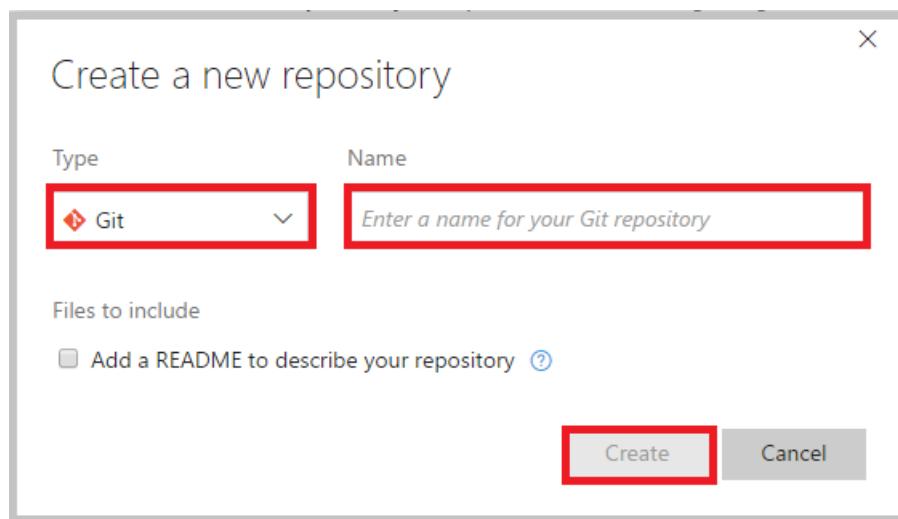
Filter repositories

- New repository**
- Manage repositories

Contents History

Name ↑	Last Change	Comments
Code	... just now	seeded the MyTeamProjectTemplate from DGADS - Hang Zhang
Data	just now	seeded the MyTeamProjectTemplate from DGADS - Hang Zhang
Docs	just now	seeded the MyTeamProjectTemplate from DGADS - Hang Zhang
README.md	just now	seeded the MyTeamProjectTemplate from DGADS - Hang Zhang

- In the **Create a new repository** window, input a name for your project git repository. Make sure that you select **Git** as the type of the repository. In this example, we use the name **DSPProject1**.



- To create your **DSPProject1** project git repository, click **Create**.

2. Seed the DSProject1 project repository

The task here is to seed the **DSPProject1** project repository (R5) from your project template repository (R3). The seeding procedure uses the directories D3 and D5 on your local DSVM as intermediate staging sites. In summary, the seeding path is: R3 -> D3 -> D5 -> R5.

If you need to customize your **DSPProject1** project repository to meet some specific project needs, you do so in the penultimate step of following procedure. Here is a summary of the steps used to seed the content of the **DSPProject1** project repository. The individual steps correspond to the subsections in the seeding procedure:

- Clone project template repository into local directory: team R3 - cloned to -> local D3.
- Clone DSProject1 repository to a local directory: team R5 - cloned to -> local D5.
- Copy cloned project template content to local clone of DSProject1 repository: D3 - contents copied to -> D5.
- (Optional) Customization local D5.
- Push local DSProject1 content to team repositories: D5 - contents add to -> team R5.

Clone your project template repository (R3) to a directory (D3) on your local machine.

On your local machine, create a directory:

- C:\GitRepos\MyTeamCommon* for Windows
- \$home/GitRepos/MyTeamCommon* for Linux

Change to that directory. Then, run the following command to clone your project template repository to your local machine.

Windows

```
git clone <the HTTPS URL of the TeamProjectTemplate repository>
```

If you are using Azure DevOps as the code-hosting platform, typically, the *HTTPS URL of your project template repository* is:

https://<Azure DevOps Services Name>.visualstudio.com/<Your project name>/_git/<Your project template repository name>

In this example, we have:

https://mysamplegroup.visualstudio.com/MyTeam/_git/MyTeamProjectTemplate.

```
PS C:\GitRepo\MyTeamCommon> git clone https://mysamplegroup.visualstudio.com/MyTeam/_git/MyTeamProjectTemplate
Cloning into 'MyTeamProjectTemplate'...
remote:
remote:           vSTS
remote:           vSTSvSTSV
remote:           vSTSvSTSVST
remote: VSTS           vSTSvSTSVSTSV
remote: VSTSvS          vSTSvSTSV STSVS
remote: VSTSvSTSVsTSVSTSVS   TSVST
remote: VS   tSVSTSVSTSV    STSVS
remote: VS   tSVSTSVST    SVSTS
remote: VS   tSVSTSVSTSVsts  VSTSv
remote: VSTSvST         SVSTSvSTS VSTSv
remote: VSTSv          STSVSTSVSTSV
remote:           VSTSvSTSVST
remote:           VSTSvSTS
remote:           VSTS   (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Unpacking objects: 100% (32/32), done.
Checking connectivity... done.
PS C:\GitRepo\MyTeamCommon>
```

Linux

```
git clone <the SSH URL of the TeamProjectTemplate repository>
```

```
benchmark@benchmark-vm-l:~/GitRepos/MyTeamCommon$ git clone ssh://mysamplegroup@mysamplegroup.visualstudio.com:22/MyTeam/_git/MyTeamProjectTemplate
Cloning into 'MyTeamProjectTemplate'...
remote:
remote:           vSTS
remote:           vSTSvSTSV
remote:           vSTSvSTSVST
remote: VSTS           vSTSvSTSVSTSV
remote: VSTSvS          vSTSvSTSV STSVS
remote: VSTSvSTSVsTSVSTSVS   TSVST
remote: VS   tSVSTSVSTSV    STSVS
remote: VS   tSVSTSVST    SVSTS
remote: VS   tSVSTSVSTSVsts  VSTSv
remote: VSTSvST         SVSTSvSTS VSTSv
remote: VSTSv          STSVSTSVSTSV
remote:           VSTSvSTSVST
remote:           VSTSvSTS
remote:           VSTS   (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Receiving objects: 100% (32/32), 14.35 KiB | 0 bytes/s, done.
Resolving deltas: 100% (1/1), done.
Checking connectivity... done.
```

If you are using Azure DevOps as the code-hosting platform, typically, the *SSH URL of the project template repository* is:

ssh://<Azure DevOps Services Name>@<Azure DevOps Services Name>.visualstudio.com:22/<Your Project Name>/_git/<Your project template repository name>.

In this example, we have:

ssh://mysamplegroup@mysamplegroup.visualstudio.com:22/MyTeam/_git/MyTeamProjectTemplate.

Clone DSProject1 repository (R5) to a directory (D5) on your local machine

Change directory to **GitRepos**, and run the following command to clone your project repository to your local machine.

Windows

```
git clone <the HTTPS URL of the Project repository>
```

```
PS C:\GitRepo> git clone https://mysamplegroup.visualstudio.com/MyTeam/_git/DSProject1
Cloning into 'DSProject1'...
warning: You appear to have cloned an empty repository.
Checking connectivity... done.
```

If you are using Azure DevOps as the code-hosting platform, typically, the *HTTPS URL of the Project repository* is

https://<Azure DevOps Services Name>.visualstudio.com/<Your Project Name>/_git/<Your project repository name>. In this example, we have

repository name. In this example, we have
https://mysamplegroup.visualstudio.com/MyTeam/_git/DSProject1

Linux

```
git clone <the SSH URL of the Project repository>
```

```
benchmark@benchmark-vm-1:~/GitRepos$ git clone ssh://mysamplegroup@mysamplegroup.visualstudio.com:22/MyTeam/_git/DSPProject1
Cloning into 'DSPProject1'...
warning: You appear to have cloned an empty repository.
Checking connectivity... done.
```

If you are using Azure DevOps as the code-hosting platform, typically, the *SSH URL of the project repository* is `_ssh://<Azure DevOps Services Name>@<Azure DevOps Services Name>.visualstudio.com:22//_git/<Your project repository name>`. In this example, we have

`ssh://mysamplegroup@mysamplegroup.visualstudio.com:22/MyTeam/_git/DSProject1`

Copy contents of D3 to D5

Now in your local machine, you need to copy the content of *D3* to *D5*, except the git metadata in .git directory. The following scripts will do the job. Make sure to type in the correct and full paths to the directories. Source folder is the one for your team (*D3*); destination folder is the one for your project (*D5*).

Windows

```
 wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-  
DataScience/master/Misc/TDSP/tdsp_local_copy_win.ps1" -outfile "tdsp_local_copy_win.ps1"  
.\\tdsp_local_copy_win.ps1 -role 3
```

```
PS D:\AML_Projects\TDSP-Linux\test0912> .\tdsp_local_copy_win.ps1 3  
Please input the full path to Your TeamProjectTemplate repository (source directory): D:\AML_Projects\TDSP-Linux\test0912\DS_Team_1_ProjectTemplate  
Please input the full path to Your Project repository (destination directory): D:\AML_Projects\TDSP-Linux\test0912\DSProject1  
Start copying files (except files in .git directory) from D:\AML_Projects\TDSP-Linux\test0912\DS_Team_1_ProjectTemplate to D:\AML_Projects\TDSP-Linux\test0912\DSProject1...
```

Now you can see in *DSProject1* folder, all the files (excluding the .git) are copied from *MyTeamProjectTemplate*.

```
PS C:\GitRepo\DSProject1> ls

Directory: C:\GitRepo\DSProject1

Mode          LastWriteTime    Length Name
----          -----        ---- 
d----
```

Linux

```
 wget "https://raw.githubusercontent.com/Azure/Azure-MachineLearning-  
DataScience/master/Misc/TDSP/tdsp_local_copy_linux.sh"  
 bash tdsp local copy linux.sh 3
```

```
[ds1@linuxdsvm6 test0912]$ bash tdsp_local_copy_linux.sh
Please input the full path to Your TeamProjectTemplate repository (source directory): /home/ds1/test0912/DS5_Team_1_ProjectTemplate/
Please input the full path to Your Project repository (destination directory): /home/ds1/test0912/DSProject1/
Code/
Code/DataPrep/
Code/DataPrep/prep/dataPrep.py
Code/Model/
Code/Model/model.R
Code/Operationalization/
Code/Operationalization/operationalization.py
Data/
Data/Modeling/
Data/Modeling/modelling.nd
Data/Processed/
Data/Processed/processed.nd
Data/Row/
Data/Row/rawData.nd
Docs/
Docs/DataDictionaries/
Docs/DataDictionaries/ReadMe.md
Docs/DataDictionaries/dict1.csv
```

Now you can see in *DSProject1* folder, all the files (except the metadata in .git) are copied from *MyTeamProjectTemplate*.

```
[ds1@linuxdsvm6 test0912]$ ll ./DSProject1/
total 4
drwxrwxr-x. 5 ds1 ds1 58 Sep 12 20:51 Code
drwxrwxr-x. 5 ds1 ds1 47 Sep 12 20:51 Data
drwxrwxr-x. 6 ds1 ds1 72 Sep 12 20:51 Docs
-rw-rw-r--. 1 ds1 ds1 80 Sep 12 20:51 README.md
```

Customize D5 if you need to (Optional)

If your project needs some specific directories or documents, other than the ones you get from your project template (copied to your D5 directory in the previous step), you can customize the content of D5 now.

Add contents of *DSProject1* in D5 to R5 on your group Azure DevOps Services

You now need to push contents in ***DSProject1*** to **R5** repository in your project on your group's Azure DevOps Services.

- Change to directory **D5**.
- Use the following git commands to add the contents in **D5** to **R5**. The commands are the same for both Windows and Linux systems.

```
git status
git add .
git commit -m"push from win DSVM"
git push
```

- Commit the change and push.

NOTE

If this is the first time you commit to a Git repository, you need to configure global parameters *user.name* and *user.email* before you run the `git commit` command. Run the following two commands:

```
git config --global user.name <your name>
git config --global user.email <your email address>
```

If you are committing to multiple Git repositories, use the same name and email address across all of them. Using the same name and email address proves convenient later on when you build PowerBI dashboards to track your Git activities on multiple repositories.

```
[ds1@linuxdsvm6 ~]$
[ds1@linuxdsvm6 ~]$ git config --global user.name "weig"
[ds1@linuxdsvm6 ~]$ git config --global user.email "weig@microsoft.com"
[ds1@linuxdsvm6 ~]$
```

6. Create and mount Azure file storage as project resources (Optional)

If you want to create Azure file storage to share data, such as the project raw data or the features generated for your project, so that all project members have access to the same datasets from multiple DSVMs, follow the instructions in sections 3 and 4 of [Team Lead tasks for a data science team](#).

Next steps

Here are links to the more detailed descriptions of the roles and tasks defined by the Team Data Science Process:

- [Group Manager tasks for a data science team](#)
- [Team Lead tasks for a data science team](#)
- [Project Lead tasks for a data science team](#)

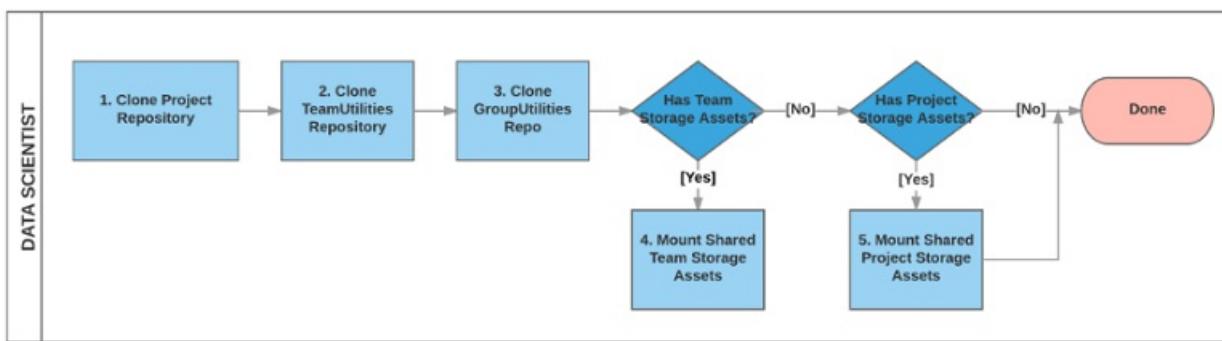
- Project Individual Contributors for a data science team

Tasks for an individual contributor in the Team Data Science Process

1/30/2019 • 4 minutes to read

This topic outlines the tasks that an individual contributor is expected to complete for their data science team. The objective is to establish collaborative team environment that standardizes on the [Team Data Science Process](#) (TDSP). For an outline of the personnel roles and their associated tasks that are handled by a data science team standardizing on this process, see [Team Data Science Process roles and tasks](#).

The tasks of project individual contributors (data scientists) to set up the TDSP environment for the project are depicted as follows:



- **GroupUtilities** is the repository that your group is maintaining to share useful utilities across the entire group.
- **TeamUtilities** is the repository that your team is maintaining specifically for your team.

For instructions on how to execute a data science project under TDSP, see [Execution of Data Science Projects](#).

[AZURE.NOTE] We outline the steps needed to set up a TDSP team environment using Azure DevOps in the following instructions. We specify how to accomplish these tasks with Azure DevOps because that is how we implement TDSP at Microsoft. If another code-hosting platform is used for your group, the tasks that need to be completed by the team lead generally do not change. But the way to complete these tasks is going to be different.

Repositories and directories

This tutorial uses abbreviated names for repositories and directories. These names make it easier to follow the operations between the repositories and directories. This notation (**R** for Git repositories and **D** for local directories on your DSVM) is used in the following sections:

- **R2**: The GroupUtilities repository on Git that your group manager has set up on your Azure DevOps group server.
- **R4**: The TeamUtilities repository on Git that your team lead has set up.
- **R5**: The Project repository on Git that has been set up by your project lead.
- **D2**: The local directory cloned from R2.
- **D4**: The local directory cloned from R4.
- **D5**: The local directory cloned from R5.

Step-0: Prerequisites

The prerequisites are satisfied by completing the tasks assigned to your group manager outlined in [Group Manager tasks for a data science team](#). To summarize here, the following requirements need to be met before you begin the team lead tasks:

- Your group manager has set up the **GroupUtilities** repository (if any).
- Your team lead has set up the **TeamUtilities** repository (if any).
- Your project lead has set up the project repository.
- You have been added to your project repository by your project lead with the privilege to clone from and push back to the project repository.

The second, **TeamUtilities** repository, prerequisite is optional, depending on whether your team has a team-specific utility repository. If any of other three prerequisites has not been completed, contact your team lead, your project lead, or their delegates to set it up by following the instructions for [Team Lead tasks for a data science team](#) or for [Project Lead tasks for a data science team](#).

- Git must be installed on your machine. If you are using a Data Science Virtual Machine (DSVM), Git has been pre-installed and you are good to go. Otherwise, see the [Platforms and tools appendix](#).
- If you are using a **Windows DSVM**, you need to have [Git Credential Manager \(GCM\)](#) installed on your machine. In the README.md file, scroll down to the **Download and Install** section and click the *latest installer*. This takes you to the latest installer page. Download the .exe installer from here and run it.
- If you are using **Linux DSVM**, create an SSH public key on your DSVM and add it to your group Azure DevOps Services. For more information about SSH, see the **Create SSH public key** section in the [Platforms and tools appendix](#).
- If your team and/or project lead has created some Azure file storage that you need to mount to your DSVM, you should get the Azure file storage information from them.

Step 1-3: Clone group, team, and project repositories to local machine

This section provides instructions on completing the first three tasks of project individual contributors:

- Clone the **GroupUtilities** repository R2 to D2
- Clone the **TeamUtilities** repository R4 to D4
- Clone the **Project** repository R5 to D5.

On your local machine, create a directory **C:\GitRepos** (for Windows) or **\$home/GitRepos** (for Linux), and then change to that directory.

Run the one of the following commands (as appropriate for your OS) to clone your **GroupUtilities**, **TeamUtilities**, and **Project** repositories to directories on your local machine:

Windows

```
git clone <the URL of the GroupUtilities repository>
git clone <the URL of the TeamUtilities repository>
git clone <the URL of the Project repository>
```

```

PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\Project_1> git clone https://weig-ds.visualstudio.com/GroupCommon/_git/GroupUtilities
Cloning into 'GroupUtilities'...
remote:
remote:           VSTS
remote:           VSTSvSVTSV
remote:           VSTSvSVSvST
remote: VSTS       VSTSvSVTSvSVS
remote: VSTSvS     VSTSvSVTSV STSVS
remote: VSTSvTSvSvSVSvSVS   TSvST
remote: VS    tSVSVTSvSVS   STSVS
remote: VS    tSVSVTSvST   SVSTS
remote: VS    tSVSVTSvSVsts  VSTSv
remote: VSTSvST   SVSTSvSVsts VSTSv
remote: VSTSvSV  STSVSVSVSvSVS
remote:           VSTSvSVSvST
remote:           VSTSvSVTs
remote:           VSTS   (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Receiving objects: 100% (272/272), 38.20 MiB | 1.29 MiB/s, done.
Resolving deltas: 100% (41/41), done.
Checking connectivity... done.
Checking out files: 100% (213/213), done.
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\Project_1> git clone https://weig-ds.visualstudio.com/DS_Team_1/_git/DS_Team_1_Utils
Cloning into 'DS_Team_1_Utils'...
remote:
remote:           VSTS
remote:           VSTSvSVTSV
remote:           VSTSvSVSvST
remote: VSTS       VSTSvSVTSvSVS
remote: VSTSvS     VSTSvSVTSV STSVS
remote: VSTSvTSvSvSVSvSVS   TSvST
remote: VS    tSVSVTSvSVS   STSVS
remote: VS    tSVSVTSvST   SVSTS
remote: VS    tSVSVTSvSVsts  VSTSv
remote: VSTSvST   SVSTSvSVsts VSTSv
remote: VSTSvSV  STSVSVSVSvSVS
remote:           VSTSvSVSvST
remote:           VSTS   (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Receiving objects: 100% (243/243), 38.12 MiB | 1.82 MiB/s, done.
Resolving deltas: 100% (21/21), done.
Checking connectivity... done.
Checking out files: 100% (213/213), done.
PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\Project_1> git clone https://weig-ds.visualstudio.com/DS_Team_1/_git/Project_1
Cloning into 'Project_1'...
remote:
remote:           VSTS
remote:           VSTSvSVTSV
remote:           VSTSvSVSvST
remote: VSTS       VSTSvSVTSvSVS
remote: VSTSvS     VSTSvSVTSV STSVS
remote: VSTSvTSvSvSVSvSVS   TSvST
remote: VS    tSVSVTSvSVS   STSVS
remote: VS    tSVSVTSvST   SVSTS
remote: VS    tSVSVTSvSVsts  VSTSv
remote: VSTSvST   SVSTSvSVsts VSTSv
remote: VSTSvSV  STSVSVSVSvSVS
remote:           VSTSvSVSvST
remote:           VSTS   (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Unpacking objects: 100% (34/34), done.
Checking connectivity... done.

```

Confirm that you see the three folders under your project directory.

```

PS D:\AML_Projects\TDSP-Linux\Tutorial_testing\Project_1> ls

Directory: D:\AML_Projects\TDSP-Linux\Tutorial_testing\Project_1

Mode                LastWriteTime          Length Name
----              - - - - - - - - - - - - - - - - - - -
d-----        8/11/2016   3:57 PM           0 DS_Team_1_Utils
d-----        8/11/2016   3:55 PM           0 GroupUtilities
d-----        8/11/2016   3:57 PM           0 Project_1

```

Linux

```

git clone <the SSH URL of the GroupUtilities repository>
git clone <the SSH URL of the TeamUtilities repository>
git clone <the SSH URL of the Project repository>

```

```

Cloning into 'GroupUtilities'...
The authenticity of host 'weig-ds.visualstudio.com (23.98.150.230)' can't be established.
RSA key fingerprint is 97:70:33:82:fd:29:3a:73:39:af:6a:07:ad:f8:80:49.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'weig-ds.visualstudio.com,23.98.150.230' (RSA) to the list of known hosts.
remote:
remote:           vSTs
remote:           vSTSVSTSv
remote:           vSTSVSTSvST
remote: VSTS       vSTSVSTSvSTvSv
remote: VSTSvS     vSTSVSTSvSvSv
remote: VSTSvSTSvTsVSTSvSv   TSVST
remote: VS   tSVSTSvSTSv       STSvS
remote: VS   tSVSTSvST       SVSTS
remote: VS   tSVSTSvSTSvSts  VSTSv
remote: VSTSvST   SVSTSvSTSv
remote: VSTSv       STSvSTSvSTSv
remote:           VSTSvSTSvST
remote:           VSTS       (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Receiving objects: 100% (272/272), 38.20 MiB | 6.92 MiB/s, done.
Resolving deltas: 100% (41/41), done.
Checking out files: 100% (213/213), done.
[dsl@weiglinuxdsvm5 Project_1]$ git clone ssh://weig-ds@weig-ds.visualstudio.com:22/DS_Team_1/_git/DS_Team_1.Utilities
Cloning into 'DS_Team_1.Utilities'...
remote:
remote:           vSTs
remote:           vSTSVSTSv
remote:           vSTSVSTSvST
remote: VSTS       vSTSVSTSvSTvSv
remote: VSTSvS     vSTSVSTSvSvSv
remote: VSTSvSTSvTsVSTSvSv   TSVST
remote: VS   tSVSTSvSTSv       STSvS
remote: VS   tSVSTSvST       SVSTS
remote: VS   tSVSTSvSTSvSts  VSTSv
remote: VSTSvST   SVSTSvSTSv
remote: VSTSv       STSvSTSvSTSv
remote:           VSTSvSTSvST
remote:           VSTSvSTSs
remote:           VSTS       (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Receiving objects: 100% (243/243), 38.12 MiB | 6.24 MiB/s, done.
Resolving deltas: 100% (21/21), done.
Checking out files: 100% (213/213), done.
[dsl@weiglinuxdsvm5 Project_1]$ git clone ssh://weig-ds@weig-ds.visualstudio.com:22/DS_Team_1/_git/Project_1
Cloning into 'Project_1'...
remote:
remote:           vSTs
remote:           vSTSVSTSv
remote:           vSTSVSTSvST
remote: VSTS       vSTSVSTSvSTvSv
remote: VSTSvS     vSTSVSTSvSvSv
remote: VSTSvSTSvTsVSTSvSv   TSVST
remote: VS   tSVSTSvSTSv       STSvS
remote: VS   tSVSTSvST       SVSTS
remote: VS   tSVSTSvSTSvSts  VSTSv
remote: VSTSvST   SVSTSvSTSv
remote: VSTSv       STSvSTSvSTSv
remote:           VSTSvSTSvST
remote:           VSTSvSTSs
remote:           VSTS       (TM)
remote:
remote: Microsoft (R) Visual Studio (R) Team Services
remote:
Receiving objects: 100% (34/34), 14.65 KiB | 0 bytes/s, done.
Resolving deltas: 100% (2/2), done.

```

Confirm that you see the three folders under your project directory.

```
[dsl@weiglinuxdsvm5 Project_1]$ ll
total 0
drwxrwxr-x. 7 dsl dsl 80 Aug 11 23:09 DS_Team_1.Utilities
drwxrwxr-x. 7 dsl dsl 80 Aug 11 23:08 GroupUtilities
drwxrwxr-x. 6 dsl dsl 66 Aug 11 23:09 Project_1
```

Step 4-5: Mount Azure file storage to your DSVM (Optional)

To mount Azure file storage to your DSVM, see the instructions in Section 4 of the [Team lead tasks for a data science team](#)

Next steps

Here are links to the more detailed descriptions of the roles and tasks defined by the Team Data Science Process:

- Group Manager tasks for a data science team
- Team Lead tasks for a data science team
- Project Lead tasks for a data science team
- Project Individual Contributors for a data science team

Team Data Science Process project planning

1/30/2019 • 2 minutes to read

The Team Data Science Process (TDSP) provides a lifecycle to structure the development of your data science projects. This article provides links to Microsoft Project and Excel templates that help you plan and manage these project stages.

The lifecycle outlines the major stages that projects typically execute, often iteratively:

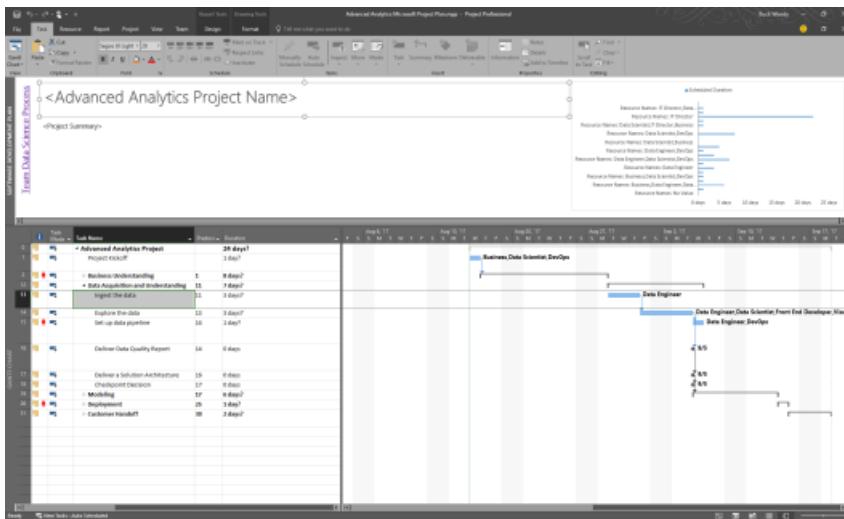
- Business Understanding
- Data Acquisition and Understanding
- Modeling
- Deployment
- Customer Acceptance

For descriptions of each of these stages, see [The Team Data Science Process lifecycle](#).

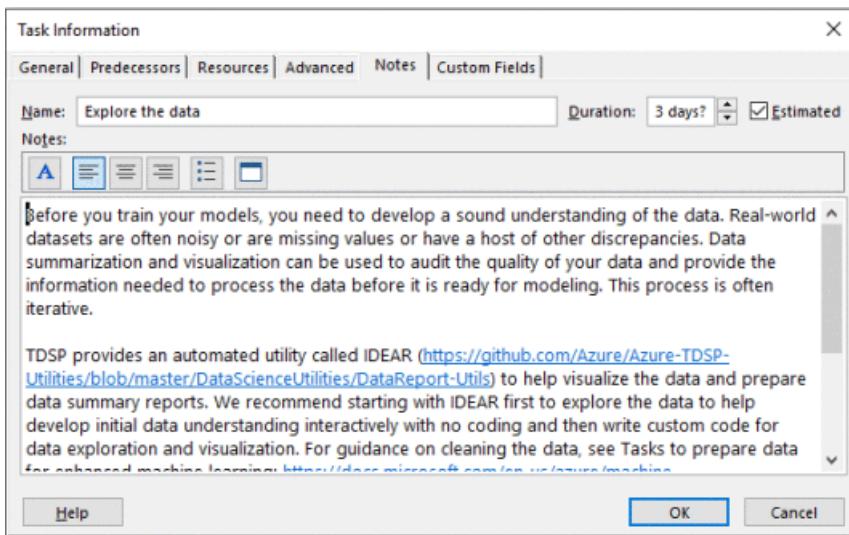
Microsoft Project template

The Microsoft Project template for the Team Data Science Process is available from here: [Microsoft Project template](#)

When you open the plan, click the link to the far left for the TDSP. Change the name and description and then add in any other team resources you need. Estimate the dates required from your experience.



Each task has a note. Open those tasks to see what resources have already been created for you.



Excel template

If you don't have access to Microsoft Project, an Excel worksheet with all the same data is also available for download here: [Excel template](#). You can pull it in to whatever tool you prefer to use.

Use these templates at your own risk. The [usual disclaimers](#) apply.

Repository template

Use this [project template repository](#) to support efficient project execution and collaboration. This repository gives you a standardized directory structure and document templates you can use for your own TDSP project.

Next steps

[Agile development of data science projects](#) This document describes how to execute a data science project in a systematic, version controlled, and collaborative way within a project team by using the Team Data Science Process.

Walkthroughs that demonstrate all the steps in the process for **specific scenarios** are also provided. They are listed and linked with thumbnail descriptions in the [Example walkthroughs](#) article. They illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

Agile development of data science projects

1/30/2019 • 6 minutes to read

This document describes how developers can execute a data science project in a systematic, version controlled, and collaborative way within a project team by using the [Team Data Science Process](#) (TDSP). The TDSP is a framework developed by Microsoft that provides a structured sequence of activities to execute cloud-based, predictive analytics solutions efficiently. For an outline of the personnel roles, and their associated tasks that are handled by a data science team standardizing on this process, see [Team Data Science Process roles and tasks](#).

This article includes instructions on how to:

1. do **sprint planning** for work items involved in a project.

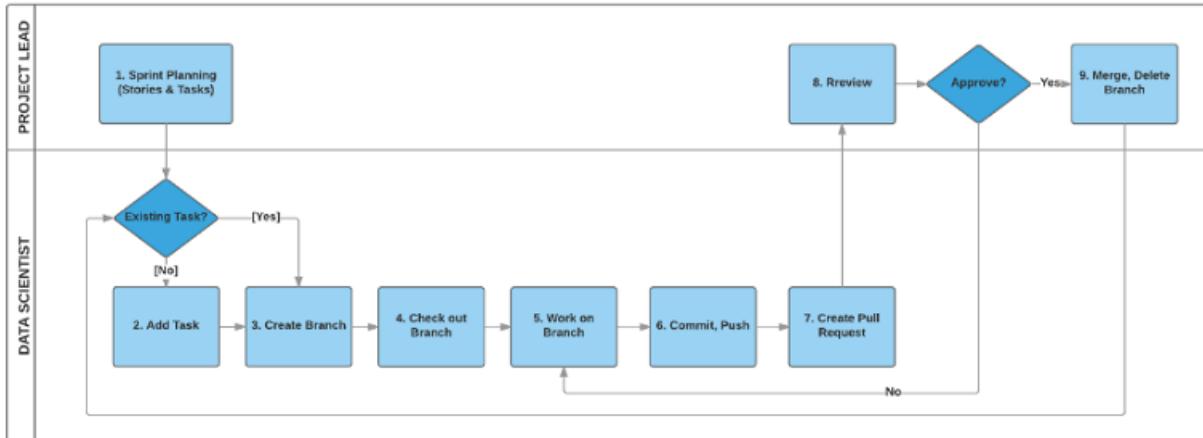
If you are unfamiliar with sprint planning, you can find details and general information [here](#).

2. **add work items** to sprints.

NOTE

The steps needed to set up a TDSP team environment using Azure DevOps Services are outlined in the following set of instructions. They specify how to accomplish these tasks with Azure DevOps Services because that is how to implement TDSP at Microsoft. If you choose to use Azure DevOps Services, items (3) and (4) in the previous list are benefits that you get naturally. If another code hosting platform is used for your group, the tasks that need to be completed by the team lead generally do not change. But the way to complete these tasks is going to be different. For example, the item in section six, **Link a work item with a Git branch**, might not be as easy as it is on Azure DevOps Services.

The following figure illustrates a typical sprint planning, coding, and source-control workflow involved in implementing a data science project:



1. Terminology

In the TDSP sprint planning framework, there are four frequently used types of **work items**: **Feature**, **User Story**, **Task**, and **Bug**. Each project maintains a single backlog for all work items. There is no backlog at the Git repository level under a project. Here are their definitions:

- **Feature**: A feature corresponds to a project engagement. Different engagements with a client are considered different features. Similarly, it is best to consider different phases of a project with a client as different features. If you choose a schema such as **ClientName-EngagementName** to name your features, then you can easily recognize the context of the project/engagement from the names themselves.

- **Story:** Stories are different work items that are needed to complete a feature (project) end-to-end. Examples of stories include:
 - Getting Data
 - Exploring Data
 - Generating Features
 - Building Models
 - Operationalizing Models
 - Retraining Models
- **Task:** Tasks are assignable code or document work items or other activities that need to be done to complete a specific story. For example, tasks in the story *Getting Data* could be:
 - Getting Credentials of SQL Server
 - Uploading Data to SQL Data Warehouse.
- **Bug:** Bugs usually refer to fixes that are needed for an existing code or document that are done when completing a task. If the bug is caused by missing stages or tasks respectively, it can escalate to being a story or a task.

NOTE

Concepts are borrowed of features, stories, tasks, and bugs from software code management (SCM) to be used in data science. They might differ slightly from their conventional SCM definitions.

NOTE

Data scientists may feel more comfortable using an agile template that specifically aligns with the TDSP lifecycle stages. With that in mind, an Agile-derived sprint planning template has been created, where Epics, Stories etc. are replaced by TDSP lifecycle stages or substages. For instructions on how to create an agile template, see [Set up agile data science process in Visual Studio Online](#).

2. Sprint planning

Sprint planning is useful for project prioritization, and resource planning and allocation. Many data scientists are engaged with multiple projects, each of which can take months to complete. Projects often proceed at different paces. On the Azure DevOps Services, you can easily create, manage, and track work items in your project and conduct sprint planning to ensure that your projects are moving forward as expected.

Follow [this link](#) for the step-by-step instructions on sprint planning in Azure DevOps Services.

3. Add a feature

After your project repository is created under a project, go to the team **Overview** page and click **Manage work**.

Welcome

Get started using Visual Studio Team Services to make the most of your team dashboard.

Manage Work
Add work to your board

Collaborate on code
Add code to your repository

Continuously integrate
Automate your builds

Visualize progress
Learn how to add charts

Work assigned to Wei Guo (0)

All done with the work assigned to you? Go to your team's backlog to find new work.

To include a feature in the backlog, click **Backlogs** --> **Features** --> **New**, type in the feature **Title** (usually your project name), and then click **Add**.

Backlogs

Features

Type Feature

Title Client Project 1

Add

Double-click the feature you created. Fill in the descriptions, assign team members for this feature, and set planning parameters for this feature.

You can also link this feature to the project repository. Click **Add link** under the **Development** section. After you have finished editing the feature, click **Save & Close** to exit.

4. Add Story under feature

Under the feature, stories can be added to describe major steps needed to finish the (feature) project. To add a new story, click the + sign to the left of the feature in backlog view.

The screenshot shows the 'Backlogs' section of Microsoft Team Services for the project 'Client_Project_1'. A new 'Feature' item is being created. The 'Title' field contains 'Client Project 1'. The 'Order' column has a red box around the '+' sign, indicating where to click to add more items.

You can edit the details of the story, such as the status, description, comments, planning, and priority In the pop-up window.

The screenshot shows a 'User Story' titled 'Data Ingestion' being edited. The 'Development' section on the right side of the screen has a red box around the '+ Add link' button, which is used to link this story to an existing repository.

You can link this story to an existing repository by clicking **+ Add link** under **Development**.

The screenshot shows the 'Add Link to User Story 27: Feature Engineering' dialog box. It allows selecting a link type and details. The 'Repository' dropdown is set to 'Client_Project_1' and the 'Branch' dropdown is set to 'master'. A red box highlights the 'Branch' dropdown.

5. Add a task to a story

Tasks are specific detailed steps that are needed to complete each story. After all tasks of a story are completed, the story should be completed too.

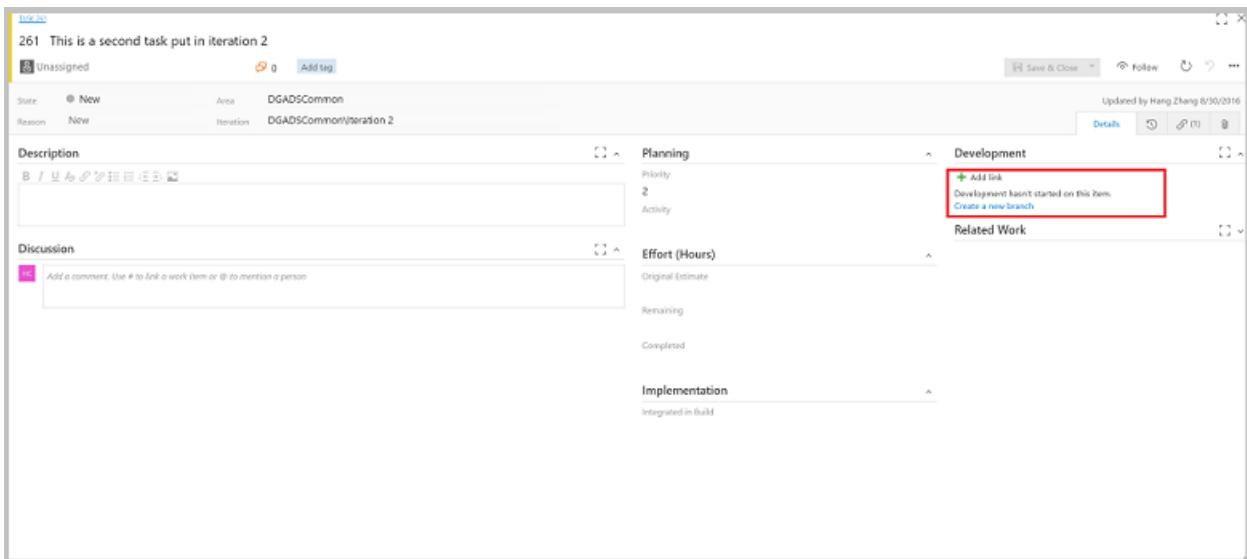
To add a task to a story, click the **+** sign next to the story item, select **Task**, and then fill in the detailed information

of this task in the pop-up window.

The screenshot shows the Microsoft Team Services interface for a project named "Client_Project_1". The "WORK" tab is selected. On the left, there's a navigation pane with "Backlogs" and "Queries" tabs, and sections for "Features", "Stories", "Current", "Iteration 1", "Future", "Iteration 2", and "Iteration 3". The main area is titled "Features" and shows a "Backlog" view. A modal dialog is open for adding a new item, with "Type" set to "Feature". The "Title" field is empty. Below the title, there are buttons for "Task" and "Bug", with "Task" highlighted and a red box drawn around it. The backlog table has columns for "Order", "Work Item Type", "Title", and "State". It lists one feature item under "Client Project 1" and several user stories under "Data Ingestion", all marked as "New".

After the features, stories, and tasks are created, you can view them in the **Backlog** or **Board** views to track their status.

This screenshot shows the same Microsoft Team Services interface after some items have been added. The "Backlog" view is still active. The backlog table now includes a feature item for "Data Ingestion" with two tasks: "Get access to project database" and "Data Exploration". Both tasks are listed under the "User Story" row for "Data Ingestion" and are marked as "New". The other user stories and feature items from the previous screenshot remain in the backlog.

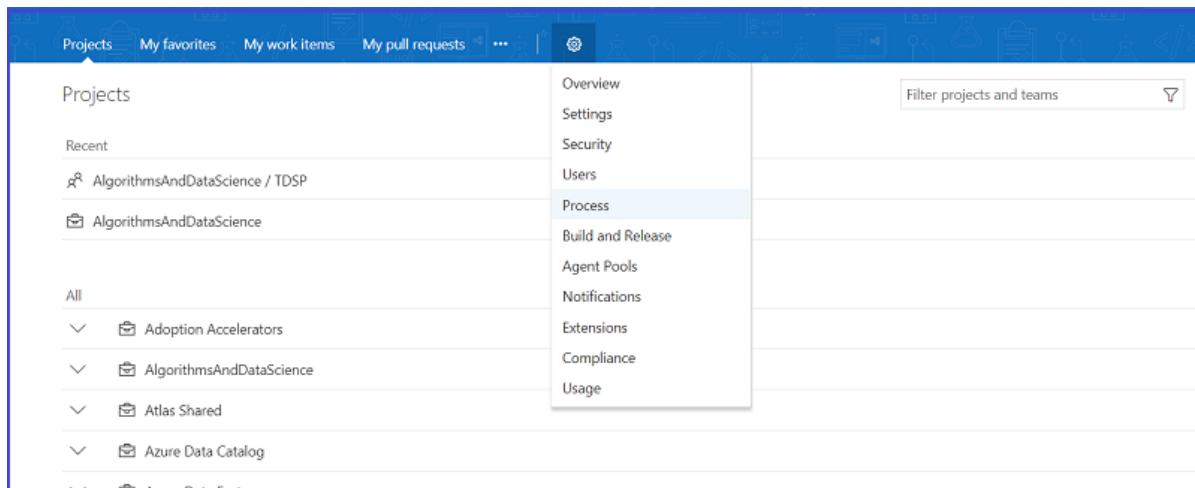


6. Set up an Agile TDSP work template in Visual Studio Online

This article explains how to set up an agile data science process template that uses the TDSP data science lifecycle stages and tracks work items with Visual Studio Online (vso). The steps below walk through an example of setting up the data science-specific agile process template *AgileDataScienceProcess* and show how to create data science work items based on the template.

Agile Data Science Process Template Setup

1. Navigate to server homepage, **Configure** -> **Process**.



2. Navigate to **All processes** -> **Processes**, under **Agile** and click on **Create inherited process**. Then put the process name "AgileDataScienceProcess" and click **Create process**.

Create inherited process from Agile

Create a new inherited process to enable customizations.

Agile [system process]

AgileDataScienceProcess

Description

Learn more

Create process **Cancel**

3. Under the **AgileDataScienceProcess** -> **Work item types** tab, disable **Epic**, **Feature**, **User Story**, and **Task** work item types by **Configure** -> **Disable**

Work item types	Backlog levels	Projects
New work item type		
Name	Description	
Bug	Describes a divergence betw...	
Epic	... Epics help teams effectively...	
Feature	... will be...	
Issue		
Task	Tracks work that needs to b...	

The screenshot shows the 'Work item types' tab of the AgileDataScienceProcess configuration. It lists several work item types: Bug, Epic, Feature, Issue, and Task. The 'Epic' row is currently selected. A context menu is open over the 'Epic' row, with the 'Disable' option highlighted.

4. Navigate to **AgileDataScienceProcess** -> **Backlog levels** tab. Rename "Epics" to "TDSP Projects" by clicking on the **Configure** -> **Edit/Rename**. In the same dialog box, click **+New work item type** in "Data

Science Project" and set the value of **Default work item type** to "TDSP Project"

Edit backlog level

The following fields are automatically added to all work item types on the Portfolio backlogs: Stack Rank

Name

TDSP Projects

Work item types on this backlog level

 Epic (disabled)

 TDSP Project

+ New work item type

Default work item type

Save Cancel

5. Similarly, change Backlog name "Features" to "TDSP Stages" and add the following to the **New work item type**:
 - Business Understanding
 - Data Acquisition
 - Modeling
 - Deployment
6. Rename "User Story" to "TDSP Substages" with default work item type set to newly created "TDSP Substage" type.
7. Set the "Tasks" to newly created Work item type "TDSP Task"
8. After these steps, the Backlog levels should look like this:

Work item types	Backlog levels	Projects
Portfolio backlog		
Portfolio backlogs provide a way to group related items into a hierarchical structure. You can rename and edit any portfolio backlog level.		
	+ New top level portfolio backlog	
Backlog	Work item types	
TDSP Projects	Epic (disabled)	
	TDSP Project	
TDSP Stages	Business Understanding	
	Data Aquisition	
	Deployment	
	Feature (disabled)	
	Modeling	
Requirement backlog		
The requirement backlog level contains your base level work items. There is only one requirement backlog and it cannot be removed, but can be renamed and edited.		
Backlog	Work item types	
TDSP Substages	TDSP Substage	
	User Story (disabled)	
Iteration backlog		
The iteration backlog contains your task work items. There is only one level of iteration backlog and it cannot be removed. The iteration backlog does not have an associated color.		
Backlog	Work item types	
Tasks	Task (disabled)	
	TDSP Task	

Create Data Science Work Items

After the data science process template is created, you can create and track your data science work items that correspond to the TDSP lifecycle.

- When you create a new project, select "Agile\AgileDataScienceProcess" as the **Work item process**:



Create new project

Projects contain your source code, work items, automated builds and more.

Project name *

TDSP CustomerX Project



Description

Version control

Git



Work item process

Agile\AgileDataScienceProcess



Show description

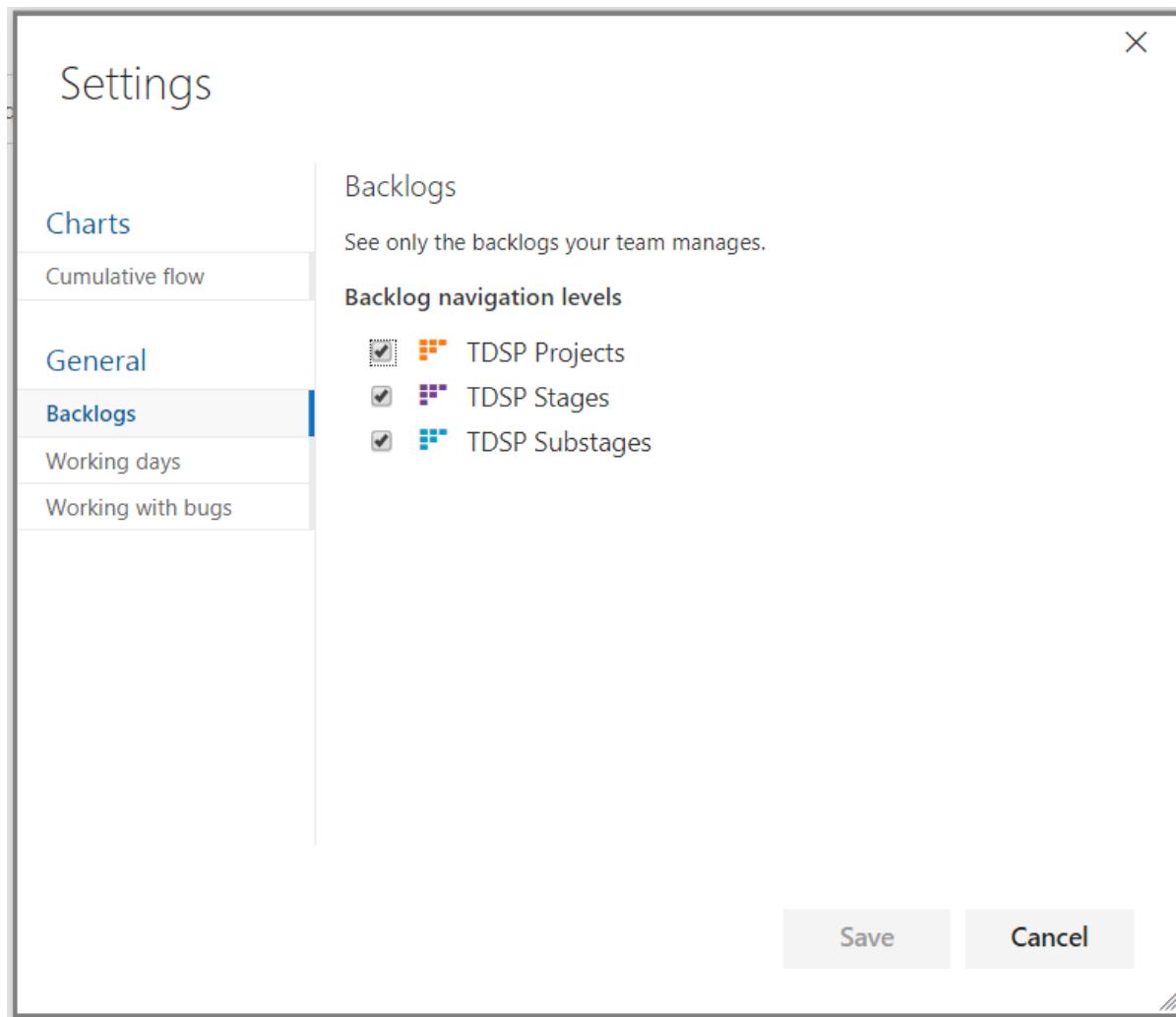
Share with

Everyone in Microsoft

Create

Cancel

2. Navigate to the newly created project, and click on **Work -> Backlogs**.
3. Make "TDSP Projects" visible by clicking on **Configure team settings** and check "TDSP Projects"; then save.



4. Now you can start creating the data science-specific work items.

The screenshot shows the 'TDSP Projects' backlog board. The top navigation bar has tabs for 'Backlogs' (selected) and 'Queries'. The left sidebar shows navigation levels: 'TDSP Projects' (selected), 'TDSP Stages', and 'TDSP Substages'. Under 'TDSP Projects', there are sections for 'Current' (Iteration 1) and 'Future' (Iteration 2, Iteration 3). The main board area is titled 'TDSP Projects' and shows columns for 'Backlog' and 'Board'. Below the columns are buttons for 'New', '+', '[-]', 'Create query', 'Column options', and an envelope icon. The board itself is currently empty.

5. Here is an example of how the data science project work items should appear:

The screenshot shows the Azure DevOps Backlog interface for a project named "TDSP Projects". The backlog is organized into iterations: Current, Iteration 1, Future, Iteration 2, and Iteration 3. The backlog items include:

- Iteration 1:**
 - TDSP Project: Fraud Detection CompanyABC (New)
 - Business Understanding: Define Objectives (New)
 - TDSP Substage: Identify business variables (New)
 - TDSP Substage: Define success metrics (New)
 - TDSP Task: Define accuracy (New)
 - Data Aquisition: Ingest data (New)
 - Data Aquisition: Explore data (New)
 - TDSP Substage: Apply IDEAR to check data quality (New)
 - TDSP Task: Find missing data (New)
 - TDSP Task: Find numerical data distribution (New)
 - Modeling: Feature engineering (New)
 - TDSP Substage: Derive features (New)
 - TDSP Task: Compute categorical features (New)
 - TDSP Task: Compute statistical features (New)
 - Modeling: Model creation (New)
 - TDSP Substage: Traing model (New)
 - TDSP Task: Split data (New)
 - TDSP Task: sweep parameter over a model (New)
 - TDSP Substage: Evaluate model (New)
 - TDSP Task: Select model eval metrics (New)
 - TDSP Task: Generate model performance n... (New)
 - Deployment: Operationalize model (New)
 - TDSP Substage: Deploy model as a web service (New)
 - TDSP Task: Set up web service with Azure c... (New)

Next steps

[Collaborative coding with Git](#) describes how to do collaborative code development for data science projects using Git as the shared code development framework and how to link these coding activities to the work planned with the agile process.

Here are additional links to resources on agile processes.

- Agile process <https://www.visualstudio.com/en-us/docs/work/guidance/agile-process>
- Agile process work item types and workflow <https://www.visualstudio.com/en-us/docs/work/guidance/agile-process-workflow>

Walkthroughs that demonstrate all the steps in the process for **specific scenarios** are also provided. They are listed and linked with thumbnail descriptions in the [Example walkthroughs](#) article. They illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

Collaborative coding with Git

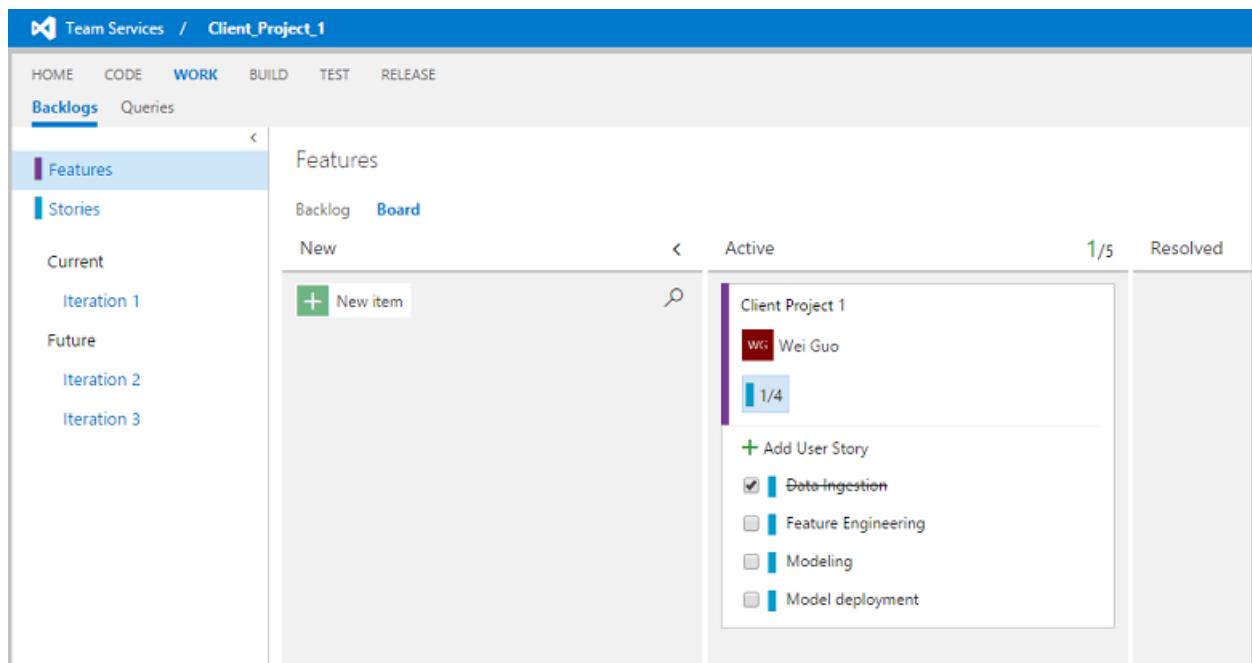
3/12/2019 • 4 minutes to read

In this article we describe how to do collaborative code development for data science projects using Git as the shared code development framework. It covers how to link these coding activities to the work planned in [Agile development](#) and how to do code reviews.

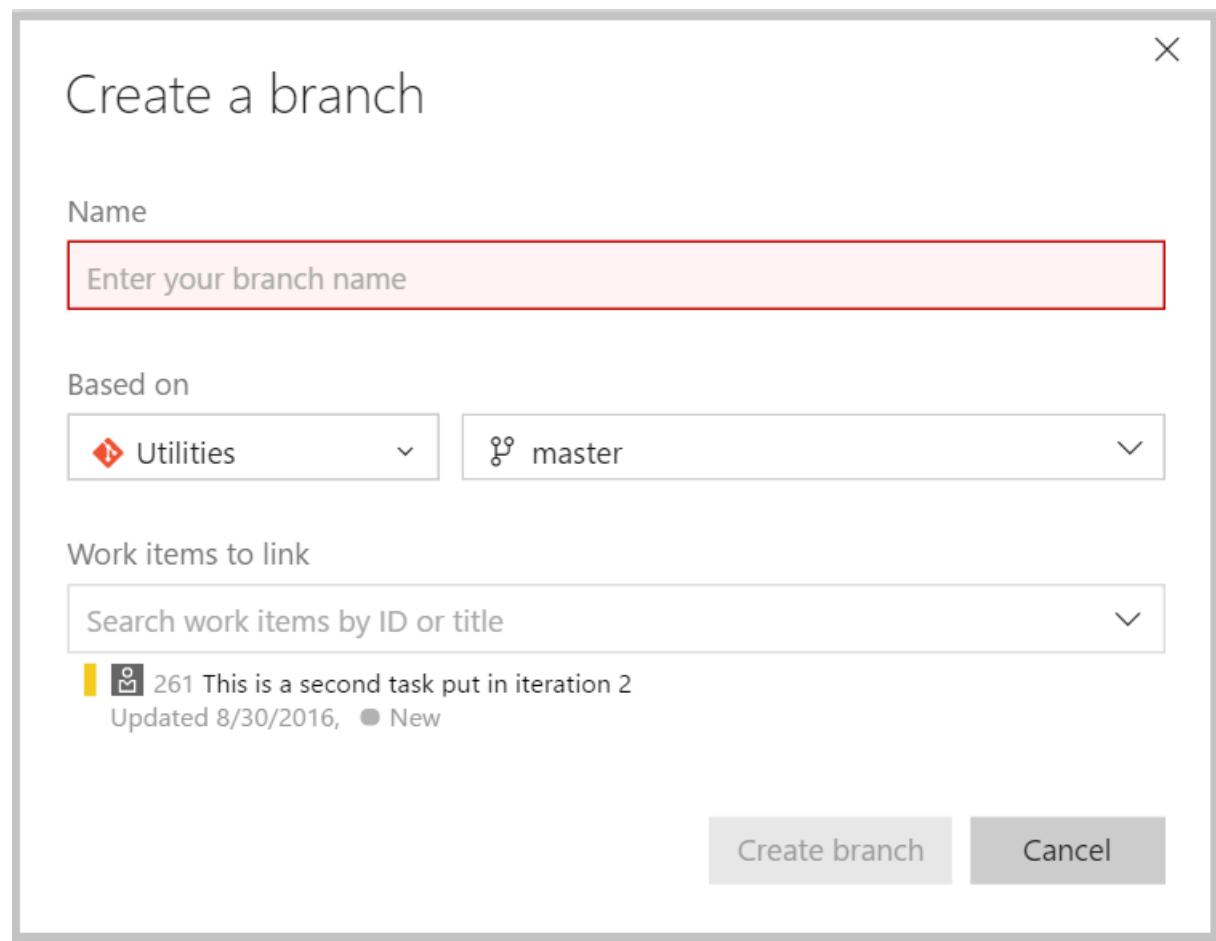
1. Link a work item with a Git branch

Azure DevOps Services provides a convenient way to connect a work item (a story or task) with a Git branch. This enables you to link your story or task directly to the code associated with it.

To connect a work item to a new branch, double-click a work item, and in the pop-up window, click **Create a new branch** under **+ Add link**.



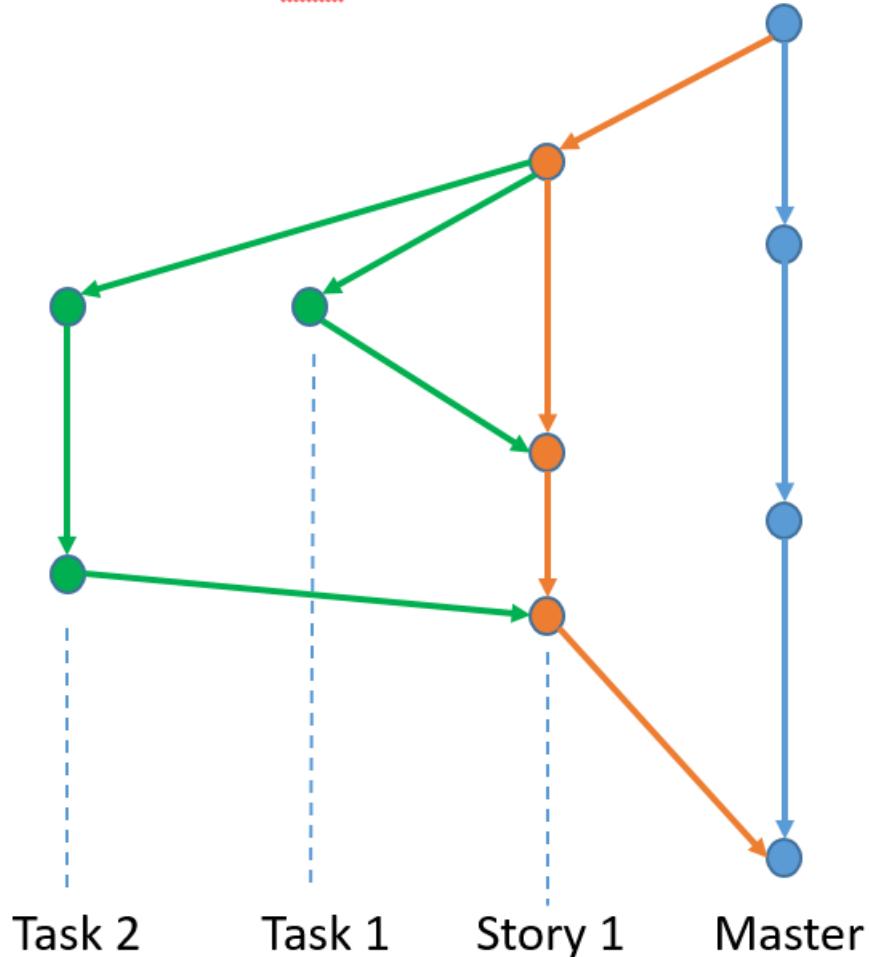
Provide the information for this new branch, such as the branch name, base Git repository, and the branch. The Git repository chosen must be the repository under the same project that the work item belongs to. The base branch can be the master branch or some other existing branch.



A good practice is to create a Git branch for each story work item. Then, for each task work item, you create a branch based on the story branch. Organizing the branches in this hierarchical way that corresponds to the story-task relationships is helpful when you have multiple people working on different stories of the same project, or you have multiple people working on different tasks of the same story. Conflicts can be minimized when each team member works on a different branch and when each member works on different codes or other artifacts when sharing a branch.

The following picture depicts the recommended branching strategy for TDSP. You might not need as many branches as are shown here, especially when you only have one or two people working on the same project, or only one person works on all tasks of a story. But separating the development branch from the master branch is always a good practice. This can help prevent the release branch from being interrupted by the development activities. More complete description of Git branch model can be found in [A Successful Git Branching Model](#).

Git Branches



To switch to the branch that you want to work on, run the following command in a shell command (Windows or Linux).

```
git checkout <branch name>
```

Changing the *<branch name>* to **master** switches you back to the **master** branch. After you switch to the working branch, you can start working on that work item, developing the code or documentation artifacts needed to complete the item.

You can also link a work item to an existing branch. In the **Detail** page of a work item, instead of clicking **Create a new branch**, you click **+ Add link**. Then, select the branch you want to link the work item to.

Add Link to Task 261: This is a second task put in it...

X

Select the link type and the details.

Link type

Branch

Repository

Utilities ▾

Branch

dev ▾

Comment

OK

Cancel

You can also create a new branch in Git Bash commands. If <base branch name> is missing, the <new branch name> is based on *master* branch.

```
git checkout -b <new branch name> <base branch name>
```

2. Work on a branch and commit the changes

Now suppose you make some change to the *data_ingestion* branch for the work item, such as adding an R file on the branch in your local machine. You can commit the R file added to the branch for this work item, provided you are in that branch in your Git shell, using the following Git commands:

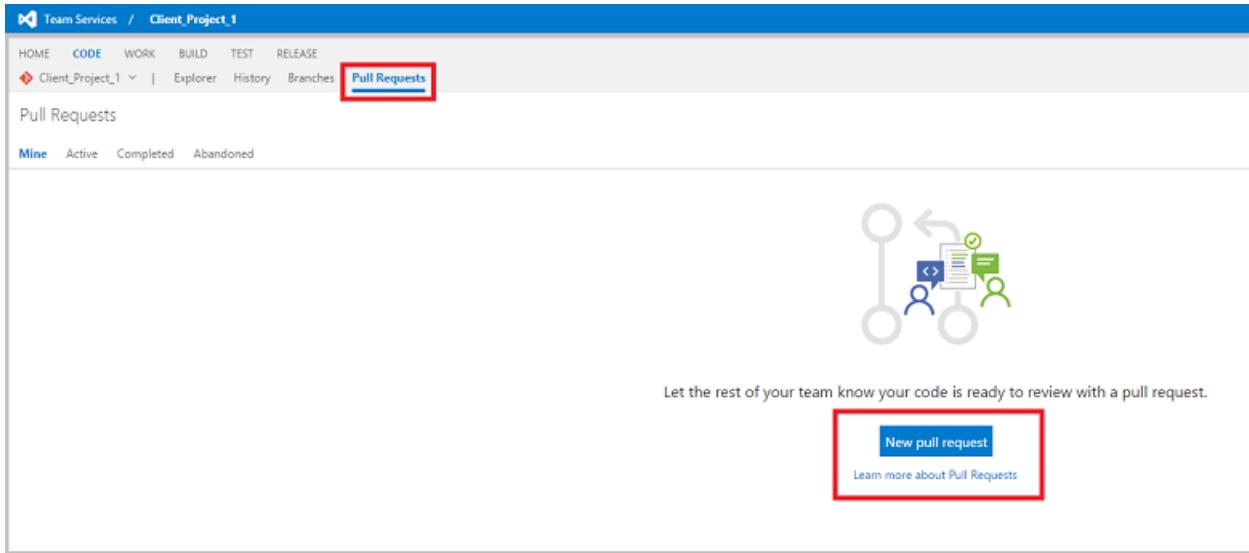
```
git status  
git add .  
git commit -m"added a R scripts"  
git push origin data_ingestion
```

```
PS D:\AML_Projects\TDSP-Linux\sprint_test\Client_Project_1\Code> git status  
On branch data_ingestion  
Untracked files:  
  (use "git add <file>..." to include in what will be committed)  
  
    process_data.R  
  
nothing added to commit but untracked files present (use "git add" to track)  
PS D:\AML_Projects\TDSP-Linux\sprint_test\Client_Project_1\Code> git add .\process_data.R  
PS D:\AML_Projects\TDSP-Linux\sprint_test\Client_Project_1\Code> git commit -m"Wei added a R script"  
[data_ingestion a5a2ff0] Wei added a R script  
 1 file changed, 5 insertions(+)  
  create mode 100644 Code/process data.R  
PS D:\AML_Projects\TDSP-Linux\sprint_test\Client_Project_1\Code> git push origin data_ingestion  
Counting objects: 4, done.  
Delta compression using up to 8 threads.  
Compressing objects: 100% (4/4), done.  
Writing objects: 100% (4/4), 393 bytes | 0 bytes/s, done.  
Total 4 (delta 2), reused 0 (delta 0)  
remote: Analyzing objects... (4/4) (12 ms)  
remote: Storing packfile... done (37 ms)  
remote: Storing index... done (66 ms)  
To https://weig-ds.visualstudio.com/_git/Client_Project_1  
  c3c5ee4..a5a2ff0  data ingestion -> data ingestion
```

3. Create a pull request on Azure DevOps Services

When you are ready after a few commits and pushes, to merge the current branch into its base branch, you can submit a **pull request** on Azure DevOps Services.

Go to the main page of your project and click **CODE**. Select the branch to be merged and the Git repository name that you want to merge the branch into. Then click **Pull Requests**, click **New pull request** to create a pull request review before the work on the branch is merged to its base branch.



Fill in some description about this pull request, add reviewers, and send it out.

Please review the R script

Description

- add a note
- Wei added a R script

Reviewers

y uso Search users and groups

Work items

Search work items by ID or title

25 Data Ingestion Updated 59 minutes ago, Closed

New pull request fewer options

COMMITS (2)

Client_Project_1

Wednesday, August 31, 2016

- Wei added a R script a5a2ff0f by Wei Guo, 21 minutes ago
- add a note c3c5ee4f by Wei Guo, an hour ago

FILES (2)

Showing: All files

Notes.txt /Notes.txt

```
1 This is a note.
```

process_data.R /Code/process_data.R

```
1 X <- rnorm(1000)
2 y <- 3*x + rnorm(1000)
3 plot(x,y)
4 lm(y~x)
5
6
```

4. Review and merge

When the pull request is created, your reviewers get an email notification to review the pull requests. The reviewers need to check whether the changes are working or not and test the changes with the requester if possible. Based on their assessment, the reviewers can approve or reject the pull request.

Discussion Files (1) Commits (1)

- weig push from local branch

WG Please review my R code,thanks!

Wei Guo - an hour ago

Wei Guo commented on the file `wei_explore_data.R`

```

^ 9
10 #work on branch
8 11 y2 <- 3*x+2+rnorm(1000)
9 12 lm(y2~x)

```

WG Can we compare the model estimates?
Wei Guo - an hour ago - reply

Status: Resolved ▾

Y it works
yuso - 55 minutes ago

✓ yuso closed the pull request with commit `c226dc97`
54 minutes ago

Wei Guo commented on the file `wei_explore_data.R`

```

1 1 x <- rnorm(1000)
2 2 y <- 3*x+2+rnorm(1000)
3 hist(x)

```

WG we can make more plots to explore
Wei Guo - 39 minutes ago - reply

Status: Active ▾

```

4 4 plot(x,y)
5 5 lm(y~x)
6 6

```

Wei Guo - save - cancel - workitems (#) - people (@)

Team Services / Client_Project_1

HOME CODE WORK BUILD TEST RELEASE

Client_Project_1 | Explorer History Branches Pull Requests

1 active 3 Please review the R script

Wei Guo IP data_ingestion into master

Overview Files Updates Comments

All changes ▾ Group ▾ All ▾ Showing All files

Client_Project_1

- Notes.txt [-]
- Code
- process_data.R [+]
- plot y vs x [last rev]

process_data.R

```

1 x <- rnorm(1000)
2 y <- 3*x + rnorm(1000)
3 plot(x,y)
4 lm(y~x)
5

```

Wei Guo 3 minutes ago
plot y vs x
done a reply

Approve Approve with suggestions Wait for author Reject Reset feedback

After the review is done, the working branch is merged to its base branch by clicking the **Complete** button. You may choose to delete the working branch after it has merged.

Team Services / Client_Project_1

HOME CODE WORK BUILD TEST RELEASE

Client_Project_1 | Explorer History Branches Pull Requests

1 active 3 Please review the R script

Wei Guo IP data_ingestion into master

Overview Files Updates Comments

All changes ▾ Group ▾ All ▾ Showing All files

Client_Project_1

- Notes.txt [-]
- Code
- process_data.R [+]
- plot y vs x [last rev]

process_data.R

```

1 x <- rnorm(1000)
2 y <- 3*x + rnorm(1000)
3 plot(x,y)
4 lm(y~x)
5

```

Wei Guo 3 minutes ago
plot y vs x
done a reply

Approve Approve with suggestions Wait for author Reject Complete

Complete pull request

PR 3: Please review the R script

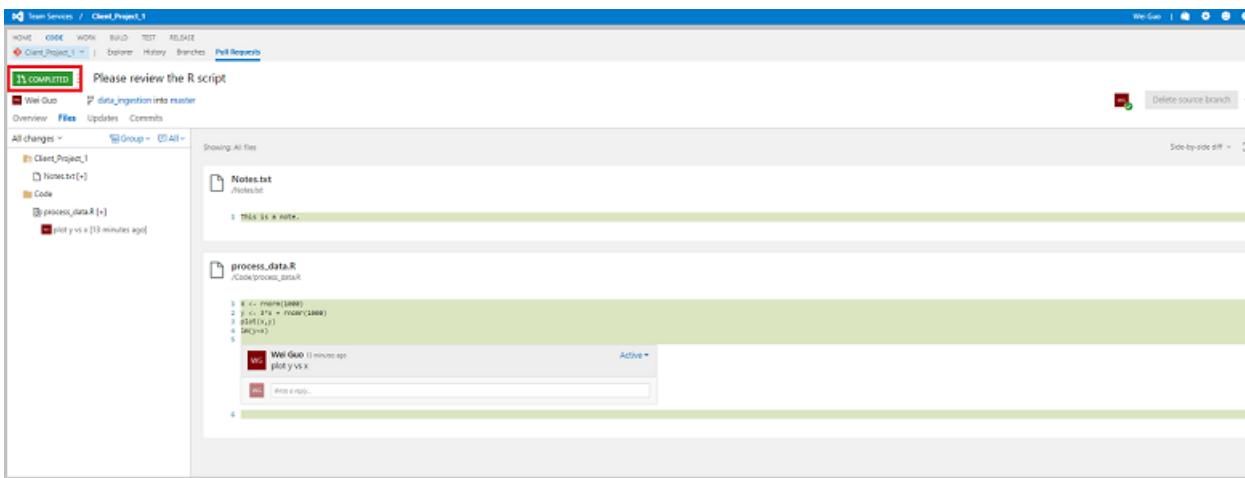
+ Add a note
+ We added a R script

Related work items #25

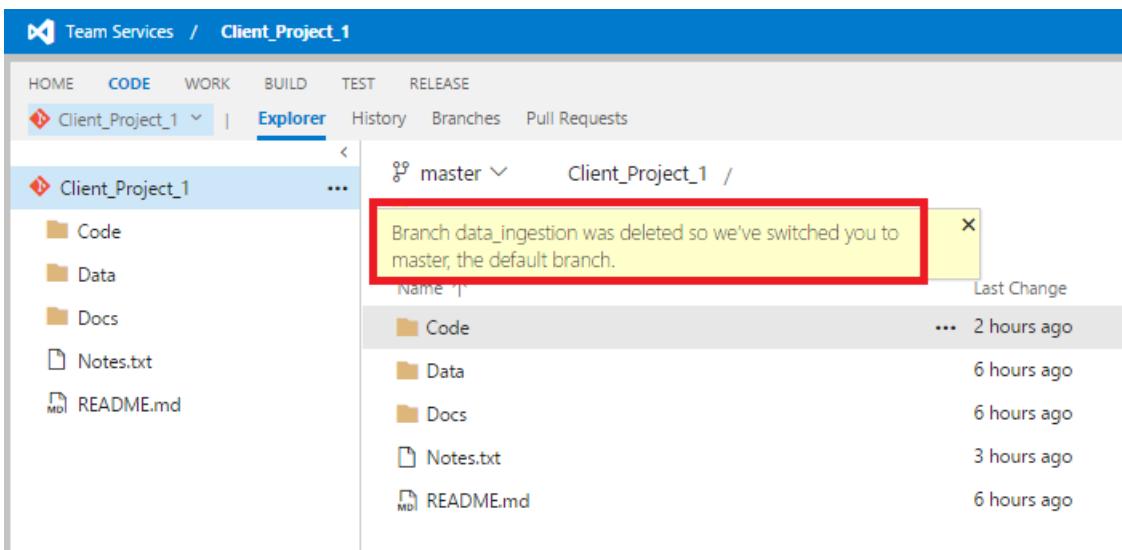
Delete data_ingestion after merging
 squash changes when merging

Complete merge Cancel

Confirm on the top left corner that the request is marked as **COMPLETED**.



When you go back to the repository under **CODE**, you are told that you have been switched to the master branch.



You can also use the following Git commands to merge your working branch to its base branch and delete the working branch after merging:

```
git checkout master
git merge data_ingestion
git branch -d data_ingestion
```

```
PS D:\AML_Projects\TDSP-Linux\sprint_test\Client_Project_1\Code> git checkout master
Switched to branch 'master'
Your branch is up-to-date with 'origin/master'.
PS D:\AML_Projects\TDSP-Linux\sprint_test\Client_Project_1\Code> git merge data_ingestion
Updating 22a62d5..82533d4
Fast-forward
 Code/process_data.R | 8 ++++++++
 Notes.txt          | 1 +
 2 files changed, 9 insertions(+)
 create mode 100644 Code/process_data.R
 create mode 100644 Notes.txt
PS D:\AML_Projects\TDSP-Linux\sprint_test\Client_Project_1\Code> git branch -d data_ingestion
Deleted branch data_ingestion (was 82533d4).
```

Next steps

[Execute of data science tasks](#) shows how to use utilities to complete several common data science tasks such as interactive data exploration, data analysis, reporting, and model creation.

Walkthroughs that demonstrate all the steps in the process for **specific scenarios** are also provided. They are listed and linked with thumbnail descriptions in the [Example walkthroughs](#) article. They illustrate how to combine

cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

Execute data science tasks: exploration, modeling, and deployment

2/4/2019 • 5 minutes to read

Typical data science tasks include data exploration, modeling, and deployment. This article shows how to use the **Interactive Data Exploration, Analysis, and Reporting (IDEAR)** and **Automated Modeling and Reporting (AMAR)** utilities to complete several common data science tasks such as interactive data exploration, data analysis, reporting, and model creation. It also outlines options for deploying a model into a production environment using a variety of toolkits and data platforms, such as the following:

- [Azure Machine Learning](#)
- [SQL-Server with ML services](#)
- [Microsoft Machine Learning Server](#)

1. Exploration

A data scientist can perform exploration and reporting in a variety of ways: by using libraries and packages available for Python (matplotlib for example) or with R (ggplot or lattice for example). Data scientists can customize such code to fit the needs of data exploration for specific scenarios. The needs for dealing with structured data are different than for unstructured data such as text or images.

Products such as Azure Machine Learning service also provide [advanced data preparation](#) for data wrangling and exploration, including feature creation. The user should decide on the tools, libraries, and packages that best suite their needs.

The deliverable at the end of this phase is a data exploration report. The report should provide a fairly comprehensive view of the data to be used for modeling and an assessment of whether the data is suitable to proceed to the modeling step. The Team Data Science Process (TDS P) utilities discussed in the following sections for semi-automated exploration, modeling, and reporting also provide standardized data exploration and modeling reports.

Interactive data exploration, analysis, and reporting using the IDEAR utility

This R markdown-based or Python notebook-based utility provides a flexible and interactive tool to evaluate and explore data sets. Users can quickly generate reports from the data set with minimal coding. Users can click buttons to export the exploration results in the interactive tool to a final report, which can be delivered to clients or used to make decisions on which variables to include in the subsequent modeling step.

At this time, the tool only works on data-frames in memory. A YAML file is needed to specify the parameters of the data-set to be explored. For more information, see [IDEAR in TDSP Data Science Utilities](#).

2. Modeling

There are numerous toolkits and packages for training models in a variety of languages. Data scientists should feel free to use which ever ones they are comfortable with, as long as performance considerations regarding accuracy and latency are satisfied for the relevant business use cases and production scenarios.

The next section shows how to use an R-based TDSP utility for semi-automated modeling. This AMAR utility can be used to generate base line models quickly as well as the parameters that need to be tuned to provide a better performing model. The following model management section shows how to have a system for registering and managing multiple models.

Model training: modeling and reporting using the AMAR utility

The [Automated Modeling and Reporting \(AMAR\) Utility](#) provides a customizable, semi-automated tool to perform model creation with hyper-parameter sweeping and to compare the accuracy of those models.

The model creation utility is an R Markdown file that can be run to produce self-contained HTML output with a table of contents for easy navigation through its different sections. Three algorithms are executed when the Markdown file is run (knit): regularized regression using the `glmnet` package, random forest using the `randomForest` package, and boosting trees using the `xgboost` package). Each of these algorithms produces a trained model. The accuracy of these models is then compared and the relative feature importance plots are reported. Currently, there are two utilities: one is for a binary classification task and one is for a regression task. The primary differences between them is the way control parameters and accuracy metrics are specified for these learning tasks.

A YAML file is used to specify:

- the data input (a SQL source or an R-Data file)
- what portion of the data is used for training and what portion for testing
- which algorithms to run
- the choice of control parameters for model optimization:
 - cross-validation
 - bootstrapping
 - folds of cross-validation
- the hyper-parameter sets for each algorithm.

The number of algorithms, the number of folds for optimization, the hyper-parameters, and the number of hyper-parameter sets to sweep over can also be modified in the Yaml file to run the models quickly. For example, they can be run with a lower number of CV folds, a lower number of parameter sets. If it is warranted, they can also be run more comprehensively with a higher number of CV folds or a larger number of parameter sets.

For more information, see [Automated Modeling and Reporting Utility in TDSP Data Science Utilities](#).

Model management

After multiple models have been built, you usually need to have a system for registering and managing the models. Typically you need a combination of scripts or APIs and a backend database or versioning system. A few options that you can consider for these management tasks are:

1. [Azure Machine Learning - model management service](#)
2. [ModelDB from MIT](#)
3. [SQL-server as a model management system](#)
4. [Microsoft Machine Learning Server](#)

3. Deployment

Production deployment enables a model to play an active role in a business. Predictions from a deployed model can be used for business decisions.

Production platforms

There are various approaches and platforms to put models into production. Here are a few options:

- [Model deployment in Azure Machine Learning service](#)
- [Deployment of a model in SQL-server](#)
- [Microsoft Machine Learning Server](#)

NOTE

Prior to deployment, one has to insure the latency of model scoring is low enough to use in production.

Further examples are available in walkthroughs that demonstrate all the steps in the process for **specific scenarios**. They are listed and linked with thumbnail descriptions in the [Example walkthroughs](#) article. They illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

NOTE

For deployment using Azure Machine Learning Studio, see [Deploy an Azure Machine Learning web service](#).

A/B testing

When multiple models are in production, it can be useful to perform [A/B testing](#) to compare performance of the models.

Next steps

[Track progress of data science projects](#) shows how a data scientist can track the progress of a data science project.

[Model operation and CI/CD](#) shows how CI/CD can be performed with developed models.

Data science code testing on Azure with the Team Data Science Process and Azure DevOps Services

1/30/2019 • 4 minutes to read

This article gives preliminary guidelines for testing code in a data science workflow. Such testing gives data scientists a systematic and efficient way to check the quality and expected outcome of their code. We use a Team Data Science Process (TDSP) [project that uses the UCI Adult Income dataset](#) that we published earlier to show how code testing can be done.

Introduction on code testing

"Unit testing" is a longstanding practice for software development. But for data science, it's often not clear what that means and how you should test code for different stages of a data science lifecycle, such as:

- Data preparation
- Data quality examination
- Modeling
- Model deployment

This article replaces the term "unit testing" with "code testing." It refers to testing as the functions that help to assess if code for a certain step of a data science lifecycle is producing results "as expected." The person who's writing the test defines what's "as expected," depending on the outcome of the function--for example, data quality check or modeling.

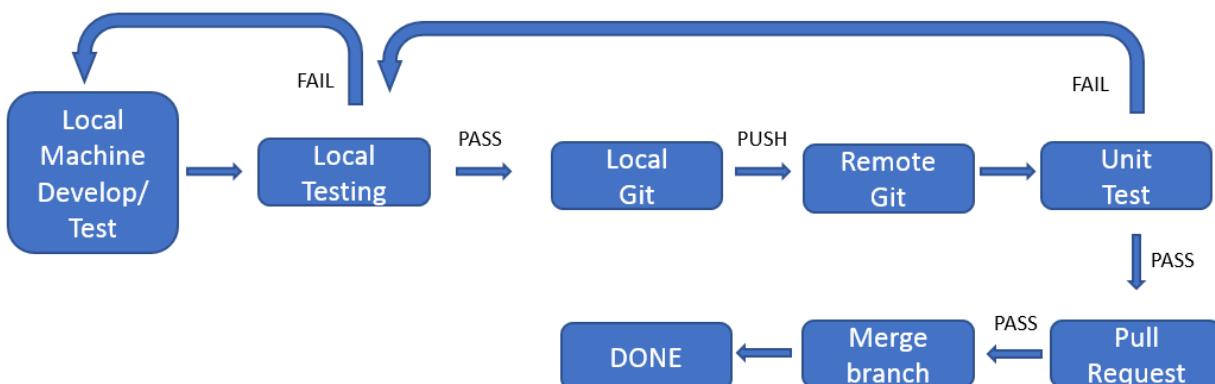
This article provides references as useful resources.

Azure DevOps for the testing framework

This article describes how to perform and automate testing by using Azure DevOps. You might decide to use alternative tools. We also show how to set up an automatic build by using Azure DevOps and build agents. For build agents, we use Azure Data Science Virtual Machines (DSVMs).

Flow of code testing

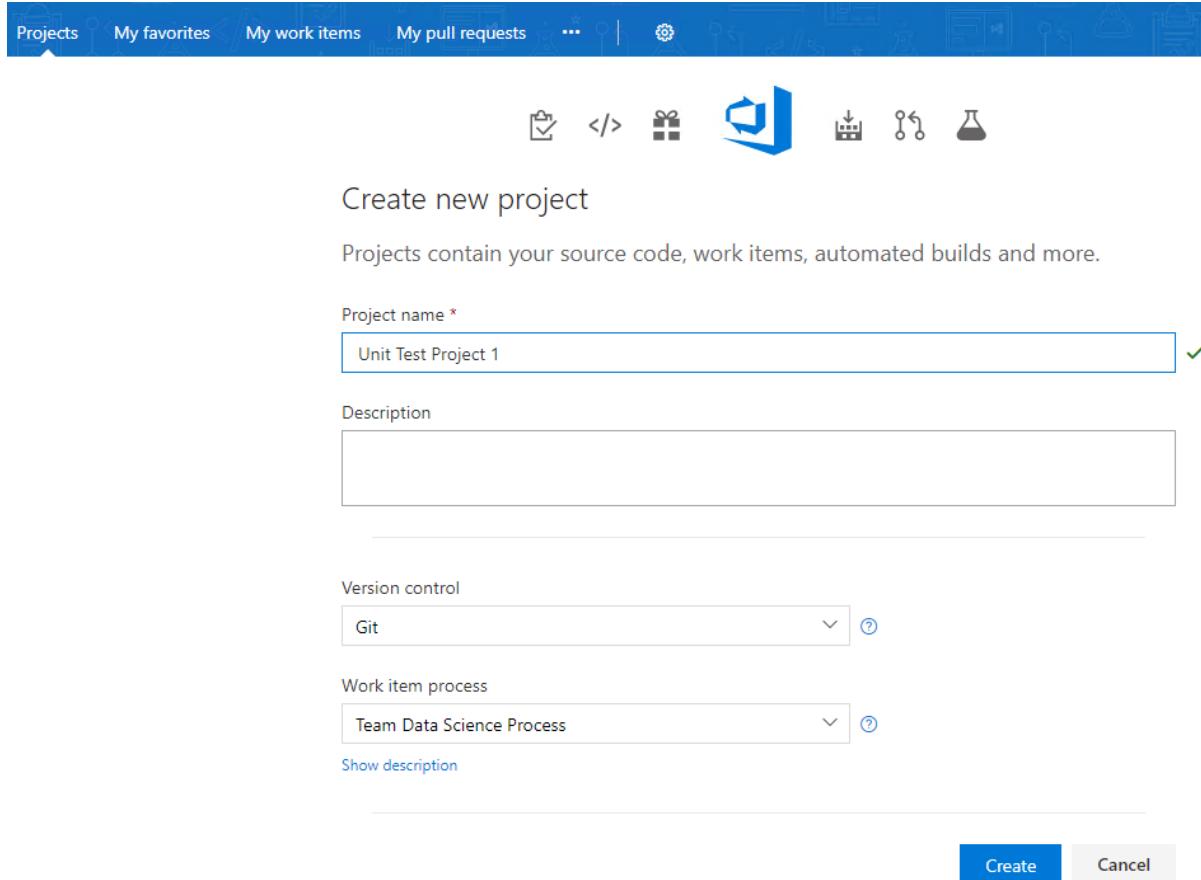
The overall workflow of testing code in a data science project looks like this:



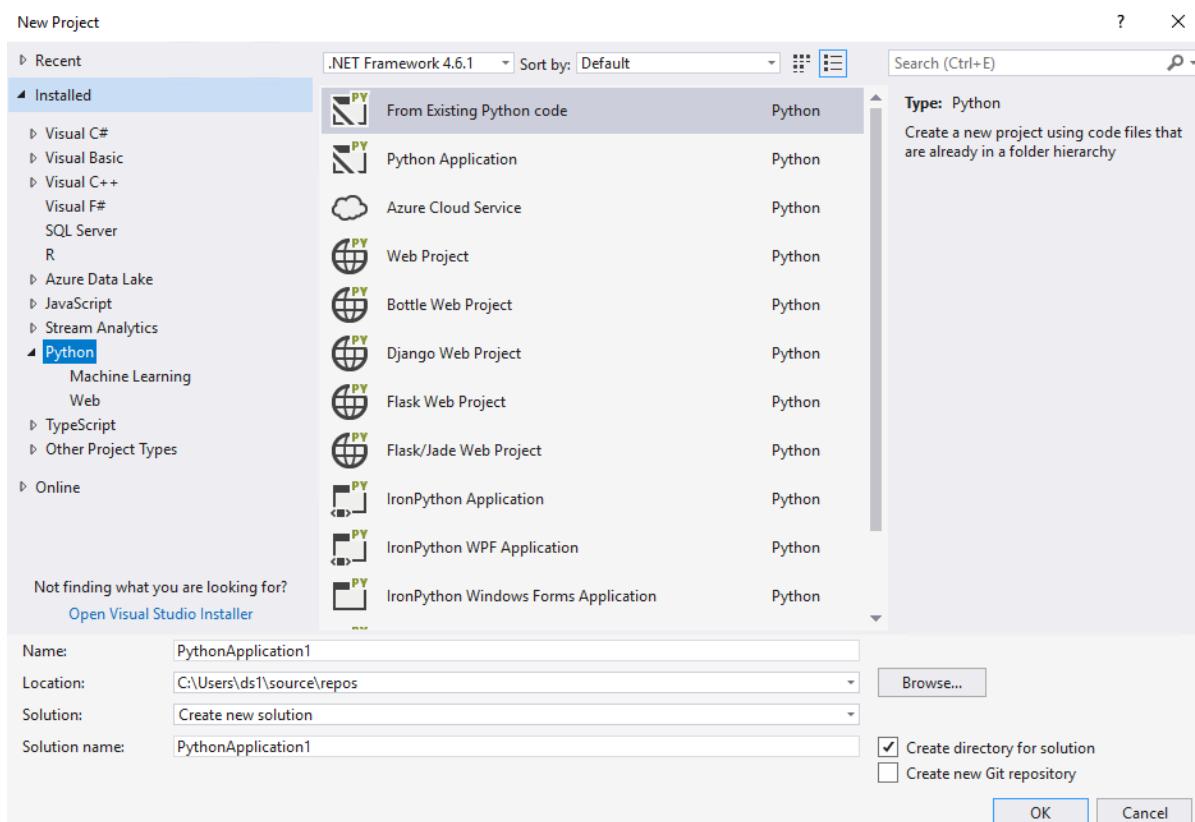
Detailed steps

Use the following steps to set up and run code testing and an automated build by using a build agent and Azure DevOps:

1. Create a project in the Visual Studio desktop application:



After you create your project, you'll find it in Solution Explorer in the right pane:



```

1 #class check_data(object):
2 #     """description of class"""
3 import pandas as pd
4 import pickle
5 import numpy as np
6
7
8 def checkColumnNames(csv_file, required_columns):
9     df = pd.read_csv(csv_file)
10    if set(df.columns) == set(required_columns):
11        print("Columns match!")
12    else:
13        print("Columns do not matched!")
14    return set(df.columns) == set(required_columns)
15
16 def checkResponseLevels(csv_file, response_column):
17     df = pd.read_csv(csv_file)
18     levels=list(df[response_column].unique())
19     if levels == [0,1]:
20         print("Levels match!")
21     else:
22         print("Levels do not match!")
23     return levels == [0,1]
24
25 def checkResponsePercent(csv_file, response_column):
26     df = pd.read_csv(csv_file)
27

```

2. Feed your project code into the Azure DevOps project code repository:

Name	Last change	Commits
__pycache__	3/14/2018	7785cf02 modified for py35 wei
.vs	3/14/2018	7785cf02 modified for py35 wei
adult_income_model.pkl	3/14/2018	ecf9822b add testing for model prediction wei
audult_income_to_score.csv	3/14/2018	0fcfa1fc7 add testing for model prediction wei
data_ingestion.py	1/11/2018	254c6c6f put some code wei
test1.py		
test1.nvc		

3. Suppose you've done some data preparation work, such as data ingestion, feature engineering, and creating label columns. You want to make sure your code is generating the results that you expect. Here's some code that you can use to test whether the data-processing code is working properly:

- Check that column names are right:

```

def checkColumnNames(csv_file, required_columns):
    df = pd.read_csv(csv_file)
    if set(df.columns) == set(required_columns):
        print("Columns match!")
    else:
        print("Columns do not matched!")
    return set(df.columns) == set(required_columns)

```

- Check that response levels are right:

```

def checkResponseLevels(csv_file, response_column):
    df = pd.read_csv(csv_file)
    levels=list(df[response_column].unique())
    if levels == [0,1]:
        print("Levels match!")
    else:
        print("Levels do not match!")
    return levels == [0,1]

```

- Check that response percentage is reasonable:

```

def checkResponsePercent(csv_file,response_column):
    df = pd.read_csv(csv_file)
    ratio1 = df.loc[df[response_column] == 1].shape[0]/df.shape[0]
    if ratio1 > 0.5:
        print("Response levels(0/1) are messed up!")
    else:
        print("Response 0/1 are OK, and OK, Happy Friday!")
    return ratio1 < 0.5

```

- Check the missing rate of each column in the data:

```

def checkMissingRate2(csv_file, threshold):
    df = pd.read_csv(csv_file)
    for x in df.columns:
        miss_rate = df[x].isnull().sum()/df.shape[0]
        if miss_rate > threshold:
            print("{} has more than {} missing values!".format(x,threshold))
            return False
    print("No column has missing values more than {}!".format(threshold))
    return True

```

- After you've done the data processing and feature engineering work, and you've trained a good model, make sure that the model you trained can score new datasets correctly. You can use the following two tests to check the prediction levels and distribution of label values:

- Check prediction levels:

```

def checkPredictionLevels(csv_file, model_file):
    df = pd.read_csv(csv_file)
    X_to_score = df[['education_num','age','hours_per_week']].values
    loaded_model = pickle.load(open(model_file, 'rb'))
    y_hat = loaded_model.predict(X_to_score)
    if list(set(y_hat)) == [0,1]:
        print("Prediction Levels match!")
    else:
        print("Prediction Levels do not match!")
    return list(set(y_hat)) == [0,1]

```

- Check the distribution of prediction values:

```

def checkPredictionPercent(csv_file,model_file):
    df = pd.read_csv(csv_file)
    X_to_score = df[['education_num','age','hours_per_week']].values
    loaded_model = pickle.load(open(model_file, 'rb'))
    y_hat = loaded_model.predict(X_to_score)
    ratio1 = sum(y_hat == 1)/len(y_hat)
    if ratio1 > 0.5:
        print("Response levels(0/1) are messed up!")
    else:
        print("Response 0/1 percent is OK, Happy prediction!")
    return ratio1 < 0.5

```

- Put all test functions together into a Python script called **test_funcs.py**:

```

test_funcs.py ✘ X

1  #class check_data(object):
2  #    """description of class"""
3  import pandas as pd
4  import pickle
5  import numpy as np
6
7
8  def checkColumnNames(csv_file, required_columns):
9      df = pd.read_csv(csv_file)
10     if set(df.columns) == set(required_columns):
11         print("Columns match!")
12     else:
13         print("Columns do not matched!")
14     return set(df.columns) == set(required_columns)
15
16  def checkResponseLevels(csv_file, response_column):
17      df = pd.read_csv(csv_file)
18      levels=list(df[response_column].unique())
19      if levels == [0,1]:
20          print("Levels match!")
21      else:
22          print("Levels do not match!")
23      return levels == [0,1]
24
25  def checkResponsePercent(csv_file, response_column):
26      df = pd.read_csv(csv_file)
27      ratio1 = df.loc[df[response_column] == 1].shape[0]/df.shape[0]
28      if ratio1 > 0.5:
29          print("Response levels(0/1) are messed up!")
30      else:
31          print("Response 0/1 are OK, and OK, Happy Friday!")
32      return ratio1 < 0.5
33

```

6. After the test codes are prepared, you can set up the testing environment in Visual Studio.

Create a Python file called **test1.py**. In this file, create a class that includes all the tests you want to do. The following example shows six tests prepared:

```

6  class Test_A(unittest.TestCase):
7
8      def checkColumnNames(self):
9          assert checkColumnNames(file_name, required_columns) == True
10
11     def checkResponseLevels(self):
12         assert checkResponseLevels(file_name, response_column) == True
13
14     def checkResponsePercent(self):
15         assert checkResponsePercent(file_name, response_column) == True
16
17     def checkMissingRate(self):
18         assert checkMissingRate2(file_name, threshold) == True
19
20     def checkPredictionLevels(self):
21         assert checkPredictionLevels(score_file_name, model_file_name) == True
22
23     def checkPredictionPercent(self):
24         assert checkPredictionPercent(score_file_name, model_file_name) == True

```

7. Those tests can be automatically discovered if you put **codetest.testCase** after your class name. Open Test Explorer in the right pane, and select **Run All**. All the tests will run sequentially and will tell you if the test is successful or not.

The screenshot shows the Visual Studio Code interface with two main panes. The left pane is the 'Test Explorer' showing test results:

- Fast < 100 ms (1)
 - ✓ test_checkPredictionLevels < 1 ms
- Not Run (5)
 - ✓ test_checkColumnNames
 - ✓ test_checkMissingRate
 - ✓ test_checkPredictionPercent
 - ✓ test_checkResponseLevels
 - ✓ test_checkResponsePercent

The right pane is the code editor with the file 'test1.py*' open, showing the following code:

```
import unittest
import pandas as pd
from test_funcs import *

class UCI_unit_test(unittest.TestCase):
    def test_checkColumnNames(self):
        assert checkColumnNames()

    def test_checkResponseLevel(self):
        assert checkResponseLevel()

    def test_checkResponsePerce(self):
        assert checkResponsePerce()

    def test_checkMissingRate(self):
        assert checkMissingRate()

    def test_checkPredictionLev(self):
        assert checkPredictionLev()

    def test_checkPredictionPer(self):
        assert checkPredictionPer()
```

8. Check in your code to the project repository by using Git commands. Your most recent work will be reflected shortly in Azure DevOps.

```
PS C:\Unit_Test_UCI_Income> git status
On branch master
Your branch is up-to-date with 'origin/master'.

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git checkout -- <file>..." to discard changes in working directory)

        modified:   .vs/unit_test_1/v15/.suo
        modified:   __pycache__/test_funcs.cpython-35.pyc
        modified:   test1.py
        modified:   test1.pyc
        modified:   test_funcs.py
        modified:   unit_test.pyproj

Untracked files:
  (use "git add <file>..." to include in what will be committed)

    __pycache__/test1.cpython-35.pyc
    __pycache__/test2.cpython-35.pyc
    __pycache__/test_model.cpython-35.pyc

no changes added to commit (use "git add" and/or "git commit -a")
PS C:\Unit_Test_UCI_Income> git add .
PS C:\Unit_Test_UCI_Income> git commit -m"revised code wei"
[master 60841e1] revised code wei
 9 files changed, 7 insertions(+), 19 deletions(-)
 rewrite .vs/unit_test_1/v15/.suo (63%)
 create mode 100644 __pycache__/test1.cpython-35.pyc
 create mode 100644 __pycache__/test2.cpython-35.pyc
 create mode 100644 __pycache__/test_model.cpython-35.pyc
 rewrite test1.pyc (64%)
PS C:\Unit_Test_UCI_Income> git push
Counting objects: 15, done.
Delta compression using up to 24 threads.
Compressing objects: 100% (13/13), done.
Writing objects: 100% (15/15), 6.16 KiB | 2.05 MiB/s, done.
Total 15 (delta 6), reused 0 (delta 0)
remote: Analyzing objects... (15/15) (26 ms)
remote: Storing packfile... done (213 ms)
remote: Storing index... done (32 ms)
To https://dg-ads.visualstudio.com/_git/Unit_Test_UCI_Income
  7785cf0..60841e1 master -> master
PS C:\Unit_Test_UCI_Income> -
```

Contents		History		
Name		Last change ↓	Commits	
e_	.vs	2 minutes ago	60841e16	revised code wei wei
ome_model.pkl	__pycache__	2 minutes ago	60841e16	revised code wei wei
income_to_score.csv	unit_test.pyproj	2 minutes ago	60841e16	revised code wei wei
estion.py	test1.py	2 minutes ago	60841e16	revised code wei wei
	test1.pyc	2 minutes ago	60841e16	revised code wei wei
	test_funcs.py	2 minutes ago	60841e16	revised code wei wei

9. Set up automatic build and test in Azure DevOps:

- a. In the project repository, select **Build and Release**, and then select **+New** to create a new build process.

```
![Selections for starting a new build process](./media/code-test/create_new_build.PNG)
```

- b. Follow the prompts to select your source code location, project name, repository, and branch information.

```
![Source, name, repository, and branch information](./media/code-test/fill_in_build_info.PNG)
```

- c. Select a template. Because there's no Python project template, start by selecting **Empty process**.

```
![List of templates and "Empty process" button](./media/code-test/start_empty_process_template.PNG)
```

- d. Name the build and select the agent. You can choose the default here if you want to use a DSVM to finish the build process. For more information about setting agents, see [Build and release agents](#).

```
![Build and agent selections](./media/code-test/select_agent.PNG)
```

- e. Select **+** in the left pane, to add a task for this build phase. Because we're going to run the Python script **test1.py** to finish all the checks, this task is using a PowerShell command to run Python code.

```
![Add tasks pane with PowerShell selected](./media/code-test/add_task_powershell.PNG)
```

- f. In the PowerShell details, fill in the required information, such as the name and version of PowerShell. Choose **Inline Script** as the type.

In the box under ****Inline Script****, you can type ****python test1.py****. Make sure the environment variable is set up correctly for Python. If you need a different version or kernel of Python, you can explicitly specify the path as shown in the figure:

```
![PowerShell details](./media/code-test/powershell_scripts.PNG)
```

- g. Select **Save & queue** to finish the build pipeline process.

```
![Save & queue button](./media/code-test/save_and_queue_build_definition.PNG)
```

Now every time a new commit is pushed to the code repository, the build process will start automatically. (Here we use master as the repository, but you can define any branch.) The process runs the **test1.py** file in the agent

machine to make sure that everything defined in the code runs correctly.

If alerts are set up correctly, you'll be notified in email when the build is finished. You can also check the build status in Azure DevOps. If it fails, you can check the details of the build and find out which piece is broken.

Reply Reply All Forward IM
Wed 3/21/2018 10:51 AM

 Visual Studio Team Services
Unit_Test_UCI_Income Build 47 succeeded
To Wei Guo

47 - Succeeded

[Open Build Report in Web Access](#)

Continuous Integration Build of Unit_Test_UCI_Income-CI (2) (Unit_Test_UCI_Income)
Ran for 0.3 minutes (Default), completed at Wed 03/21/2018 05:50 PM

Request Summary

Request 47 Wei Guo Completed

Summary

| Finalize build

0 error(s), 0 warning(s)

| Phase 1

0 error(s), 0 warning(s)

Notes:

- All dates and times are shown in UTC

We sent you this notification due to a default subscription | [Unsubscribe](#) | [View](#)

Provided by [Microsoft Visual Studio® Team Foundation Server](#)

Builds Releases Library Task Groups Deployment Groups*

✓ Build 47

✓ Phase 1

- ✓ Job
 - ✓ Initialize Job
 - ✓ Get Sources
 - ✓ PowerShell Script
 - ✓ Post Job Cleanup
- ✓ Finalize build
- ✓ Report build status

[Edit build definition](#) [Queue new build...](#) :

Build succeeded

Build 47  Ran for 16 seconds (Default), com

Summary Timeline Code coverage* Tests

Build details

Definition	Unit_Test_UCI_Income-CI (2) (edit)
Source	master
Source version	Commit 60841e16
Requested by	Microsoft.VisualStudio.Services.TFS on
Queue name	Default
Queued	Wednesday, March 21, 2018 5:50 PM
Started	Wednesday, March 21, 2018 5:50 PM
Finished	Wednesday, March 21, 2018 5:50 PM
Retained state	Build not retained

Next steps

- See the [UCI income prediction repository](#) for concrete examples of unit tests for data science scenarios.
- Follow the preceding outline and examples from the UCI income prediction scenario in your own data science projects.

References

- [Team Data Science Process](#)
- [Visual Studio Testing Tools](#)
- [Azure DevOps Testing Resources](#)
- [Data Science Virtual Machines](#)

Tracking the progress of data science projects

1/30/2019 • 2 minutes to read

Data science group managers, team leads, and project leads need to track the progress of their projects, what work has been done on them and by whom, and remains on the to-do lists.

Azure DevOps dashboards

If you are using Azure DevOps, you are able to build dashboards to track the activities and the work items associated with a given Agile project.

For more information on how to create and customize dashboards and widgets on Azure DevOps, see the following sets of instructions:

- [Add and manage dashboards](#)
- [Add widgets to a dashboard.](#)

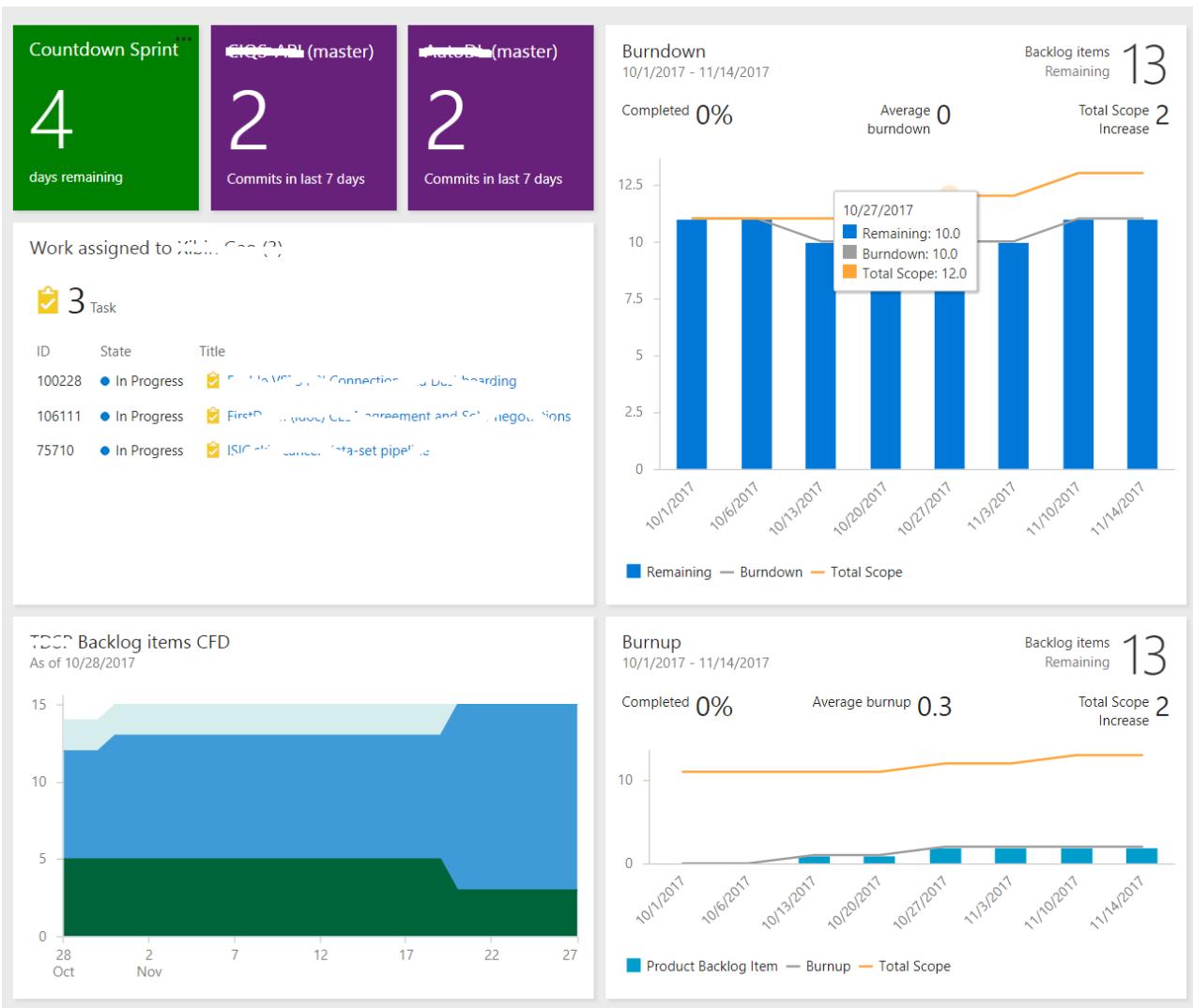
Example dashboard

Here is a simple example dashboard that is built to track the sprint activities of an Agile data science project, as well as the number of commits to associated repositories. The **top left** panel shows:

- the countdown of the current sprint,
- the number of commits for each repository in the last 7 days
- the work item for specific users.

The remaining panels show the cumulative flow diagram (CFD), burndown, and burnup for a project:

- **Bottom left:** CFD the quantity of work in a given state, showing approved in gray, committed in blue, and done in green.
- **Top right:** burndown chart the work left to complete versus the time remaining).
- **Bottom right:** burnup chart the work that has been completed versus the total amount of work.



For a description of how to build these charts, see the quickstarts and tutorials at [Dashboards](#).

Next steps

Walkthroughs that demonstrate all the steps in the process for **specific scenarios** are also provided. They are listed and linked with thumbnail descriptions in the [Example walkthroughs](#) article. They illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

Creating continuous integration pipeline on Azure using Docker, Kubernetes, and Python Flask application

1/30/2019 • 2 minutes to read

For an AI application, there are frequently two streams of work, Data Scientists building machine learning models and App developers building the application and exposing it to end users to consume. In this article, we demonstrate how to implement a Continuous Integration (CI)/Continuous Delivery (CD) pipeline for an AI application. AI application is a combination of application code embedded with a pretrained machine learning (ML) model. For this article, we are fetching a pretrained model from a private Azure blob storage account, it could be an AWS S3 account as well. We will use a simple python flask web application for the article.

NOTE

This is one of several ways CI/CD can be performed. There are alternatives to the tooling and other pre-requisites mentioned below. As we develop additional content, we will publish those.

GitHub repository with document and code

You can download the source code from [GitHub](#). A [detailed tutorial](#) is also available.

Pre-requisites

The following are the pre-requisites for following the CI/CD pipeline described below:

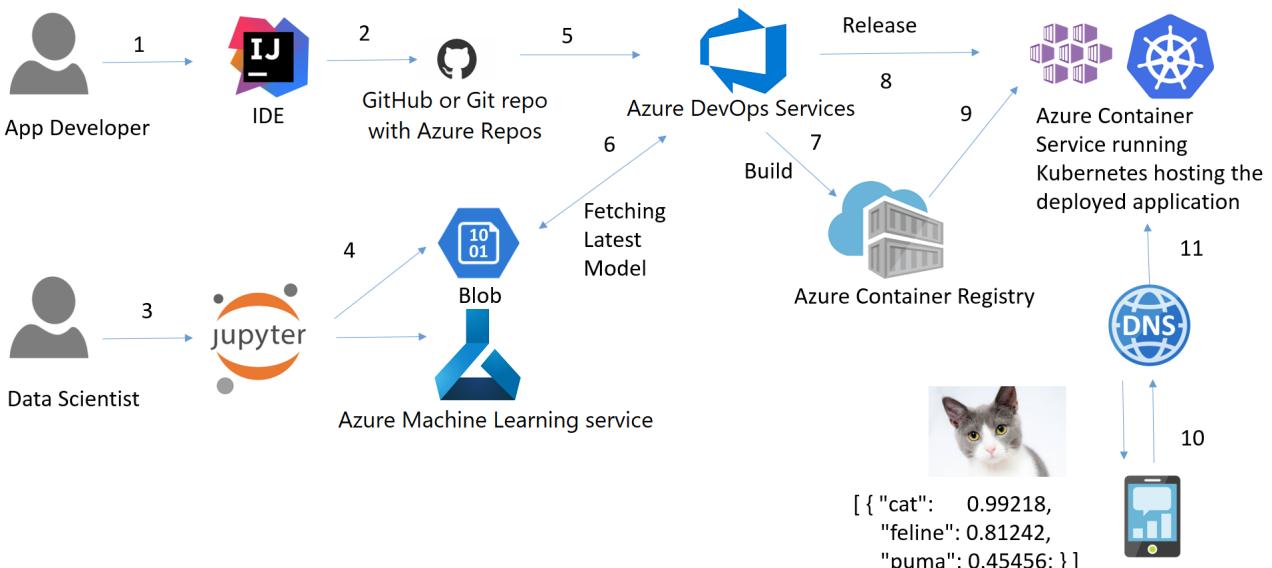
- [Azure DevOps Organization](#)
- [Azure CLI](#)
- [Azure Container Service \(AKS\) cluster running Kubernetes](#)
- [Azure Container Registry \(ACR\) account](#)
- [Install Kubectl to run commands against Kubernetes cluster](#). We will need this to fetch configuration from ACS cluster.
- Fork the repository to your GitHub account.

Description of the CI/CD pipeline

The pipeline kicks off for each new commit, run the test suite, if the test passes takes the latest build, packages it in a Docker container. The container is then deployed using Azure Container Service (ACS) and images are securely stored in Azure Container Registry (ACR). ACS is running Kubernetes for managing container cluster but you can choose Docker Swarm or Mesos.

The application securely pulls the latest model from an Azure Storage account and packages that as part of the application. The deployed application has the app code and ML model packaged as single container. This decouples the app developers and data scientists, to make sure that their production app is always running the latest code with latest ML model.

The pipeline architecture is given below.



Steps of the CI/CD pipeline

1. Developer work on the IDE of their choice on the application code.
2. They commit the code to source control of their choice (Azure DevOps has good support for various source controls)
3. Separately, the data scientist work on developing their model.
4. Once happy, they publish the model to a model repository, in this case we are using a blob storage account.
5. A build is kicked off in Azure DevOps based on the commit in GitHub.
6. Azure DevOps Build pipeline pulls the latest model from Blob container and creates a container.
7. Azure DevOps pushes the image to private image repository in Azure Container Registry
8. On a set schedule (nightly), release pipeline is kicked off.
9. Latest image from ACR is pulled and deployed across Kubernetes cluster on ACS.
10. Users request for the app goes through DNS server.
11. DNS server passes the request to load balancer and sends the response back to user.

Next steps

- Refer to the [tutorial](#) to follow the details and implement your own CI/CD pipeline for your application.

References

- [Team Data Science Process \(TDSP\)](#)
- [Azure Machine Learning \(AML\)](#)
- [Azure DevOps](#)
- [Azure Kubernetes Services \(AKS\)](#)

Walkthroughs executing the Team Data Science Process

1/30/2019 • 2 minutes to read

These **end-to-end walkthroughs** demonstrate the steps in the Team Data Science Process for specific scenarios. They illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an **intelligent application**. The walkthroughs are grouped by **platform** that they use.

Walkthrough descriptions

Here are brief descriptions of what these walkthrough examples provide on their respective platforms:

- [HDInsight Spark walkthroughs using PySpark and Scala](#) These walkthroughs use PySpark and Scala on an Azure Spark cluster to do predictive analytics.
- [HDInsight Hadoop walkthroughs using Hive](#) These walkthroughs use Hive with an HDInsight Hadoop cluster to do predictive analytics.
- [Azure Data Lake walkthroughs using U-SQL](#) These walkthroughs use U-SQL with Azure Data Lake to do predictive analytics.
- [SQL Server](#) These walkthroughs use SQL Server, SQL Server R Services, and SQL Server Python Services to do predictive analytics.
- [SQL Data Warehouse](#) These walkthroughs use SQL Data Warehouse to do predictive analytics.

Next steps

For a discussion of the key components that comprise the Team Data Science Process, see [Team Data Science Process overview](#).

For a discussion of the Team Data Science Process lifecycle that you can use to structure your data science projects, see [Team Data Science Process lifecycle](#). The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

For an overview of topics that walk you through the tasks that comprise the data science process in Azure, see [Data Science Process](#).

HDInsight Spark data science walkthroughs using PySpark and Scala on Azure

1/30/2019 • 2 minutes to read

These walkthroughs use PySpark and Scala on an Azure Spark cluster to do predictive analytics. They follow the steps outlined in the Team Data Science Process. For an overview of the Team Data Science Process, see [Data Science Process](#). For an overview of Spark on HDInsight, see [Introduction to Spark on HDInsight](#).

Additional data science walkthroughs that execute the Team Data Science Process are grouped by the **platform** that they use. See [Walkthroughs executing the Team Data Science Process](#) for an itemization of these examples.

Predict taxi tips using PySpark on Azure Spark

The [Use Spark on Azure HDInsight](#) walkthrough uses data from New York taxis to predict whether a tip is paid and the range of amounts expected to be paid. It uses the Team Data Science Process in a scenario using an [Azure HDInsight Spark cluster](#) to store, explore, and feature engineer data from the publicly available NYC taxi trip and fare dataset. This overview topic sets you up with an HDInsight Spark cluster and the Jupyter PySpark notebooks used in the rest of the walkthrough. These notebooks show you how to explore your data and then how to create and consume models. The advanced data exploration and modeling notebook shows how to include cross-validation, hyper-parameter sweeping, and model evaluation.

Data Exploration and modeling with Spark

Explore the dataset and create, score, and evaluate the machine learning models by working through the [Create binary classification and regression models for data with the Spark MLlib toolkit](#) topic.

Model consumption

To learn how to score the classification and regression models created in this topic, see [Score and evaluate Spark-built machine learning models](#).

Cross-validation and hyperparameter sweeping

See [Advanced data exploration and modeling with Spark](#) on how models can be trained using cross-validation and hyper-parameter sweeping.

Predict taxi tips using Scala on Azure Spark

The [Use Scala with Spark on Azure](#) walkthrough uses data from New York taxis to predict whether a tip is paid and the range of amounts expected to be paid. It shows how to use Scala for supervised machine learning tasks with the Spark machine learning library (MLlib) and SparkML packages on an Azure HDInsight Spark cluster. It walks you through the tasks that constitute the [Data Science Process](#): data ingestion and exploration, visualization, feature engineering, modeling, and model consumption. The models built include logistic and linear regression, random forests, and gradient boosted trees.

Next steps

For a discussion of the key components that comprise the Team Data Science Process, see [Team Data Science Process overview](#).

For a discussion of the Team Data Science Process lifecycle that you can use to structure your data science projects, see [Team Data Science Process lifecycle](#). The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

Data exploration and modeling with Spark

3/12/2019 • 31 minutes to read

This walkthrough uses HDInsight Spark to do data exploration and binary classification and regression modeling tasks on a sample of the NYC taxi trip and fare 2013 dataset. It walks you through the steps of the [Data Science Process](#), end-to-end, using an HDInsight Spark cluster for processing and Azure blobs to store the data and the models. The process explores and visualizes data brought in from an Azure Storage Blob and then prepares the data to build predictive models. These models are built using the Spark MLlib toolkit to do binary classification and regression modeling tasks.

- The **binary classification** task is to predict whether or not a tip is paid for the trip.
- The **regression** task is to predict the amount of the tip based on other tip features.

The models we use include logistic and linear regression, random forests, and gradient boosted trees:

- [Linear regression with SGD](#) is a linear regression model that uses a Stochastic Gradient Descent (SGD) method and for optimization and feature scaling to predict the tip amounts paid.
- [Logistic regression with LBFGS](#) or "logit" regression, is a regression model that can be used when the dependent variable is categorical to do data classification. LBFGS is a quasi-Newton optimization algorithm that approximates the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm using a limited amount of computer memory and that is widely used in machine learning.
- [Random forests](#) are ensembles of decision trees. They combine many decision trees to reduce the risk of overfitting. Random forests are used for regression and classification and can handle categorical features and can be extended to the multiclass classification setting. They do not require feature scaling and are able to capture non-linearities and feature interactions. Random forests are one of the most successful machine learning models for classification and regression.
- [Gradient boosted trees](#) (GBTs) are ensembles of decision trees. GBTs train decision trees iteratively to minimize a loss function. GBTs are used for regression and classification and can handle categorical features, do not require feature scaling, and are able to capture non-linearities and feature interactions. They can also be used in a multiclass-classification setting.

The modeling steps also contain code showing how to train, evaluate, and save each type of model. Python has been used to code the solution and to show the relevant plots.

NOTE

Although the Spark MLlib toolkit is designed to work on large datasets, a relatively small sample (~30 Mb using 170K rows, about 0.1% of the original NYC dataset) is used here for convenience. The exercise given here runs efficiently (in about 10 minutes) on an HDInsight cluster with 2 worker nodes. The same code, with minor modifications, can be used to process larger data-sets, with appropriate modifications for caching data in memory and changing the cluster size.

Prerequisites

You need an Azure account and a Spark 1.6 (or Spark 2.0) HDInsight cluster to complete this walkthrough. See the [Overview of Data Science using Spark on Azure HDInsight](#) for instructions on how to satisfy these requirements. That topic also contains a description of the NYC 2013 Taxi data used here and instructions on how to execute code from a Jupyter notebook on the Spark cluster.

Spark clusters and notebooks

Setup steps and code are provided in this walkthrough for using an HDInsight Spark 1.6. But Jupyter notebooks are provided for both HDInsight Spark 1.6 and Spark 2.0 clusters. A description of the notebooks and links to them are provided in the [Readme.md](#) for the GitHub repository containing them. Moreover, the code here and in the linked notebooks is generic and should work on any Spark cluster. If you are not using HDInsight Spark, the cluster setup and management steps may be slightly different from what is shown here. For convenience, here are the links to the Jupyter notebooks for Spark 1.6 (to be run in the pySpark kernel of the Jupyter Notebook server) and Spark 2.0 (to be run in the pySpark3 kernel of the Jupyter Notebook server):

Spark 1.6 notebooks

[pySpark-machine-learning-data-science-spark-data-exploration-modeling.ipynb](#): Provides information on how to perform data exploration, modeling, and scoring with several different algorithms.

Spark 2.0 notebooks

The regression and classification tasks that are implemented using a Spark 2.0 cluster are in separate notebooks and the classification notebook uses a different data set:

- [Spark2.0-pySpark3-machine-learning-data-science-spark-advanced-data-exploration-modeling.ipynb](#): This file provides information on how to perform data exploration, modeling, and scoring in Spark 2.0 clusters using the NYC Taxi trip and fare data-set described [here](#). This notebook may be a good starting point for quickly exploring the code we have provided for Spark 2.0. For a more detailed notebook analyzes the NYC Taxi data, see the next notebook in this list. See the notes following this list that compare these notebooks.
- [Spark2.0-pySpark3_NYC_Taxi_Tip_Regression.ipynb](#): This file shows how to perform data wrangling (Spark SQL and dataframe operations), exploration, modeling and scoring using the NYC Taxi trip and fare data-set described [here](#).
- [Spark2.0-pySpark3_Airline_Departure_Delay_Classification.ipynb](#): This file shows how to perform data wrangling (Spark SQL and dataframe operations), exploration, modeling and scoring using the well-known Airline On-time departure dataset from 2011 and 2012. We integrated the airline dataset with the airport weather data (e.g. windspeed, temperature, altitude etc.) prior to modeling, so these weather features can be included in the model.

NOTE

The airline dataset was added to the Spark 2.0 notebooks to better illustrate the use of classification algorithms. See the following links for information about airline on-time departure dataset and weather dataset:

- Airline on-time departure data: <https://www.transtats.bts.gov/ONTIME/>
- Airport weather data: <https://www.ncdc.noaa.gov/>

NOTE

The Spark 2.0 notebooks on the NYC taxi and airline flight delay data-sets can take 10 mins or more to run (depending on the size of your HDI cluster). The first notebook in the above list shows many aspects of the data exploration, visualization and ML model training in a notebook that takes less time to run with down-sampled NYC data set, in which the taxi and fare files have been pre-joined: [Spark2.0-pySpark3-machine-learning-data-science-spark-advanced-data-exploration-modeling.ipynb](#) This notebook takes a much shorter time to finish (2-3 mins) and may be a good starting point for quickly exploring the code we have provided for Spark 2.0.

WARNING

Billing for HDInsight clusters is prorated per minute, whether you use them or not. Be sure to delete your cluster after you finish using it. See [how to delete an HDInsight cluster](#).

NOTE

The descriptions below are related to using Spark 1.6. For Spark 2.0 versions, please use the notebooks described and linked above.

Setup: storage locations, libraries, and the preset Spark context

Spark is able to read and write to Azure Storage Blob (also known as WASB). So any of your existing data stored there can be processed using Spark and the results stored again in WASB.

To save models or files in WASB, the path needs to be specified properly. The default container attached to the Spark cluster can be referenced using a path beginning with: "wasb:///". Other locations are referenced by "wasb://".

Set directory paths for storage locations in WASB

The following code sample specifies the location of the data to be read and the path for the model storage directory to which the model output is saved:

```
# SET PATHS TO FILE LOCATIONS: DATA AND MODEL STORAGE

# LOCATION OF TRAINING DATA
taxi_train_file_loc =
"wasb://mllibwalkthroughs@cdsparksamples.blob.core.windows.net/Data/NYCTaxi/JoinedTaxiTripFare.Point1Pct.Tra
in.tsv";

# SET THE MODEL STORAGE DIRECTORY PATH
# NOTE THAT THE FINAL BACKSLASH IN THE PATH IS NEEDED.
modelDir = "wasb:///user/remoteuser/NYCTaxi/Models/"
```

Import libraries

Set up also requires importing necessary libraries. Set spark context and import necessary libraries with the following code:

```
# IMPORT LIBRARIES
import pyspark
from pyspark import SparkConf
from pyspark import SparkContext
from pyspark.sql import SQLContext
import matplotlib
import matplotlib.pyplot as plt
from pyspark.sql import Row
from pyspark.sql.functions import UserDefinedFunction
from pyspark.sql.types import *
import atexit
from numpy import array
import numpy as np
import datetime
```

Preset Spark context and PySpark magics

The PySpark kernels that are provided with Jupyter notebooks have a preset context. So you do not need to set the Spark or Hive contexts explicitly before you start working with the application you are developing. These contexts are available for you by default. These contexts are:

- sc - for Spark
- sqlContext - for Hive

The PySpark kernel provides some predefined "magics", which are special commands that you can call with %.

There are two such commands that are used in these code samples.

- **%%local** Specifies that the code in subsequent lines is to be executed locally. Code must be valid Python code.
- **%%sql -o** Executes a Hive query against the sqlContext. If the -o parameter is passed, the result of the query is persisted in the %%local Python context as a Pandas DataFrame.

For more information on the kernels for Jupyter notebooks and the predefined "magics" that they provide, see [Kernels available for Jupyter notebooks with HDInsight Spark Linux clusters on HDInsight](#).

Data ingestion from public blob

The first step in the data science process is to ingest the data to be analyzed from sources where it resides into your data exploration and modeling environment. The environment is Spark in this walkthrough. This section contains the code to complete a series of tasks:

- ingest the data sample to be modeled
- read in the input dataset (stored as a .tsv file)
- format and clean the data
- create and cache objects (RDDs or data-frames) in memory
- register it as a temp-table in SQL-context.

Here is the code for data ingestion.

```

# INGEST DATA

# RECORD START TIME
timestart = datetime.datetime.now()

# IMPORT FILE FROM PUBLIC BLOB
taxi_train_file = sc.textFile(taxi_train_file_loc)

# GET SCHEMA OF THE FILE FROM HEADER
schema_string = taxi_train_file.first()
fields = [StructField(field_name, StringType(), True) for field_name in schema_string.split('\t')]
fields[7].dataType = IntegerType() #Pickup hour
fields[8].dataType = IntegerType() # Pickup week
fields[9].dataType = IntegerType() # Weekday
fields[10].dataType = IntegerType() # Passenger count
fields[11].dataType = FloatType() # Trip time in secs
fields[12].dataType = FloatType() # Trip distance
fields[19].dataType = FloatType() # Fare amount
fields[20].dataType = FloatType() # Surcharge
fields[21].dataType = FloatType() # Mta_tax
fields[22].dataType = FloatType() # Tip amount
fields[23].dataType = FloatType() # Tolls amount
fields[24].dataType = FloatType() # Total amount
fields[25].dataType = IntegerType() # Tipped or not
fields[26].dataType = IntegerType() # Tip class
taxi_schema = StructType(fields)

# PARSE FIELDS AND CONVERT DATA TYPE FOR SOME FIELDS
taxi_header = taxi_train_file.filter(lambda l: "medallion" in l)
taxi_temp = taxi_train_file.subtract(taxi_header).map(lambda k: k.split("\t"))\
    .map(lambda p: (p[0],p[1],p[2],p[3],p[4],p[5],p[6],int(p[7]),int(p[8]),int(p[9]),int(p[10]),\
                    float(p[11]),float(p[12]),p[13],p[14],p[15],p[16],p[17],p[18],float(p[19]),\
                    float(p[20]),float(p[21]),float(p[22]),float(p[23]),float(p[24]),int(p[25]),int(p[26])))

# CREATE DATA FRAME
taxi_train_df = sqlContext.createDataFrame(taxi_temp, taxi_schema)

# CREATE A CLEANED DATA-FRAME BY DROPPING SOME UN-NECESSARY COLUMNS & FILTERING FOR UNDESIRED VALUES OR OUTLIERS
taxi_df_train_cleaned =
taxi_train_df.drop('medallion').drop('hack_license').drop('store_and_fwd_flag').drop('pickup_datetime')\
    .drop('dropoff_datetime').drop('pickup_longitude').drop('pickup_latitude').drop('dropoff_latitude')\
    .drop('dropoff_longitude').drop('tip_class').drop('total_amount').drop('tolls_amount').drop('mta_tax')\
    .drop('direct_distance').drop('surcharge')\
    .filter("passenger_count > 0 and passenger_count < 8 AND payment_type in ('CSH', 'CRD') AND tip_amount >= 0 AND tip_amount < 30 AND fare_amount >= 1 AND fare_amount < 150 AND trip_distance > 0 AND trip_distance < 100 AND trip_time_in_secs > 30 AND trip_time_in_secs < 7200" )

# CACHE DATA-FRAME IN MEMORY & MATERIALIZE DF IN MEMORY
taxi_df_train_cleaned.cache()
taxi_df_train_cleaned.count()

# REGISTER DATA-FRAME AS A TEMP-TABLE IN SQL-CONTEXT
taxi_df_train_cleaned.registerTempTable("taxi_train")

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 51.72 seconds

Data exploration & visualization

Once the data has been brought into Spark, the next step in the data science process is to gain deeper understanding of the data through exploration and visualization. In this section, we examine the taxi data using SQL queries and plot the target variables and prospective features for visual inspection. Specifically, we plot the frequency of passenger counts in taxi trips, the frequency of tip amounts, and how tips vary by payment amount and type.

Plot a histogram of passenger count frequencies in the sample of taxi trips

This code and subsequent snippets use SQL magic to query the sample and local magic to plot the data.

- **SQL magic (`%%sql`)** The HDInsight PySpark kernel supports easy inline HiveQL queries against the `sqlContext`. The `(-o VARIABLE_NAME)` argument persists the output of the SQL query as a Pandas DataFrame on the Jupyter server. This means it is available in the local mode.
- The `%%local` **magic** is used to run code locally on the Jupyter server, which is the headnode of the HDInsight cluster. Typically, you use `%%local` magic in conjunction with the `%%sql` magic with `-o` parameter. The `-o` parameter would persist the output of the SQL query locally and then `%%local` magic would trigger the next set of code snippet to run locally against the output of the SQL queries that is persisted locally

The output is automatically visualized after you run the code.

This query retrieves the trips by passenger count.

```
# PLOT FREQUENCY OF PASSENGER COUNTS IN TAXI TRIPS

# HIVEQL QUERY AGAINST THE sqlContext
%%sql -q -o sqlResults
SELECT passenger_count, COUNT(*) as trip_counts
FROM taxi_train
WHERE passenger_count > 0 and passenger_count < 7
GROUP BY passenger_count
```

This code creates a local data-frame from the query output and plots the data. The `%%local` magic creates a local data-frame, `sqlResults`, which can be used for plotting with matplotlib.

NOTE

This PySpark magic is used multiple times in this walkthrough. If the amount of data is large, you should sample to create a data-frame that can fit in local memory.

```
#CREATE LOCAL DATA-FRAME AND USE FOR MATPLOTLIB PLOTTING

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER
%%local

# USE THE JUPYTER AUTO-PLOTTING FEATURE TO CREATE INTERACTIVE FIGURES.
# CLICK ON THE TYPE OF PLOT TO BE GENERATED (E.G. LINE, AREA, BAR ETC.)
sqlResults
```

Here is the code to plot the trips by passenger counts

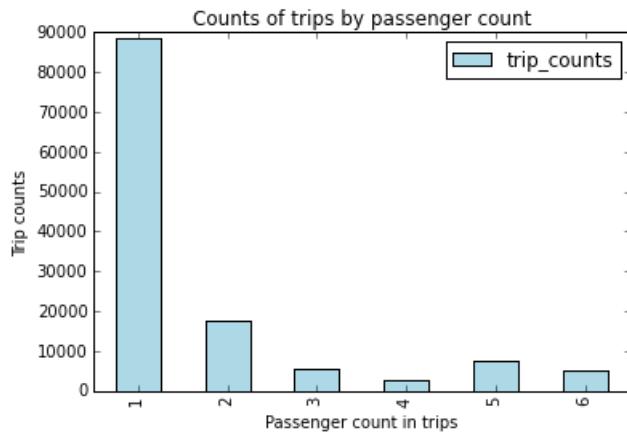
```

# PLOT PASSENGER NUMBER VS. TRIP COUNTS
%%local
import matplotlib.pyplot as plt
%matplotlib inline

x_labels = sqlResults['passenger_count'].values
fig = sqlResults[['trip_counts']].plot(kind='bar', facecolor='lightblue')
fig.set_xticklabels(x_labels)
fig.set_title('Counts of trips by passenger count')
fig.set_xlabel('Passenger count in trips')
fig.set_ylabel('Trip counts')
plt.show()

```

OUTPUT:



You can select among several different types of visualizations (Table, Pie, Line, Area, or Bar) by using the **Type** menu buttons in the notebook. The Bar plot is shown here.

Plot a histogram of tip amounts and how tip amount varies by passenger count and fare amounts.

Use a SQL query to sample data.

```

#PLOT HISTOGRAM OF TIP AMOUNTS AND VARIATION BY PASSENGER COUNT AND PAYMENT TYPE

# HIVEQL QUERY AGAINST THE sqlContext
%%sql -q -o sqlResults
SELECT fare_amount, passenger_count, tip_amount, tipped
FROM taxi_train
WHERE passenger_count > 0
AND passenger_count < 7
AND fare_amount > 0
AND fare_amount < 200
AND payment_type in ('CSH', 'CRD')
AND tip_amount > 0
AND tip_amount < 25

```

This code cell uses the SQL query to create three plots the data.

```

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER
%%local

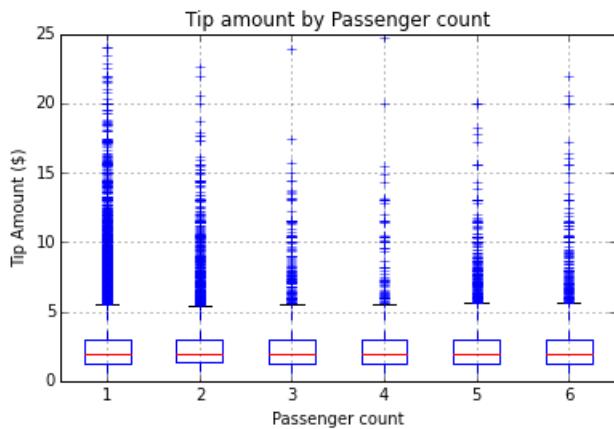
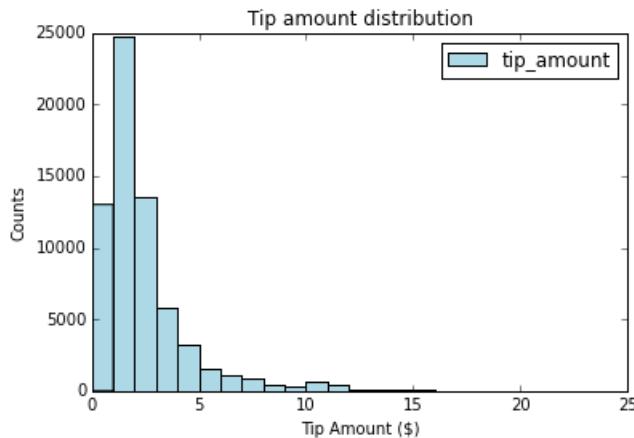
# HISTOGRAM OF TIP AMOUNTS AND PASSENGER COUNT
ax1 = sqlResults[['tip_amount']].plot(kind='hist', bins=25, facecolor='lightblue')
ax1.set_title('Tip amount distribution')
ax1.set_xlabel('Tip Amount ($)')
ax1.set_ylabel('Counts')
plt.suptitle('')
plt.show()

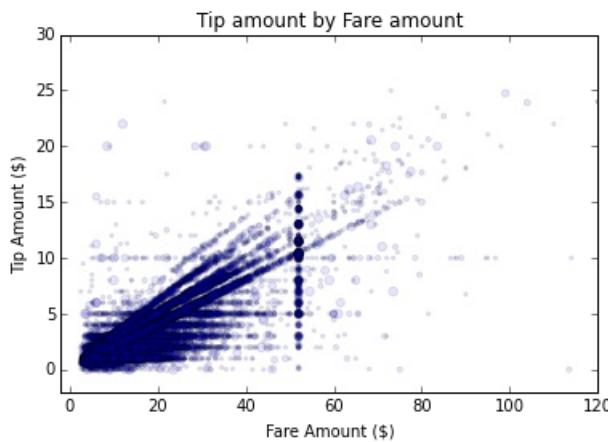
# TIP BY PASSENGER COUNT
ax2 = sqlResults.boxplot(column=['tip_amount'], by=['passenger_count'])
ax2.set_title('Tip amount by Passenger count')
ax2.set_xlabel('Passenger count')
ax2.set_ylabel('Tip Amount ($)')
plt.suptitle('')
plt.show()

# TIP AMOUNT BY FARE AMOUNT, POINTS ARE SCALED BY PASSENGER COUNT
ax = sqlResults.plot(kind='scatter', x= 'fare_amount', y = 'tip_amount', c='blue', alpha = 0.10, s=5*(sqlResults.passenger_count))
ax.set_title('Tip amount by Fare amount')
ax.set_xlabel('Fare Amount ($)')
ax.set_ylabel('Tip Amount ($)')
plt.axis([-2, 100, -2, 20])
plt.show()

```

OUTPUT:





Feature engineering, transformation and data preparation for modeling

This section describes and provides the code for procedures used to prepare data for use in ML modeling. It shows how to do the following tasks:

- Create a new feature by binning hours into traffic time buckets
- Index and encode categorical features
- Create labeled point objects for input into ML functions
- Create a random sub-sampling of the data and split it into training and testing sets
- Feature scaling
- Cache objects in memory

Create a new feature by binning hours into traffic time buckets

This code shows how to create a new feature by binning hours into traffic time buckets and then how to cache the resulting data frame in memory. Where Resilient Distributed Datasets (RDDs) and data-frames are used repeatedly, caching leads to improved execution times. Accordingly, we cache RDDs and data-frames at several stages in the walkthrough.

```
# CREATE FOUR BUCKETS FOR TRAFFIC TIMES
sqlStatement = """
    SELECT *,
    CASE
        WHEN (pickup_hour <= 6 OR pickup_hour >= 20) THEN "Night"
        WHEN (pickup_hour >= 7 AND pickup_hour <= 10) THEN "AMRush"
        WHEN (pickup_hour >= 11 AND pickup_hour <= 15) THEN "Afternoon"
        WHEN (pickup_hour >= 16 AND pickup_hour <= 19) THEN "PMRush"
    END as TrafficTimeBins
    FROM taxi_train
"""
taxi_df_train_with_newFeatures = sqlContext.sql(sqlStatement)

# CACHE DATA-FRAME IN MEMORY & MATERIALIZE DF IN MEMORY
# THE .COUNT() GOES THROUGH THE ENTIRE DATA-FRAME,
# MATERIALIZES IT IN MEMORY, AND GIVES THE COUNT OF ROWS.
taxi_df_train_with_newFeatures.cache()
taxi_df_train_with_newFeatures.count()
```

OUTPUT:

126050

Index and encode categorical features for input into modeling functions

This section shows how to index or encode categorical features for input into the modeling functions. The modeling and predict functions of MLLib require features with categorical input data to be indexed or encoded

prior to use. Depending on the model, you need to index or encode them in different ways:

- **Tree-based modeling** requires categories to be encoded as numerical values (for example, a feature with three categories may be encoded with 0, 1, 2). This is provided by MLlib's [StringIndexer](#) function. This function encodes a string column of labels to a column of label indices that are ordered by label frequencies. Although indexed with numerical values for input and data handling, the tree-based algorithms can be specified to treat them appropriately as categories.
- **Logistic and Linear Regression models** require one-hot encoding, where, for example, a feature with three categories can be expanded into three feature columns, with each containing 0 or 1 depending on the category of an observation. MLlib provides [OneHotEncoder](#) function to do one-hot encoding. This encoder maps a column of label indices to a column of binary vectors, with at most a single one-value. This encoding allows algorithms that expect numerical valued features, such as logistic regression, to be applied to categorical features.

Here is the code to index and encode categorical features:

```
# INDEX AND ENCODE CATEGORICAL FEATURES

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.ml.feature import OneHotEncoder, StringIndexer, VectorAssembler, VectorIndexer

# INDEX AND ENCODE VENDOR_ID
stringIndexer = StringIndexer(inputCol="vendor_id", outputCol="vendorIndex")
model = stringIndexer.fit(taxi_df_train_with_newFeatures) # Input data-frame is the cleaned one from above
indexed = model.transform(taxi_df_train_with_newFeatures)
encoder = OneHotEncoder(dropLast=False, inputCol="vendorIndex", outputCol="vendorVec")
encoded1 = encoder.transform(indexed)

# INDEX AND ENCODE RATE_CODE
stringIndexer = StringIndexer(inputCol="rate_code", outputCol="rateIndex")
model = stringIndexer.fit(encoded1)
indexed = model.transform(encoded1)
encoder = OneHotEncoder(dropLast=False, inputCol="rateIndex", outputCol="rateVec")
encoded2 = encoder.transform(indexed)

# INDEX AND ENCODE PAYMENT_TYPE
stringIndexer = StringIndexer(inputCol="payment_type", outputCol="paymentIndex")
model = stringIndexer.fit(encoded2)
indexed = model.transform(encoded2)
encoder = OneHotEncoder(dropLast=False, inputCol="paymentIndex", outputCol="paymentVec")
encoded3 = encoder.transform(indexed)

# INDEX AND TRAFFIC TIME BINS
stringIndexer = StringIndexer(inputCol="TrafficTimeBins", outputCol="TrafficTimeBinsIndex")
model = stringIndexer.fit(encoded3)
indexed = model.transform(encoded3)
encoder = OneHotEncoder(dropLast=False, inputCol="TrafficTimeBinsIndex", outputCol="TrafficTimeBinsVec")
encodedFinal = encoder.transform(indexed)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT:

Time taken to execute above cell: 1.28 seconds

Create labeled point objects for input into ML functions

This section contains code that shows how to index categorical text data as a labeled point data type and encode it so that it can be used to train and test MLlib logistic regression and other classification models. Labeled point objects are Resilient Distributed Datasets (RDD) formatted in a way that is needed as input data by most of ML algorithms in MLlib. A [labeled point](#) is a local vector, either dense or sparse, associated with a label/response.

This section contains code that shows how to index categorical text data as a [labeled point](#) data type and encode it so that it can be used to train and test MLlib logistic regression and other classification models. Labeled point objects are Resilient Distributed Datasets (RDD) consisting of a label (target/response variable) and feature vector. This format is needed as input by many ML algorithms in MLlib.

Here is the code to index and encode text features for binary classification.

```
# FUNCTIONS FOR BINARY CLASSIFICATION

# LOAD LIBRARIES
from pyspark.mllib.regression import LabeledPoint
from numpy import array

# INDEXING CATEGORICAL TEXT FEATURES FOR INPUT INTO TREE-BASED MODELS
def parseRowIndexingBinary(line):
    features = np.array([line.paymentIndex, line.vendorIndex, line.rateIndex, line.TrafficTimeBinsIndex,
                        line.pickup_hour, line.weekday, line.passenger_count, line.trip_time_in_secs,
                        line.trip_distance, line.fare_amount])
    labPt = LabeledPoint(line.tipped, features)
    return labPt

# ONE-HOT ENCODING OF CATEGORICAL TEXT FEATURES FOR INPUT INTO LOGISTIC REGRESSION MODELS
def parseRowOneHotBinary(line):
    features = np.concatenate((np.array([line.pickup_hour, line.weekday, line.passenger_count,
                                         line.trip_time_in_secs, line.trip_distance, line.fare_amount]),
                               line.vendorVec.toArray(), line.rateVec.toArray(),
                               line.paymentVec.toArray(), line.TrafficTimeBinsVec.toArray()), axis=0)
    labPt = LabeledPoint(line.tipped, features)
    return labPt
```

Here is the code to encode and index categorical text features for linear regression analysis.

```
# FUNCTIONS FOR REGRESSION WITH TIP AMOUNT AS TARGET VARIABLE

# ONE-HOT ENCODING OF CATEGORICAL TEXT FEATURES FOR INPUT INTO TREE-BASED MODELS
def parseRowIndexingRegression(line):
    features = np.array([line.paymentIndex, line.vendorIndex, line.rateIndex, line.TrafficTimeBinsIndex,
                        line.pickup_hour, line.weekday, line.passenger_count, line.trip_time_in_secs,
                        line.trip_distance, line.fare_amount])

    labPt = LabeledPoint(line.tip_amount, features)
    return labPt

# INDEXING CATEGORICAL TEXT FEATURES FOR INPUT INTO LINEAR REGRESSION MODELS
def parseRowOneHotRegression(line):
    features = np.concatenate((np.array([line.pickup_hour, line.weekday, line.passenger_count,
                                         line.trip_time_in_secs, line.trip_distance, line.fare_amount]),
                               line.vendorVec.toArray(), line.rateVec.toArray(),
                               line.paymentVec.toArray(), line.TrafficTimeBinsVec.toArray()), axis=0)
    labPt = LabeledPoint(line.tip_amount, features)
    return labPt
```

Create a random sub-sampling of the data and split it into training and testing sets

This code creates a random sampling of the data (25% is used here). Although it is not required for this example due to the size of the dataset, we demonstrate how you can sample here so you know how to use it for your own problem when needed. When samples are large, this can save significant time while training models. Next we split

the sample into a training part (75% here) and a testing part (25% here) to use in classification and regression modeling.

```
# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.sql.functions import rand

# SPECIFY SAMPLING AND SPLITTING FRACTIONS
samplingFraction = 0.25;
trainingFraction = 0.75; testingFraction = (1-trainingFraction);
seed = 1234;
encodedFinalSampled = encodedFinal.sample(False, samplingFraction, seed=seed)

# SPLIT SAMPLED DATA-FRAME INTO TRAIN/TEST
# INCLUDE RAND COLUMN FOR CREATING CROSS-VALIDATION FOLDS (FOR USE LATER IN AN ADVANCED TOPIC)
dfTmpRand = encodedFinalSampled.select("*", rand(0).alias("rand"));
trainData, testData = dfTmpRand.randomSplit([trainingFraction, testingFraction], seed=seed);

# FOR BINARY CLASSIFICATION TRAINING AND TESTING
indexedTRAINbinary = trainData.map(parseRowIndexingBinary)
indexedTESTbinary = testData.map(parseRowIndexingBinary)
oneHotTRAINbinary = trainData.map(parseRowIndexingBinary)
oneHotTESTbinary = testData.map(parseRowIndexingBinary)

# FOR REGRESSION TRAINING AND TESTING
indexedTRAINreg = trainData.map(parseRowIndexingRegression)
indexedTESTreg = testData.map(parseRowIndexingRegression)
oneHotTRAINreg = trainData.map(parseRowIndexingRegression)
oneHotTESTreg = testData.map(parseRowIndexingRegression)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT:

Time taken to execute above cell: 0.24 seconds

Feature scaling

Feature scaling, also known as data normalization, insures that features with widely disbursed values are not given excessive weigh in the objective function. The code for feature scaling uses the [StandardScaler](#) to scale the features to unit variance. It is provided by MLlib for use in linear regression with Stochastic Gradient Descent (SGD), a popular algorithm for training a wide range of other machine learning models such as regularized regressions or support vector machines (SVM).

NOTE

We have found the LinearRegressionWithSGD algorithm to be sensitive to feature scaling.

Here is the code to scale variables for use with the regularized linear SGD algorithm.

```

# FEATURE SCALING

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.regression import LabeledPoint
from pyspark.mllib.linalg import Vectors
from pyspark.mllib.feature import StandardScaler, StandardScalerModel
from pyspark.mllib.util import MLUtils

# SCALE VARIABLES FOR REGULARIZED LINEAR SGD ALGORITHM
label = oneHotTRAINreg.map(lambda x: x.label)
features = oneHotTRAINreg.map(lambda x: x.features)
scaler = StandardScaler(withMean=False, withStd=True).fit(features)
dataTMP = label.zip(scaler.transform(features.map(lambda x: Vectors.dense(x.toArray()))))
oneHotTRAINregScaled = dataTMP.map(lambda x: LabeledPoint(x[0], x[1]))

label = oneHotTESTreg.map(lambda x: x.label)
features = oneHotTESTreg.map(lambda x: x.features)
scaler = StandardScaler(withMean=False, withStd=True).fit(features)
dataTMP = label.zip(scaler.transform(features.map(lambda x: Vectors.dense(x.toArray()))))
oneHotTESTregScaled = dataTMP.map(lambda x: LabeledPoint(x[0], x[1]))

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 13.17 seconds

Cache objects in memory

The time taken for training and testing of ML algorithms can be reduced by caching the input data frame objects used for classification, regression, and scaled features.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# FOR BINARY CLASSIFICATION TRAINING AND TESTING
indexedTRAINbinary.cache()
indexedTESTbinary.cache()
oneHotTRAINbinary.cache()
oneHotTESTbinary.cache()

# FOR REGRESSION TRAINING AND TESTING
indexedTRAINreg.cache()
indexedTESTreg.cache()
oneHotTRAINreg.cache()
oneHotTESTreg.cache()

# SCALED FEATURES
oneHotTRAINregScaled.cache()
oneHotTESTregScaled.cache()

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 0.15 seconds

Predict whether or not a tip is paid with binary classification models

This section shows how use three models for the binary classification task of predicting whether or not a tip is paid for a taxi trip. The models presented are:

- Regularized logistic regression
- Random forest model
- Gradient Boosting Trees

Each model building code section is split into steps:

1. **Model training** data with one parameter set
2. **Model evaluation** on a test data set with metrics
3. **Saving model** in blob for future consumption

Classification using logistic regression

The code in this section shows how to train, evaluate, and save a logistic regression model with [LBFGS](#) that predicts whether or not a tip is paid for a trip in the NYC taxi trip and fare dataset.

Train the logistic regression model using CV and hyperparameter sweeping

```
# LOGISTIC REGRESSION CLASSIFICATION WITH CV AND HYPERPARAMETER SWEEPING

# GET ACCURACY FOR HYPERPARAMETERS BASED ON CROSS-VALIDATION IN TRAINING DATA-SET

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD LIBRARIES
from pyspark.mllib.classification import LogisticRegressionWithLBFGS
from sklearn.metrics import roc_curve, auc
from pyspark.mllib.evaluation import BinaryClassificationMetrics
from pyspark.mllib.evaluation import MulticlassMetrics

# CREATE MODEL WITH ONE SET OF PARAMETERS
logitModel = LogisticRegressionWithLBFGS.train(oneHotTRAINbinary, iterations=20, initialWeights=None,
                                                regParam=0.01, regType='l2', intercept=True, corrections=10,
                                                tolerance=0.0001, validateData=True, numClasses=2)

# PRINT COEFFICIENTS AND INTERCEPT OF THE MODEL
# NOTE: There are 20 coefficient terms for the 10 features,
#       and the different categories for features: vendorVec (2), rateVec, paymentVec (6), TrafficTimeBinsVec (4)
print("Coefficients: " + str(logitModel.weights))
print("Intercept: " + str(logitModel.intercept))

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT:

Coefficients: [0.0082065285375, -0.0223675576104, -0.0183812028036, -3.48124578069e-05, -0.00247646947233, -0.00165897881503, 0.0675394837328, -0.111823113101, -0.324609912762, -0.204549780032, -1.36499216354, 0.591088507921, -0.664263411392, -1.00439726852, 3.46567827545, -3.51025855172, -0.0471341112232, -0.043521833294, 0.000243375810385, 0.054518719222]

Intercept: -0.0111216486893

Time taken to execute above cell: 14.43 seconds

Evaluate the binary classification model with standard metrics

```
#EVALUATE LOGISTIC REGRESSION MODEL WITH LBFGS

# RECORD START TIME
timestart = datetime.datetime.now()

# PREDICT ON TEST DATA WITH MODEL
predictionAndLabels = oneHotTESTbinary.map(lambda lp: (float(logitModel.predict(lp.features)), lp.label))

# INstantiate METRICS OBJECT
metrics = BinaryClassificationMetrics(predictionAndLabels)

# AREA UNDER PRECISION-RECALL CURVE
print("Area under PR = %s" % metrics.areaUnderPR)

# AREA UNDER ROC CURVE
print("Area under ROC = %s" % metrics.areaUnderROC)
metrics = MulticlassMetrics(predictionAndLabels)

# OVERALL STATISTICS
precision = metrics.precision()
recall = metrics.recall()
f1Score = metrics.fMeasure()
print("Summary Stats")
print("Precision = %s" % precision)
print("Recall = %s" % recall)
print("F1 Score = %s" % f1Score)

## SAVE MODEL WITH DATE-STAMP
datestamp = unicode(datetime.datetime.now()).replace(' ', '_').replace(':', '_');
logisticregressionfilename = "LogisticRegressionWithLBFGS_" + datestamp;
dirfilename = modelDir + logisticregressionfilename;
logitModel.save(sc, dirfilename);

# OUTPUT PROBABILITIES AND REGISTER TEMP TABLE
logitModel.clearThreshold(); # This clears threshold for classification (0.5) and outputs probabilities
predictionAndLabelsDF = predictionAndLabels.toDF()
predictionAndLabelsDF.registerTempTable("tmp_results");

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT:

Area under PR = 0.985297691373

Area under ROC = 0.983714670256

Summary Stats

Precision = 0.984304060189

Recall = 0.984304060189

F1 Score = 0.984304060189

Time taken to execute above cell: 57.61 seconds

Plot the ROC curve.

The `predictionAndLabelsDF` is registered as a table, `tmp_results`, in the previous cell. `tmp_results` can be used to do queries and output results into the `sqlResults` data-frame for plotting. Here is the code.

```
# QUERY RESULTS
%%sql -q -o sqlResults
SELECT * from tmp_results
```

Here is the code to make predictions and plot the ROC-curve.

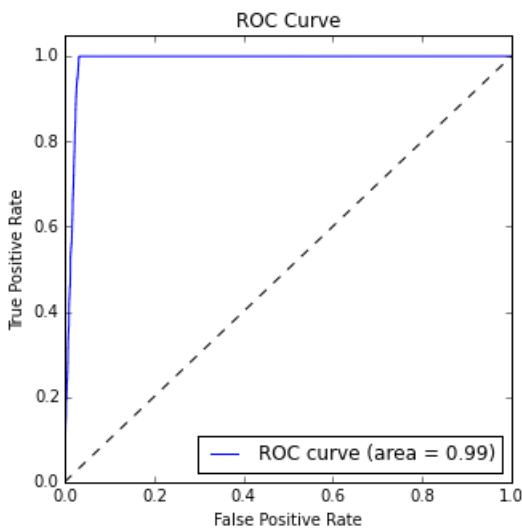
```
# MAKE PREDICTIONS AND PLOT ROC-CURVE

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
%matplotlib inline
from sklearn.metrics import roc_curve,auc

# MAKE PREDICTIONS
predictions_pddf = test_predictions.rename(columns={'_1': 'probability', '_2': 'label'})
prob = predictions_pddf["probability"]
fpr, tpr, thresholds = roc_curve(predictions_pddf['label'], prob, pos_label=1);
roc_auc = auc(fpr, tpr)

# PLOT ROC CURVE
plt.figure(figsize=(5,5))
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()
```

OUTPUT:



Random forest classification

The code in this section shows how to train, evaluate, and save a random forest model that predicts whether or not a tip is paid for a trip in the NYC taxi trip and fare dataset.

```

#PREDICT WHETHER A TIP IS PAID OR NOT USING RANDOM FOREST

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import RandomForest, RandomForestModel
from pyspark.mllib.util import MLUtils
from pyspark.mllib.evaluation import BinaryClassificationMetrics
from pyspark.mllib.evaluation import MulticlassMetrics

# SPECIFY NUMBER OF CATEGORIES FOR CATEGORICAL FEATURES. FEATURE #0 HAS 2 CATEGORIES, FEATURE #2 HAS 2 CATEGORIES, AND SO ON
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}

# TRAIN RANDOMFOREST MODEL
rfModel = RandomForest.trainClassifier(indexedTRAINbinary, numClasses=2,
                                         categoricalFeaturesInfo=categoricalFeaturesInfo,
                                         numTrees=25, featureSubsetStrategy="auto",
                                         impurity='gini', maxDepth=5, maxBins=32)

## UN-COMMENT IF YOU WANT TO PRINT TREES
#print('Learned classification forest model:')
#print(rfModel.toDebugString())

# PREDICT ON TEST DATA AND EVALUATE
predictions = rfModel.predict(indexedTESTbinary.map(lambda x: x.features))
predictionAndLabels = indexedTESTbinary.map(lambda lp: lp.label).zip(predictions)

# AREA UNDER ROC CURVE
metrics = BinaryClassificationMetrics(predictionAndLabels)
print("Area under ROC = %s" % metrics.areaUnderROC)

# PERSIST MODEL IN BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ', '').replace(':', '_');
rfclassificationfilename = "RandomForestClassification_" + datestamp;
dirfilename = modelDir + rfclassificationfilename;

rfModel.save(sc, dirfilename);

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Area under ROC = 0.985297691373

Time taken to execute above cell: 31.09 seconds

Gradient boosting trees classification

The code in this section shows how to train, evaluate, and save a gradient boosting trees model that predicts whether or not a tip is paid for a trip in the NYC taxi trip and fare dataset.

```

#PREDICT WHETHER A TIP IS PAID OR NOT USING GRADIENT BOOSTING TREES

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import GradientBoostedTrees, GradientBoostedTreesModel

# SPECIFY NUMBER OF CATEGORIES FOR CATEGORICAL FEATURES. FEATURE #0 HAS 2 CATEGORIES, FEATURE #2 HAS 2 CATEGORIES, AND SO ON
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}

gbtModel = GradientBoostedTrees.trainClassifier(indexedTRAINbinary,
categoricalFeaturesInfo=categoricalFeaturesInfo, numIterations=5)
## UNCOMMENT IF YOU WANT TO PRINT TREE DETAILS
#print('Learned classification GBT model:')
#print(bgtModel.toDebugString())

# PREDICT ON TEST DATA AND EVALUATE
predictions = gbtModel.predict(indexedTESTbinary.map(lambda x: x.features))
predictionAndLabels = indexedTESTbinary.map(lambda lp: lp.label).zip(predictions)

# AREA UNDER ROC CURVE
metrics = BinaryClassificationMetrics(predictionAndLabels)
print("Area under ROC = %s" % metrics.areaUnderROC)

# PERSIST MODEL IN A BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
btclassificationfilename = "GradientBoostingTreeClassification_" + datestamp;
dirfilename = modelDir + btclassificationfilename;

gbtModel.save(sc, dirfilename)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Area under ROC = 0.985297691373

Time taken to execute above cell: 19.76 seconds

Predict tip amounts for taxi trips with regression models

This section shows how use three models for the regression task of predicting the amount of the tip paid for a taxi trip based on other tip features. The models presented are:

- Regularized linear regression
- Random forest
- Gradient Boosting Trees

These models were described in the introduction. Each model building code section is split into steps:

1. **Model training** data with one parameter set
2. **Model evaluation** on a test data set with metrics
3. **Saving model** in blob for future consumption

Linear regression with SGD

The code in this section shows how to use scaled features to train a linear regression that uses stochastic gradient descent (SGD) for optimization, and how to score, evaluate, and save the model in Azure Blob Storage (WASB).

TIP

In our experience, there can be issues with the convergence of LinearRegressionWithSGD models, and parameters need to be changed/optimized carefully for obtaining a valid model. Scaling of variables significantly helps with convergence.

```
#PREDICT TIP AMOUNTS USING LINEAR REGRESSION WITH SGD

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD LIBRARIES
from pyspark.mllib.regression import LabeledPoint, LinearRegressionWithSGD, LinearRegressionModel
from pyspark.mllib.evaluation import RegressionMetrics
from scipy import stats

# USE SCALED FEATURES TO TRAIN MODEL
linearModel = LinearRegressionWithSGD.train(oneHotTRAINregScaled, iterations=100, step = 0.1, regType='l2',
regParam=0.1, intercept = True)

# PRINT COEFFICIENTS AND INTERCEPT OF THE MODEL
# NOTE: There are 20 coefficient terms for the 10 features,
#       and the different categories for features: vendorVec (2), rateVec, paymentVec (6), TrafficTimeBinsVec (4)
print("Coefficients: " + str(linearModel.weights))
print("Intercept: " + str(linearModel.intercept))

# SCORE ON SCALED TEST DATA-SET & EVALUATE
predictionAndLabels = oneHotTESTregScaled.map(lambda lp: (float(linearModel.predict(lp.features)), lp.label))
testMetrics = RegressionMetrics(predictionAndLabels)

# PRINT TEST METRICS
print("RMSE = %s" % testMetrics.rootMeanSquaredError)
print("R-sqr = %s" % testMetrics.r2)

# SAVE MODEL WITH DATE-STAMP IN THE DEFAULT BLOB FOR THE CLUSTER
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
linearregressionfilename = "LinearRegressionWithSGD_" + datestamp;
dirfilename = modelDir + linearregressionfilename;

linearModel.save(sc, dirfilename)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT:

Coefficients: [0.00457675809917, -0.0226314167349, -0.0191910355236, 0.246793409578, 0.312047890459, 0.359634405999, 0.00928692253981, -0.000987181489428, -0.0888306617845, 0.0569376211553, 0.115519551711, 0.149250164995, -0.00990211159703, -0.00637410344522, 0.545083566179, -0.536756072402, 0.0105762393099, -0.0130117577055, 0.0129304737772, -0.00171065945959]

Intercept: 0.853872718283

RMSE = 1.24190115863

R-sqr = 0.608017146081

Time taken to execute above cell: 58.42 seconds

Random Forest regression

The code in this section shows how to train, evaluate, and save a random forest regression that predicts tip amount for the NYC taxi trip data.

```
#PREDICT TIP AMOUNTS USING RANDOM FOREST

# RECORD START TIME
timestart= datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import RandomForest, RandomForestModel
from pyspark.mllib.util import MLUtils
from pyspark.mllib.evaluation import RegressionMetrics

## TRAIN MODEL
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}
rfModel = RandomForest.trainRegressor(indexedTRAINreg, categoricalFeaturesInfo=categoricalFeaturesInfo,
                                      numTrees=25, featureSubsetStrategy="auto",
                                      impurity='variance', maxDepth=10, maxBins=32)
## UN-COMMENT IF YOU WANT TO PRING TREES
#print('Learned classification forest model:')
#print(rfModel.toDebugString())

## PREDICT AND EVALUATE ON TEST DATA-SET
predictions = rfModel.predict(indexedTESTreg.map(lambda x: x.features))
predictionAndLabels = oneHotTESTreg.map(lambda lp: lp.label).zip(predictions)

# TEST METRICS
testMetrics = RegressionMetrics(predictionAndLabels)
print("RMSE = %s" % testMetrics.rootMeanSquaredError)
print("R-sqr = %s" % testMetrics.r2)

# SAVE MODEL IN BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
rfregressionfilename = "RandomForestRegression_" + datestamp;
dirfilename = modelDir + rfregressionfilename;

rfModel.save(sc, dirfilename);

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT:

RMSE = 0.891209218139

R-sqr = 0.759661334921

Time taken to execute above cell: 49.21 seconds

Gradient boosting trees regression

The code in this section shows how to train, evaluate, and save a gradient boosting trees model that predicts tip amount for the NYC taxi trip data.

Train and evaluate

```

#PREDICT TIP AMOUNTS USING GRADIENT BOOSTING TREES

# RECORD START TIME
timestart= datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import GradientBoostedTrees, GradientBoostedTreesModel
from pyspark.mllib.util import MLUtils

## TRAIN MODEL
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}
gbtModel = GradientBoostedTrees.trainRegressor(indexedTRAINreg,
categoricalFeaturesInfo=categoricalFeaturesInfo,
numIterations=10, maxBins=32, maxDepth = 4, learningRate=0.1)

## EVALUATE A TEST DATA-SET
predictions = gbtModel.predict(indexedTESTreg.map(lambda x: x.features))
predictionAndLabels = indexedTESTreg.map(lambda lp: lp.label).zip(predictions)

# TEST METRICS
testMetrics = RegressionMetrics(predictionAndLabels)
print("RMSE = %s" % testMetrics.rootMeanSquaredError)
print("R-sqr = %s" % testMetrics.r2)

# SAVE MODEL IN BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
btregressionfilename = "GradientBoostingTreeRegression_" + datestamp;
dirfilename = modelDir + btregressionfilename;
gbtModel.save(sc, dirfilename)

# CONVERT RESULTS TO DF AND REGISTER TEMP TABLE
test_predictions = sqlContext.createDataFrame(predictionAndLabels)
test_predictions.registerTempTable("tmp_results");

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

RMSE = 0.908473148639

R-sqr = 0.753835096681

Time taken to execute above cell: 34.52 seconds

Plot

tmp_results is registered as a Hive table in the previous cell. Results from the table are output into the *sqlResults* data-frame for plotting. Here is the code

```

# PLOT SCATTER-PLOT BETWEEN ACTUAL AND PREDICTED TIP VALUES

# SELECT RESULTS
%%sql -q -o sqlResults
SELECT * from tmp_results

```

Here is the code to plot the data using the Jupyter server.

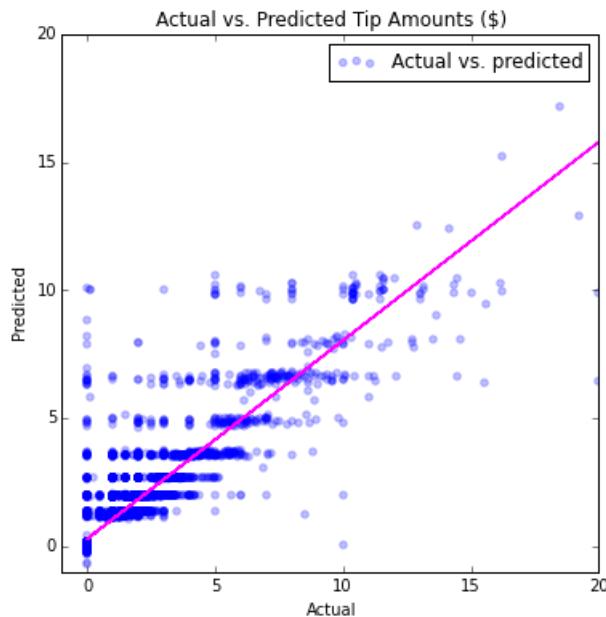
```

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
%matplotlib inline
import numpy as np

# PLOT
ax = test_predictions_pddf.plot(kind='scatter', figsize = (6,6), x='_1', y='_2', color='blue', alpha = 0.25,
label='Actual vs. predicted');
fit = np.polyfit(test_predictions_pddf['_1'], test_predictions_pddf['_2'], deg=1)
ax.set_title('Actual vs. Predicted Tip Amounts ($)')
ax.set_xlabel("Actual")
ax.set_ylabel("Predicted")
ax.plot(test_predictions_pddf['_1'], fit[0] * test_predictions_pddf['_1'] + fit[1], color='magenta')
plt.axis([-1, 20, -1, 20])
plt.show(ax)

```

OUTPUT:



Clean up objects from memory

Use `unpersist()` to delete objects cached in memory.

```

# REMOVE ORIGINAL DFS
taxi_df_train_cleaned.unpersist()
taxi_df_train_with_newFeatures.unpersist()

# FOR BINARY CLASSIFICATION TRAINING AND TESTING
indexedTRAINbinary.unpersist()
indexedTESTbinary.unpersist()
oneHotTRAINbinary.unpersist()
oneHotTESTbinary.unpersist()

# FOR REGRESSION TRAINING AND TESTING
indexedTRAINreg.unpersist()
indexedTESTreg.unpersist()
oneHotTRAINreg.unpersist()
oneHotTESTreg.unpersist()

# SCALED FEATURES
oneHotTRAINregScaled.unpersist()
oneHotTESTregScaled.unpersist()

```

Record storage locations of the models for consumption and scoring

To consume and score an independent dataset described in the [Score and evaluate Spark-built machine learning models](#) topic, you need to copy and paste these file names containing the saved models created here into the Consumption Jupyter notebook. Here is the code to print out the paths to model files you need there.

```
# MODEL FILE LOCATIONS FOR CONSUMPTION
print "logisticRegFileLoc = modelDir + \"" + logisticregressionfilename + """;
print "linearRegFileLoc = modelDir + \"" + linearregressionfilename + """;
print "randomForestClassificationFileLoc = modelDir + \"" + rfclassificationfilename + """;
print "randomForestRegFileLoc = modelDir + \"" + rfregressionfilename + """;
print "BoostedTreeClassificationFileLoc = modelDir + \"" + btclassificationfilename + """;
print "BoostedTreeRegressionFileLoc = modelDir + \"" + btregressionfilename + """;
```

OUTPUT

```
logisticRegFileLoc = modelDir + "LogisticRegressionWithLBFGS_2016-05-0317_03_23.516568"
linearRegFileLoc = modelDir + "LinearRegressionWithSGD_2016-05-0317_05_21.577773"
randomForestClassificationFileLoc = modelDir + "RandomForestClassification_2016-05-0317_04_11.950206"
randomForestRegFileLoc = modelDir + "RandomForestRegression_2016-05-0317_06_08.723736"
BoostedTreeClassificationFileLoc = modelDir + "GradientBoostingTreeClassification_2016-05-0317_04_36.346583"
BoostedTreeRegressionFileLoc = modelDir + "GradientBoostingTreeRegression_2016-05-0317_06_51.737282"
```

What's next?

Now that you have created regression and classification models with the Spark MLlib, you are ready to learn how to score and evaluate these models. The advanced data exploration and modeling notebook dives deeper into including cross-validation, hyper-parameter sweeping, and model evaluation.

Model consumption: To learn how to score and evaluate the classification and regression models created in this topic, see [Score and evaluate Spark-built machine learning models](#).

Cross-validation and hyperparameter sweeping: See [Advanced data exploration and modeling with Spark](#) on how models can be trained using cross-validation and hyper-parameter sweeping

Advanced data exploration and modeling with Spark

3/12/2019 • 37 minutes to read

This walkthrough uses HDInsight Spark to do data exploration and train binary classification and regression models using cross-validation and hyperparameter optimization on a sample of the NYC taxi trip and fare 2013 dataset. It walks you through the steps of the [Data Science Process](#), end-to-end, using an HDInsight Spark cluster for processing and Azure blobs to store the data and the models. The process explores and visualizes data brought in from an Azure Storage Blob and then prepares the data to build predictive models. Python has been used to code the solution and to show the relevant plots. These models are built using the Spark MLlib toolkit to do binary classification and regression modeling tasks.

- The **binary classification** task is to predict whether or not a tip is paid for the trip.
- The **regression** task is to predict the amount of the tip based on other tip features.

The modeling steps also contain code showing how to train, evaluate, and save each type of model. The topic covers some of the same ground as the [Data exploration and modeling with Spark](#) topic. But it is more "advanced" in that it also uses cross-validation with hyperparameter sweeping to train optimally accurate classification and regression models.

Cross-validation (CV) is a technique that assesses how well a model trained on a known set of data generalizes to predicting the features of datasets on which it has not been trained. A common implementation used here is to divide a dataset into K folds and then train the model in a round-robin fashion on all but one of the folds. The ability of the model to predict accurately when tested against the independent dataset in this fold not used to train the model is assessed.

Hyperparameter optimization is the problem of choosing a set of hyperparameters for a learning algorithm, usually with the goal of optimizing a measure of the algorithm's performance on an independent data set.

Hyperparameters are values that must be specified outside of the model training procedure. Assumptions about these values can impact the flexibility and accuracy of the models. Decision trees have hyperparameters, for example, such as the desired depth and number of leaves in the tree. Support Vector Machines (SVMs) require setting a misclassification penalty term.

A common way to perform hyperparameter optimization used here is a grid search, or a **parameter sweep**. This consists of performing an exhaustive search through the values a specified subset of the hyperparameter space for a learning algorithm. Cross validation can supply a performance metric to sort out the optimal results produced by the grid search algorithm. CV used with hyperparameter sweeping helps limit problems like overfitting a model to training data so that the model retains the capacity to apply to the general set of data from which the training data was extracted.

The models we use include logistic and linear regression, random forests, and gradient boosted trees:

- [Linear regression with SGD](#) is a linear regression model that uses a Stochastic Gradient Descent (SGD) method and for optimization and feature scaling to predict the tip amounts paid.
- [Logistic regression with LBFGS](#) or "logit" regression, is a regression model that can be used when the dependent variable is categorical to do data classification. LBFGS is a quasi-Newton optimization algorithm that approximates the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm using a limited amount of computer memory and that is widely used in machine learning.
- [Random forests](#) are ensembles of decision trees. They combine many decision trees to reduce the risk of overfitting. Random forests are used for regression and classification and can handle categorical features and can be extended to the multiclass classification setting. They do not require feature scaling and are able to capture non-linearities and feature interactions. Random forests are one of the most successful machine

learning models for classification and regression.

- **Gradient boosted trees** (GBTs) are ensembles of decision trees. GBTs train decision trees iteratively to minimize a loss function. GBTs are used for regression and classification and can handle categorical features, do not require feature scaling, and are able to capture non-linearities and feature interactions. They can also be used in a multiclass-classification setting.

Modeling examples using CV and Hyperparameter sweep are shown for the binary classification problem. Simpler examples (without parameter sweeps) are presented in the main topic for regression tasks. But in the appendix, validation using elastic net for linear regression and CV with parameter sweep using for random forest regression are also presented. The **elastic net** is a regularized regression method for fitting linear regression models that linearly combines the L1 and L2 metrics as penalties of the **lasso** and **ridge** methods.

NOTE

Although the Spark MLlib toolkit is designed to work on large datasets, a relatively small sample (~30 Mb using 170K rows, about 0.1% of the original NYC dataset) is used here for convenience. The exercise given here runs efficiently (in about 10 minutes) on an HDInsight cluster with 2 worker nodes. The same code, with minor modifications, can be used to process larger data-sets, with appropriate modifications for caching data in memory and changing the cluster size.

Setup: Spark clusters and notebooks

Setup steps and code are provided in this walkthrough for using an HDInsight Spark 1.6. But Jupyter notebooks are provided for both HDInsight Spark 1.6 and Spark 2.0 clusters. A description of the notebooks and links to them are provided in the [Readme.md](#) for the GitHub repository containing them. Moreover, the code here and in the linked notebooks is generic and should work on any Spark cluster. If you are not using HDInsight Spark, the cluster setup and management steps may be slightly different from what is shown here. For convenience, here are the links to the Jupyter notebooks for Spark 1.6 and 2.0 to be run in the pyspark kernel of the Jupyter Notebook server:

Spark 1.6 notebooks

[pySpark-machine-learning-data-science-spark-advanced-data-exploration-modeling.ipynb](#): Includes topics in notebook #1, and model development using hyperparameter tuning and cross-validation.

Spark 2.0 notebooks

[Spark2.0-pySpark3-machine-learning-data-science-spark-advanced-data-exploration-modeling.ipynb](#): This file provides information on how to perform data exploration, modeling, and scoring in Spark 2.0 clusters.

WARNING

Billing for HDInsight clusters is prorated per minute, whether you use them or not. Be sure to delete your cluster after you finish using it. See [how to delete an HDInsight cluster](#).

Setup: storage locations, libraries, and the preset Spark context

Spark is able to read and write to Azure Storage Blob (also known as WASB). So any of your existing data stored there can be processed using Spark and the results stored again in WASB.

To save models or files in WASB, the path needs to be specified properly. The default container attached to the Spark cluster can be referenced using a path beginning with: "wasb://". Other locations are referenced by "wasb://".

Set directory paths for storage locations in WASB

The following code sample specifies the location of the data to be read and the path for the model storage

directory to which the model output is saved:

```
# SET PATHS TO FILE LOCATIONS: DATA AND MODEL STORAGE

# LOCATION OF TRAINING DATA
taxi_train_file_loc =
"wasb://mllibwalkthroughs@cdsparksamples.blob.core.windows.net/Data/NYCTaxi/JoinedTaxiTripFare.Point1Pct.Tra
in.tsv";

# SET THE MODEL STORAGE DIRECTORY PATH
# NOTE THAT THE FINAL BACKSLASH IN THE PATH IS NEEDED.
modelDir = "wasb:///user/remoteuser/NYCTaxi/Models/";

# PRINT START TIME
import datetime
datetime.datetime.now()
```

OUTPUT

```
datetime.datetime(2016, 4, 18, 17, 36, 27, 832799)
```

Import libraries

Import necessary libraries with the following code:

```
# LOAD PYSPARK LIBRARIES
import pyspark
from pyspark import SparkConf
from pyspark import SparkContext
from pyspark.sql import SQLContext
import matplotlib
import matplotlib.pyplot as plt
from pyspark.sql import Row
from pyspark.sql.functions import UserDefinedFunction
from pyspark.sql.types import *
import atexit
from numpy import array
import numpy as np
import datetime
```

Preset Spark context and PySpark magics

The PySpark kernels that are provided with Jupyter notebooks have a preset context. So you do not need to set the Spark or Hive contexts explicitly before you start working with the application you are developing. These contexts are available for you by default. These contexts are:

- sc - for Spark
- sqlContext - for Hive

The PySpark kernel provides some predefined "magics", which are special commands that you can call with %. There are two such commands that are used in these code samples.

- **%%local** Specifies that the code in subsequent lines is to be executed locally. Code must be valid Python code.
- **%%sql -o** Executes a Hive query against the sqlContext. If the -o parameter is passed, the result of the query is persisted in the %%local Python context as a Pandas DataFrame.

For more information on the kernels for Jupyter notebooks and the predefined "magics" that they provide, see [Kernels available for Jupyter notebooks with HDInsight Spark Linux clusters on HDInsight](#).

Data ingestion from public blob:

The first step in the data science process is to ingest the data to be analyzed from sources where it resides into your data exploration and modeling environment. This environment is Spark in this walkthrough. This section contains the code to complete a series of tasks:

- ingest the data sample to be modeled
- read in the input dataset (stored as a .tsv file)
- format and clean the data
- create and cache objects (RDDs or data-frames) in memory
- register it as a temp-table in SQL-context.

Here is the code for data ingestion.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# IMPORT FILE FROM PUBLIC BLOB
taxi_train_file = sc.textFile(taxi_train_file_loc)

# GET SCHEMA OF THE FILE FROM HEADER
schema_string = taxi_train_file.first()
fields = [StructField(field_name, StringType(), True) for field_name in schema_string.split('\t')]
fields[7].dataType = IntegerType() #Pickup hour
fields[8].dataType = IntegerType() # Pickup week
fields[9].dataType = IntegerType() # Weekday
fields[10].dataType = IntegerType() # Passenger count
fields[11].dataType = FloatType() # Trip time in secs
fields[12].dataType = FloatType() # Trip distance
fields[19].dataType = FloatType() # Fare amount
fields[20].dataType = FloatType() # Surcharge
fields[21].dataType = FloatType() # Mta_tax
fields[22].dataType = FloatType() # Tip amount
fields[23].dataType = FloatType() # Tolls amount
fields[24].dataType = FloatType() # Total amount
fields[25].dataType = IntegerType() # Tipped or not
fields[26].dataType = IntegerType() # Tip class
taxi_schema = StructType(fields)

# PARSE FIELDS AND CONVERT DATA TYPE FOR SOME FIELDS
taxi_header = taxi_train_file.filter(lambda l: "medallion" in l)
taxi_temp = taxi_train_file.subtract(taxi_header).map(lambda k: k.split("\t"))\
    .map(lambda p: (p[0],p[1],p[2],p[3],p[4],p[5],p[6],int(p[7]),int(p[8]),int(p[9]),int(p[10]),\
        float(p[11]),float(p[12]),p[13],p[14],p[15],p[16],p[17],p[18],float(p[19]),\
        float(p[20]),float(p[21]),float(p[22]),float(p[23]),float(p[24]),int(p[25]),int(p[26])))

# CREATE DATA FRAME
taxi_train_df = sqlContext.createDataFrame(taxi_temp, taxi_schema)

# CREATE A CLEANED DATA-FRAME BY DROPPING SOME UN-NECESSARY COLUMNS & FILTERING FOR UNDESIRED VALUES OR OUTLIERS
taxi_df_train_cleaned =
taxi_train_df.drop('medallion').drop('hack_license').drop('store_and_fwd_flag').drop('pickup_datetime')\
    .drop('dropoff_datetime').drop('pickup_longitude').drop('pickup_latitude').drop('dropoff_latitude')\
    .drop('dropoff_longitude').drop('tip_class').drop('total_amount').drop('tolls_amount').drop('mta_tax')\
    .drop('direct_distance').drop('surcharge')\
    .filter("passenger_count > 0 and passenger_count < 8 AND payment_type in ('CSH', 'CRD') AND tip_amount >= 0 AND tip_amount < 30 AND fare_amount >= 1 AND fare_amount < 150 AND trip_distance > 0 AND trip_distance < 100 AND trip_time_in_secs > 30 AND trip_time_in_secs < 7200" )

# CACHE & MATERIALIZE DATA-FRAME IN MEMORY. GOING THROUGH AND COUNTING NUMBER OF ROWS MATERIALIZES THE DATA-FRAME IN MEMORY
taxi_df_train_cleaned.cache()
taxi_df_train_cleaned.count()

# REGISTER DATA-FRAME AS A TEMP-TABLE IN SQL-CONTEXT
taxi_df_train_cleaned.registerTempTable("taxi_train")

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Time taken to execute above cell: 276.62 seconds

Data exploration & visualization

Once the data has been brought into Spark, the next step in the data science process is to gain deeper understanding of the data through exploration and visualization. In this section, we examine the taxi data using SQL queries and plot the target variables and prospective features for visual inspection. Specifically, we plot the frequency of passenger counts in taxi trips, the frequency of tip amounts, and how tips vary by payment amount and type.

Plot a histogram of passenger count frequencies in the sample of taxi trips

This code and subsequent snippets use SQL magic to query the sample and local magic to plot the data.

- **SQL magic (`%%sql`)** The HDInsight PySpark kernel supports easy inline HiveQL queries against the `sqlContext`. The `(-o VARIABLE_NAME)` argument persists the output of the SQL query as a Pandas DataFrame on the Jupyter server. This means it is available in the local mode.
- The `%%local` **magic** is used to run code locally on the Jupyter server, which is the headnode of the HDInsight cluster. Typically, you use `%%local` magic after the `%%sql -o` magic is used to run a query. The `-o` parameter would persist the output of the SQL query locally. Then the `%%local` magic triggers the next set of code snippets to run locally against the output of the SQL queries that has been persisted locally. The output is automatically visualized after you run the code.

This query retrieves the trips by passenger count.

```
# PLOT FREQUENCY OF PASSENGER COUNTS IN TAXI TRIPS

# SQL QUERY
%%sql -q -o sqlResults
SELECT passenger_count, COUNT(*) as trip_counts FROM taxi_train WHERE passenger_count > 0 and passenger_count < 7 GROUP BY passenger_count
```

This code creates a local data-frame from the query output and plots the data. The `%%local` magic creates a local data-frame, `sqlResults`, which can be used for plotting with matplotlib.

NOTE

This PySpark magic is used multiple times in this walkthrough. If the amount of data is large, you should sample to create a data-frame that can fit in local memory.

```
# RUN THE CODE LOCALLY ON THE JUPYTER SERVER
%%local

# USE THE JUPYTER AUTO-PLOTTING FEATURE TO CREATE INTERACTIVE FIGURES.
# CLICK ON THE TYPE OF PLOT TO BE GENERATED (E.G. LINE, AREA, BAR ETC.)
sqlResults
```

Here is the code to plot the trips by passenger counts

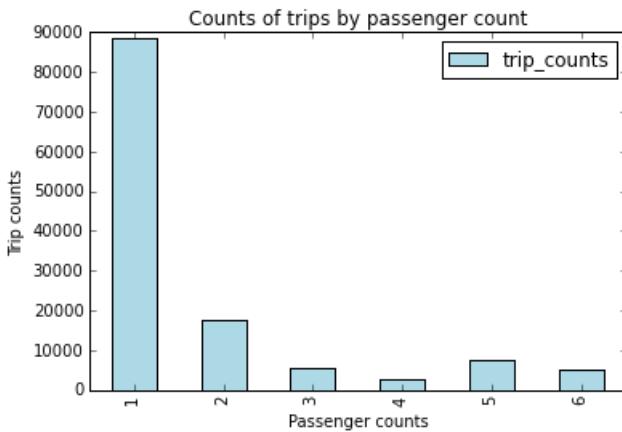
```

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
import matplotlib.pyplot as plt
%matplotlib inline

# PLOT PASSENGER NUMBER VS TRIP COUNTS
x_labels = sqlResults['passenger_count'].values
fig = sqlResults[['trip_counts']].plot(kind='bar', facecolor='lightblue')
fig.set_xticklabels(x_labels)
fig.set_title('Counts of trips by passenger count')
fig.set_xlabel('Passenger count in trips')
fig.set_ylabel('Trip counts')
plt.show()

```

OUTPUT



You can select among several different types of visualizations (Table, Pie, Line, Area, or Bar) by using the **Type** menu buttons in the notebook. The Bar plot is shown here.

Plot a histogram of tip amounts and how tip amount varies by passenger count and fare amounts.

Use a SQL query to sample data..

```

# SQL QUERY
%%sql -q -o sqlResults
SELECT fare_amount, passenger_count, tip_amount, tipped
FROM taxi_train
WHERE passenger_count > 0
AND passenger_count < 7
AND fare_amount > 0
AND fare_amount < 200
AND payment_type in ('CSH', 'CRD')
AND tip_amount > 0
AND tip_amount < 25

```

This code cell uses the SQL query to create three plots the data.

```

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
%matplotlib inline

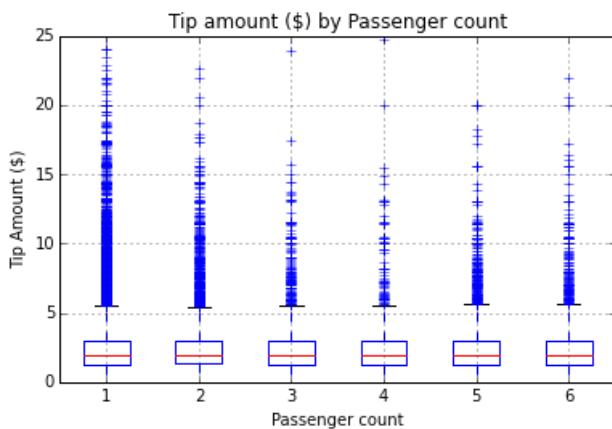
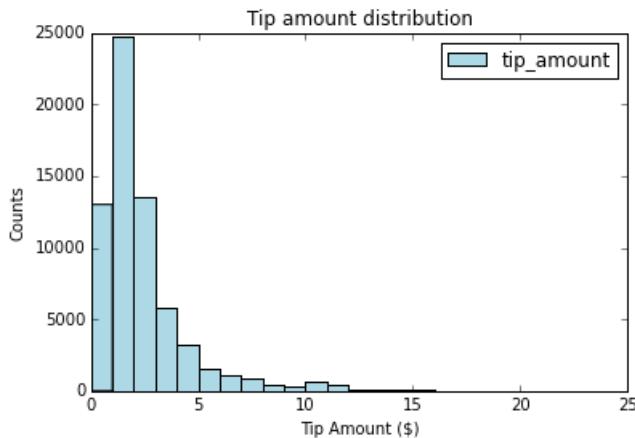
# TIP BY PAYMENT TYPE AND PASSENGER COUNT
ax1 = resultsPDDF[['tip_amount']].plot(kind='hist', bins=25, facecolor='lightblue')
ax1.set_title('Tip amount distribution')
ax1.set_xlabel('Tip Amount ($)')
ax1.set_ylabel('Counts')
plt.suptitle('')
plt.show()

# TIP BY PASSENGER COUNT
ax2 = resultsPDDF.boxplot(column=['tip_amount'], by=['passenger_count'])
ax2.set_title('Tip amount ($) by Passenger count')
ax2.set_xlabel('Passenger count')
ax2.set_ylabel('Tip Amount ($)')
plt.suptitle('')
plt.show()

# TIP AMOUNT BY FARE AMOUNT, POINTS ARE SCALED BY PASSENGER COUNT
ax = resultsPDDF.plot(kind='scatter', x= 'fare_amount', y = 'tip_amount', c='blue', alpha = 0.10, s=5*(resultsPDDF.passenger_count))
ax.set_title('Tip amount by Fare amount ($)')
ax.set_xlabel('Fare Amount')
ax.set_ylabel('Tip Amount')
plt.axis([-2, 120, -2, 30])
plt.show()

```

OUTPUT:





Feature engineering, transformation, and data preparation for modeling

This section describes and provides the code for procedures used to prepare data for use in ML modeling. It shows how to do the following tasks:

- Create a new feature by partitioning hours into traffic time bins
- Index and on-hot encode categorical features
- Create labeled point objects for input into ML functions
- Create a random sub-sampling of the data and split it into training and testing sets
- Feature scaling
- Cache objects in memory

Create a new feature by partitioning traffic times into bins

This code shows how to create a new feature by partitioning traffic times into bins and then how to cache the resulting data frame in memory. Caching leads to improved execution time where Resilient Distributed Datasets (RDDs) and data-frames are used repeatedly. So, we cache RDDs and data-frames at several stages in this walkthrough.

```
# CREATE FOUR BUCKETS FOR TRAFFIC TIMES
sqlStatement = """
SELECT *,
CASE
    WHEN (pickup_hour <= 6 OR pickup_hour >= 20) THEN "Night"
    WHEN (pickup_hour >= 7 AND pickup_hour <= 10) THEN "AMRush"
    WHEN (pickup_hour >= 11 AND pickup_hour <= 15) THEN "Afternoon"
    WHEN (pickup_hour >= 16 AND pickup_hour <= 19) THEN "PMRush"
END as TrafficTimeBins
FROM taxi_train
"""

taxi_df_train_with_newFeatures = sqlContext.sql(sqlStatement)

# CACHE DATA-FRAME IN MEMORY & MATERIALIZE DF IN MEMORY
# THE .COUNT() GOES THROUGH THE ENTIRE DATA-FRAME,
# MATERIALIZES IT IN MEMORY, AND GIVES THE COUNT OF ROWS.
taxi_df_train_with_newFeatures.cache()
taxi_df_train_with_newFeatures.count()
```

OUTPUT

126050

Index and one-hot encode categorical features

This section shows how to index or encode categorical features for input into the modeling functions. The

modeling and predict functions of MLlib require that features with categorical input data be indexed or encoded prior to use.

Depending on the model, you need to index or encode them in different ways. For example, Logistic and Linear Regression models require one-hot encoding, where, for example, a feature with three categories can be expanded into three feature columns, with each containing 0 or 1 depending on the category of an observation. MLlib provides [OneHotEncoder](#) function to do one-hot encoding. This encoder maps a column of label indices to a column of binary vectors, with at most a single one-value. This encoding allows algorithms that expect numerical valued features, such as logistic regression, to be applied to categorical features.

Here is the code to index and encode categorical features:

```
# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.ml.feature import OneHotEncoder, StringIndexer, VectorAssembler, OneHotEncoder, VectorIndexer

# INDEX AND ENCODE VENDOR_ID
stringIndexer = StringIndexer(inputCol="vendor_id", outputCol="vendorIndex")
model = stringIndexer.fit(taxi_df_train_with_newFeatures) # Input data-frame is the cleaned one from above
indexed = model.transform(taxi_df_train_with_newFeatures)
encoder = OneHotEncoder(dropLast=False, inputCol="vendorIndex", outputCol="vendorVec")
encoded1 = encoder.transform(indexed)

# INDEX AND ENCODE RATE_CODE
stringIndexer = StringIndexer(inputCol="rate_code", outputCol="rateIndex")
model = stringIndexer.fit(encoded1)
indexed = model.transform(encoded1)
encoder = OneHotEncoder(dropLast=False, inputCol="rateIndex", outputCol="rateVec")
encoded2 = encoder.transform(indexed)

# INDEX AND ENCODE PAYMENT_TYPE
stringIndexer = StringIndexer(inputCol="payment_type", outputCol="paymentIndex")
model = stringIndexer.fit(encoded2)
indexed = model.transform(encoded2)
encoder = OneHotEncoder(dropLast=False, inputCol="paymentIndex", outputCol="paymentVec")
encoded3 = encoder.transform(indexed)

# INDEX AND TRAFFIC TIME BINS
stringIndexer = StringIndexer(inputCol="TrafficTimeBins", outputCol="TrafficTimeBinsIndex")
model = stringIndexer.fit(encoded3)
indexed = model.transform(encoded3)
encoder = OneHotEncoder(dropLast=False, inputCol="TrafficTimeBinsIndex", outputCol="TrafficTimeBinsVec")
encodedFinal = encoder.transform(indexed)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT

Time taken to execute above cell: 3.14 seconds

Create labeled point objects for input into ML functions

This section contains code that shows how to index categorical text data as a labeled point data type and how to encode it. This prepares it to be used to train and test MLlib logistic regression and other classification models. Labeled point objects are Resilient Distributed Datasets (RDD) formatted in a way that is needed as input data by most of ML algorithms in MLlib. A [labeled point](#) is a local vector, either dense or sparse, associated with a label/response.

Here is the code to index and encode text features for binary classification.

```
# FUNCTIONS FOR BINARY CLASSIFICATION

# LOAD LIBRARIES
from pyspark.mllib.regression import LabeledPoint
from numpy import array

# INDEXING CATEGORICAL TEXT FEATURES FOR INPUT INTO TREE-BASED MODELS
def parseRowIndexingBinary(line):
    features = np.array([line.paymentIndex, line.vendorIndex, line.rateIndex, line.pickup_hour, line.weekday,
                        line.passenger_count, line.trip_time_in_secs, line.trip_distance, line.fare_amount])
    labPt = LabeledPoint(line.tipped, features)
    return labPt

# ONE-HOT ENCODING OF CATEGORICAL TEXT FEATURES FOR INPUT INTO LOGISTIC REGRESSION MODELS
def parseRowOneHotBinary(line):
    features = np.concatenate((np.array([line.pickup_hour, line.weekday, line.passenger_count,
                                         line.trip_time_in_secs, line.trip_distance, line.fare_amount]),
                               line.vendorVec.toArray(), line.rateVec.toArray(), line.paymentVec.toArray()),
                              axis=0)
    labPt = LabeledPoint(line.tipped, features)
    return labPt
```

Here is the code to encode and index categorical text features for linear regression analysis.

```
# FUNCTIONS FOR REGRESSION WITH TIP AMOUNT AS TARGET VARIABLE

# ONE-HOT ENCODING OF CATEGORICAL TEXT FEATURES FOR INPUT INTO TREE-BASED MODELS
def parseRowIndexingRegression(line):
    features = np.array([line.paymentIndex, line.vendorIndex, line.rateIndex, line.TrafficTimeBinsIndex,
                        line.pickup_hour, line.weekday, line.passenger_count, line.trip_time_in_secs,
                        line.trip_distance, line.fare_amount])
    labPt = LabeledPoint(line.tip_amount, features)
    return labPt

# INDEXING CATEGORICAL TEXT FEATURES FOR INPUT INTO LINEAR REGRESSION MODELS
def parseRowOneHotRegression(line):
    features = np.concatenate((np.array([line.pickup_hour, line.weekday, line.passenger_count,
                                         line.trip_time_in_secs, line.trip_distance, line.fare_amount]),
                               line.vendorVec.toArray(), line.rateVec.toArray(),
                               line.paymentVec.toArray(), line.TrafficTimeBinsVec.toArray()), axis=0)
    labPt = LabeledPoint(line.tip_amount, features)
    return labPt
```

Create a random sub-sampling of the data and split it into training and testing sets

This code creates a random sampling of the data (25% is used here). Although it is not required for this example due to the size of the dataset, we demonstrate how you can sample the data here. Then you know how to use it for your own problem if needed. When samples are large, this can save significant time while training models. Next we split the sample into a training part (75% here) and a testing part (25% here) to use in classification and regression modeling.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# SPECIFY SAMPLING AND SPLITTING FRACTIONS
from pyspark.sql.functions import rand

samplingFraction = 0.25;
trainingFraction = 0.75; testingFraction = (1-trainingFraction);
seed = 1234;
encodedFinalSampled = encodedFinal.sample(False, samplingFraction, seed=seed)

# SPLIT SAMPLED DATA-FRAME INTO TRAIN/TEST, WITH A RANDOM COLUMN ADDED FOR DOING CV (SHOWN LATER)
# INCLUDE RAND COLUMN FOR CREATING CROSS-VALIDATION FOLDS
dfTmpRand = encodedFinalSampled.select("*", rand(0).alias("rand"));
trainData, testData = dfTmpRand.randomSplit([trainingFraction, testingFraction], seed=seed);

# CACHE TRAIN AND TEST DATA
trainData.cache()
testData.cache()

# FOR BINARY CLASSIFICATION TRAINING AND TESTING
indexedTRAINbinary = trainData.map(parseRowIndexingBinary)
indexedTESTbinary = testData.map(parseRowIndexingBinary)
oneHotTRAINbinary = trainData.map(parseRowIndexingBinary)
oneHotTESTbinary = testData.map(parseRowIndexingBinary)

# FOR REGRESSION TRAINING AND TESTING
indexedTRAINreg = trainData.map(parseRowOneHotRegression)
indexedTESTreg = testData.map(parseRowOneHotRegression)
oneHotTRAINreg = trainData.map(parseRowOneHotRegression)
oneHotTESTreg = testData.map(parseRowOneHotRegression)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Time taken to execute above cell: 0.31 seconds

Feature scaling

Feature scaling, also known as data normalization, insures that features with widely disbursed values are not given excessive weigh in the objective function. The code for feature scaling uses the [StandardScaler](#) to scale the features to unit variance. It is provided by MLlib for use in linear regression with Stochastic Gradient Descent (SGD). SGD is a popular algorithm for training a wide range of other machine learning models such as regularized regressions or support vector machines (SVM).

TIP

We have found the `LinearRegressionWithSGD` algorithm to be sensitive to feature scaling.

Here is the code to scale variables for use with the regularized linear SGD algorithm.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.regression import LabeledPoint
from pyspark.mllib.linalg import Vectors
from pyspark.mllib.feature import StandardScaler, StandardScalerModel
from pyspark.mllib.util import MLUtils

# SCALE VARIABLES FOR REGULARIZED LINEAR SGD ALGORITHM
label = oneHotTRAINreg.map(lambda x: x.label)
features = oneHotTRAINreg.map(lambda x: x.features)
scaler = StandardScaler(withMean=False, withStd=True).fit(features)
dataTMP = label.zip(scaler.transform(features.map(lambda x: Vectors.dense(x.toArray()))))
oneHotTRAINregScaled = dataTMP.map(lambda x: LabeledPoint(x[0], x[1]))

label = oneHotTESTreg.map(lambda x: x.label)
features = oneHotTESTreg.map(lambda x: x.features)
scaler = StandardScaler(withMean=False, withStd=True).fit(features)
dataTMP = label.zip(scaler.transform(features.map(lambda x: Vectors.dense(x.toArray()))))
oneHotTESTregScaled = dataTMP.map(lambda x: LabeledPoint(x[0], x[1]))

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Time taken to execute above cell: 11.67 seconds

Cache objects in memory

The time taken for training and testing of ML algorithms can be reduced by caching the input data frame objects used for classification, regression and, scaled features.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# FOR BINARY CLASSIFICATION TRAINING AND TESTING
indexedTRAINbinary.cache()
indexedTESTbinary.cache()
oneHotTRAINbinary.cache()
oneHotTESTbinary.cache()

# FOR REGRESSION TRAINING AND TESTING
indexedTRAINreg.cache()
indexedTESTreg.cache()
oneHotTRAINreg.cache()
oneHotTESTreg.cache()

# SCALED FEATURES
oneHotTRAINregScaled.cache()
oneHotTESTregScaled.cache()

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Time taken to execute above cell: 0.13 seconds

Predict whether or not a tip is paid with binary classification models

This section shows how to use three models for the binary classification task of predicting whether or not a tip is paid for a taxi trip. The models presented are:

- Logistic regression
- Random forest
- Gradient Boosting Trees

Each model building code section is split into steps:

1. **Model training** data with one parameter set
2. **Model evaluation** on a test data set with metrics
3. **Saving model** in blob for future consumption

We show how to do cross-validation (CV) with parameter sweeping in two ways:

1. Using **generic** custom code which can be applied to any algorithm in MLlib and to any parameter sets in an algorithm.
2. Using the **pySpark CrossValidator pipeline function**. Note that CrossValidator has a few limitations for Spark 1.5.0:
 - Pipeline models cannot be saved/persisted for future consumption.
 - Cannot be used for every parameter in a model.
 - Cannot be used for every MLlib algorithm.

Generic cross validation and hyperparameter sweeping used with the logistic regression algorithm for binary classification

The code in this section shows how to train, evaluate, and save a logistic regression model with [LBFGS](#) that predicts whether or not a tip is paid for a trip in the NYC taxi trip and fare dataset. The model is trained using cross validation (CV) and hyperparameter sweeping implemented with custom code that can be applied to any of the learning algorithms in MLlib.

NOTE

The execution of this custom CV code can take several minutes.

Train the logistic regression model using CV and hyperparameter sweeping

```
# LOGISTIC REGRESSION CLASSIFICATION WITH CV AND HYPERPARAMETER SWEEPING

# GET ACCURACY FOR HYPERPARAMETERS BASED ON CROSS-VALIDATION IN TRAINING DATA-SET

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD LIBRARIES
from pyspark.mllib.classification import LogisticRegressionWithLBFGS
from pyspark.mllib.evaluation import BinaryClassificationMetrics

# CREATE PARAMETER GRID FOR LOGISTIC REGRESSION PARAMETER SWEEP
from sklearn.grid_search import ParameterGrid
grid = [{'regParam': [0.01, 0.1], 'iterations': [5, 10], 'regType': ["l1", "l2"], 'tolerance': [1e-3, 1e-4]}]
paramGrid = list(ParameterGrid(grid))
numModels = len(paramGrid)

# SET NUM FOLDS AND NUM PARAMETER SETS TO SWEEP ON
nFolds = 3;
h = 1.0 / nFolds;
```

```

metricSum = np.zeros(numModels);

# BEGIN CV WITH PARAMETER SWEEP
for i in range(nFolds):
    # Create training and x-validation sets
    validateLB = i * h
    validateUB = (i + 1) * h
    condition = (trainData["rand"] >= validateLB) & (trainData["rand"] < validateUB)
    validation = trainData.filter(condition)
    # Create LabeledPoints from data-frames
    if i > 0:
        trainCVLabPt.unpersist()
        validationLabPt.unpersist()
    trainCV = trainData.filter(~condition)
    trainCVLabPt = trainCV.map(parseRowOneHotBinary)
    trainCVLabPt.cache()
    validationLabPt = validation.map(parseRowOneHotBinary)
    validationLabPt.cache()
    # For parameter sets compute metrics from x-validation
    for j in range(numModels):
        regt = paramGrid[j]['regType']
        regp = paramGrid[j]['regParam']
        iters = paramGrid[j]['iterations']
        tol = paramGrid[j]['tolerance']
        # Train logistic regression model with hyperparameter set
        model = LogisticRegressionWithLBFGS.train(trainCVLabPt, regType=regt, iterations=iters,
                                                   regParam=regp, tolerance = tol, intercept=True)
        predictionAndLabels = validationLabPt.map(lambda lp: (float(model.predict(lp.features)), lp.label))
        # Use ROC-AUC as accuracy metrics
        validMetrics = BinaryClassificationMetrics(predictionAndLabels)
        metric = validMetrics.areaUnderROC
        metricSum[j] += metric

    avgAcc = metricSum / nFolds;
    bestParam = paramGrid[np.argmax(avgAcc)];

    # UNPERSIST OBJECTS
    trainCVLabPt.unpersist()
    validationLabPt.unpersist()

    # TRAIN ON FULL TRAIING SET USING BEST PARAMETERS FROM CV/PARAMETER SWEEP
    logitBest = LogisticRegressionWithLBFGS.train(oneHotTRAINbinary, regType=bestParam['regType'],
                                                 iterations=bestParam['iterations'],
                                                 regParam=bestParam['regParam'], tolerance =
    bestParam['tolerance'],
                                                 intercept=True)

    # PRINT COEFFICIENTS AND INTERCEPT OF THE MODEL
    # NOTE: There are 20 coefficient terms for the 10 features,
    #       and the different categories for features: vendorVec (2), rateVec, paymentVec (6), TrafficTimeBinsVec
    (4)
    print("Coefficients: " + str(logitBest.weights))
    print("Intercept: " + str(logitBest.intercept))

    # PRINT ELAPSED TIME
    timeend = datetime.datetime.now()
    timedelta = round((timeend-timestart).total_seconds(), 2)
    print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Coefficients: [0.0082065285375, -0.0223675576104, -0.0183812028036, -3.48124578069e-05, -0.00247646947233, -0.00165897881503, 0.0675394837328, -0.111823113101, -0.324609912762, -0.204549780032, -1.36499216354, 0.591088507921, -0.664263411392, -1.00439726852, 3.46567827545, -3.51025855172, -0.0471341112232, -0.043521833294, 0.000243375810385, 0.054518719222]

Intercept: -0.0111216486893

Time taken to execute above cell: 14.43 seconds

Evaluate the binary classification model with standard metrics

The code in this section shows how to evaluate a logistic regression model against a test data-set, including a plot of the ROC curve.

```
# RECORD START TIME
timestart = datetime.datetime.now()

#IMPORT LIBRARIES
from sklearn.metrics import roc_curve,auc
from pyspark.mllib.evaluation import BinaryClassificationMetrics
from pyspark.mllib.evaluation import MulticlassMetrics

# PREDICT ON TEST DATA WITH BEST/FINAL MODEL
predictionAndLabels = oneHotTESTbinary.map(lambda lp: (float(logitBest.predict(lp.features)), lp.label))

# INSTANTIATE METRICS OBJECT
metrics = BinaryClassificationMetrics(predictionAndLabels)

# AREA UNDER PRECISION-RECALL CURVE
print("Area under PR = %s" % metrics.areaUnderPR)

# AREA UNDER ROC CURVE
print("Area under ROC = %s" % metrics.areaUnderROC)
metrics = MulticlassMetrics(predictionAndLabels)

# OVERALL STATISTICS
precision = metrics.precision()
recall = metrics.recall()
f1Score = metrics.fMeasure()
print("Summary Stats")
print("Precision = %s" % precision)
print("Recall = %s" % recall)
print("F1 Score = %s" % f1Score)

# OUTPUT PROBABILITIES AND REGISTER TEMP TABLE
logitBest.clearThreshold(); # This clears threshold for classification (0.5) and outputs probabilities
predictionAndLabelsDF = predictionAndLabels.toDF()
predictionAndLabelsDF.registerTempTable("tmp_results");

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT

Area under PR = 0.985336538462

Area under ROC = 0.983383274312

Summary Stats

Precision = 0.984174341679

Recall = 0.984174341679

F1 Score = 0.984174341679

Time taken to execute above cell: 2.67 seconds

Plot the ROC curve.

The `predictionAndLabelsDF` is registered as a table, `tmp_results`, in the previous cell. `tmp_results` can be used to do queries and output results into the `sqlResults` data-frame for plotting. Here is the code.

```
# QUERY RESULTS
%%sql -q -o sqlResults
SELECT * from tmp_results
```

Here is the code to make predictions and plot the ROC-curve.

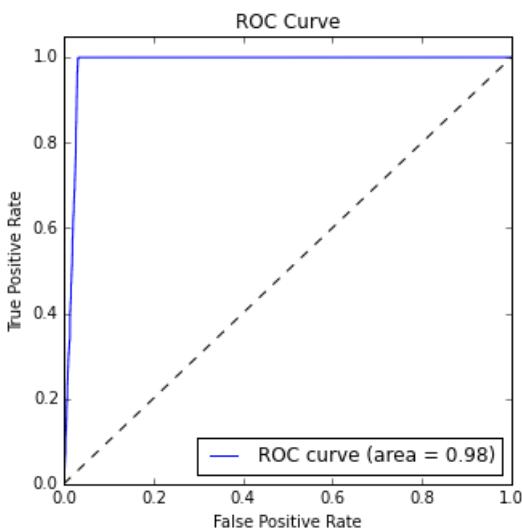
```
# MAKE PREDICTIONS AND PLOT ROC-CURVE

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
%matplotlib inline
from sklearn.metrics import roc_curve,auc

#PREDICTIONS
predictions_pddf = sqlResults.rename(columns={'_1': 'probability', '_2': 'label'})
prob = predictions_pddf["probability"]
fpr, tpr, thresholds = roc_curve(predictions_pddf['label'], prob, pos_label=1);
roc_auc = auc(fpr, tpr)

# PLOT ROC CURVES
plt.figure(figsize=(5,5))
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()
```

OUTPUT



Persist model in a blob for future consumption

The code in this section shows how to save the logistic regression model for consumption.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.classification import LogisticRegressionModel

# PERSIST MODEL
datestamp = unicode(datetime.datetime.now()).replace(' ', '').replace(':', '_')
logisticregressionfilename = "LogisticRegressionWithLBFGS_" + datestamp;
dirfilename = modelDir + logisticregressionfilename;

logitBest.save(sc, dirfilename);

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Time taken to execute above cell: 34.57 seconds

Use MLlib's CrossValidator pipeline function with logistic regression (Elastic regression) model

The code in this section shows how to train, evaluate, and save a logistic regression model with [LBFGS](#) that predicts whether or not a tip is paid for a trip in the NYC taxi trip and fare dataset. The model is trained using cross validation (CV) and hyperparameter sweeping implemented with the MLlib CrossValidator pipeline function for CV with parameter sweep.

NOTE

The execution of this MLlib CV code can take several minutes.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline
from pyspark.ml.evaluation import BinaryClassificationEvaluator
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder
from sklearn.metrics import roc_curve,auc

# DEFINE ALGORITHM / MODEL
lr = LogisticRegression()

# DEFINE GRID PARAMETERS
paramGrid = ParamGridBuilder().addGrid(lr.regParam, (0.01, 0.1))\
    .addGrid(lr.maxIter, (5, 10))\
    .addGrid(lr.tol, (1e-4, 1e-5))\
    .addGrid(lr.elasticNetParam, (0.25,0.75))\
    .build()

# DEFINE CV WITH PARAMETER SWEEP
cv = CrossValidator(estimator= lr,
                     estimatorParamMaps=paramGrid,
                     evaluator=BinaryClassificationEvaluator(),
                     numFolds=3)

# CONVERT TO DATA-FRAME: THIS DOES NOT RUN ON RDDS
trainDataFrame = sqlContext.createDataFrame(oneHotTRAINbinary, ["features", "label"])

# TRAIN WITH CROSS-VALIDATION
cv_model = cv.fit(trainDataFrame)

## PREDICT AND EVALUATE ON TEST DATA-SET

# USE TEST DATASET FOR PREDICTION
testDataFrame = sqlContext.createDataFrame(oneHotTESTbinary, ["features", "label"])
test_predictions = cv_model.transform(testDataFrame)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Time taken to execute above cell: 107.98 seconds

Plot the ROC curve.

The *predictionAndLabelsDF* is registered as a table, *tmp_results*, in the previous cell. *tmp_results* can be used to do queries and output results into the *sqlResults* data-frame for plotting. Here is the code.

```

# QUERY RESULTS
%%sql -q -o sqlResults
SELECT label, prediction, probability from tmp_results

```

Here is the code to plot the ROC curve.

```

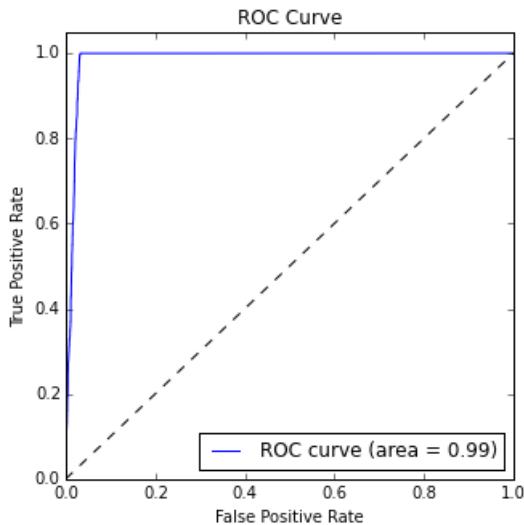
# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
from sklearn.metrics import roc_curve,auc

# ROC CURVE
prob = [x["values"][1] for x in sqlResults["probability"]]
fpr, tpr, thresholds = roc_curve(sqlResults['label'], prob, pos_label=1);
roc_auc = auc(fpr, tpr)

#PLOT
plt.figure(figsize=(5,5))
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()

```

OUTPUT



Random forest classification

The code in this section shows how to train, evaluate, and save a random forest regression that predicts whether or not a tip is paid for a trip in the NYC taxi trip and fare dataset.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import RandomForest, RandomForestModel
from pyspark.mllib.util import MLUtils
from pyspark.mllib.evaluation import BinaryClassificationMetrics
from pyspark.mllib.evaluation import MulticlassMetrics

# SPECIFY NUMBER OF CATEGORIES FOR CATEGORICAL FEATURES. FEATURE #0 HAS 2 CATEGORIES, FEATURE #2 HAS 2 CATEGORIES, AND SO ON
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}

# TRAIN RANDOMFOREST MODEL
rfModel = RandomForest.trainClassifier(indexedTRAINbinary, numClasses=2,
                                         categoricalFeaturesInfo=categoricalFeaturesInfo,
                                         numTrees=25, featureSubsetStrategy="auto",
                                         impurity='gini', maxDepth=5, maxBins=32)

## UN-COMMENT IF YOU WANT TO PRING TREES
#print('Learned classification forest model:')
#print(rfModel.toDebugString())

# PREDICT ON TEST DATA AND EVALUATE
predictions = rfModel.predict(indexedTESTbinary.map(lambda x: x.features))
predictionAndLabels = indexedTESTbinary.map(lambda lp: lp.label).zip(predictions)

# AREA UNDER ROC CURVE
metrics = BinaryClassificationMetrics(predictionAndLabels)
print("Area under ROC = %s" % metrics.areaUnderROC)

# PERSIST MODEL IN BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ', '').replace(':', '_');
rfclassificationfilename = "RandomForestClassification_" + datestamp;
dirfilename = modelDir + rfclassificationfilename;

rfModel.save(sc, dirfilename);

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Area under ROC = 0.985336538462

Time taken to execute above cell: 26.72 seconds

Gradient boosting trees classification

The code in this section shows how to train, evaluate, and save a gradient boosting trees model that predicts whether or not a tip is paid for a trip in the NYC taxi trip and fare dataset.

```

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import GradientBoostedTrees, GradientBoostedTreesModel

# SPECIFY NUMBER OF CATEGORIES FOR CATEGORICAL FEATURES. FEATURE #0 HAS 2 CATEGORIES, FEATURE #2 HAS 2 CATEGORIES, AND SO ON
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}

gbtModel = GradientBoostedTrees.trainClassifier(indexedTRAINbinary,
categoricalFeaturesInfo=categoricalFeaturesInfo,
                                         numIterations=10)
## UNCOMMENT IF YOU WANT TO PRINT TREE DETAILS
#print('Learned classification GBT model:')
#print(bgtModel.toDebugString())

# PREDICT ON TEST DATA AND EVALUATE
predictions = gbtModel.predict(indexedTESTbinary.map(lambda x: x.features))
predictionAndLabels = indexedTESTbinary.map(lambda lp: lp.label).zip(predictions)

# Area under ROC curve
metrics = BinaryClassificationMetrics(predictionAndLabels)
print("Area under ROC = %s" % metrics.areaUnderROC)

# PERSIST MODEL IN A BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
btclassificationfilename = "GradientBoostingTreeClassification_" + datestamp;
dirfilename = modelDir + btclassificationfilename;

gbtModel.save(sc, dirfilename)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Area under ROC = 0.985336538462

Time taken to execute above cell: 28.13 seconds

Predict tip amount with regression models (not using CV)

This section shows how use three models for the regression task: predict the tip amount paid for a taxi trip based on other tip features. The models presented are:

- Regularized linear regression
- Random forest
- Gradient Boosting Trees

These models were described in the introduction. Each model building code section is split into steps:

1. **Model training** data with one parameter set
2. **Model evaluation** on a test data set with metrics
3. **Saving model** in blob for future consumption

AZURE NOTE: Cross-validation is not used with the three regression models in this section, since this was shown in detail for the logistic regression models. An example showing how to use CV with Elastic Net for linear regression is provided in the Appendix of this topic.

AZURE NOTE: In our experience, there can be issues with convergence of LinearRegressionWithSGD models, and parameters need to be changed/optimized carefully for obtaining a valid model. Scaling of variables significantly helps with convergence. Elastic net regression, shown in the Appendix to this topic, can also be used instead of LinearRegressionWithSGD.

Linear regression with SGD

The code in this section shows how to use scaled features to train a linear regression that uses stochastic gradient descent (SGD) for optimization, and how to score, evaluate, and save the model in Azure Blob Storage (WASB).

TIP

In our experience, there can be issues with the convergence of LinearRegressionWithSGD models, and parameters need to be changed/optimized carefully for obtaining a valid model. Scaling of variables significantly helps with convergence.

```
# LINEAR REGRESSION WITH SGD

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD LIBRARIES
from pyspark.mllib.regression import LabeledPoint, LinearRegressionWithSGD, LinearRegressionModel
from pyspark.mllib.evaluation import RegressionMetrics
from scipy import stats

# USE SCALED FEATURES TO TRAIN MODEL
linearModel = LinearRegressionWithSGD.train(oneHotTRAINregScaled, iterations=100, step = 0.1, regType='l2',
                                             regParam=0.1, intercept = True)

# PRINT COEFFICIENTS AND INTERCEPT OF THE MODEL
# NOTE: There are 20 coefficient terms for the 10 features,
#       and the different categories for features: vendorVec (2), rateVec, paymentVec (6), TrafficTimeBinsVec (4)
print("Coefficients: " + str(linearModel.weights))
print("Intercept: " + str(linearModel.intercept))

# SCORE ON SCALED TEST DATA-SET & EVALUATE
predictionAndLabels = oneHotTESTregScaled.map(lambda lp: (float(linearModel.predict(lp.features)), lp.label))
testMetrics = RegressionMetrics(predictionAndLabels)

print("RMSE = %s" % testMetrics.rootMeanSquaredError)
print("R-sqr = %s" % testMetrics.r2)

# SAVE MODEL IN BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
linearregressionfilename = "LinearRegressionWithSGD_" + datestamp;
dirfilename = modelDir + linearregressionfilename;

linearModel.save(sc, dirfilename)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT

Coefficients: [0.0141707753435, -0.0252930927087, -0.0231442517137, 0.247070902996, 0.312544147152, 0.360296120645, 0.0122079566092, -0.00456498588241, -0.0898228505177, 0.0714046248793, 0.102171263868, 0.100022455632, -0.00289545676449, -0.00791124681938, 0.54396316518, -0.536293513569, 0.0119076553369, -0.0173039244582, 0.0119632796147, 0.00146764882502]

Intercept: 0.854507624459

RMSE = 1.23485131376

R-sqr = 0.597963951127

Time taken to execute above cell: 38.62 seconds

Random Forest regression

The code in this section shows how to train, evaluate, and save a random forest model that predicts tip amount for the NYC taxi trip data.

NOTE

Cross-validation with parameter sweeping using custom code is provided in the appendix.

```
#PREDICT TIP AMOUNTS USING RANDOM FOREST

# RECORD START TIME
timestart= datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import RandomForest, RandomForestModel
from pyspark.mllib.util import MLUtils
from pyspark.mllib.evaluation import RegressionMetrics

# TRAIN MODEL
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}
rfModel = RandomForest.trainRegressor(indexedTRAINreg, categoricalFeaturesInfo=categoricalFeaturesInfo,
                                       numTrees=25, featureSubsetStrategy="auto",
                                       impurity='variance', maxDepth=10, maxBins=32)

# UN-COMMENT IF YOU WANT TO PRING TREES
#print('Learned classification forest model:')
#print(rfModel.toDebugString())

# PREDICT AND EVALUATE ON TEST DATA-SET
predictions = rfModel.predict(indexedTESTreg.map(lambda x: x.features))
predictionAndLabels = oneHotTESTreg.map(lambda lp: lp.label).zip(predictions)

testMetrics = RegressionMetrics(predictionAndLabels)
print("RMSE = %s" % testMetrics.rootMeanSquaredError)
print("R-sqr = %s" % testMetrics.r2)

# SAVE MODEL IN BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
rfregressionfilename = "RandomForestRegression_" + datestamp;
dirfilename = modelDir + rfregressionfilename;

rfModel.save(sc, dirfilename);

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT

RMSE = 0.931981967875

R-sqr = 0.733445485802

Time taken to execute above cell: 25.98 seconds

Gradient boosting trees regression

The code in this section shows how to train, evaluate, and save a gradient boosting trees model that predicts tip amount for the NYC taxi trip data.

Train and evaluate

```
#PREDICT TIP AMOUNTS USING GRADIENT BOOSTING TREES

# RECORD START TIME
timestart= datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.mllib.tree import GradientBoostedTrees, GradientBoostedTreesModel
from pyspark.mllib.util import MLUtils

# TRAIN MODEL
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}
gbtModel = GradientBoostedTrees.trainRegressor(indexedTRAINreg,
categoricalFeaturesInfo=categoricalFeaturesInfo,
                           numIterations=10, maxBins=32, maxDepth = 4, learningRate=0.1)

# EVALUATE A TEST DATA-SET
predictions = gbtModel.predict(indexedTESTreg.map(lambda x: x.features))
predictionAndLabels = indexedTESTreg.map(lambda lp: lp.label).zip(predictions)

testMetrics = RegressionMetrics(predictionAndLabels)
print("RMSE = %s" % testMetrics.rootMeanSquaredError)
print("R-sqr = %s" % testMetrics.r2)

# PLOT SCATTER-PLOT BETWEEN ACTUAL AND PREDICTED TIP VALUES
test_predictions= sqlContext.createDataFrame(predictionAndLabels)
test_predictions_pddf = test_predictions.toPandas()

# SAVE MODEL IN BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
btregressionfilename = "GradientBoostingTreeRegression_" + datestamp;
dirfilename = modelDir + btregressionfilename;
gbtModel.save(sc, dirfilename)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT

RMSE = 0.928172197114

R-sqr = 0.732680354389

Time taken to execute above cell: 20.9 seconds

Plot

tmp_results is registered as a Hive table in the previous cell. Results from the table are output into the *sqlResults* data-frame for plotting. Here is the code

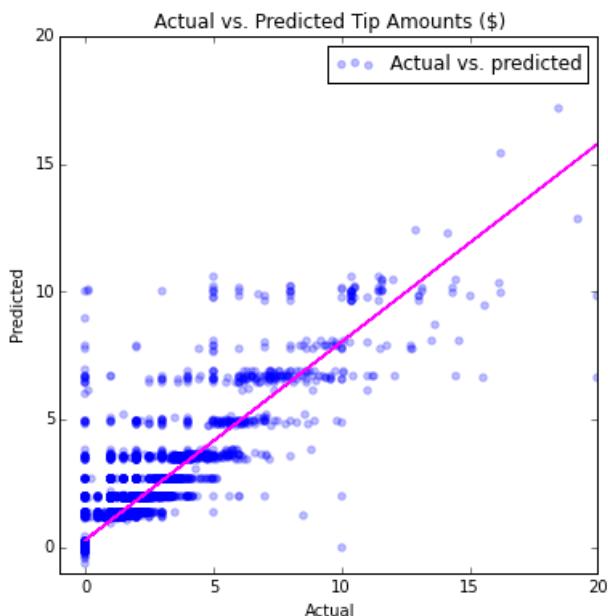
```
# PLOT SCATTER-PLOT BETWEEN ACTUAL AND PREDICTED TIP VALUES

# SELECT RESULTS
%%sql -q -o sqlResults
SELECT * from tmp_results
```

Here is the code to plot the data using the Jupyter server.

```
# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
import numpy as np

# PLOT
ax = sqlResults.plot(kind='scatter', figsize = (6,6), x='_1', y='_2', color='blue', alpha = 0.25,
label='Actual vs. predicted');
fit = np.polyfit(sqlResults['_1'], sqlResults['_2'], deg=1)
ax.set_title('Actual vs. Predicted Tip Amounts ($)')
ax.set_xlabel("Actual")
ax.set_ylabel("Predicted")
ax.plot(sqlResults['_1'], fit[0] * sqlResults['_1'] + fit[1], color='magenta')
plt.axis([-1, 15, -1, 15])
plt.show(ax)
```



Appendix: Additional regression tasks using cross validation with parameter sweeps

This appendix contains code showing how to do CV using Elastic net for linear regression and how to do CV with parameter sweep using custom code for random forest regression.

Cross validation using Elastic net for linear regression

The code in this section shows how to do cross validation using Elastic net for linear regression and how to evaluate the model against test data.

```

### CV USING ELASTIC NET FOR LINEAR REGRESSION

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.ml.regression import LinearRegression
from pyspark.ml import Pipeline
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder

# DEFINE ALGORITHM/MODEL
lr = LinearRegression()

# DEFINE GRID PARAMETERS
paramGrid = ParamGridBuilder().addGrid(lr.regParam, (0.01, 0.1))\
    .addGrid(lr.maxIter, (5, 10))\
    .addGrid(lr.tol, (1e-4, 1e-5))\
    .addGrid(lr.elasticNetParam, (0.25, 0.75))\
    .build()

# DEFINE PIPELINE
# SIMPLY THE MODEL HERE, WITHOUT TRANSFORMATIONS
pipeline = Pipeline(stages=[lr])

# DEFINE CV WITH PARAMETER SWEEP
cv = CrossValidator(estimator= lr,
                     estimatorParamMaps=paramGrid,
                     evaluator=RegressionEvaluator(),
                     numFolds=3)

# CONVERT TO DATA FRAME, AS CROSSVALIDATOR WON'T RUN ON RDDS
trainDataFrame = sqlContext.createDataFrame(oneHotTRAINreg, ["features", "label"])

# TRAIN WITH CROSS-VALIDATION
cv_model = cv.fit(trainDataFrame)

# EVALUATE MODEL ON TEST SET
testDataFrame = sqlContext.createDataFrame(oneHotTESTreg, ["features", "label"])

# MAKE PREDICTIONS ON TEST DOCUMENTS
# cvModel uses the best model found (lrModel).
predictionAndLabels = cv_model.transform(testDataFrame)

# CONVERT TO DF AND SAVE REGISTER DF AS TABLE
predictionAndLabels.registerTempTable("tmp_results");

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

Time taken to execute above cell: 161.21 seconds

Evaluate with R-SQR metric

tmp_results is registered as a Hive table in the previous cell. Results from the table are output into the *sqlResults* data-frame for plotting. Here is the code

```
# SELECT RESULTS
%%sql -q -o sqlResults
SELECT label,prediction from tmp_results
```

Here is the code to calculate R-sqr.

```
# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
from scipy import stats

#R-SQR TEST METRIC
corstats = stats.linregress(sqlResults['label'],sqlResults['prediction'])
r2 = (corstats[2]*corstats[2])
print("R-sqr = %s" % r2)
```

OUTPUT

R-sqr = 0.619184907088

Cross validation with parameter sweep using custom code for random forest regression

The code in this section shows how to do cross validation with parameter sweep using custom code for random forest regression and how to evaluate the model against test data.

```
# RECORD START TIME
timestart= datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
# GET ACCURACY FOR HYPERPARAMETERS BASED ON CROSS-VALIDATION IN TRAINING DATA-SET
from pyspark.mllib.tree import RandomForest, RandomForestModel
from pyspark.mllib.util import MLUtils
from pyspark.mllib.evaluation import RegressionMetrics
from sklearn.grid_search import ParameterGrid

## CREATE PARAMETER GRID
grid = [{"maxDepth": [5,10], 'numTrees': [25,50]}]
paramGrid = list(ParameterGrid(grid))

## SPECIFY LEVELS OF CATEGORICAL VARIABLES
categoricalFeaturesInfo={0:2, 1:2, 2:6, 3:4}

# SPECIFY NUMFOLDS AND ARRAY TO HOLD METRICS
nFolds = 3;
numModels = len(paramGrid)
h = 1.0 / nFolds;
metricSum = np.zeros(numModels);

for i in range(nFolds):
    # Create training and x-validation sets
    validateLB = i * h
    validateUB = (i + 1) * h
    condition = (trainData["rand"] >= validateLB) & (trainData["rand"] < validateUB)
    validation = trainData.filter(condition)
    # Create labeled points from data-frames
    if i > 0:
        trainCVLabPt.unpersist()
        validationLabPt.unpersist()
    trainCV = trainData.filter(~condition)
    trainCVLabPt = trainCV.map(parseRowIndexingRegression)
    trainCVLabPt.cache()
    validationLabPt = validation.map(parseRowIndexingRegression)
    validationLabPt.cache()
    # For parameter sets compute metrics from x-validation
    for j in range(numModels):
```

```

maxD = paramGrid[j]['maxDepth']
numT = paramGrid[j]['numTrees']
# Train logistic regression model with hyperparameter set
rfModel = RandomForest.trainRegressor(trainCVLabPt, categoricalFeaturesInfo=categoricalFeaturesInfo,
                                       numTrees=numT, featureSubsetStrategy="auto",
                                       impurity='variance', maxDepth=maxD, maxBins=32)
predictions = rfModel.predict(validationLabPt.map(lambda x: x.features))
predictionAndLabels = validationLabPt.map(lambda lp: lp.label).zip(predictions)
# Use ROC-AUC as accuracy metrics
validMetrics = RegressionMetrics(predictionAndLabels)
metric = validMetrics.rootMeanSquaredError
metricSum[j] += metric

avgAcc = metricSum/nFolds;
bestParam = paramGrid[np.argmin(avgAcc)];

# UNPERSIST OBJECTS
trainCVLabPt.unpersist()
validationLabPt.unpersist()

## TRAIN FINAL MODEL WITH BEST PARAMETERS
rfModel = RandomForest.trainRegressor(indexedTRAINreg, categoricalFeaturesInfo=categoricalFeaturesInfo,
                                       numTrees=bestParam['numTrees'], featureSubsetStrategy="auto",
                                       impurity='variance', maxDepth=bestParam['maxDepth'], maxBins=32)

# EVALUATE MODEL ON TEST DATA
predictions = rfModel.predict(indexedTESTreg.map(lambda x: x.features))
predictionAndLabels = indexedTESTreg.map(lambda lp: lp.label).zip(predictions)

#PRINT TEST METRICS
testMetrics = RegressionMetrics(predictionAndLabels)
print("RMSE = %s" % testMetrics.rootMeanSquaredError)
print("R-sqr = %s" % testMetrics.r2)

# PRINT ELAPSED TIME
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT

RMSE = 0.906972198262

R-sqr = 0.740751197012

Time taken to execute above cell: 69.17 seconds

Clean up objects from memory and print model locations

Use `unpersist()` to delete objects cached in memory.

```

# UNPERSIST OBJECTS CACHED IN MEMORY

# REMOVE ORIGINAL DFS
taxi_df_train_cleaned.unpersist()
taxi_df_train_with_newFeatures.unpersist()
trainData.unpersist()
trainData.unpersist()

# FOR BINARY CLASSIFICATION TRAINING AND TESTING
indexedTRAINbinary.unpersist()
indexedTESTbinary.unpersist()
oneHotTRAINbinary.unpersist()
oneHotTESTbinary.unpersist()

# FOR REGRESSION TRAINING AND TESTING
indexedTRAINreg.unpersist()
indexedTESTreg.unpersist()
oneHotTRAINreg.unpersist()
oneHotTESTreg.unpersist()

# SCALED FEATURES
oneHotTRAINregScaled.unpersist()
oneHotTESTregScaled.unpersist()

```

OUTPUT

PythonRDD[122] at RDD at PythonRDD.scala:43

**Printout path to model files to be used in the consumption notebook. ** To consume and score an independent data-set, you need to copy and paste these file names in the "Consumption notebook".

```

# PRINT MODEL FILE LOCATIONS FOR CONSUMPTION
print "logisticRegFileLoc = modelDir + \\" + logisticregressionfilename + "\\";
print "linearRegFileLoc = modelDir + \\" + linearregressionfilename + "\\";
print "randomForestClassificationFileLoc = modelDir + \\" + rfclassificationfilename + "\\";
print "randomForestRegFileLoc = modelDir + \\" + rfregressionfilename + "\\";
print "BoostedTreeClassificationFileLoc = modelDir + \\" + btclassificationfilename + "\\";
print "BoostedTreeRegressionFileLoc = modelDir + \\" + btregressionfilename + "\\";

```

OUTPUT

```

logisticRegFileLoc = modelDir + "LogisticRegressionWithLBFGS_2016-05-0316_47_30.096528"
linearRegFileLoc = modelDir + "LinearRegressionWithSGD_2016-05-0316_51_28.433670"
randomForestClassificationFileLoc = modelDir + "RandomForestClassification_2016-05-0316_50_17.454440"
randomForestRegFileLoc = modelDir + "RandomForestRegression_2016-05-0316_51_57.331730"
BoostedTreeClassificationFileLoc = modelDir + "GradientBoostingTreeClassification_2016-05-0316_50_40.138809"
BoostedTreeRegressionFileLoc = modelDir + "GradientBoostingTreeRegression_2016-05-0316_52_18.827237"

```

What's next?

Now that you have created regression and classification models with the Spark MLlib, you are ready to learn how to score and evaluate these models.

Model consumption: To learn how to score and evaluate the classification and regression models created in this topic, see [Score and evaluate Spark-built machine learning models](#).

Operationalize Spark-built machine learning models

3/12/2019 • 17 minutes to read

This topic shows how to operationalize a saved machine learning model (ML) using Python on HDInsight Spark clusters. It describes how to load machine learning models that have been built using Spark MLlib and stored in Azure Blob Storage (WASB), and how to score them with datasets that have also been stored in WASB. It shows how to pre-process the input data, transform features using the indexing and encoding functions in the MLlib toolkit, and how to create a labeled point data object that can be used as input for scoring with the ML models. The models used for scoring include Linear Regression, Logistic Regression, Random Forest Models, and Gradient Boosting Tree Models.

Spark clusters and Jupyter notebooks

Setup steps and the code to operationalize an ML model are provided in this walkthrough for using an HDInsight Spark 1.6 cluster as well as a Spark 2.0 cluster. The code for these procedures is also provided in Jupyter notebooks.

Notebook for Spark 1.6

The [pySpark-machine-learning-data-science-spark-model-consumption.ipynb](#) Jupyter notebook shows how to operationalize a saved model using Python on HDInsight clusters.

Notebook for Spark 2.0

To modify the Jupyter notebook for Spark 1.6 to use with an HDInsight Spark 2.0 cluster, replace the Python code file with [this file](#). This code shows how to consume the models created in Spark 2.0.

Prerequisites

1. You need an Azure account and a Spark 1.6 (or Spark 2.0) HDInsight cluster to complete this walkthrough. See the [Overview of Data Science using Spark on Azure HDInsight](#) for instructions on how to satisfy these requirements. That topic also contains a description of the NYC 2013 Taxi data used here and instructions on how to execute code from a Jupyter notebook on the Spark cluster.
2. You must also create the machine learning models to be scored here by working through the [Data exploration and modeling with Spark](#) topic for the Spark 1.6 cluster or the Spark 2.0 notebooks.
3. The Spark 2.0 notebooks use an additional data set for the classification task, the well-known Airline On-time departure dataset from 2011 and 2012. A description of the notebooks and links to them are provided in the [Readme.md](#) for the GitHub repository containing them. Moreover, the code here and in the linked notebooks is generic and should work on any Spark cluster. If you are not using HDInsight Spark, the cluster setup and management steps may be slightly different from what is shown here.

WARNING

Billing for HDInsight clusters is prorated per minute, whether you use them or not. Be sure to delete your cluster after you finish using it. See [how to delete an HDInsight cluster](#).

Setup: storage locations, libraries, and the preset Spark context

Spark is able to read and write to an Azure Storage Blob (WASB). So any of your existing data stored there can be processed using Spark and the results stored again in WASB.

To save models or files in WASB, the path needs to be specified properly. The default container attached to the Spark cluster can be referenced using a path beginning with: "wasb///". The following code sample specifies the location of the data to be read and the path for the model storage directory to which the model output is saved.

Set directory paths for storage locations in WASB

Models are saved in: "wasb:///user/remoteuser/NYCTaxi/Models". If this path is not set properly, models are not loaded for scoring.

The scored results have been saved in: "wasb:///user/remoteuser/NYCTaxi/ScoredResults". If the path to folder is incorrect, results are not saved in that folder.

NOTE

The file path locations can be copied and pasted into the placeholders in this code from the output of the last cell of the **machine-learning-data-science-spark-data-exploration-modeling.ipynb** notebook.

Here is the code to set directory paths:

```
# LOCATION OF DATA TO BE SCORED (TEST DATA)
taxi_test_file_loc =
"wasb://mllibwalkthroughs@cdsparksamples.blob.core.windows.net/Data/NYCTaxi/JoinedTaxiTripFare.Point1Pct.Test.tsv";

# SET THE MODEL STORAGE DIRECTORY PATH
# NOTE THE LAST BACKSLASH IN THIS PATH IS NEEDED
modelDir = "wasb:///user/remoteuser/NYCTaxi/Models/"

# SET SCORDED RESULT DIRECTORY PATH
# NOTE THE LAST BACKSLASH IN THIS PATH IS NEEDED
scoredResultDir = "wasb:///user/remoteuser/NYCTaxi/ScoredResults/";

# FILE LOCATIONS FOR THE MODELS TO BE SCORED
logisticRegFileLoc = modelDir + "LogisticRegressionWithLBFGS_2016-04-1817_40_35.796789"
linearRegFileLoc = modelDir + "LinearRegressionWithSGD_2016-04-1817_44_00.993832"
randomForestClassificationFileLoc = modelDir + "RandomForestClassification_2016-04-1817_42_58.899412"
randomForestRegFileLoc = modelDir + "RandomForestRegression_2016-04-1817_44_27.204734"
BoostedTreeClassificationFileLoc = modelDir + "GradientBoostingTreeClassification_2016-04-1817_43_16.354770"
BoostedTreeRegressionFileLoc = modelDir + "GradientBoostingTreeRegression_2016-04-1817_44_46.206262"

# RECORD START TIME
import datetime
datetime.datetime.now()
```

OUTPUT:

```
datetime.datetime(2016, 4, 25, 23, 56, 19, 229403)
```

Import libraries

Set spark context and import necessary libraries with the following code

```
#IMPORT LIBRARIES
import pyspark
from pyspark import SparkConf
from pyspark import SparkContext
from pyspark.sql import SQLContext
import matplotlib
import matplotlib.pyplot as plt
from pyspark.sql import Row
from pyspark.sql.functions import UserDefinedFunction
from pyspark.sql.types import *
import atexit
from numpy import array
import numpy as np
import datetime
```

Preset Spark context and PySpark magics

The PySpark kernels that are provided with Jupyter notebooks have a preset context. So you do not need to set the Spark or Hive contexts explicitly before you start working with the application you are developing. These are available for you by default. These contexts are:

- sc - for Spark
- sqlContext - for Hive

The PySpark kernel provides some predefined "magics", which are special commands that you can call with %. There are two such commands that are used in these code samples.

- **%%local** Specified that the code in subsequent lines is executed locally. Code must be valid Python code.
- **%%sql -o**
- Executes a Hive query against the sqlContext. If the -o parameter is passed, the result of the query is persisted in the %%local Python context as a Pandas dataframe.

For more information on the kernels for Jupyter notebooks and the predefined "magics" that they provide, see [Kernels available for Jupyter notebooks with HDInsight Spark Linux clusters on HDInsight](#).

Ingest data and create a cleaned data frame

This section contains the code for a series of tasks required to ingest the data to be scored. Read in a joined 0.1% sample of the taxi trip and fare file (stored as a .tsv file), format the data, and then creates a clean data frame.

The taxi trip and fare files were joined based on the procedure provided in the: [The Team Data Science Process in action: using HDInsight Hadoop clusters](#) topic.

```

# INGEST DATA AND CREATE A CLEANED DATA FRAME

# RECORD START TIME
timestart = datetime.datetime.now()

# IMPORT FILE FROM PUBLIC BLOB
taxi_test_file = sc.textFile(taxi_test_file_loc)

# GET SCHEMA OF THE FILE FROM HEADER
taxi_header = taxi_test_file.filter(lambda l: "medallion" in l)

# PARSE FIELDS AND CONVERT DATA TYPE FOR SOME FIELDS
taxi_temp = taxi_test_file.subtract(taxi_header).map(lambda k: k.split("\t"))\
    .map(lambda p: (p[0],p[1],p[2],p[3],p[4],p[5],p[6],int(p[7]),int(p[8]),int(p[9]),int(p[10]),\
        float(p[11]),float(p[12]),p[13],p[14],p[15],p[16],p[17],p[18],float(p[19]),\
        float(p[20]),float(p[21]),float(p[22]),float(p[23]),float(p[24]),int(p[25]),int(p[26])))

# GET SCHEMA OF THE FILE FROM HEADER
schema_string = taxi_test_file.first()
fields = [StructField(field_name, StringType(), True) for field_name in schema_string.split('\t')]
fields[7].dataType = IntegerType() # Pickup hour
fields[8].dataType = IntegerType() # Pickup week
fields[9].dataType = IntegerType() # Weekday
fields[10].dataType = IntegerType() # Passenger count
fields[11].dataType = FloatType() # Trip time in secs
fields[12].dataType = FloatType() # Trip distance
fields[19].dataType = FloatType() # Fare amount
fields[20].dataType = FloatType() # Surcharge
fields[21].dataType = FloatType() # Mta_tax
fields[22].dataType = FloatType() # Tip amount
fields[23].dataType = FloatType() # Tolls amount
fields[24].dataType = FloatType() # Total amount
fields[25].dataType = IntegerType() # Tipped or not
fields[26].dataType = IntegerType() # Tip class
taxi_schema = StructType(fields)

# CREATE DATA FRAME
taxi_df_test = sqlContext.createDataFrame(taxi_temp, taxi_schema)

# CREATE A CLEANED DATA-FRAME BY DROPPING SOME UN-NECESSARY COLUMNS & FILTERING FOR UNDESIRED VALUES OR
# OUTLIERS
taxi_df_test_cleaned =
taxi_df_test.drop('medallion').drop('hack_license').drop('store_and_fwd_flag').drop('pickup_datetime')\
    .drop('dropoff_datetime').drop('pickup_longitude').drop('pickup_latitude').drop('dropoff_latitude')\
    .drop('dropoff_longitude').drop('tip_class').drop('total_amount').drop('tolls_amount').drop('mta_tax')\
    .drop('direct_distance').drop('surcharge')\
    .filter("passenger_count > 0 and passenger_count < 8 AND payment_type in ('CSH', 'CRD') AND tip_amount >=\
0 AND tip_amount < 30 AND fare_amount >= 1 AND fare_amount < 150 AND trip_distance > 0 AND trip_distance <\
100 AND trip_time_in_secs > 30 AND trip_time_in_secs < 7200" )

# CACHE DATA-FRAME IN MEMORY & MATERIALIZE DF IN MEMORY
taxi_df_test_cleaned.cache()
taxi_df_test_cleaned.count()

# REGISTER DATA-FRAME AS A TEMP-TABLE IN SQL-CONTEXT
taxi_df_test_cleaned.registerTempTable("taxi_test")

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 46.37 seconds

Prepare data for scoring in Spark

This section shows how to index, encode, and scale categorical features to prepare them for use in MLlib supervised learning algorithms for classification and regression.

Feature transformation: index and encode categorical features for input into models for scoring

This section shows how to index categorical data using a `StringIndexer` and encode features with `OneHotEncoder` input into the models.

The `StringIndexer` encodes a string column of labels to a column of label indices. The indices are ordered by label frequencies.

The `OneHotEncoder` maps a column of label indices to a column of binary vectors, with at most a single one-value. This encoding allows algorithms that expect continuous valued features, such as logistic regression, to be applied to categorical features.

```

#INDEX AND ONE-HOT ENCODE CATEGORICAL FEATURES

# RECORD START TIME
timestart = datetime.datetime.now()

# LOAD PYSPARK LIBRARIES
from pyspark.ml.feature import OneHotEncoder, StringIndexer, VectorAssembler, VectorIndexer

# CREATE FOUR BUCKETS FOR TRAFFIC TIMES
sqlStatement = """
    SELECT *,
    CASE
        WHEN (pickup_hour <= 6 OR pickup_hour >= 20) THEN "Night"
        WHEN (pickup_hour >= 7 AND pickup_hour <= 10) THEN "AMRush"
        WHEN (pickup_hour >= 11 AND pickup_hour <= 15) THEN "Afternoon"
        WHEN (pickup_hour >= 16 AND pickup_hour <= 19) THEN "PMRush"
    END as TrafficTimeBins
    FROM taxi_test
"""
taxi_df_test_with_newFeatures = sqlContext.sql(sqlStatement)

# CACHE DATA-FRAME IN MEMORY & MATERIALIZE DF IN MEMORY
taxi_df_test_with_newFeatures.cache()
taxi_df_test_with_newFeatures.count()

# INDEX AND ONE-HOT ENCODING
stringIndexer = StringIndexer(inputCol="vendor_id", outputCol="vendorIndex")
model = stringIndexer.fit(taxi_df_test_with_newFeatures) # Input data-frame is the cleaned one from above
indexed = model.transform(taxi_df_test_with_newFeatures)
encoder = OneHotEncoder(dropLast=False, inputCol="vendorIndex", outputCol="vendorVec")
encoded1 = encoder.transform(indexed)

# INDEX AND ENCODE RATE_CODE
stringIndexer = StringIndexer(inputCol="rate_code", outputCol="rateIndex")
model = stringIndexer.fit(encoded1)
indexed = model.transform(encoded1)
encoder = OneHotEncoder(dropLast=False, inputCol="rateIndex", outputCol="rateVec")
encoded2 = encoder.transform(indexed)

# INDEX AND ENCODE PAYMENT_TYPE
stringIndexer = StringIndexer(inputCol="payment_type", outputCol="paymentIndex")
model = stringIndexer.fit(encoded2)
indexed = model.transform(encoded2)
encoder = OneHotEncoder(dropLast=False, inputCol="paymentIndex", outputCol="paymentVec")
encoded3 = encoder.transform(indexed)

# INDEX AND ENCODE TRAFFIC TIME BINS
stringIndexer = StringIndexer(inputCol="TrafficTimeBins", outputCol="TrafficTimeBinsIndex")
model = stringIndexer.fit(encoded3)
indexed = model.transform(encoded3)
encoder = OneHotEncoder(dropLast=False, inputCol="TrafficTimeBinsIndex", outputCol="TrafficTimeBinsVec")
encodedFinal = encoder.transform(indexed)

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 5.37 seconds

Create RDD objects with feature arrays for input into models

This section contains code that shows how to index categorical text data as an RDD object and one-hot encode it so it can be used to train and test MLlib logistic regression and tree-based models. The indexed data is stored in

Resilient Distributed Dataset (RDD) objects. These are the basic abstraction in Spark. An RDD object represents an immutable, partitioned collection of elements that can be operated on in parallel with Spark.

It also contains code that shows how to scale data with the `StandardScalar` provided by MLlib for use in linear regression with Stochastic Gradient Descent (SGD), a popular algorithm for training a wide range of machine learning models. The `StandardScaler` is used to scale the features to unit variance. Feature scaling, also known as data normalization, insures that features with widely disbursed values are not given excessive weigh in the objective function.

```

# CREATE RDD OBJECTS WITH FEATURE ARRAYS FOR INPUT INTO MODELS

# RECORD START TIME
timestart = datetime.datetime.now()

# IMPORT LIBRARIES
from pyspark.mllib.linalg import Vectors
from pyspark.mllib.feature import StandardScaler, StandardScalerModel
from pyspark.mllib.util import MLUtils
from numpy import array

# INDEXING CATEGORICAL TEXT FEATURES FOR INPUT INTO TREE-BASED MODELS
def parseRowIndexingBinary(line):
    features = np.array([line.paymentIndex, line.vendorIndex, line.rateIndex, line.TrafficTimeBinsIndex,
                        line.pickup_hour, line.weekday, line.passenger_count, line.trip_time_in_secs,
                        line.trip_distance, line.fare_amount])
    return features

# ONE-HOT ENCODING OF CATEGORICAL TEXT FEATURES FOR INPUT INTO LOGISTIC REGRESSION MODELS
def parseRowOneHotBinary(line):
    features = np.concatenate(([line.pickup_hour, line.weekday, line.passenger_count,
                               line.trip_time_in_secs, line.trip_distance, line.fare_amount],
                               line.vendorVec.toArray(), line.rateVec.toArray(),
                               line.paymentVec.toArray(), line.TrafficTimeBinsVec.toArray()),
                             axis=0)
    return features

# ONE-HOT ENCODING OF CATEGORICAL TEXT FEATURES FOR INPUT INTO TREE-BASED MODELS
def parseRowIndexingRegression(line):
    features = np.array([line.paymentIndex, line.vendorIndex, line.rateIndex, line.TrafficTimeBinsIndex,
                        line.pickup_hour, line.weekday, line.passenger_count, line.trip_time_in_secs,
                        line.trip_distance, line.fare_amount])
    return features

# INDEXING CATEGORICAL TEXT FEATURES FOR INPUT INTO LINEAR REGRESSION MODELS
def parseRowOneHotRegression(line):
    features = np.concatenate(([line.pickup_hour, line.weekday, line.passenger_count,
                               line.trip_time_in_secs, line.trip_distance, line.fare_amount],
                               line.vendorVec.toArray(), line.rateVec.toArray(),
                               line.paymentVec.toArray(), line.TrafficTimeBinsVec.toArray()),
                             axis=0)
    return features

# FOR BINARY CLASSIFICATION TRAINING AND TESTING
indexedTESTbinary = encodedFinal.map(parseRowIndexingBinary)
oneHotTESTbinary = encodedFinal.map(parseRowOneHotBinary)

# FOR REGRESSION CLASSIFICATION TRAINING AND TESTING
indexedTESTreg = encodedFinal.map(parseRowIndexingRegression)
oneHotTESTreg = encodedFinal.map(parseRowOneHotRegression)

# SCALING FEATURES FOR LINEARREGRESSIONWITHSGD MODEL
scaler = StandardScaler(withMean=False, withStd=True).fit(oneHotTESTreg)
oneHotTESTregScaled = scaler.transform(oneHotTESTreg)

# CACHE RDDS IN MEMORY
indexedTESTbinary.cache();
oneHotTESTbinary.cache();
indexedTESTreg.cache();
oneHotTESTreg.cache();
oneHotTESTregScaled.cache();

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 11.72 seconds

Score with the Logistic Regression Model and save output to blob

The code in this section shows how to load a Logistic Regression Model that has been saved in Azure blob storage and use it to predict whether or not a tip is paid on a taxi trip, score it with standard classification metrics, and then save and plot the results to blob storage. The scored results are stored in RDD objects.

```
# SCORE AND EVALUATE LOGISTIC REGRESSION MODEL

# RECORD START TIME
timestart = datetime.datetime.now()

# IMPORT LIBRARIES
from pyspark.mllib.classification import LogisticRegressionModel

## LOAD SAVED MODEL
savedModel = LogisticRegressionModel.load(sc, logisticRegFileLoc)
predictions = oneHotTESTbinary.map(lambda features: (float(savedModel.predict(features)))))

## SAVE SCORED RESULTS (RDD) TO BLOB
datestamp = unicode(datetime.datetime.now()).replace(' ', '').replace(':', '_');
logisticregressionfilename = "LogisticRegressionWithLBFGS_" + datestamp + ".txt";
dirfilename = scoredResultDir + logisticregressionfilename;
predictions.saveAsTextFile(dirfilename)

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";
```

OUTPUT:

Time taken to execute above cell: 19.22 seconds

Score a Linear Regression Model

We used [LinearRegressionWithSGD](#) to train a linear regression model using Stochastic Gradient Descent (SGD) for optimization to predict the amount of tip paid.

The code in this section shows how to load a Linear Regression Model from Azure blob storage, score using scaled variables, and then save the results back to the blob.

```

#SCORE LINEAR REGRESSION MODEL

# RECORD START TIME
timestart = datetime.datetime.now()

#LOAD LIBRARIES
from pyspark.mllib.regression import LinearRegressionWithSGD, LinearRegressionModel

# LOAD MODEL AND SCORE USING ** SCALED VARIABLES **
savedModel = LinearRegressionModel.load(sc, linearRegFileLoc)
predictions = oneHotTESTregScaled.map(lambda features: (float(savedModel.predict(features)))))

# SAVE RESULTS
datestamp = unicode(datetime.datetime.now()).replace(' ', '').replace(':', '_');
linearregressionfilename = "LinearRegressionWithSGD_" + datestamp;
dirfilename = scoredResultDir + linearregressionfilename;
predictions.saveAsTextFile(dirfilename)

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 16.63 seconds

Score classification and regression Random Forest Models

The code in this section shows how to load the saved classification and regression Random Forest Models saved in Azure blob storage, score their performance with standard classifier and regression measures, and then save the results back to blob storage.

[Random forests](#) are ensembles of decision trees. They combine many decision trees to reduce the risk of overfitting. Random forests can handle categorical features, extend to the multiclass classification setting, do not require feature scaling, and are able to capture non-linearities and feature interactions. Random forests are one of the most successful machine learning models for classification and regression.

[spark.mllib](#) supports random forests for binary and multiclass classification and for regression, using both continuous and categorical features.

```

# SCORE RANDOM FOREST MODELS FOR CLASSIFICATION AND REGRESSION

# RECORD START TIME
timestart = datetime.datetime.now()

#IMPORT MLLIB LIBRARIES
from pyspark.mllib.tree import RandomForest, RandomForestModel

# CLASSIFICATION: LOAD SAVED MODEL, SCORE AND SAVE RESULTS BACK TO BLOB
savedModel = RandomForestModel.load(sc, randomForestClassificationFileLoc)
predictions = savedModel.predict(indexedTESTbinary)

# SAVE RESULTS
datestamp = unicode(datetime.datetime.now()).replace(' ', '').replace(':', '_');
rfclassificationfilename = "RandomForestClassification_" + datestamp + ".txt";
dirfilename = scoredResultDir + rfclassificationfilename;
predictions.saveAsTextFile(dirfilename)

# REGRESSION: LOAD SAVED MODEL, SCORE AND SAVE RESULTS BACK TO BLOB
savedModel = RandomForestModel.load(sc, randomForestRegFileLoc)
predictions = savedModel.predict(indexedTESTreg)

# SAVE RESULTS
datestamp = unicode(datetime.datetime.now()).replace(' ', '').replace(':', '_');
rfregressionfilename = "RandomForestRegression_" + datestamp + ".txt";
dirfilename = scoredResultDir + rfregressionfilename;
predictions.saveAsTextFile(dirfilename)

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 31.07 seconds

Score classification and regression Gradient Boosting Tree Models

The code in this section shows how to load classification and regression Gradient Boosting Tree Models from Azure blob storage, score their performance with standard classifier and regression measures, and then save the results back to blob storage.

spark.mllib supports GBTs for binary classification and for regression, using both continuous and categorical features.

Gradient Boosting Trees (GBTs) are ensembles of decision trees. GBTs train decision trees iteratively to minimize a loss function. GBTs can handle categorical features, do not require feature scaling, and are able to capture non-linearities and feature interactions. They can also be used in a multiclass-classification setting.

```

# SCORE GRADIENT BOOSTING TREE MODELS FOR CLASSIFICATION AND REGRESSION

# RECORD START TIME
timestart = datetime.datetime.now()

#IMPORT MLLIB LIBRARIES
from pyspark.mllib.tree import GradientBoostedTrees, GradientBoostedTreesModel

# CLASSIFICATION: LOAD SAVED MODEL, SCORE AND SAVE RESULTS BACK TO BLOB

#LOAD AND SCORE THE MODEL
savedModel = GradientBoostedTreesModel.load(sc, BoostedTreeClassificationFileLoc)
predictions = savedModel.predict(indexedTESTbinary)

# SAVE RESULTS
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
btclassificationfilename = "GradientBoostingTreeClassification_" + datestamp + ".txt";
dirfilename = scoredResultDir + btclassificationfilename;
predictions.saveAsTextFile(dirfilename)

# REGRESSION: LOAD SAVED MODEL, SCORE AND SAVE RESULTS BACK TO BLOB

# LOAD AND SCORE MODEL
savedModel = GradientBoostedTreesModel.load(sc, BoostedTreeRegressionFileLoc)
predictions = savedModel.predict(indexedTESTreg)

# SAVE RESULTS
datestamp = unicode(datetime.datetime.now()).replace(' ','').replace(':', '_');
btregressionfilename = "GradientBoostingTreeRegression_" + datestamp + ".txt";
dirfilename = scoredResultDir + btregressionfilename;
predictions.saveAsTextFile(dirfilename)

# PRINT HOW MUCH TIME IT TOOK TO RUN THE CELL
timeend = datetime.datetime.now()
timedelta = round((timeend-timestart).total_seconds(), 2)
print "Time taken to execute above cell: " + str(timedelta) + " seconds";

```

OUTPUT:

Time taken to execute above cell: 14.6 seconds

Clean up objects from memory and print scored file locations

```

# UNPERSIST OBJECTS CACHED IN MEMORY
taxi_df_test_cleaned.unpersist()
indexedTESTbinary.unpersist();
oneHotTESTbinary.unpersist();
indexedTESTreg.unpersist();
oneHotTESTreg.unpersist();
oneHotTESTregScaled.unpersist();

# PRINT OUT PATH TO SCORED OUTPUT FILES
print "logisticRegFileLoc: " + logisticregressionfilename;
print "linearRegFileLoc: " + linearregressionfilename;
print "randomForestClassificationFileLoc: " + rfclassificationfilename;
print "randomForestRegFileLoc: " + rfregressionfilename;
print "BoostedTreeClassificationFileLoc: " + btclassificationfilename;
print "BoostedTreeRegressionFileLoc: " + btregressionfilename;

```

OUTPUT:

logisticRegFileLoc: LogisticRegressionWithLBFGS_2016-05-0317_22_38.953814.txt

linearRegFileLoc: LinearRegressionWithSGD_2016-05-0317_22_58.878949

randomForestClassificationFileLoc: RandomForestClassification_2016-05-0317_23_15.939247.txt

randomForestRegFileLoc: RandomForestRegression_2016-05-0317_23_31.459140.txt

BoostedTreeClassificationFileLoc: GradientBoostingTreeClassification_2016-05-0317_23_49.648334.txt

BoostedTreeRegressionFileLoc: GradientBoostingTreeRegression_2016-05-0317_23_56.860740.txt

Consume Spark Models through a web interface

Spark provides a mechanism to remotely submit batch jobs or interactive queries through a REST interface with a component called Livy. Livy is enabled by default on your HDInsight Spark cluster. For more information on Livy, see: [Submit Spark jobs remotely using Livy](#).

You can use Livy to remotely submit a job that batch scores a file that is stored in an Azure blob and then writes the results to another blob. To do this, you upload the Python script from

[GitHub](#) to the blob of the Spark cluster. You can use a tool like **Microsoft Azure Storage Explorer** or **AzCopy** to copy the script to the cluster blob. In our case we uploaded the script to
`wasb:///example/python/ConsumeGBNYCReg.py`.

NOTE

The access keys that you need can be found on the portal for the storage account associated with the Spark cluster.

Once uploaded to this location, this script runs within the Spark cluster in a distributed context. It loads the model and runs predictions on input files based on the model.

You can invoke this script remotely by making a simple HTTPS/REST request on Livy. Here is a curl command to construct the HTTP request to invoke the Python script remotely. Replace CLUSTERLOGIN, CLUSTERPASSWORD, CLUSTERNAME with the appropriate values for your Spark cluster.

```
# CURL COMMAND TO INVOKE PYTHON SCRIPT WITH HTTP REQUEST

curl -k --user "CLUSTERLOGIN:CLUSTERPASSWORD" -X POST --data "{\"file\":
\"wasb:///example/python/ConsumeGBNYCReg.py\"}" -H "Content-Type: application/json"
https://CLUSTERNAME.azurehdinsight.net/livy/batches
```

You can use any language on the remote system to invoke the Spark job through Livy by making a simple HTTPS call with Basic Authentication.

NOTE

It would be convenient to use the Python Requests library when making this HTTP call, but it is not currently installed by default in Azure Functions. So older HTTP libraries are used instead.

Here is the Python code for the HTTP call:

```

#MAKE AN HTTPS CALL ON LIVY.

import os

# OLDER HTTP LIBRARIES USED HERE INSTEAD OF THE REQUEST LIBRARY AS THEY ARE AVAILABLE BY DEFAULT
import httplib, urllib, base64

# REPLACE VALUE WITH ONES FOR YOUR SPARK CLUSTER
host = '<spark cluster name>.azurehdinsight.net:443'
username='<username>'
password='<password>'

#AUTHORIZATION
conn = httplib.HTTPSConnection(host)
auth = base64.encodestring('%s:%s' % (username, password)).replace('\n', '')
headers = {'Content-Type': 'application/json', 'Authorization': 'Basic %s' % auth}

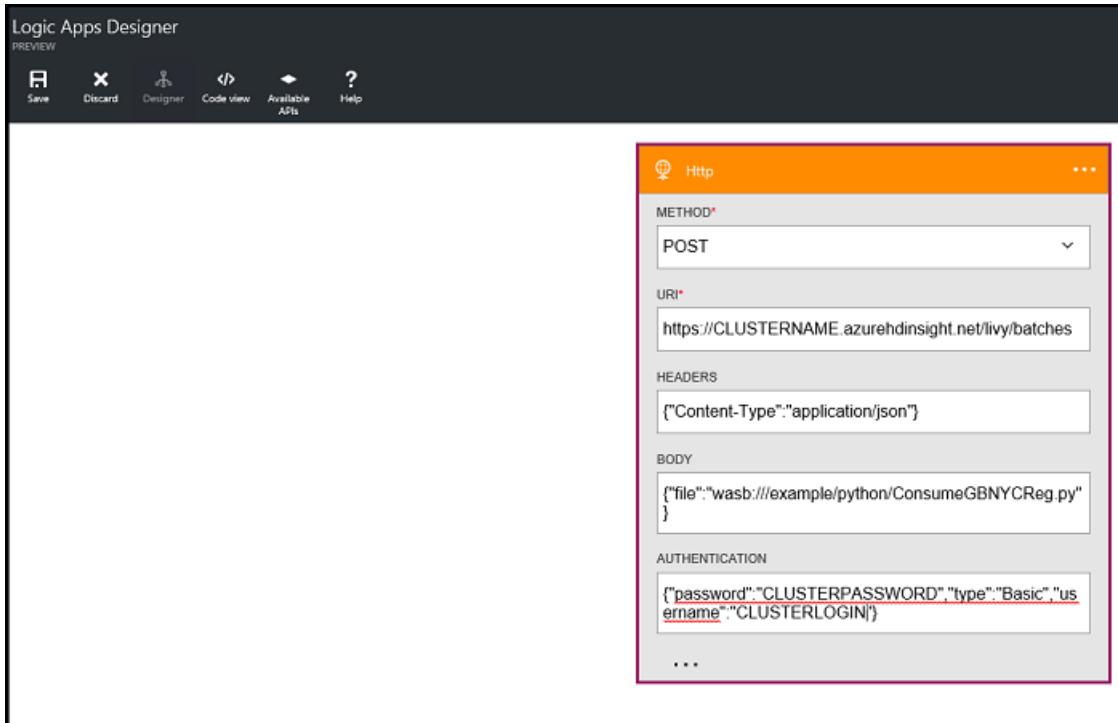
# SPECIFY THE PYTHON SCRIPT TO RUN ON THE SPARK CLUSTER
# IN THE FILE PARAMETER OF THE JSON POST REQUEST BODY
r=conn.request("POST", '/livy/batches', '{"file": "wasb:///example/python/ConsumeGBNYCReg.py"}', headers )
response = conn.getresponse().read()
print(response)
conn.close()

```

You can also add this Python code to [Azure Functions](#) to trigger a Spark job submission that scores a blob based on various events like a timer, creation, or update of a blob.

If you prefer a code free client experience, use the [Azure Logic Apps](#) to invoke the Spark batch scoring by defining an HTTP action on the **Logic Apps Designer** and setting its parameters.

- From Azure portal, create a new Logic App by selecting **+New -> Web + Mobile -> Logic App**.
- To bring up the **Logic Apps Designer**, enter the name of the Logic App and App Service Plan.
- Select an HTTP action and enter the parameters shown in the following figure:



What's next?

Cross-validation and hyperparameter sweeping: See [Advanced data exploration and modeling with Spark](#) on how models can be trained using cross-validation and hyper-parameter sweeping.

HDInsight Hadoop data science walkthroughs using Hive on Azure

3/12/2019 • 2 minutes to read

These walkthroughs use Hive with an HDInsight Hadoop cluster to do predictive analytics. They follow the steps outlined in the Team Data Science Process. For an overview of the Team Data Science Process, see [Data Science Process](#). For an introduction to Azure HDInsight, see [Introduction to Azure HDInsight, the Hadoop technology stack, and Hadoop clusters](#).

Additional data science walkthroughs that execute the Team Data Science Process are grouped by the **platform** that they use. See [Walkthroughs executing the Team Data Science Process](#) for an itemization of these examples.

Predict taxi tips using Hive with HDInsight Hadoop

The [Use HDInsight Hadoop clusters](#) walkthrough uses data from New York taxis to predict:

- Whether a tip is paid
- The distribution of tip amounts

The scenario is implemented using Hive with an [Azure HDInsight Hadoop cluster](#). You learn how to store, explore, and feature engineer data from a publicly available NYC taxi trip and fare dataset. You also use Azure Machine Learning to build and deploy the models.

Predict advertisement clicks using Hive with HDInsight Hadoop

The [Use Azure HDInsight Hadoop Clusters on a 1-TB dataset](#) walkthrough uses a publicly available [Criteo](#) click dataset to predict whether a tip is paid and the range of amounts expected. The scenario is implemented using Hive with an [Azure HDInsight Hadoop cluster](#) to store, explore, feature engineer, and down sample data. It uses Azure Machine Learning to build, train, and score a binary classification model predicting whether a user clicks on an advertisement. The walkthrough concludes showing how to publish one of these models as a Web service.

Next steps

For a discussion of the key components that comprise the Team Data Science Process, see [Team Data Science Process overview](#).

For a discussion of the Team Data Science Process lifecycle that you can use to structure your data science projects, see [Team Data Science Process lifecycle](#). The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

Azure Data Lake data science walkthroughs using U-SQL

1/30/2019 • 2 minutes to read

These walkthroughs use U-SQL with Azure Data Lake to do predictive analytics. They follow the steps outlined in the Team Data Science Process. For an overview of the Team Data Science Process, see [Data Science Process](#). For an introduction to Azure Data Lake, see [Overview of Azure Data Lake Store](#).

Additional data science walkthroughs that execute the Team Data Science Process are grouped by the **platform** that they use. See [Walkthroughs executing the Team Data Science Process](#) for an itemization of these examples.

Predict taxi tips using U-SQL with Azure Data Lake

The [Use Azure Data Lake for data science](#) walkthrough shows how to use Azure Data Lake to do data exploration and binary classification tasks on a sample of the NYC taxi dataset to predict whether or not a tip is paid by a customer.

Next steps

For a discussion of the key components that comprise the Team Data Science Process, see [Team Data Science Process overview](#).

For a discussion of the Team Data Science Process lifecycle that you can use to structure your data science projects, see [Team Data Science Process lifecycle](#). The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

SQL Server data science walkthroughs using R, Python and T-SQL

1/30/2019 • 2 minutes to read

These walkthroughs use SQL Server, SQL Server R Services, and SQL Server Python Services to do predictive analytics. R and Python code is deployed in stored procedures. They follow the steps outlined in the Team Data Science Process. For an overview of the Team Data Science Process, see [Data Science Process](#).

Additional data science walkthroughs that execute the Team Data Science Process are grouped by the **platform** that they use. See [Walkthroughs executing the Team Data Science Process](#) for an itemization of these examples.

Predict taxi tips using Python and SQL queries with SQL Server

The [Use SQL Server](#) walkthrough shows how you build and deploy machine learning classification and regression models using SQL Server and a publicly available NYC taxi trip and fare dataset.

Predict taxi tips using Microsoft R with SQL Server

The [Use SQL Server R Services](#) walkthrough provides data scientists with a combination of R code, SQL Server data, and custom SQL functions to build and deploy an R model to SQL Server. The walkthrough is designed to introduce R developers to R Services (In-Database).

Predict taxi tips using R from T-SQL or stored procedures with SQL Server

The [Data science walkthrough for R and SQL Server](#) provides SQL programmers with experience building an advanced analytics solution with Transact-SQL using SQL Server R Services to operationalize an R solution.

Predict taxi tips using Python in SQL Server stored procedures

The [Use T-SQL with SQL Server Python Services](#) walkthrough provides SQL programmers with experience building a machine learning solution in SQL Server. It demonstrates how to incorporate Python into an application by adding Python code to stored procedures.

Next steps

For a discussion of the key components that comprise the Team Data Science Process, see [Team Data Science Process overview](#).

For a discussion of the Team Data Science Process lifecycle that you can use to structure your data science projects, see [Team Data Science Process lifecycle](#). The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

SQL Data Warehouse data science walkthroughs using T-SQL and Python on Azure

1/30/2019 • 2 minutes to read

These walkthroughs use of SQL Data Warehouse to do predictive analytics. They follow the steps outlined in the Team Data Science Process. For an overview of the Team Data Science Process, see [Data Science Process](#). For an introduction to SQL Data Warehouse, see [What is Azure SQL Data Warehouse?](#)

Additional data science walkthroughs that execute the Team Data Science Process are grouped by the **platform** that they use. See [Walkthroughs executing the Team Data Science Process](#) for an itemization of these examples.

Predict taxi tips using T-SQL and IPython notebooks with SQL Data Warehouse

The [Use SQL Data Warehouse walkthrough](#) shows you how to build and deploy machine learning classification and regression models using SQL Data Warehouse (SQL DW) for a publicly available NYC taxi trip and fare dataset.

Next steps

For a discussion of the key components that comprise the Team Data Science Process, see [Team Data Science Process overview](#).

For a discussion of the Team Data Science Process lifecycle that you can use to structure your data science projects, see [Team Data Science Process lifecycle](#). The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

Team Data Science Process for data scientists

3/12/2019 • 8 minutes to read

This article provides guidance to a set of objectives that are typically used to implement comprehensive data science solutions with Azure technologies. You are guided through:

- understanding an analytics workload
- using the Team Data Science Process
- using Azure Machine Learning
- the foundations of data transfer and storage
- providing data source documentation
- using tools for analytics processing

These training materials are related to the Team Data Science Process (TDSP) and Microsoft and open-source software and toolkits, which are helpful for envisioning, executing and delivering data science solutions.

Lesson Path

You can use the items in the following table to guide your own self-study. Read the *Description* column to follow the path, click on the *Topic* links for study references, and check your skills using the *Knowledge Check* column.

OBJECTIVE	TOPIC	DESCRIPTION	KNOWLEDGE CHECK
Understand the processes for developing analytic projects	An introduction to the Team Data Science Process	We begin by covering an overview of the Team Data Science Process – the TDSP. This process guides you through each step of an analytics project. Read through each of these sections to learn more about the process and how you can implement it.	Review and download the TDSP Project Structure artifacts to your local machine for your project.
	Agile Development	The Team Data Science Process works well with many different programming methodologies. In this Learning Path, we use Agile software development. Read through the "What is Agile Development?" and "Building Agile Culture" articles, which cover the basics of working with Agile. There are also other references at this site where you can learn more.	Explain Continuous Integration and Continuous Delivery to a colleague.

OBJECTIVE	TOPIC	DESCRIPTION	KNOWLEDGE CHECK
	DevOps for Data Science	Developer Operations (DevOps) involves people, processes, and platforms you can use to work through a project and integrate your solution into an organization's standard IT. This integration is essential for adoption, safety, and security. In this online course, you learn about DevOps practices as well as understand some of the toolchain options you have.	Prepare a 30-minute presentation to a technical audience on how DevOps is essential for analytics projects.
Understand the Technologies for Data Storage and Processing	Microsoft Business Analytics and AI	We focus on a few technologies in this Learning Path that you can use to create an analytics solution, but Microsoft has many more. To understand the options you have, it's important to review the platforms and features available in Microsoft Azure, the Azure Stack, and on-premises options. Review this resource to learn the various tools you have available to answer analytics question.	Download and review the presentation materials from this workshop.
Setup and Configure your training, development, and production environments	Microsoft Azure	Now let's create an account in Microsoft Azure for training and learn how to create development and test environments. These free training resources get you started. Complete the "Beginner" and "Intermediate" paths.	If you do not have an Azure Account, create one. Log in to the Microsoft Azure portal and create one Resource Group for training.
	The Microsoft Azure Command-Line Interface (CLI)	There are multiple ways of working with Microsoft Azure – from graphical tools like VSCode and Visual Studio, to Web interfaces such as the Azure portal, and from the command line, such as Azure PowerShell commands and functions. In this article, we cover the Command-Line Interface (CLI), which you can use locally on your workstation, in Windows and other Operating Systems, as well as in the Azure portal.	Set your default subscription with the Azure CLI.

OBJECTIVE	TOPIC	DESCRIPTION	KNOWLEDGE CHECK
	Microsoft Azure Storage	You need a place to store your data. In this article, you learn about Microsoft Azure's storage options, how to create a storage account, and how to copy or move data to the cloud. Read through this introduction to learn more.	Create a Storage Account in your training Resource Group, create a container for a Blob object, and upload and download data.
	Microsoft Azure Active Directory	Microsoft Azure Active Directory (AAD) forms the basis of securing your application. In this article, you learn more about accounts, rights, and permissions. Active Directory and security are complex topics, so just read through this resource to understand the fundamentals.	Add one user to Azure Active Directory. NOTE: You may not have permissions for this action if you are not the administrator for the subscription. If that's the case, simply review this tutorial to learn more .
	The Microsoft Azure Data Science Virtual Machine	You can install the tools for working with Data Science locally on multiple operating systems. But the Microsoft Azure Data Science Virtual Machine (DSVM) contains all of the tools you need and plenty of project samples to work with. In this article, you learn more about the DVSM and how to work through its examples. This resource explains the Data Science Virtual Machine, how you can create one, and a few options for developing code with it. It also contains all the software you need to complete this learning path – so make sure you complete the Knowledge Path for this topic.	Create a Data Science Virtual Machine and work through at least one lab.
Install and Understand the tools and technologies for working with Data Science solutions			

OBJECTIVE	TOPIC	DESCRIPTION	KNOWLEDGE CHECK
Working with git	To follow our DevOps process with the TDSP, we need to have a version-control system. Microsoft Azure Machine Learning service uses git, a popular open-source distributed repository system. In this article, you learn more about how to install, configure, and work with git and a central repository – GitHub.	Clone this GitHub project for your learning path project structure.	
	VSCode	VSCode is a cross-platform Integrated Development Environment (IDE) that you can use with multiple languages and Azure tools. You can use this single environment to create your entire solution. Watch these introductory videos to get started.	Install VSCode, and work through the VS Code features in the Interactive Editor Playground.
	Programming with Python	In this solution we use Python, one of the most popular languages in Data Science. This article covers the basics of writing analytic code with Python, and resources to learn more. Work through sections 1-9 of this reference, then check your knowledge.	Add one entity to an Azure Table using Python.
	Working with Notebooks	Notebooks are a way of introducing text and code in the same document. Azure Machine Learning service work with Notebooks, so it is beneficial to understand how to use them. Read through this tutorial and give it a try in the Knowledge Check section.	Open this page, and click on the "Welcome to Python.ipynb" link. Work through the examples on that page.
	Machine Learning		

OBJECTIVE	TOPIC	DESCRIPTION	KNOWLEDGE CHECK
Creating advanced Analytic solutions involves working with data, using Machine Learning, which also forms the basis of working with Artificial Intelligence and Deep Learning. This course teaches you more about Machine Learning. For a comprehensive course on Data Science, check out this certification.	Locate a resource on Machine Learning Algorithms. (Hint: Search on "azure machine learning algorithm cheat sheet")		
	scikit-learn	The scikit-learn set of tools allows you to perform data science tasks in Python. We use this framework in our solution. This article covers the basics and explains where you can learn more.	Using the Iris dataset, persist an SVM model using Pickle.
	Working with Docker	Docker is a distributed platform used to build, ship, and run applications, and is used frequently in Azure Machine Learning service. This article covers the basics of this technology and explains where you can go to learn more.	Open Visual Studio Code, and install the Docker Extension . Create a simple Node Docker container.
	HDInsight	HDInsight is the Hadoop open-source infrastructure, available as a service in Microsoft Azure. Your Machine Learning algorithms may involve large sets of data, and HDInsight has the ability to store, transfer and process data at large scale. This article covers working with HDInsight.	Create a small HDInsight cluster . Use HiveQL statements to project columns onto an /example/data/sample.log file . Alternatively, you can complete this knowledge check on your local system .

OBJECTIVE	TOPIC	DESCRIPTION	KNOWLEDGE CHECK
Create a Data Processing Flow from Business Requirements	Determining the Question, following the TDSP	With the development environment installed and configured, and the understanding of the technologies and processes in place, it's time to put everything together using the TDSP to perform an analysis. We need to start by defining the question, selecting the data sources, and the rest of the steps in the Team Data Science Process. Keep in mind the DevOps process as we work through this process. In this article, you learn how to take the requirements from your organization and create a data flow map through your application to define your solution using the Team Data Science Process.	
Locate a resource on " The 5 data science questions " and describe one question your organization might have in these areas. Which algorithms should you focus on for that question?			
Use Azure Machine Learning service to create a predictive solution	Azure Machine Learning service	Microsoft Azure Machine Learning includes working with data sources, using AI for data wrangling and feature engineering, creating experiments, and tracking model runs. All of this works in a single environment and most functions can run locally or in Azure. You can use the PyTorch, TensorFlow, and other frameworks to create your experiments. In this article, we focus on a complete example of this process, using everything you've learned so far.	

OBJECTIVE	TOPIC	DESCRIPTION	KNOWLEDGE CHECK
Use Power BI to visualize results	Power BI	Power BI is Microsoft's data visualization tool. It is available on multiple platforms from Web to mobile devices and desktop computers. In this article you learn how to work with the output of the solution you've created by accessing the results from Azure storage and creating visualizations using Power BI.	Complete this tutorial on Power BI . Then connect Power BI to the Blob CSV created in an experiment run.
Monitor your Solution	Application Insights	There are multiple tools you can use to monitor your end solution. Azure Application Insights makes it easy to integrate built-in monitoring into your solution.	Set up Application Insights to monitor an Application .
	Azure Monitor logs	Another method to monitor your application is to integrate it into your DevOps process. The Azure Monitor logs system provides a rich set of features to help you watch your analytic solutions after you deploy them.	Complete this tutorial on using Azure Monitor logs .
Complete this Learning Path		Congratulations! You've completed this learning path. There is a lot more to learn.	

Next steps

[Team Data Science Process for Developer Operations](#) This article explores the Developer Operations (DevOps) functions that are specific to an Advanced Analytics and Cognitive Services solution implementation.

Team Data Science Process for Developer Operations

2/27/2019 • 9 minutes to read

This article explores the Developer Operations (DevOps) functions that are specific to an Advanced Analytics and Cognitive Services solution implementation. These training materials are related to the Team Data Science Process (TDSP) and Microsoft and open-source software and toolkits, which are helpful for envisioning, executing and delivering data science solutions. It references topics that cover the DevOps Toolchain that is specific to Data Science and AI projects and solutions.

Lesson Path

The following table provides guidance at specified levels to help complete the DevOps objectives that are needed to implement data science solutions with Azure technologies.

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
Understand Advanced Analytics	The Team Data Science Process Lifecycle	This technical walkthrough describes the Team Data Science Process	Data Science	Intermediate	General technology background, familiarity with data solutions, Familiarity with IT projects and solution implementation
Understand the Microsoft Azure Platform for Advanced Analytics	Information Management				
This reference gives and overview of Azure Data Factory to build pipelines to collect and orchestrate data from the services you use for analysis	Microsoft Azure Data Factory	Experienced	General technology background, familiarity with data solutions, Familiarity with IT projects and solution implementation		

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
This reference covers an overview of the Azure Data Catalog which you can use to document and manage metadata on your data sources	Microsoft Azure Data Catalog	Intermediate	General technology background, familiarity with data solutions, familiarity with Relational Database Management Systems (RDBMS) and NoSQL data sources		
This reference covers an overview of the Azure Event Hubs system and how you and use it to ingest data into your solution	Azure Event Hubs	Intermediate	General technology background, familiarity with data solutions, familiarity with Relational Database Management Systems (RDBMS) and NoSQL data sources, familiarity with the Internet of Things (IoT) terminology and use		
	Big Data Stores				
This reference covers an overview of using the Azure SQL Data Warehouse to store and process large amounts of data	Azure SQL Data Warehouse	Experienced	General technology background, familiarity with data solutions, familiarity with Relational Database Management Systems (RDBMS) and NoSQL data sources, familiarity with HDFS terminology and use		

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
		This reference covers an overview of using Azure Data Lake to capture data of any size, type, and ingestion speed in one single place for operational and exploratory analytics	Azure Data Lake Store	Intermediate	General technology background, familiarity with data solutions, familiarity with NoSQL data sources, familiarity with HDFS
	Machine Learning and Analytics	This reference covers an introduction to Machine Learning, predictive analytics, and Artificial Intelligence systems	Azure Machine Learning	Intermediate	General technology background, familiarity with data solutions, familiarity with Data Science terms, familiarity with Machine Learning and Artificial Intelligence terms
		This article provides an introduction to Azure HDInsight, a cloud distribution of the Hadoop technology stack. It also covers what a Hadoop cluster is and when you would use it	Azure HDInsight	Intermediate	General technology background, familiarity with data solutions, familiarity with NoSQL data sources
		This reference covers an overview of the Azure Data Lake Analytics job service	Azure Data Lake Analytics	Intermediate	General technology background, familiarity with data solutions, familiarity with NoSQL data sources
		This overview covers using Azure Stream Analytics as a fully-managed event-processing engine to perform real-time analytic computations on streaming data	Azure Stream Analytics	Intermediate	General technology background, familiarity with data solutions, familiarity with structured and unstructured data concepts

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
	Intelligence	This reference covers an overview of the available Cognitive Services (such as vision, text, and search) and how to get started using them	Cognitive Services	Experienced	General technology background, familiarity with data solutions, software development
		This reference covers and introduction to the Microsoft Bot Framework and how to get started using it	Bot Framework	Experienced	General technology background, familiarity with data solutions
	Visualization	This self-paced, online course covers the Power BI system, and how to create and publish reports	Microsoft Power BI	Beginner	General technology background, familiarity with data solutions
	Solutions	This resource page covers multiple applications you can review, test and implement to see a complete solution from start to finish	Microsoft Azure, Azure Machine Learning, Cognitive Services, Microsoft R, Azure Search, Python, Azure Data Factory, Power BI, Azure Document DB, Application Insights, Azure SQL DB, Azure SQL Data Warehouse, Microsoft SQL Server, Azure Data Lake, Cognitive Services, Bot Framework, Azure Batch,	Intermediate	General technology background, familiarity with data solutions

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
Understand and Implement DevOps Processes	DevOps Fundamentals	This video series explains the covers the fundamentals of DevOps and helps you understand how they map to DevOps practices, and how they can be implemented by a variety of products and tools	DevOps, Microsoft Azure Platform, Azure DevOps	Experienced	Used an SDLC, familiarity with Agile and other Development Frameworks, IT Operations Familiarity
Use the DevOps Toolchain for Data Science	Configure	This reference covers the basics of choosing the proper visualization in Visio to communicate your project design	Visio	Intermediate	General technology background, familiarity with data solutions
		This reference describes the Azure Resource Manager, terms, and serves as the primary root source for samples, getting started, and other references	Azure Resource Manager, Azure PowerShell, Azure CLI	Intermediate	General technology background, familiarity with data solutions
		This reference explains the Azure Data Science Virtual Machines for Linux and Windows	Data Science Virtual Machine	Experienced	Familiarity with Data Science Workloads, Linux
		This walkthrough explains configuring Azure cloud service roles with Visual Studio - pay close attention to the connection strings specifically for storage accounts	Visual Studio	Intermediate	Software Development

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
		This series teaches you how to use Microsoft Project to schedule time, resources and goals for an Advanced Analytics project	Microsoft Project	Intermediate	Understand Project Management Fundamentals
		This Microsoft Project template provides a time, resources and goals tracking for an Advanced Analytics project	Microsoft Project	Intermediate	Understand Project Management Fundamentals
		This tutorial helps you get started with Azure Data Catalog, a fully managed cloud service that serves as a system of registration and system of discovery for enterprise data assets	Azure Data Catalog	Beginner	Familiarity with Data Sources and Structures
		This Microsoft Virtual Academy course explains how to set up Dev/Test with Visual Studio Online and Microsoft Azure	Visual Studio Online	Experienced	Software Development, familiarity with Dev/Test environments
		This Management Pack download for Microsoft System Center contains a Guidelines Document to assist in working with Azure assets	System Center	Intermediate	Experience with System Center for IT Management

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
		This document is intended for developer and operations teams to understand the benefits of PowerShell Desired State Configuration	PowerShell DSC	Intermediate	Experience with PowerShell coding, enterprise architectures, scripting
	Code	This download also contains documentation on using Visual Studio Online Code for creating Data Science and AI applications	Visual Studio Online	Intermediate	Software Development
		This getting started site teaches you about DevOps and Visual Studio	Visual Studio	Beginner	Software Development
		You can write code directly from the Azure Portal using the App Service Editor. Learn more at this resource about Continuous Integration with this tool	Azure Portal	Highly Experienced	Data Science background - but read this anyway
		This resource explains how to code and create Predictive Analytics experiments using the web-based Azure ML Studio tool	Azure ML Studio	Experienced	Software Development
		This reference contains a list and a study link to all of the development tools on the Data Science Virtual Machine in Azure	Data Science Virtual Machine	Experienced	Software Development, Data Science

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
		Read and understand each of the references in this Azure Security Trust Center for Security, Privacy, and Compliance - VERY important	Azure Security	Intermediate	System Architecture Experience, Security Development experience
	Build	This course teaches you about enabling DevOps Practices with Visual Studio Online Build	Visual Studio Online	Experienced	Software Development, Familiarity with an SDLC
		This reference explains compiling and building using Visual Studio	Visual Studio	Intermediate	Software Development, Familiarity with an SDLC
		This reference explains how to orchestrate processes such as software builds with Runbooks	System Center	Experienced	Experience with System Center Orchestrator
	Test	Use this reference to understand how to use Visual Studio Online for Test Case Management	Visual Studio Online	Experienced	Software Development, Familiarity with an SDLC
		Use this previous reference for Runbooks to automate tests using System Center	System Center	Experienced	Experience with System Center Orchestrator
		As part of not only testing but development, you should build in Security. The Microsoft SDL Threat Modeling Tool can help in all phases. Learn more and download it here	Threat Monitoring Tool	Experienced	Familiarity with security concepts, software development

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
		This article explains how to use the Microsoft Attack Surface Analyzer to test your Advanced Analytics solution	Attack Surface Analyzer	Experienced	Familiarity with security concepts, software development
	Package	This reference explains the concepts of working with Packages in TFS and VSO	Visual Studio Online	Experienced	Software development, familiarity with an SDLC
		Use this previous reference for Runbooks to automate packaging using System Center	System Center	Experienced	Experience with System Center Orchestrator
		This reference explains how to create a data pipeline for your solution, which you can save as a JSON template as a "package"	Azure Data Factory	Intermediate	General computing background, data project experience
		This topic describes the structure of an Azure Resource Manager template	Azure Resource Manager	Intermediate	Familiarity with the Microsoft Azure Platform
		DSC is a management platform in PowerShell that enables you to manage your IT and development infrastructure with configuration as code, saved as a package. This reference is an overview for that topic	PowerShell Desired State Configuration	Intermediate	PowerShell coding, familiarity with enterprise architectures, scripting

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
	Release	This head-reference article contains concepts for build, test, and release for CI/CD environments	Visual Studio Online	Experienced	Software development, familiarity with CI/CD environments, familiarity with an SDLC
		Use this previous reference for Runbooks to automate release management using System Center	System Center	Experienced	Experience with System Center Orchestrator
		This article helps you determine the best option to deploy the files for your web app, mobile app backend, or API app to Azure App Service, and then guides you to appropriate resources with instructions specific to your preferred option	Microsoft Azure Deployment	Intermediate	Software development, experience with the Microsoft Azure platform
	Monitor	This reference explains Application Insights and how you can add it to your Advanced Analytics Solutions	Application Insights	Intermediate	Software Development, familiarity with the Microsoft Azure platform
		This topic explains basic concepts about Operations Manager for the administrator who manages the Operations Manager infrastructure and the operator who monitors and supports the Advanced Analytics Solution	System Center	Experienced	Familiarity with enterprise monitoring, System Center Operations Manager

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
		This blog entry explains how to use the Azure Data Factory to monitor and manage the Advanced Analytics pipeline	Azure Data Factory	Intermediate	Familiarity with Azure Data Factory
		This video shows how to monitor a log with Azure Monitor logs	Azure Logs, PowerShell	Experienced	Familiarity with the Azure Platform
Understand how to use Open Source DevOps Tools with DevOps on Azure	Open Source DevOps Tools and Azure	This reference page contains two videos and a whitepaper on using Chef with Azure deployments	Chef	Experienced	Familiarity with the Azure Platform, Familiarity with DevOps
		This site has a toolchain selection path	DevOps, Microsoft Azure Platform, Azure DevOps, Open Source Software	Experienced	Used an SDLC, familiarity with Agile and other Development Frameworks, IT Operations Familiarity
		This tutorial explains how to automate the build and test phase of application development, using a continuous integration and deployment CI/CD pipeline	Jenkins	Experienced	Familiarity with the Azure Platform, Familiarity with DevOps, Familiarity with Jenkins
		This contains an overview of working with Docker and Azure as well as additional references for implementation for Data Science applications	Docker	Intermediate	Familiarity with the Azure Platform, Familiarity with Server Operating Systems

OBJECTIVE	TOPIC	RESOURCE	TECHNOLOGIES	LEVEL	PREREQUISITES
		This installation and explanation explains how to use Visual Studio Code with Azure assets	VSCODE	Intermediate	Software Development, familiarity with the Microsoft Azure Platform
		This blog entry explains how to use R Studio with Microsoft R	R Studio	Intermediate	R Language experience
		This blog entry shows how to use continuous integration with Azure and GitHub	Git, GitHub	Intermediate	Software Development

Next steps

[Team Data Science Process for data scientists](#) This article provides guidance to a set of objectives that are typically used to implement comprehensive data science solutions with Azure technologies.

Set up data science environments for use in the Team Data Science Process

1/30/2019 • 2 minutes to read

The Team Data Science Process uses various data science environments for the storage, processing, and analysis of data. They include Azure Blob Storage, several types of Azure virtual machines, HDInsight (Hadoop) clusters, and Azure Machine Learning workspaces. The decision about which environment to use depends on the type and quantity of data to be modeled and the target destination for that data in the cloud.

- For guidance on questions to consider when making this decision, see [Plan Your Azure Machine Learning Data Science Environment](#).
- For a catalog of some of the scenarios you might encounter when doing advanced analytics, see [Scenarios for the Team Data Science Process](#)

The following articles describe how to set up the various data science environments used by the Team Data Science Process.

- [Azure storage-account](#)
- [HDInsight \(Hadoop\) cluster](#)
- [Azure Machine Learning Studio workspace](#)

The **Microsoft Data Science Virtual Machine (DSVM)** is also available as an Azure virtual machine (VM) image. This VM is pre-installed and configured with several popular tools that are commonly used for data analytics and machine learning. The DSVM is available on both Windows and Linux. For more information, see [Introduction to the cloud-based Data Science Virtual Machine for Linux and Windows](#).

Learn how to create:

- [Windows DSVM](#)
- [Ubuntu DSVM](#)
- [CentOS DSVM](#)
- [Deep Learning VM](#)

2 minutes to read

Platforms and tools for data science projects

1/30/2019 • 9 minutes to read

Microsoft provides a full spectrum of data and analytics services and resources for both cloud or on-premises platforms. They can be deployed to make the execution of your data science projects efficient and scalable. Guidance for teams implementing data science projects in a trackable, version controlled, and collaborative way is provided by the [Team Data Science Process](#) (TDSP). For an outline of the personnel roles, and their associated tasks that are handled by a data science team standardizing on this process, see [Team Data Science Process roles and tasks](#).

The data and analytics services available to data science teams using the TDSP include:

- Data Science Virtual Machines (both Windows and Linux CentOS)
- HDInsight Spark Clusters
- SQL Data Warehouse
- Azure Data Lake
- HDInsight Hive Clusters
- Azure File Storage
- SQL Server 2016 R Services

In this document, we briefly describe the resources and provide links to the tutorials and walkthroughs the TDSP teams have published. They can help you learn how to use them step by step and start using them to build your intelligent applications. More information on these resources is available on their product pages.

Data Science Virtual Machine (DSVM)

The data science virtual machine offered on both Windows and Linux by Microsoft, contains popular tools for data science modeling and development activities. It includes tools such as:

- Microsoft R Server Developer Edition
- Anaconda Python distribution
- Jupyter notebooks for Python and R
- Visual Studio Community Edition with Python and R Tools on Windows / Eclipse on Linux
- Power BI desktop for Windows
- SQL Server 2016 Developer Edition on Windows / Postgres on Linux

It also includes **ML and AI tools** like CNTK (an Open Source Deep Learning toolkit from Microsoft), xgboost, mxnet and Vowpal Wabbit.

Currently DSVM is available in **Windows** and **Linux CentOS** operating systems. Choose the size of your DSVM (number of CPU cores and the amount of memory) based on the needs of the data science projects that you are planning to execute on it.

For more information on Windows edition of DSVM, see [Microsoft Data Science Virtual Machine](#) on the Azure marketplace. For the Linux edition of the DSVM, see [Linux Data Science Virtual Machine](#).

To learn how to execute some of the common data science tasks on the DSVM efficiently, see [Ten things you can do on the Data science Virtual Machine](#)

Azure HDInsight Spark clusters

Apache Spark is an open-source parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications. The Spark processing engine is built for speed, ease of use, and sophisticated analytics. Spark's in-memory computation capabilities make it a good choice for iterative algorithms in machine learning and for graph computations. Spark is also compatible with Azure Blob storage (WASB), so your existing data stored in Azure can easily be processed using Spark.

When you create a Spark cluster in HDInsight, you create Azure compute resources with Spark installed and configured. It takes about 10 minutes to create a Spark cluster in HDInsight. Store the data to be processed in Azure Blob storage. For information on using Azure Blob Storage with a cluster, see [Use HDFS-compatible Azure Blob storage with Hadoop in HDInsight](#).

TDSP team from Microsoft has published two end-to-end walkthroughs on how to use Azure HDInsight Spark Clusters to build data science solutions, one using Python and the other Scala. For more information on Azure HDInsight **Spark Clusters**, see [Overview: Apache Spark on HDInsight Linux](#). To learn how to build a data science solution using **Python** on an Azure HDInsight Spark Cluster, see [Overview of Data Science using Spark on Azure HDInsight](#). To learn how to build a data science solution using **Scala** on an Azure HDInsight Spark Cluster, see [Data Science using Scala and Spark on Azure](#).

Azure SQL Data Warehouse

Azure SQL Data Warehouse allows you to scale compute resources easily and in seconds, without over-provisioning or over-paying. It also offers the unique option to pause the use of compute resources, giving you the freedom to better manage your cloud costs. The ability to deploy scalable compute resources makes it possible to bring all your data into Azure SQL Data Warehouse. Storage costs are minimal and you can run compute only on the parts of datasets that you want to analyze.

For more information on Azure SQL Data Warehouse, see the [SQL Data Warehouse](#) website. To learn how to build end-to-end advanced analytics solutions with SQL Data Warehouse, see [The Team Data Science Process in action: using SQL Data Warehouse](#).

Azure Data Lake

Azure data lake is as an enterprise-wide repository of every type of data collected in a single location, prior to any formal requirements or schema being imposed. This flexibility allows every type of data to be kept in a data lake, regardless of its size or structure or how fast it is ingested. Organizations can then use Hadoop or advanced analytics to find patterns in these data lakes. Data lakes can also serve as a repository for lower-cost data preparation before curating the data and moving it into a data warehouse.

For more information on Azure Data Lake, see [Introducing Azure Data Lake](#). To learn how to build a scalable end-to-end data science solution with Azure Data Lake, see [Scalable Data Science in Azure Data Lake: An end-to-end Walkthrough](#)

Azure HDInsight Hive (Hadoop) clusters

Apache Hive is a data warehouse system for Hadoop, which enables data summarization, querying, and the analysis of data using HiveQL, a query language similar to SQL. Hive can be used to interactively explore your data or to create reusable batch processing jobs.

Hive allows you to project structure on largely unstructured data. After you define the structure, you can use Hive to query that data in a Hadoop cluster without having to use, or even know, Java or MapReduce. HiveQL (the Hive query language) allows you to write queries with statements that are similar to T-SQL.

For data scientists, Hive can run Python User-Defined Functions (UDFs) in Hive queries to process records. This ability extends the capability of Hive queries in data analysis considerably. Specifically, it allows data scientists to conduct scalable feature engineering in languages they are mostly familiar with: the SQL-like HiveQL and Python.

For more information on Azure HDInsight Hive Clusters, see [Use Hive and HiveQL with Hadoop in HDInsight](#). To learn how to build a scalable end-to-end data science solution with Azure HDInsight Hive Clusters, see [The Team Data Science Process in action: using HDInsight Hadoop clusters](#).

Azure File Storage

Azure File Storage is a service that offers file shares in the cloud using the standard Server Message Block (SMB) Protocol. Both SMB 2.1 and SMB 3.0 are supported. With Azure File storage, you can migrate legacy applications that rely on file shares to Azure quickly and without costly rewrites. Applications running in Azure virtual machines or cloud services or from on-premises clients can mount a file share in the cloud, just as a desktop application mounts a typical SMB share. Any number of application components can then mount and access the File storage share simultaneously.

Especially useful for data science projects is the ability to create an Azure file store as the place to share project data with your project team members. Each of them then has access to the same copy of the data in the Azure file storage. They can also use this file storage to share feature sets generated during the execution of the project. If the project is a client engagement, your clients can create an Azure file storage under their own Azure subscription to share the project data and features with you. In this way, the client has full control of the project data assets. For more information on Azure File Storage, see [Get started with Azure File storage on Windows](#) and [How to use Azure File Storage with Linux](#).

SQL Server 2016 R Services

R Services (In-database) provide a platform for developing and deploying intelligent applications that can uncover new insights. You can use the rich and powerful R language, including the many packages provided by the R community, to create models and generate predictions from your SQL Server data. Because R Services (In-database) integrate the R language with SQL Server, analytics are kept close to the data, which eliminates the costs and security risks associated with moving data.

R Services (In-database) support the open source R language with a comprehensive set of SQL Server tools and technologies. They offer superior performance, security, reliability, and manageability. You can deploy R solutions using convenient and familiar tools. Your production applications can call the R runtime and retrieve predictions and visuals using Transact-SQL. You also use the ScaleR libraries to improve the scale and performance of your R solutions. For more information, see [SQL Server R Services](#).

The TDSP team from Microsoft has published two end-to-end walkthroughs that show how to build data science solutions in SQL Server 2016 R Services: one for R programmers and one for SQL developers. For **R Programmers**, see [Data Science End-to-End Walkthrough](#). For **SQL Developers**, see [In-Database Advanced Analytics for SQL Developers \(Tutorial\)](#).

Appendix: Tools to set up data science projects

Install Git Credential Manager on Windows

If you are following the TDSP on **Windows**, you need to install the **Git Credential Manager (GCM)** to communicate with the Git repositories. To install GCM, you first need to install **Chocolatey**. To install Chocolatey and the GCM, run the following commands in Windows PowerShell as an **Administrator**:

```
iwr https://chocolatey.org/install.ps1 -UseBasicParsing | iex  
choco install git-credential-manager-for-windows -y
```

Install Git on Linux (CentOS) machines

Run the following bash command to install Git on Linux (CentOS) machines:

```
sudo yum install git
```

Generate public SSH key on Linux (CentOS) machines

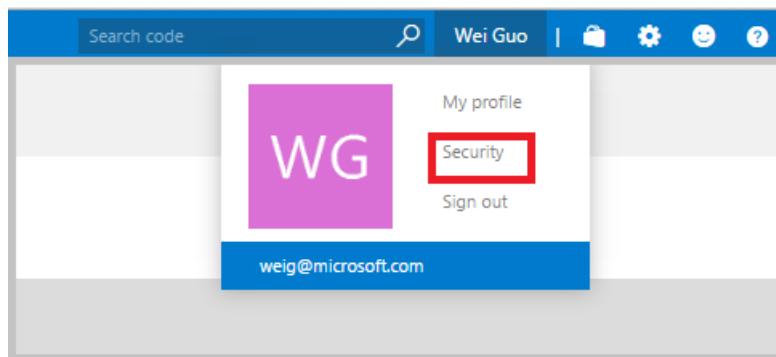
If you are using Linux (CentOS) machines to run the git commands, you need to add the public SSH key of your machine to your Azure DevOps Services, so that this machine is recognized by the Azure DevOps Services. First, you need to generate a public SSH key and add the key to SSH public keys in your Azure DevOps Services security setting page.

- To generate the SSH key, run the following two commands:

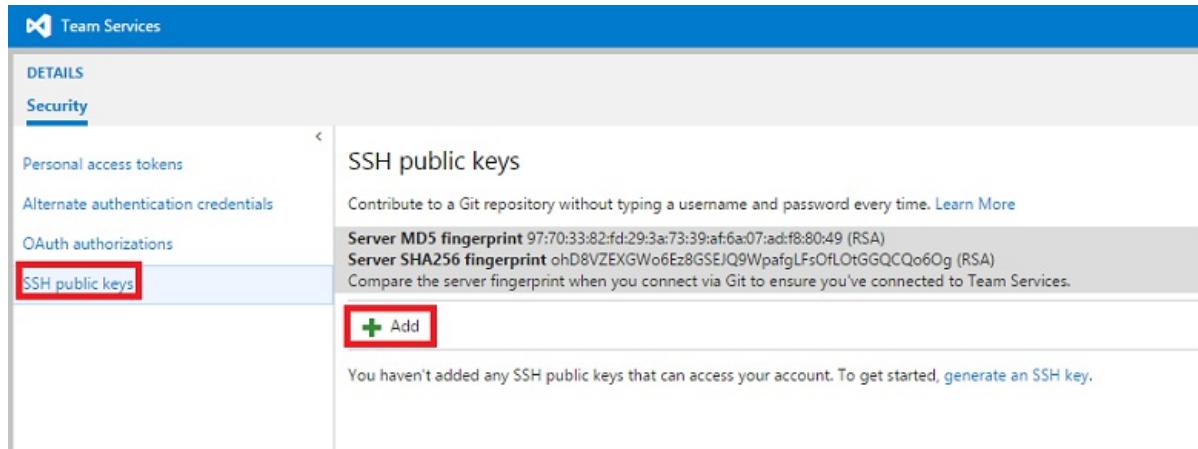
```
ssh-keygen  
cat .ssh/id_rsa.pub
```

```
[dsl@weiglinuxdsv3 ~]$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/dsl/.ssh/id_rsa):
Created directory '/home/dsl/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/dsl/.ssh/id_rsa.
Your public key has been saved in /home/dsl/.ssh/id_rsa.pub.
The key fingerprint is:
04:0f:bf:29:79:0e:7b:f3:99:c9:ff:6f:92:bb:75 dsl@weiglinuxdsv3
The key's randomart image is:
+---[ RSA 2048]----+
|          .         |
|          .         |
|          .         |
|          .         |
|          .         |
|          .         |
|          .         |
|          .         |
|          .         |
|          .         |
+---[ RSA 2048]----+
[dsl@weiglinuxdsv3 ~]$ cat .ssh/id_rsa.pub
ssh-rsa AAAQABAAQdAMMwAQB5hJLwvSnlWtAtQo0nFGCbgpuP0VQkqAf1z5SSSLgZjFpmUJ3dUh0iX2R0B0lfcGraIVdPgdtNC4/OY4jhx2ldx1dxEWpSLUBGcAE2xR1BMy0fz2In18cfJbZqgAHObt0H9pSPR3seUa/PmGt
iZLP1mNC7cN1S10l0KBMs/BNeJrz454tNU+sB025XNyvHc148hR8Udj129kOZg935eRu09YTgj8x2EdkpuEPxFaEzjssbeiuT5cUFreiE+JldBFdx4Lcok8yfZoENMcISGHrrcB3Jxqy05ibSMh4o4og17eb1SP18Fc5oM0PfqTlIEip2bcvA8
3HoX[dsl@weiglinuxdsv3 ~]$
```

- Copy the entire ssh key including `ssh-rsa`.
 - Log in to your Azure DevOps Services.
 - Click **<Your Name>** at the top right corner of the page and click **security**.



- Click **SSH public keys**, and click **+Add**.



- Paste the ssh key just copied into the text box and save.

Next steps

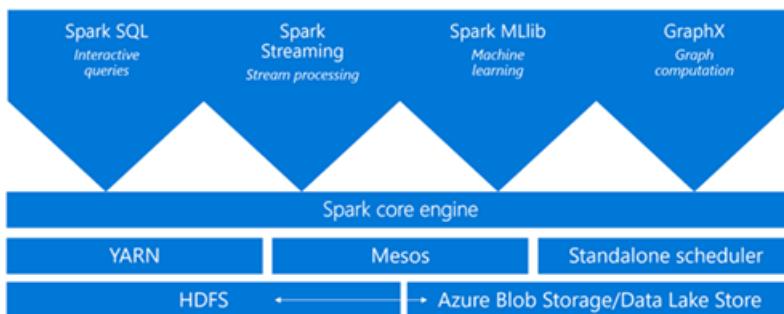
Full end-to-end walkthroughs that demonstrate all the steps in the process for **specific scenarios** are also provided. They are listed and linked with thumbnail descriptions in the [Example walkthroughs](#) topic. They illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

For examples executing steps in the Team Data Science Process that use Azure Machine Learning Studio, see the [With Azure ML](#) learning path.

What is Apache Spark in Azure HDInsight

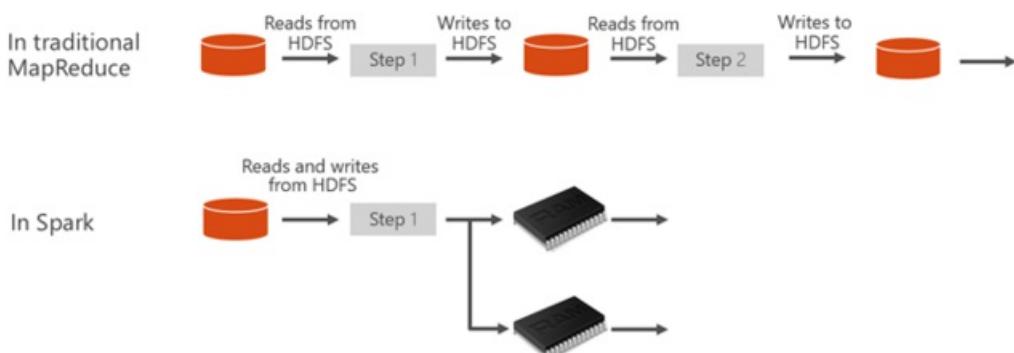
2/27/2019 • 6 minutes to read

Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications. Apache Spark in Azure HDInsight is the Microsoft implementation of Apache Spark in the cloud. HDInsight makes it easier to create and configure a Spark cluster in Azure. Spark clusters in HDInsight are compatible with Azure Storage and Azure Data Lake Storage. So you can use HDInsight Spark clusters to process your data stored in Azure. For the components and the versioning information, see [Apache Hadoop components and versions in Azure HDInsight](#).



What is Apache Spark?

Spark provides primitives for in-memory cluster computing. A Spark job can load and cache data into memory and query it repeatedly. In-memory computing is much faster than disk-based applications, such as Hadoop, which shares data through Hadoop distributed file system (HDFS). Spark also integrates into the Scala programming language to let you manipulate distributed data sets like local collections. There's no need to structure everything as map and reduce operations.



Spark clusters in HDInsight offer a fully managed Spark service. Benefits of creating a Spark cluster in HDInsight are listed here.

FEATURE	DESCRIPTION
Ease creation	You can create a new Spark cluster in HDInsight in minutes using the Azure portal, Azure PowerShell, or the HDInsight .NET SDK. See Get started with Apache Spark cluster in HDInsight .

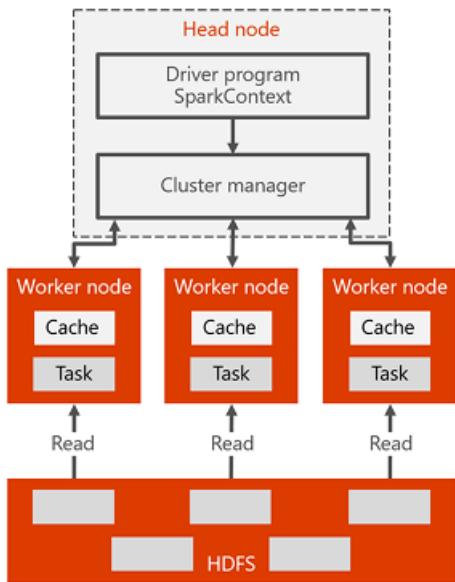
FEATURE	DESCRIPTION
Ease of use	Spark cluster in HDInsight include Jupyter and Apache Zeppelin notebooks. You can use these notebooks for interactive data processing and visualization.
REST APIs	Spark clusters in HDInsight include Apache Livy , a REST API-based Spark job server to remotely submit and monitor jobs. See Use Apache Spark REST API to submit remote jobs to an HDInsight Spark cluster .
Support for Azure Data Lake Storage	Spark clusters in HDInsight can use Azure Data Lake Storage as both the primary storage or additional storage. For more information on Data Lake Storage, see Overview of Azure Data Lake Storage .
Integration with Azure services	Spark cluster in HDInsight comes with a connector to Azure Event Hubs. You can build streaming applications using the Event Hubs, in addition to Apache Kafka , which is already available as part of Spark.
Support for ML Server	Support for ML Server in HDInsight is provided as the ML Services cluster type. You can set up an ML Services cluster to run distributed R computations with the speeds promised with a Spark cluster. For more information, see Get started using ML Server in HDInsight .
Integration with third-party IDEs	HDInsight provides several IDE plugins that are useful to create and submit applications to an HDInsight Spark cluster. For more information, see Use Azure Toolkit for IntelliJ IDEA , Use HDInsight for VSCode , and Use Azure Toolkit for Eclipse .
Concurrent Queries	Spark clusters in HDInsight support concurrent queries. This capability enables multiple queries from one user or multiple queries from various users and applications to share the same cluster resources.
Caching on SSDs	You can choose to cache data either in memory or in SSDs attached to the cluster nodes. Caching in memory provides the best query performance but could be expensive. Caching in SSDs provides a great option for improving query performance without the need to create a cluster of a size that is required to fit the entire dataset in memory.
Integration with BI Tools	Spark clusters in HDInsight provide connectors for BI tools such as Power BI for data analytics.
Pre-loaded Anaconda libraries	Spark clusters in HDInsight come with Anaconda libraries pre-installed. Anaconda provides close to 200 libraries for machine learning, data analysis, visualization, etc.
Scalability	HDInsight allow you to change the number of cluster nodes. Also, Spark clusters can be dropped with no loss of data since all the data is stored in Azure Storage or Data Lake Storage.
SLA	Spark clusters in HDInsight come with 24/7 support and an SLA of 99.9% up-time.

Apache Spark clusters in HDInsight include the following components that are available on the clusters by default.

- [Spark Core](#). Includes Spark Core, Spark SQL, Spark streaming APIs, GraphX, and MLlib.
- [Anaconda](#)
- [Apache Livy](#)
- [Jupyter notebook](#)
- [Apache Zeppelin notebook](#)

Spark clusters in HDInsight also provide an [ODBC driver](#) for connectivity to Spark clusters in HDInsight from BI tools such as Microsoft Power BI.

Spark cluster architecture



It is easy to understand the components of Spark by understanding how Spark runs on HDInsight clusters.

Spark applications run as independent sets of processes on a cluster, coordinated by the `SparkContext` object in your main program (called the driver program).

The `SparkContext` can connect to several types of cluster managers, which allocate resources across applications. These cluster managers include [Apache Mesos](#), [Apache Hadoop YARN](#), or the Spark cluster manager. In HDInsight, Spark runs using the YARN cluster manager. Once connected, Spark acquires executors on workers nodes in the cluster, which are processes that run computations and store data for your application. Next, it sends your application code (defined by JAR or Python files passed to `SparkContext`) to the executors. Finally, `SparkContext` sends tasks to the executors to run.

The `SparkContext` runs the user's main function and executes the various parallel operations on the worker nodes. Then, the `SparkContext` collects the results of the operations. The worker nodes read and write data from and to the Hadoop distributed file system. The worker nodes also cache transformed data in-memory as Resilient Distributed Datasets (RDDs).

The `SparkContext` connects to the Spark master and is responsible for converting an application to a directed graph (DAG) of individual tasks that get executed within an executor process on the worker nodes. Each application gets its own executor processes, which stay up for the duration of the whole application and run tasks in multiple threads.

Spark in HDInsight use cases

Spark clusters in HDInsight enable the following key scenarios:

- Interactive data analysis and BI

Apache Spark in HDInsight stores data in Azure Storage or Azure Data Lake Storage. Business experts and key decision makers can analyze and build reports over that data and use Microsoft Power BI to build interactive reports from the analyzed data. Analysts can start from unstructured/semi structured data in cluster storage, define a schema for the data using notebooks, and then build data models using Microsoft Power BI. Spark clusters in HDInsight also support a number of third-party BI tools such as Tableau making it easier for data analysts, business experts, and key decision makers.

[Tutorial: Visualize Spark data using Power BI](#)

- Spark Machine Learning

Apache Spark comes with [MLlib](#), a machine learning library built on top of Spark that you can use from a Spark cluster in HDInsight. Spark cluster in HDInsight also includes Anaconda, a Python distribution with a variety of packages for machine learning. Couple this with a built-in support for Jupyter and Zeppelin notebooks, and you have an environment for creating machine learning applications.

[Tutorial: Predict building temperatures using HVAC data](#)

[Tutorial: Predict food inspection results](#)

- Spark streaming and real-time data analysis

Spark clusters in HDInsight offer a rich support for building real-time analytics solutions. While Spark already has connectors to ingest data from many sources like Kafka, Flume, Twitter, ZeroMQ, or TCP sockets, Spark in HDInsight adds first-class support for ingesting data from Azure Event Hubs. Event Hubs is the most widely used queuing service on Azure. Having an out-of-the-box support for Event Hubs makes Spark clusters in HDInsight an ideal platform for building real-time analytics pipeline.

Where do I start?

You can use the following articles to learn more about Apache Spark in HDInsight:

- [QuickStart: Create an Apache Spark cluster in HDInsight and run interactive query using Jupyter](#)
- [Tutorial: Run an Apache Spark job using Jupyter](#)
- [Tutorial: Analyze data using BI tools](#)
- [Tutorial: Machine learning using Apache Spark](#)
- [Tutorial: Create a Scala Maven application using IntelliJ](#)

Next Steps

In this overview, you get some basic understanding of Apache Spark in Azure HDInsight. Advance to the next article to learn how to create an HDInsight Spark cluster and run some Spark SQL queries:

- [Create an Apache Spark cluster in HDInsight](#)

Quickstart: Create an Apache Spark cluster in HDInsight using template

2/6/2019 • 5 minutes to read

Learn how to create an [Apache Spark](#) cluster in Azure HDInsight, and how to run Spark SQL queries against [Apache Hive](#) tables. Apache Spark enables fast data analytics and cluster computing using in-memory processing. For information on Spark on HDInsight, see [Overview: Apache Spark on Azure HDInsight](#).

In this quickstart, you use a Resource Manager template to create an HDInsight Spark cluster. Similar templates can be viewed at [Azure Quickstart Templates](#). The template reference can be found [here](#).

The cluster uses Azure Storage Blobs as the cluster storage. For more information on using Data Lake Storage Gen2, see [Quickstart: Set up clusters in HDInsight](#).

IMPORTANT

Billing for HDInsight clusters is prorated per minute, whether you are using them or not. Be sure to delete your cluster after you have finished using it. For more information, see the [Clean up resources](#) section of this article.

If you don't have an Azure subscription, [create a free account](#) before you begin.

Create an HDInsight Spark cluster

Create an HDInsight Spark cluster using an Azure Resource Manager template. The template can be found in [GitHub](#). For the JSON syntax and properties of the cluster, see [Microsoft.HDInsight/clusters](#).

1. Select the following link to open the template in the Azure portal in a new browser tab:

[Deploy to Azure](#)

2. Enter the following values:

PROPERTY	VALUE
Subscription	Select your Azure subscription used for creating this cluster. The subscription used for this quickstart is < Azure subscription name >.
Resource group	Create a resource group or select an existing one. Resource group is used to manage Azure resources for your projects. The new resource group name used for this quickstart is myspark20180403rg .
Location	Select a location for the resource group. The template uses this location for creating the cluster as well as for the default cluster storage. The location used for this quickstart is East US 2 .
ClusterName	Enter a name for the HDInsight cluster that you want to create. The new cluster name used for this quickstart is myspark20180403 .

PROPERTY	VALUE
Cluster login name and password	The default login name is admin. Choose a password for the cluster login. The login name used for this quickstart is admin .
SSH user name and password	Choose a password for the SSH user. The SSH user name used for this quickstart is sshuser .

The screenshot shows the Azure portal interface for deploying a Spark cluster. On the left, there's a sidebar with various service icons like All services, Favorites, and specific Azure services like HDInsight clusters, Data Lake Storage Gen1, and App Services. The main area is titled 'Deploy a Spark cluster in Azure HDInsight' and shows a template named '101-hdinsight-spark-linux' which creates 2 resources. It's set to a subscription 'BDHadoopTeamPMTestDemo', resource group 'myspark20180403rg', and location 'East US 2'. In the 'SETTINGS' section, the 'Cluster Name' is set to 'myspark20180403'. Under 'TERMS AND CONDITIONS', there's a checkbox for agreeing to terms and conditions, which is checked. At the bottom, there's a large red 'Purchase' button.

3. Select **I agree to the terms and conditions stated above**, select **Pin to dashboard**, and then select **Purchase**. You can see a new tile titled **Deploying Template deployment**. It takes about 20 minutes to create the cluster. The cluster must be created before you can proceed to the next session.

If you run into an issue with creating HDInsight clusters, it could be that you do not have the right permissions to do so. For more information, see [Access control requirements](#).

Install IntelliJ/Eclipse for Spark application

Use the Azure Toolkit for IntelliJ/Eclipse plug-in to develop Spark applications written in [Scala](#), and then submit them to an Azure HDInsight Spark cluster directly from the IntelliJ/Eclipse integrated development environment (IDE). For more information, see [Use IntelliJ to author/submit Spark application](#) and [Use Eclipse to author/submit Spark application](#).

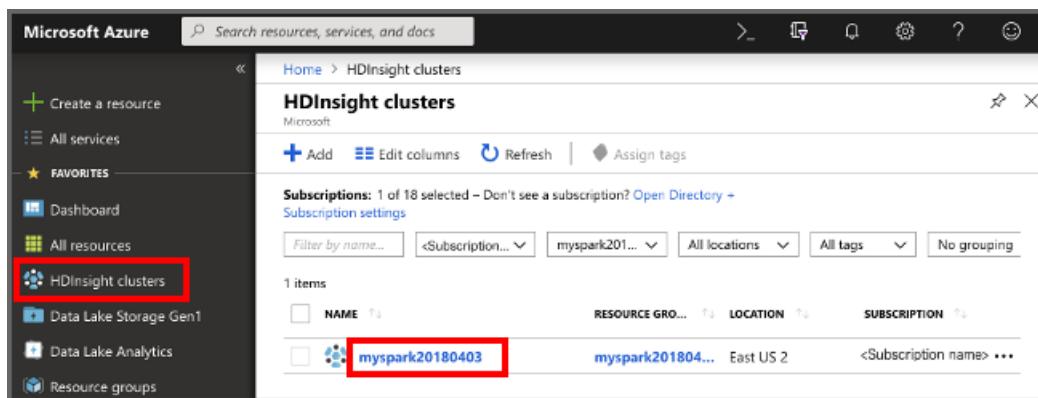
Install VSCode for PySpark/hive applications

Learn how to use the Azure HDInsight Tools for Visual Studio Code (VSCode) to create and submit Hive batch jobs, interactive Hive queries, PySpark batch, and PySpark interactive scripts. The Azure HDInsight Tools can be installed on the platforms that are supported by VSCode. These include Windows, Linux, and macOS. For more information, see [Use VSCode to author/submit PySpark application](#).

Create a Jupyter notebook

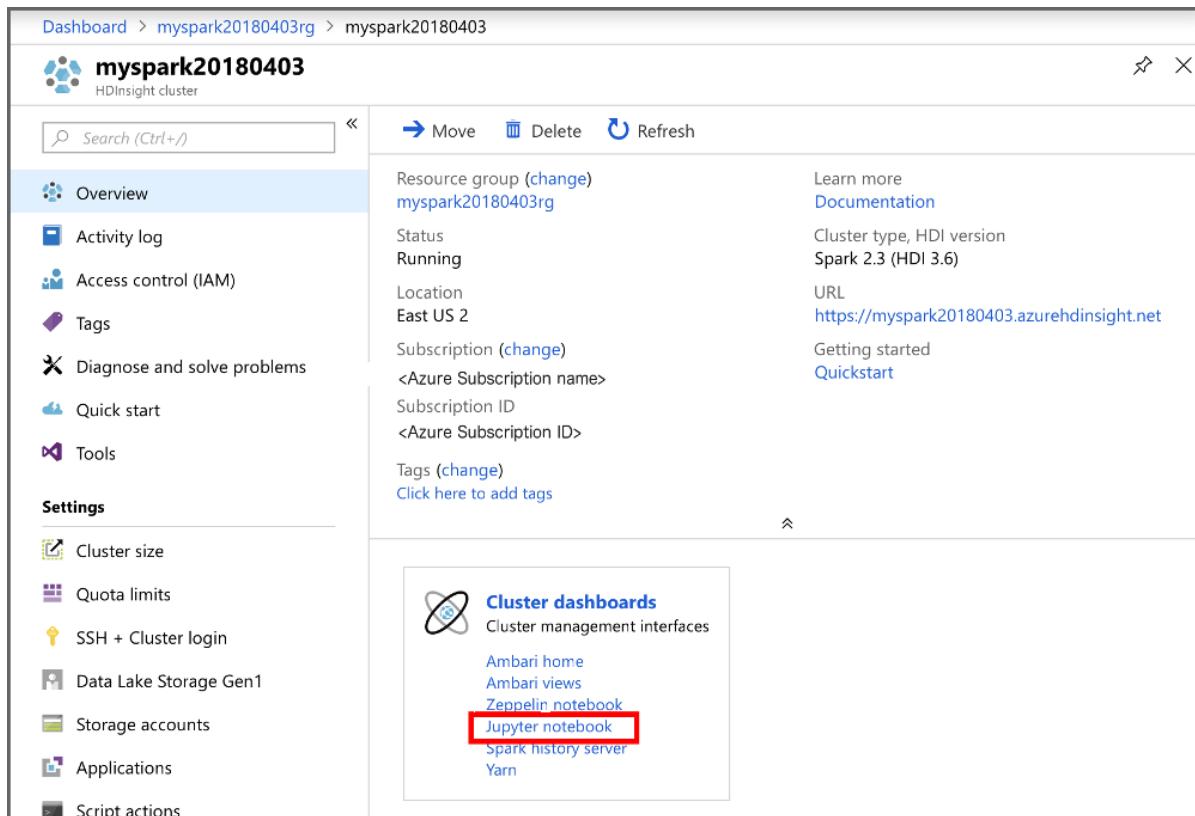
[Jupyter Notebook](#) is an interactive notebook environment that supports various programming languages. The notebook allows you to interact with your data, combine code with markdown text and perform simple visualizations.

1. Open the [Azure portal](#).
2. Select **HDInsight clusters**, and then select the cluster you created.



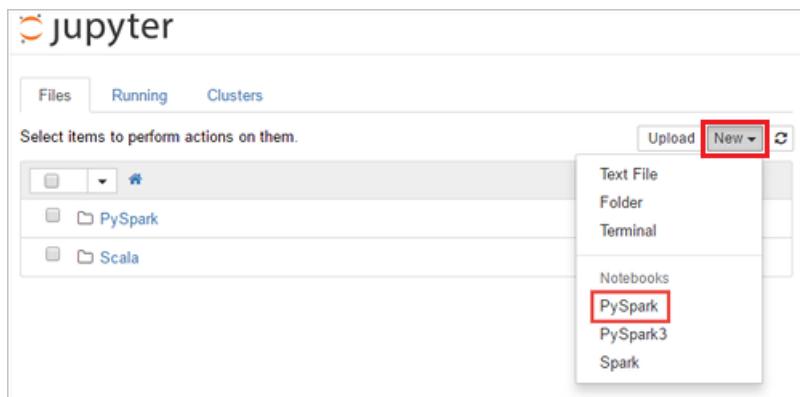
NAME	RESOURCE GRO...	LOCATION	SUBSCRIPTION
myspark20180403	myspark201804...	East US 2	<Subscription name> ***

3. From the portal, in **Cluster dashboards** section, click on **Jupyter Notebook**. If prompted, enter the cluster login credentials for the cluster.



The screenshot shows the Azure portal interface for managing an HDInsight cluster named 'myspark20180403'. The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Tools, Settings, Cluster size, Quota limits, SSH + Cluster login, Data Lake Storage Gen1, Storage accounts, Applications, and Script actions. The 'Overview' link is currently selected. In the main content area, there is a summary of the cluster's status: Resource group (change) 'myspark20180403rg', Status 'Running', Location 'East US 2', and a URL 'https://myspark20180403.azurehdinsight.net'. Below this, there are links for Documentation, Cluster type, HDI version (Spark 2.3 (HDI 3.6)), URL, Getting started, and Quickstart. Under the 'Cluster dashboards' heading, there is a list of management interfaces: Ambari home, Ambari views, Zeppelin notebook, Jupyter notebook (which is highlighted with a red box), Spark history server, and Yarn.

4. Select **New > PySpark** to create a notebook.

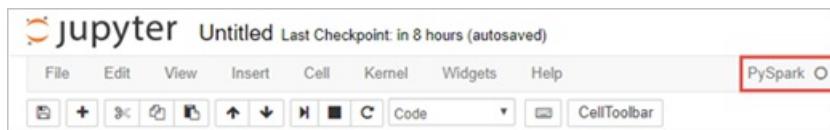


A new notebook is created and opened with the name Untitled(Untitled.ipynb).

Run Spark SQL statements

SQL (Structured Query Language) is the most common and widely used language for querying and transforming data. Spark SQL functions as an extension to Apache Spark for processing structured data, using the familiar SQL syntax.

1. Verify the kernel is ready. The kernel is ready when you see a hollow circle next to the kernel name in the notebook. Solid circle denotes that the kernel is busy.



When you start the notebook for the first time, the kernel performs some tasks in the background. Wait for the kernel to be ready.

2. Paste the following code in an empty cell, and then press **SHIFT + ENTER** to run the code. The command lists the Hive tables on the cluster:

```
%%sql  
SHOW TABLES
```

When you use a Jupyter Notebook with your HDInsight Spark cluster, you get a preset `spark` session that you can use to run Hive queries using Spark SQL. `%%sql` tells Jupyter Notebook to use the preset `spark` session to run the Hive query. The query retrieves the top 10 rows from a Hive table (**hivesamptable**) that comes with all HDInsight clusters by default. The first time you submit the query Jupyter will create Spark Application for the notebook. It takes about 30 seconds to complete. Once the spark application is ready the query is executed in about a second and produces the results. The output looks like:

Every time you run a query in Jupyter, your web browser window title shows a (**Busy**) status along with the notebook title. You also see a solid circle next to the **PySpark** text in the top-right corner.

- Run another query to see the data in `hivesamptable`.

```
%%sql
SELECT * FROM hivesamptable LIMIT 10
```

The screen shall refresh to show the query output.

clientid	querytime	market	deviceplatform	devicemake	devicemodel	state	country	querydweltime	sessionid	sessionpagevieworder
0	2018-04-05 05:51:45	en-US	Android	Samsung	SCH-I500	California	United States	31.423273	1	41
1	2018-04-05 05:51:33	en-US	Android	Samsung	SCH-I500	California	United States	11.878175	1	40
2	2018-04-05 05:51:03	en-US	Android	Samsung	SCH-I500	California	United States	30.195784	1	39
3	2018-04-05 05:51:03	en-US	Android	Samsung	SCH-I500	California	United States	0.129492	1	38
4	2018-04-05 05:50:44	en-US	Android	Samsung	SCH-I500	California	United States	19.026435	1	37
5	2018-04-05 05:50:27	en-US	Android	Samsung	SCH-I500	California	United States	16.411516	1	36
6	2018-04-05 05:50:13	en-US	Android	Samsung	SCH-I500	California	United States	13.761518	1	35
7	2018-04-05 05:50:04	en-US	Android	Samsung	SCH-I500	California	United States	9.091954	1	34
8	2018-04-05 05:49:57	en-US	Android	Samsung	SCH-I500	California	United States	7.709461	1	33
9	2018-04-05 05:54:21	en-US	Android	Samsung	SCH-I500	California	United States	0.899126	1	48

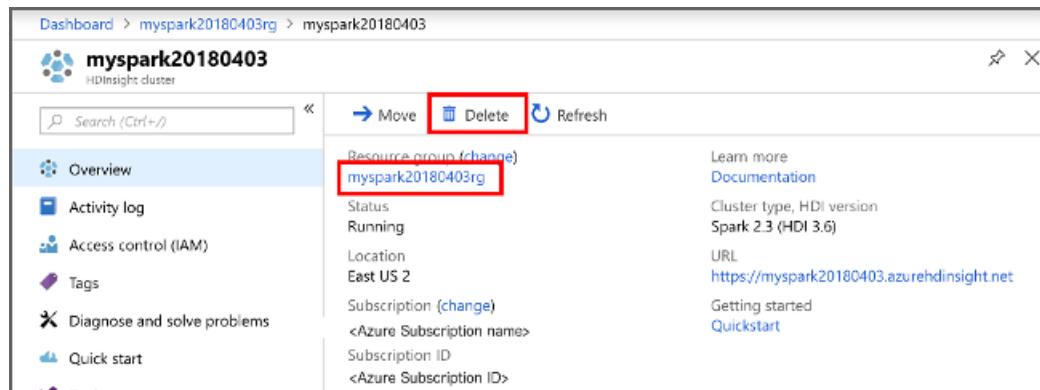
- From the **File** menu on the notebook, select **Close and Halt**. Shutting down the notebook releases the cluster resources, including Spark Application.

Clean up resources

HDInsight saves your data and Jupyter notebooks in Azure Storage or Azure Data Lake Store, so you can safely delete a cluster when it is not in use. You are also charged for an HDInsight cluster, even when it is not in use. Since the charges for the cluster are many times more than the charges for storage, it makes economic sense to delete clusters when they are not in use. If you plan to work on the tutorial listed in [Next steps](#) immediately, you might

want to keep the cluster.

Switch back to the Azure portal, and select **Delete**.



The screenshot shows the Azure portal's 'HDInsight cluster' details page for 'myspark20180403'. The 'Delete' button in the top right is highlighted with a red box. The left sidebar lists navigation options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, and Quick start. The main content area displays cluster details: Resource group (change) to 'myspark20180403rg' (also highlighted with a red box), Status 'Running', Location 'East US 2', Subscription (change) to '<Azure Subscription name>', and Subscription ID to '<Azure Subscription ID>'. On the right, there are links for 'Learn more', 'Documentation', 'Cluster type, HDI version' (Spark 2.3 (HDI 3.6)), 'URL' (<https://myspark20180403.azurehdinsight.net>), 'Getting started', and 'Quickstart'.

You can also select the resource group name to open the resource group page, and then select **Delete resource group**. By deleting the resource group, you delete both the HDInsight Spark cluster, and the default storage account.

Next steps

In this quickstart, you learned how to create an HDInsight Spark cluster and run a basic Spark SQL query. Advance to the next tutorial to learn how to use an HDInsight Spark cluster to run interactive queries on sample data.

[Run interactive queries on Apache Spark](#)

Kernels for Jupyter notebook on Apache Spark clusters in Azure HDInsight

12/27/2018 • 8 minutes to read

HDInsight Spark clusters provide kernels that you can use with the Jupyter notebook on [Apache Spark](#) for testing your applications. A kernel is a program that runs and interprets your code. The three kernels are:

- **PySpark** - for applications written in Python2.
- **PySpark3** - for applications written in Python3.
- **Spark** - for applications written in Scala.

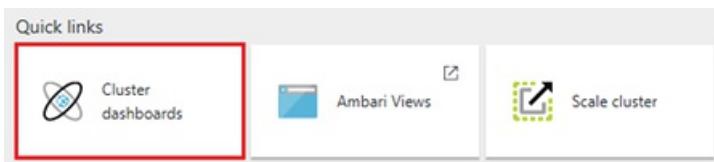
In this article, you learn how to use these kernels and the benefits of using them.

Prerequisites

- An Apache Spark cluster in HDInsight. For instructions, see [Create Apache Spark clusters in Azure HDInsight](#).

Create a Jupyter notebook on Spark HDInsight

1. From the [Azure portal](#), open your cluster. See [List and show clusters](#) for the instructions. The cluster is opened in a new portal blade.
2. From the **Quick links** section, click **Cluster dashboards** to open the **Cluster dashboards** blade. If you don't see **Quick Links**, click **Overview** from the left menu on the blade.



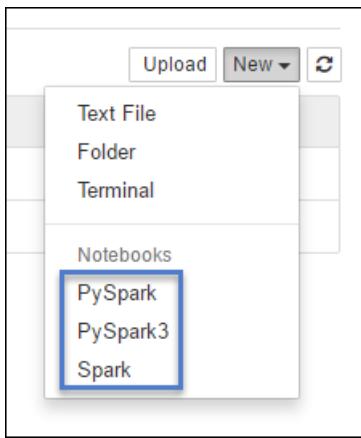
3. Click **Jupyter Notebook**. If prompted, enter the admin credentials for the cluster.

NOTE

You may also reach the Jupyter notebook on Spark cluster by opening the following URL in your browser. Replace **CLUSTERNAME** with the name of your cluster:

```
https://CLUSTERNAME.azurehdinsight.net/jupyter
```

4. Click **New**, and then click either **Pyspark**, **PySpark3**, or **Spark** to create a notebook. Use the Spark kernel for Scala applications, PySpark kernel for Python2 applications, and PySpark3 kernel for Python3 applications.



5. A notebook opens with the kernel you selected.

Benefits of using the kernels

Here are a few benefits of using the new kernels with Jupyter notebook on Spark HDInsight clusters.

- **Preset contexts.** With **PySpark**, **PySpark3**, or the **Spark** kernels, you do not need to set the Spark or Hive contexts explicitly before you start working with your applications. These are available by default. These contexts are:
 - **sc** - for Spark context
 - **sqlContext** - for Hive context

So, you don't have to run statements like the following to set the contexts:

```
sc = SparkContext('yarn-client')
sqlContext = HiveContext(sc)
```

Instead, you can directly use the preset contexts in your application.

- **Cell magics.** The PySpark kernel provides some predefined "magics", which are special commands that you can call with `%` (for example, `%%MAGIC`). The magic command must be the first word in a code cell and allow for multiple lines of content. The magic word should be the first word in the cell. Adding anything before the magic, even comments, causes an error. For more information on magics, see [here](#).

The following table lists the different magics available through the kernels.

MAGIC	EXAMPLE	DESCRIPTION
help	<code>%%help</code>	Generates a table of all the available magics with example and description
info	<code>%%info</code>	Outputs session information for the current Livy endpoint

MAGIC	EXAMPLE	DESCRIPTION
configure	<pre>%%configure -f {"executorMemory": "1000M", "executorCores": 4}</pre>	Configures the parameters for creating a session. The force flag (-f) is mandatory if a session has already been created, which ensures that the session is dropped and recreated. Look at Livy's POST /sessions Request Body for a list of valid parameters. Parameters must be passed in as a JSON string and must be on the next line after the magic, as shown in the example column.
sql	<pre>%%sql -o <variable name> SHOW TABLES</pre>	Executes a Hive query against the sqlContext. If the <code>-o</code> parameter is passed, the result of the query is persisted in the %%local Python context as a Pandas dataframe.
local	<pre>%%local a=1</pre>	All the code in subsequent lines is executed locally. Code must be valid Python2 code even irrespective of the kernel you are using. So, even if you selected PySpark3 or Spark kernels while creating the notebook, if you use the <code>%%local</code> magic in a cell, that cell must only have valid Python2 code..
logs	<pre>%%logs</pre>	Outputs the logs for the current Livy session.
delete	<pre>%%delete -f -s <session number></pre>	Deletes a specific session of the current Livy endpoint. You cannot delete the session that is initiated for the kernel itself.
cleanup	<pre>%%cleanup -f</pre>	Deletes all the sessions for the current Livy endpoint, including this notebook's session. The force flag -f is mandatory.

NOTE

In addition to the magics added by the PySpark kernel, you can also use the [built-in IPython magics](#), including `%%sh`. You can use the `%%sh` magic to run scripts and block of code on the cluster headnode.

1. **Auto visualization.** The **Pyspark** kernel automatically visualizes the output of Hive and SQL queries. You can choose between several different types of visualizations including Table, Pie, Line, Area, Bar.

Parameters supported with the %%sql magic

The `%%sql` magic supports different parameters that you can use to control the kind of output that you receive when you run queries. The following table lists the output.

PARAMETER	EXAMPLE	DESCRIPTION
-o	-o <VARIABLE NAME>	Use this parameter to persist the result of the query, in the %%local Python context, as a Pandas dataframe. The name of the dataframe variable is the variable name you specify.
-q	-q	Use this to turn off visualizations for the cell. If you don't want to auto-visualize the content of a cell and just want to capture it as a dataframe, then use -q -o <VARIABLE>. If you want to turn off visualizations without capturing the results (for example, for running a SQL query, like a CREATE TABLE statement), use -q without specifying a -o argument.
-m	-m <METHOD>	Where METHOD is either take or sample (default is take). If the method is take , the kernel picks elements from the top of the result data set specified by MAXROWS (described later in this table). If the method is sample , the kernel randomly samples elements of the data set according to -r parameter, described next in this table.
-r	-r <FRACTION>	Here FRACTION is a floating-point number between 0.0 and 1.0. If the sample method for the SQL query is sample , then the kernel randomly samples the specified fraction of the elements of the result set for you. For example, if you run a SQL query with the arguments -m sample -r 0.01, then 1% of the result rows are randomly sampled.
-n	-n <MAXROWS>	MAXROWS is an integer value. The kernel limits the number of output rows to MAXROWS . If MAXROWS is a negative number such as -1, then the number of rows in the result set is not limited.

Example:

```
%%sql -q -m sample -r 0.1 -n 500 -o query2
SELECT * FROM hivesampletable
```

The statement above does the following:

- Selects all records from **hivesampletable**.
- Because we use -q, it turns off auto-visualization.
- Because we use -m sample -r 0.1 -n 500 it randomly samples 10% of the rows in the hivesampletable and limits the size of the result set to 500 rows.

- Finally, because we used `-o query2` it also saves the output into a dataframe called **query2**.

Considerations while using the new kernels

Whichever kernel you use, leaving the notebooks running consumes the cluster resources. With these kernels, because the contexts are preset, simply exiting the notebooks does not kill the context and hence the cluster resources continue to be in use. A good practice is to use the **Close and Halt** option from the notebook's **File** menu when you are finished using the notebook, which kills the context and then exits the notebook.

Show me some examples

When you open a Jupyter notebook, you see two folders available at the root level.

- The **PySpark** folder has sample notebooks that use the new **Python** kernel.
- The **Scala** folder has sample notebooks that use the new **Spark** kernel.

You can open the **00 - [READ ME FIRST] Spark Magic Kernel Features** notebook from the **PySpark** or **Spark** folder to learn about the different magics available. You can also use the other sample notebooks available under the two folders to learn how to achieve different scenarios using Jupyter notebooks with HDInsight Spark clusters.

Where are the notebooks stored?

If your cluster uses Azure Storage as the default storage account, Jupyter notebooks are saved to storage account under the **/HdiNotebooks** folder. Notebooks, text files, and folders that you create from within Jupyter are accessible from the storage account. For example, if you use Jupyter to create a folder **myfolder** and a notebook **myfolder/mynotebook.ipynb**, you can access that notebook at `/HdiNotebooks/myfolder/mynotebook.ipynb` within the storage account. The reverse is also true, that is, if you upload a notebook directly to your storage account at `/HdiNotebooks/mynotebook1.ipynb`, the notebook is visible from Jupyter as well. Notebooks remain in the storage account even after the cluster is deleted.

NOTE

HDInsight clusters with Azure Data Lake Storage as the default storage do not store notebooks in associated storage.

The way notebooks are saved to the storage account is compatible with [Apache Hadoop HDFS](#). So, if you SSH into the cluster you can use file management commands as shown in the following snippet:

```
hdfs dfs -ls /HdiNotebooks          # List everything at the root directory - everything
in this directory is visible to Jupyter from the home page
hdfs dfs -copyToLocal /HdiNotebooks    # Download the contents of the HdiNotebooks folder
hdfs dfs -copyFromLocal example.ipynb /HdiNotebooks  # Upload a notebook example.ipynb to the root folder so
it's visible from Jupyter
```

Irrespective of whether the cluster uses Azure Storage or Azure Data Lake Storage as the default storage account, the notebooks are also saved on the cluster headnode at `/var/lib/jupyter`.

Supported browser

Jupyter notebooks on Spark HDInsight clusters are supported only on Google Chrome.

Feedback

The new kernels are in evolving stage and will mature over time. This could also mean that APIs could change as these kernels mature. We would appreciate any feedback that you have while using these new kernels. This is

useful in shaping the final release of these kernels. You can leave your comments/feedback under the **Comments** section at the bottom of this article.

See also

- [Overview: Apache Spark on Azure HDInsight](#)

Scenarios

- [Apache Spark with BI](#): Perform interactive data analysis using Spark in HDInsight with BI tools
- [Apache Spark with Machine Learning](#): Use Spark in HDInsight for analyzing building temperature using HVAC data
- [Apache Spark with Machine Learning](#): Use Spark in HDInsight to predict food inspection results
- [Website log analysis using Apache Spark in HDInsight](#)

Create and run applications

- [Create a standalone application using Scala](#)
- [Run jobs remotely on an Apache Spark cluster using Apache Livy](#)

Tools and extensions

- [Use HDInsight Tools Plugin for IntelliJ IDEA to create and submit Spark Scala applications](#)
- [Use HDInsight Tools Plugin for IntelliJ IDEA to debug Apache Spark applications remotely](#)
- [Use Apache Zeppelin notebooks with an Apache Spark cluster on HDInsight](#)
- [Use external packages with Jupyter notebooks](#)
- [Install Jupyter on your computer and connect to an HDInsight Spark cluster](#)

Manage resources

- [Manage resources for the Apache Spark cluster in Azure HDInsight](#)
- [Track and debug jobs running on an Apache Spark cluster in HDInsight](#)

Introduction to ML Services and open-source R capabilities on HDInsight

12/18/2018 • 10 minutes to read

NOTE

In September 2017, Microsoft R Server was released under the new name of **Microsoft Machine Learning Server** or ML Server. Consequently, R Server cluster on HDInsight is now called **Machine Learning Services** or **ML Services** cluster on HDInsight. For more information on the R Server name change, see [Microsoft R Server is now Microsoft Machine Learning Server](#).

Microsoft Machine Learning Server is available as a deployment option when you create HDInsight clusters in Azure. The cluster type that provides this option is called **ML Services**. This capability provides data scientists, statisticians, and R programmers with on-demand access to scalable, distributed methods of analytics on HDInsight.

NOTE

[Learn more about upcoming enhancements and capabilities.](#)

ML Services on HDInsight provides the latest capabilities for R-based analytics on datasets of virtually any size, loaded to either Azure Blob or Data Lake storage. Since ML Services cluster is built on open-source R, the R-based applications you build can leverage any of the 8000+ open-source R packages. The routines in ScaleR, Microsoft's big data analytics package are also available.

The edge node of a cluster provides a convenient place to connect to the cluster and to run your R scripts. With an edge node, you have the option of running the parallelized distributed functions of ScaleR across the cores of the edge node server. You can also run them across the nodes of the cluster by using ScaleR's Hadoop Map Reduce or Apache Spark compute contexts.

The models or predictions that result from analysis can be downloaded for on-premises use. They can also be operationalized elsewhere in Azure, in particular through [Azure Machine Learning Studio web service](#).

Get started with ML Services on HDInsight

To create an ML Services cluster in Azure HDInsight, select the **ML Services** cluster type when creating an HDInsight cluster using the Azure portal. The ML Services cluster type includes ML Server on the data nodes of the cluster and on an edge node, which serves as a landing zone for ML Services-based analytics. See [Getting Started with ML Services on HDInsight](#) for a walkthrough on how to create the cluster.

Why choose ML Services in HDInsight?

ML Services in HDInsight provides the following benefits:

AI innovation from Microsoft and open-source

ML Services includes highly scalable, distributed set of algorithms such as [RevoscaleR](#), [revoscalepy](#), and [microsoftML](#) that can work on data sizes larger than the size of physical memory, and run on a wide variety of platforms in a distributed manner. Learn more about the collection of Microsoft's custom [R packages](#) and [Python packages](#) included with the product.

ML Services bridges these Microsoft innovations and contributions coming from the open-source community (R, Python, and AI toolkits) all on top of a single enterprise-grade platform. Any R or Python open-source machine learning package can work side by side with any proprietary innovation from Microsoft.

Simple, secure, and high-scale operationalization and administration

Enterprises relying on traditional paradigms and environments invest much time and effort towards operationalization. This results in inflated costs and delays including the translation time for models, iterations to keep them valid and current, regulatory approval, and managing permissions through operationalization.

ML Services offers enterprise grade [operationalization](#), in that, after a machine learning model is completed, it takes just a few clicks to generate web services APIs. These [web services](#) are hosted on a server grid in the cloud and can be integrated with line-of-business applications. The ability to deploy to an elastic grid lets you scale seamlessly with the needs of your business, both for batch and real-time scoring. For instructions, see [Operationalize ML Services on HDInsight](#).

Key features of ML Services on HDInsight

The following features are included in ML Services on HDInsight.

FEATURE CATEGORY	DESCRIPTION
R-enabled	R packages for solutions written in R, with an open source distribution of R, and run-time infrastructure for script execution.
Python-enabled	Python modules for solutions written in Python, with an open source distribution of Python, and run-time infrastructure for script execution.
Pre-trained models	For visual analysis and text sentiment analysis, ready to score data you provide.
Deploy and consume	Operationalize your server and deploy solutions as a web service.
Remote execution	Start remote sessions on ML Services cluster on your network from your client workstation.

Data storage options for ML Services on HDInsight

Default storage for the HDFS file system of HDInsight clusters can be associated with either an Azure Storage account or an Azure Data Lake Storage. This association ensures that whatever data is uploaded to the cluster storage during analysis is made persistent and the data is available even after the cluster is deleted. There are various tools for handling the data transfer to the storage option that you select, including the portal-based upload facility of the storage account and the [AzCopy](#) utility.

You have the option of enabling access to additional Blob and Data lake stores during the cluster provisioning process regardless of the primary storage option in use. See [Getting started with ML Services on HDInsight](#) for information on adding access to additional accounts. See [Azure Storage options for ML Services on HDInsight](#) article to learn more about using multiple storage accounts.

You can also use [Azure Files](#) as a storage option for use on the edge node. Azure Files enables you to mount a file share that was created in Azure Storage to the Linux file system. For more information about these data storage options for ML Services on HDInsight cluster, see [Azure Storage options for ML Services on HDInsight](#).

Access ML Services edge node

You can connect to Microsoft ML Server on the edge node using a browser. It is installed by default during cluster creation. For more information, see [Get started with ML Services on HDInsight](#). You can also connect to the cluster edge node from the command line by using SSH/PuTTY to access the R console.

Develop and run R scripts

The R scripts you create and run can use any of the 8000+ open-source R packages in addition to the parallelized and distributed routines available in the ScaleR library. In general, a script that is run with ML Services on the edge node runs within the R interpreter on that node. The exceptions are those steps that need to call a ScaleR function with a compute context that is set to Hadoop Map Reduce (RxHadoopMR) or Spark (RxSpark). In this case, the function runs in a distributed fashion across those data (task) nodes of the cluster that are associated with the data referenced. For more information about the different compute context options, see [Compute context options for ML Services on HDInsight](#).

Operationalize a model

When your data modeling is complete, you can operationalize the model to make predictions for new data either from Azure or on-premises. This process is known as scoring. Scoring can be done in HDInsight, Azure Machine Learning, or on-premises.

Score in HDInsight

To score in HDInsight, write an R function that calls your model to make predictions for a new data file that you've loaded to your storage account. Then, save the predictions back to the storage account. You can run this routine on-demand on the edge node of your cluster or by using a scheduled job.

Score in Azure Machine Learning (AML)

To score using Azure Machine Learning, use the open-source Azure Machine Learning R package known as [AzureML](#) to publish your model as an Azure web service. For convenience, this package is pre-installed on the edge node. Next, use the facilities in Azure Machine Learning to create a user interface for the web service, and then call the web service as needed for scoring.

If you choose this option, you must convert any ScaleR model objects to equivalent open-source model objects for use with the web service. Use ScaleR coercion functions, such as `as.randomForest()` for ensemble-based models, for this conversion.

Score on-premises

To score on-premises after creating your model, you can serialize the model in R, download it, de-serialize it, and then use it for scoring new data. You can score new data by using the approach described earlier in [Score in HDInsight](#) or by using [web services](#).

Maintain the cluster

Install and maintain R packages

Most of the R packages that you use are required on the edge node since most steps of your R scripts run there. To install additional R packages on the edge node, you can use the `install.packages()` method in R.

If you are just using routines from the ScaleR library across the cluster, you do not usually need to install additional R packages on the data nodes. However, you might need additional packages to support the use of **rxExec** or **RxDatapartition** execution on the data nodes.

In such cases, the additional packages can be installed with a script action after you create the cluster. For more information, see [Manage ML Services in HDInsight cluster](#).

Change Apache Hadoop MapReduce memory settings

A cluster can be modified to change the amount of memory that is available to ML Services when it is running a MapReduce job. To modify a cluster, use the Apache Ambari UI that's available through the Azure portal blade for your cluster. For instructions about how to access the Ambari UI for your cluster, see [Manage HDInsight clusters using the Ambari Web UI](#).

It is also possible to change the amount of memory that is available to ML Services by using Hadoop switches in the call to **RxHadoopMR** as follows:

```
hadoopSwitches = "-libjars /etc/hadoop/conf -Dmapred.job.map.memory.mb=6656"
```

Scale your cluster

An existing ML Services cluster on HDInsight can be scaled up or down through the portal. By scaling up, you can gain the additional capacity that you might need for larger processing tasks, or you can scale back a cluster when it is idle. For instructions about how to scale a cluster, see [Manage HDInsight clusters](#).

Maintain the system

Maintenance to apply OS patches and other updates is performed on the underlying Linux VMs in an HDInsight cluster during off-hours. Typically, maintenance is done at 3:30 AM (based on the local time for the VM) every Monday and Thursday. Updates are performed in such a way that they don't impact more than a quarter of the cluster at a time.

Since the head nodes are redundant and not all data nodes are impacted, any jobs that are running during this time might slow down. However, they should still run to completion. Any custom software or local data that you have is preserved across these maintenance events unless a catastrophic failure occurs that requires a cluster rebuild.

IDE options for ML Services on HDInsight

The Linux edge node of an HDInsight cluster is the landing zone for R-based analysis. Recent versions of HDInsight provide a default installation of RStudio Server on the edge node as a browser-based IDE. Use of RStudio Server as an IDE for the development and execution of R scripts can be considerably more productive than just using the R console.

Additionally, you can install a desktop IDE and use it to access the cluster through use of a remote MapReduce or Spark compute context. Options include Microsoft's [R Tools for Visual Studio](#) (RTVS), RStudio, and Walware's Eclipse-based [StatET](#).

Additionally, you can access the R console on the edge node by typing **R** at the Linux command prompt after connecting via SSH or PuTTY. When using the console interface, it is convenient to run a text editor for R script development in another window, and cut and paste sections of your script into the R console as needed.

Pricing

The prices that are associated with an ML Services HDInsight cluster are structured similarly to the prices for other HDInsight cluster types. They are based on the sizing of the underlying VMs across the name, data, and edge nodes, with the addition of a core-hour uplift. For more information, see [HDInsight pricing](#).

Next steps

To learn more about how to use ML Services on HDInsight clusters, see the following topics:

- [Get started with ML Services cluster on HDInsight](#)
- [Compute context options for ML Services cluster on HDInsight](#)
- [Storage options for ML Services cluster on HDInsight](#)

Get started with ML Services on Azure HDInsight

3/5/2019 • 6 minutes to read

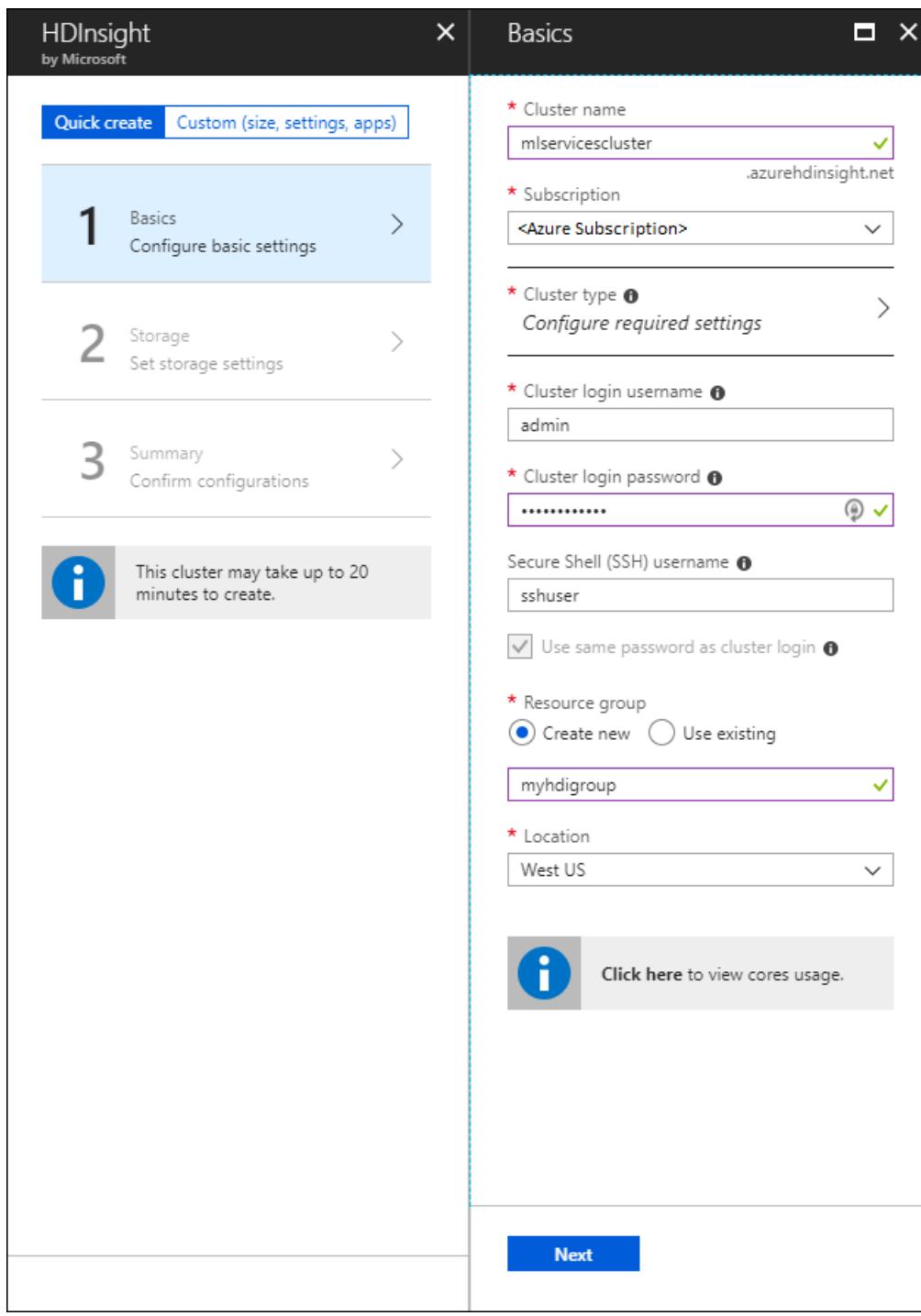
Azure HDInsight enables you to create an ML Services cluster. This option allows R scripts to use [Apache Spark](#) and [Apache Hadoop MapReduce](#) to run distributed computations. In this article, you learn how to create an ML Service cluster on HDInsight and how to run an R script that demonstrates using Spark for distributed R computations.

Prerequisites

- **An Azure subscription:** Before you begin this tutorial, you must have an Azure subscription. For more information, see [Get Microsoft Azure free trial](#).
- **A Secure Shell (SSH) client:** An SSH client is used to remotely connect to the HDInsight cluster and run commands directly on the cluster. For more information, see [Use SSH with HDInsight](#).

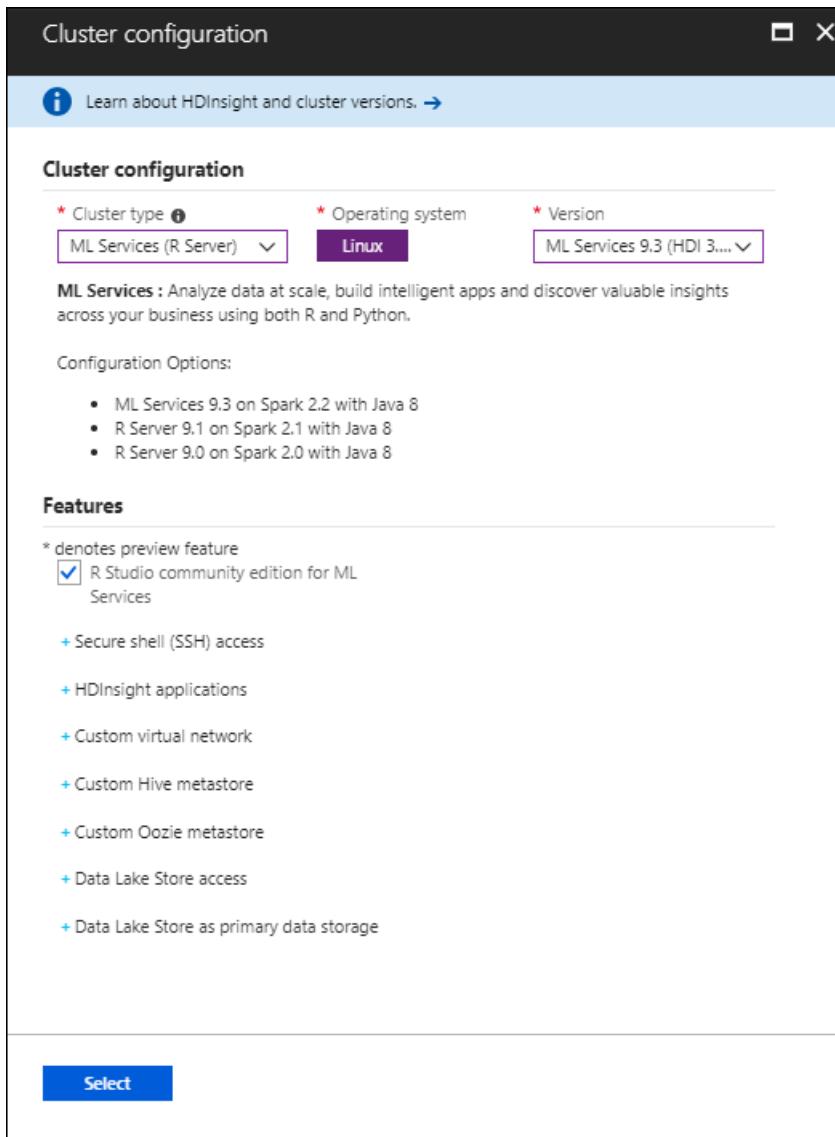
Create the cluster using the Azure portal

1. Sign in to the [Azure portal](#).
2. Navigate to + **Create a resource > Analytics > HDInsight**.
3. From **Basics**, enter the following information:
 - **Cluster Name:** The name of the HDInsight cluster.
 - **Subscription:** Select the subscription to use.
 - **Cluster login username** and **Cluster login password:** The login when accessing the cluster over HTTPS. You use these credentials to access services such as the Apache Ambari Web UI or REST API.
 - **Secure Shell (SSH) username:** The login used when accessing the cluster over SSH. By default the password is the same as the cluster login password.
 - **Resource Group:** The resource group to create the cluster in.
 - **Location:** The Azure region to create the cluster in.



4. Select **Cluster type**, and then set the following values in the **Cluster configuration** section:

- **Cluster Type:** ML Services
- **Operating system:** Linux
- **Version:** ML Server 9.3 (HDI 3.6). Release notes for ML Server 9.3 are available on docs.microsoft.com.
- **R Studio community edition for ML Server:** This browser-based IDE is installed by default on the edge node. Clear the check box if you prefer to not have it installed. If you choose to have it installed, the URL for accessing the RStudio Server login is available on the portal application blade for your cluster once it's been created.



5. After selecting the cluster type, use the **Select** button to set the cluster type. Next, use the **Next** button to finish basic configuration.
6. From the **Storage** section, select or create a Storage account. For the steps in this document, leave the other fields in this section at the default values. Use the **Next** button to save storage configuration.

Storage

The cluster will use this data source as the primary location for most data access, such as job input and log output.

Storage Account Settings

* Primary storage type
 Azure Storage Data Lake Store

* Selection method [i](#)
 My subscriptions Access key

* Create a new Storage account
myhdiclusterstorage 

[Select existing](#)

* Default container [i](#)
mlservicescluster-2018-06-20t22-49-46-117z

Additional storage accounts > Optional

Data Lake Store access [i](#) > Optional

Metastore Settings (optional)
Filtered to location and subscription of cluster.

To preserve your metadata outside this cluster, link a SQL database to this account.

Select a SQL database for Hive
No database in westus for subscription. 

Select a SQL database for Oozie
No database in westus for subscription. 

Next

7. From the **Summary** section, review the configuration for the cluster. Use the **Edit** links to change any settings that are incorrect. Finally, use the **Create** button to create the cluster.

Cluster summary

Click '**Create**' below to deploy your cluster. Once created you will be billed until your cluster is deleted.

Basics (Edit)

Cluster name	miservicescluster
Subscription	<Azure Subscription>
Cluster type	ML Services 9.3 on Linux (HDI 3.6)
Cluster login username	admin
SSH username	sshuser
Resource group	myhdigroup
Location	West US

Storage (Edit)

Azure Storage account	myhdiclusclusterstorage (new)
Additional Storage accounts	---
Metastores	---

Applications (optional) (Edit)

Applications (optional)	---
-------------------------	-----

Cluster size (Edit)

Nodes	Head (2 x D12 v2), Worker (4 x D4 v2), Zookeeper (3 x A2) MLServices (1 x D4 v2)
-------	---

Advanced settings (Edit)

Script actions	---
Virtual network	---

5.02 USD
USD/HOUR (ESTIMATED)

10 NODES (2 HEAD + 4 WORKER + 3 ZOOKEEPER + 1 RSERVER) 54 CORES - --
R SERVER 9.1 ON LINUX (HDI 3.6) MYHDISTORAGEACC (EAST US 2)

Create [Download template and parameters](#)

NOTE

It can take up to 20 minutes to create the cluster.

Connect to RStudio Server

If you chose to install RStudio Server Community Edition as part of your HDInsight cluster, then you can access the RStudio login using one of the following two methods:

- **Option 1** - Go to the following URL (where **CLUSTERNAME** is the name of the ML Services cluster you created):

```
https://CLUSTERNAME.azurehdinsight.net/rstudio/
```

- **Option 2** - Use the Azure portal. From the portal:

1. Select **All services** from the left menu.

2. Under **ANALYTICS**, select **HDInsight clusters**.
3. Select your cluster name from the **HDInsight clusters** page.
4. From **ML Services dashboards**, select **R Studio server**.

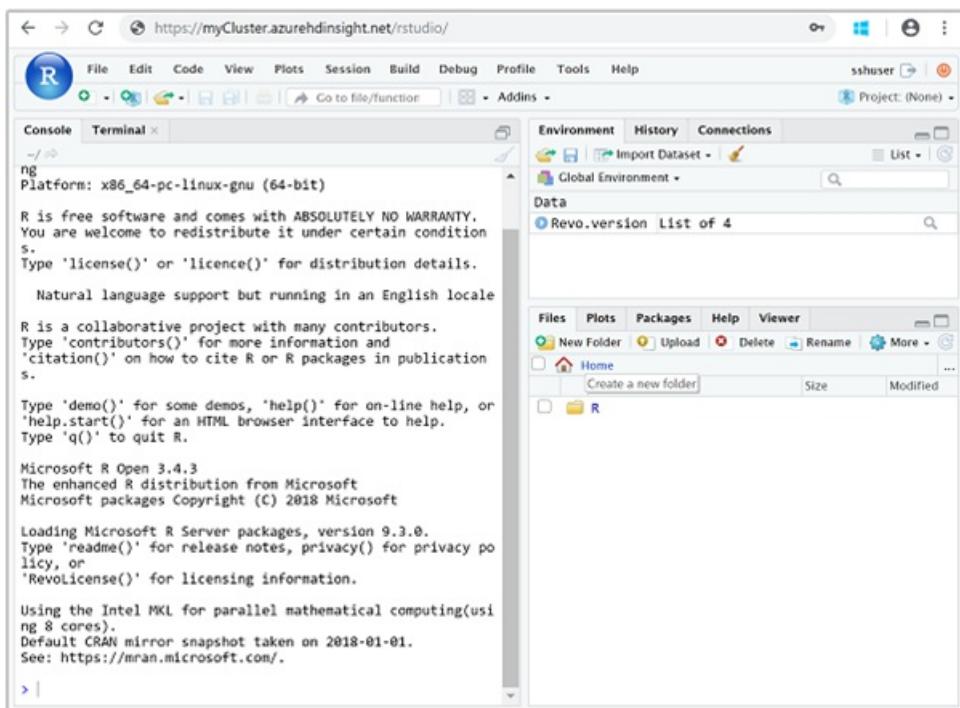
ML Services dashboards
Cluster management interfaces

- [Ambari home](#)
- [Zeppelin notebook](#)
- [Jupyter notebook](#)
- [R Studio server](#) (highlighted with a red box)
- [Spark history server](#)
- [Yarn](#)

IMPORTANT

Regardless of the method used, the first time you log in you need to authenticate twice. For the first authentication prompt, provide the *cluster Admin userid* and *password*. For the second authentication prompt, provide the *SSH userid* and *password*. Subsequent log-ins only require the SSH credentials.

Once you are connected, your screen should resemble the following screenshot:



Run a sample job

You can submit a job using ScaleR functions. Here is an example of the commands used to run a job:

```
# Set the HDFS (WASB) location of example data.
bigDataDirRoot <- "/example/data"

# Create a local folder for storing data temporarily.
source <- "/tmp/AirOnTimeCSV2012"
dir.create(source)
```

```

# Download data to the tmp folder.
remoteDir <- "https://packages.revolutionanalytics.com/datasets/AirOnTimeCSV2012"
download.file(file.path(remoteDir, "airOT201201.csv"), file.path(source, "airOT201201.csv"))
download.file(file.path(remoteDir, "airOT201202.csv"), file.path(source, "airOT201202.csv"))
download.file(file.path(remoteDir, "airOT201203.csv"), file.path(source, "airOT201203.csv"))
download.file(file.path(remoteDir, "airOT201204.csv"), file.path(source, "airOT201204.csv"))
download.file(file.path(remoteDir, "airOT201205.csv"), file.path(source, "airOT201205.csv"))
download.file(file.path(remoteDir, "airOT201206.csv"), file.path(source, "airOT201206.csv"))
download.file(file.path(remoteDir, "airOT201207.csv"), file.path(source, "airOT201207.csv"))
download.file(file.path(remoteDir, "airOT201208.csv"), file.path(source, "airOT201208.csv"))
download.file(file.path(remoteDir, "airOT201209.csv"), file.path(source, "airOT201209.csv"))
download.file(file.path(remoteDir, "airOT201210.csv"), file.path(source, "airOT201210.csv"))
download.file(file.path(remoteDir, "airOT201211.csv"), file.path(source, "airOT201211.csv"))
download.file(file.path(remoteDir, "airOT201212.csv"), file.path(source, "airOT201212.csv"))

# Set directory in bigDataDirRoot to load the data.
inputDir <- file.path(bigDataDirRoot,"AirOnTimeCSV2012")

# Create the directory.
rxHadoopMakeDir(inputDir)

# Copy the data from source to input.
rxHadoopCopyFromLocal(source, bigDataDirRoot)

# Define the HDFS (WASB) file system.
hdfsFS <- RxHdfsFileSystem()

# Create info list for the airline data.
airlineColInfo <- list(
  DAY_OF_WEEK = list(type = "factor"),
  ORIGIN = list(type = "factor"),
  DEST = list(type = "factor"),
  DEP_TIME = list(type = "integer"),
  ARR_DEL15 = list(type = "logical"))

# Get all the column names.
varNames <- names(airlineColInfo)

# Define the text data source in HDFS.
airOnTimeData <- RxTextData(inputDir, colInfo = airlineColInfo, varsToKeep = varNames, fileSystem = hdfsFS)

# Define the text data source in local system.
airOnTimeDataLocal <- RxTextData(source, colInfo = airlineColInfo, varsToKeep = varNames)

# Specify the formula to use.
formula = "ARR_DEL15 ~ ORIGIN + DAY_OF_WEEK + DEP_TIME + DEST"

# Define the Spark compute context.
mySparkCluster <- RxSpark()

# Set the compute context.
rxSetComputeContext(mySparkCluster)

# Run a logistic regression.
system.time(
  modelSpark <- rxLogit(formula, data = airOnTimeData)
)

# Display a summary.
summary(modelSpark)

```

Connect to the cluster edge node

In this section, you learn how to connect to the edge node of an ML Services HDInsight cluster using SSH. For familiarity on using SSH, see [Use SSH with HDInsight](#).

The SSH command to connect to the ML Services cluster edge node is:

```
ssh USERNAME@CLUSTERNAME-ed-ssh.azurehdinsight.net
```

To find the SSH command for your cluster, from the Azure portal click the cluster name, click **SSH + Cluster login**, and then for **Hostname**, select the edge node. This displays the SSH Endpoint information for the edge node.

The screenshot shows the Azure HDInsight cluster dashboard for a cluster named "myrcluster". The left sidebar has a "SSH + Cluster login" section highlighted with a red box. The main content area shows the "Connect to cluster using secure shell (SSH)" section, which includes a "Hostname" dropdown set to "myrcluster-ed-ssh.azurehdinsight.net" and a "ssh sshuser@myrcluster-ed-ssh.azurehdinsight.net" text input field. Below this is the "Connect to cluster using Cluster Login" section, which includes a "Cluster login username" input field set to "admin". A "Reset credential" button is at the bottom.

If you used a password to secure your SSH user account, you are prompted to enter it. If you used a public key, you may have to use the `-i` parameter to specify the matching private key. For example:

```
ssh -i ~/.ssh/id_rsa USERNAME@CLUSTERNAME-ed-ssh.azurehdinsight.net
```

Once connected, you get at a prompt similar to the following:

```
sshuser@ed00-myrcclu:~$
```

Use the R console

- From the SSH session, use the following command to start the R console:

```
R
```

- You should see an output with the version of ML Server, in addition to other information.

- From the `>` prompt, you can enter R code. ML Services on HDInsight includes packages that allow you to easily interact with Hadoop and run distributed computations. For example, use the following command to view the root of the default file system for the HDInsight cluster:

```
rxHadoopListFiles("/")
```

- You can also use the WASB style addressing.

```
rxHadoopListFiles("wasb:///")
```

5. To quit the R console, use the following command:

```
quit()
```

Automated cluster creation

You can automate the creation of ML Services cluster for HDInsight by using the SDK and the PowerShell.

- To create an ML Services cluster using the .NET SDK, see [Create Linux-based clusters in HDInsight using the .NET SDK](#).
- To create an ML Services cluster using powershell, see the article on [Create HDInsight clusters using Azure PowerShell](#).

Delete the cluster

WARNING

Billing for HDInsight clusters is prorated per minute, whether you use them or not. Be sure to delete your cluster after you finish using it. See [how to delete an HDInsight cluster](#).

Troubleshoot

If you run into issues with creating HDInsight clusters, see [access control requirements](#).

Next steps

In this article, you learned how to create a new ML Services cluster in Azure HDInsight and the basics of using the R console from an SSH session. The following articles explain other ways of managing and working with ML Services on HDInsight:

- [Submit jobs from R Tools for Visual Studio](#)
- [Manage ML Services cluster on HDInsight](#)
- [Operationalize ML Services cluster on HDInsight](#)
- [Compute context options for ML Services cluster on HDInsight](#)
- [Azure Storage options for ML Services cluster on HDInsight](#)

Create and share an Azure Machine Learning Studio workspace

2/25/2019 • 2 minutes to read

To use Azure Machine Learning Studio, you need to have a Machine Learning Studio workspace. This workspace contains the tools you need to create, manage, and publish experiments.

Create a Studio workspace

1. Sign in to the [Azure portal](#)

NOTE

To sign in and create a Studio workspace, you need to be an Azure subscription administrator.

2. Click **+New**
3. In the search box, type **Machine Learning Studio Workspace** and select the matching item. Then, select click **Create** at the bottom of the page.
4. Enter your workspace information:
 - The *workspace name* may be up to 260 characters, not ending in a space. The name can't include these characters: < > * % & : \ ? + /
 - The *web service plan* you choose (or create), along with the associated *pricing tier* you select, is used if you deploy web services from this workspace.

Machine Learning Workspace □ X

Machine Learning Workspace

* Workspace name
My-workspace ✓

* Subscription
My-subscription

* Resource group ⓘ
 Create new Use existing
My-resource-group ✓

* Location
South Central US

* Storage account ⓘ
 Create new Use existing
storageformyworkspace ✓

Workspace pricing tier ⓘ
Standard

* Web service plan ⓘ
 Create new Use existing
My-web-service-plan ✓

* Web service plan pricing tier ⓘ >
S1 Standard

5. Click **Create**.

NOTE

Machine Learning Studio relies on an Azure storage account that you provide to save intermediary data when it executes the workflow. After the workspace is created, if the storage account is deleted, or if the access keys are changed, the workspace will stop functioning and all experiments in that workspace will fail. If you accidentally delete the storage account, recreate the storage account with the same name in the same region as the deleted storage account and resync the access key. If you changed storage account access keys, resync the access keys in the workspace by using the Azure portal.

Once the workspace is deployed, you can open it in Machine Learning Studio.

1. Browse to Machine Learning Studio at <https://studio.azureml.net/>.
2. Select your workspace in the upper-right-hand corner.



3. Click **my experiments**.

Welcome back luisa!

MY RECENT WORKSPACES:

 My-workspace

MY RECENT EXPERIMENTS:

my experiments 

For information about managing your Studio workspace, see [Manage an Azure Machine Learning Studio workspace](#). If you encounter a problem creating your workspace, see [Troubleshooting guide: Create and connect to a Machine Learning Studio workspace](#).

Share an Azure Machine Learning Studio workspace

Once a Machine Learning Studio workspace is created, you can invite users to your workspace to share access to your workspace and all its experiments, datasets, notebooks, etc. You can add users in one of two roles:

- **User** - A workspace user can create, open, modify, and delete experiments, datasets, etc. in the workspace.
- **Owner** - An owner can invite and remove users in the workspace, in addition to what a user can do.

NOTE

The administrator account that creates the workspace is automatically added to the workspace as workspace Owner. However, other administrators or users in that subscription are not automatically granted access to the workspace - you need to invite them explicitly.

To share a Studio workspace

1. Sign in to Machine Learning Studio at <https://studio.azureml.net/Home>
2. In the left panel, click **SETTINGS**
3. Click the **USERS** tab
4. Click **INVITE MORE USERS** at the bottom of the page

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a sidebar with icons for Projects, Experiments, Web Services, Notebooks, Datasets, Trained Models, and Settings. The 'SETTINGS' button is highlighted with a red box. At the bottom of the sidebar, there's a 'NEW' button with a plus sign and an 'INVITE MORE USERS' button with a person icon, also highlighted with a red box. The main area is titled 'settings' and has tabs for NAME, AUTHORIZATION TOKENS, USERS (which is selected and highlighted with a red box), and DATA GATEWAYS. Below these tabs is a table with columns for NAME, EMAIL, ROLE, and STATUS. One row is shown: 'luisa' with email 'luisa@contoso.com', role 'Owner', and status 'Active'. There's also a magnifying glass icon for search.

5. Enter one or more email addresses. The users need a valid Microsoft account or an organizational account (from Azure Active Directory).
6. Select whether you want to add the users as Owner or User.
7. Click the **OK** checkmark button.

Each user you add will receive an email with instructions on how to sign in to the shared workspace.

NOTE

For users to be able to deploy or manage web services in this workspace, they must be a contributor or administrator in the Azure subscription.

How to identify scenarios and plan for advanced analytics data processing

3/12/2019 • 4 minutes to read

What resources are required for you to create an environment that can perform advanced analytics processing on a dataset? This article suggests a series of questions to ask that can help identify tasks and resources relevant to your scenario.

To learn about the order of high-level steps for predictive analytics, see [What is the Team Data Science Process \(TDSP\)](#). Each step requires specific resources for the tasks relevant to your particular scenario.

Answer key questions in the following areas to identify your scenario:

- data logistics
- data characteristics
- dataset quality
- preferred tools and languages

Try [Azure Machine Learning Studio](#), available in paid or free options.

Logistic questions: data locations and movement

The logistic questions cover the following items:

- data source location
- target destination in Azure
- requirements for moving the data, including the schedule, amount, and resources involved

You may need to move the data several times during the analytics process. A common scenario is to move local data into some form of storage on Azure and then into Machine Learning Studio.

What is your data source?

Is your data local or in the cloud? Possible locations include:

- a publicly available HTTP address
- a local or network file location
- a SQL Server database
- an Azure storage container

What is the Azure destination?

Where does your data need to be for processing or modeling?

- Azure Blob Storage
- SQL Azure databases
- SQL Server on Azure VM
- HDInsight (Hadoop on Azure) or Hive tables
- Azure Machine Learning
- Mountable Azure virtual hard disks

How are you going to move the data?

For procedures and resources to ingest or load data into a variety of different storage and processing environments, see:

- [Load data into storage environments for analytics](#)
- [Import your training data into Azure Machine Learning Studio from various data sources](#)

Does the data need to be moved on a regular schedule or modified during migration?

Consider using Azure Data Factory (ADF) when data needs to be continually migrated. ADF can be helpful for:

- a hybrid scenario that involves both on-premises and cloud resources
- a scenario where the data is transacted, modified, or changed by business logic in the course of being migrated

For further information, see [Move data from an on-premises SQL server to SQL Azure with Azure Data Factory](#).

How much of the data is to be moved to Azure?

Extremely large datasets may exceed the storage capacity of certain environments. For an example, see the discussion of size limits for Machine Learning Studio in the next section. In such cases, you might use a sample of the data during the analysis. For details of how to down-sample a dataset in various Azure environments, see [Sample data in the Team Data Science Process](#).

Data characteristics questions: type, format, and size

These questions are key to planning your storage and processing environments. They will help you choose the appropriate scenario for your data type and understand any restrictions.

What are the data types?

- Numerical
- Categorical
- Strings
- Binary

How is your data formatted?

- Comma-separated (CSV) or tab-separated (TSV) flat files
- Compressed or uncompressed
- Azure blobs
- Hadoop Hive tables
- SQL Server tables

How large is your data?

- Small: Less than 2 GB
- Medium: Greater than 2 GB and less than 10 GB
- Large: Greater than 10 GB

Take the Azure Machine Learning Studio environment for example:

- For a list of the data formats and types supported by Azure Machine Learning Studio, see [Data formats and data types supported](#) section.
- For information on the limitations of other Azure services used in the analytics process, see [Azure Subscription and Service Limits, Quotas, and Constraints](#).

Data quality questions: exploration and pre-processing

What do you know about your data?

Understand the basic characteristics about your data:

- What patterns or trends it exhibits
- What outliers it has
- How many values are missing

This step is important to help you:

- Determine how much pre-processing is needed
- Formulate hypotheses that suggest the most appropriate features or type of analysis
- Formulate plans for additional data collection

Useful techniques for data inspection include descriptive statistics calculation and visualization plots. For details of how to explore a dataset in various Azure environments, see [Explore data in the Team Data Science Process](#).

Does the data require preprocessing or cleaning?

You might need to preprocess and clean your data before you can use the dataset effectively for machine learning. Raw data is often noisy and unreliable. It might be missing values. Using such data for modeling can produce misleading results. For a description, see [Tasks to prepare data for enhanced machine learning](#).

Tools and languages questions

There are many options for languages, development environments, and tools. Be aware of your needs and preferences.

What languages do you prefer to use for analysis?

- R
- Python
- SQL

What tools should you use for data analysis?

- [Microsoft Azure Powershell](#) - a script language used to administer your Azure resources in a script language
- [Azure Machine Learning Studio](#)
- [Revolution Analytics](#)
- [RStudio](#)
- [Python Tools for Visual Studio](#)
- [Anaconda](#)
- [Jupyter notebooks](#)
- [Microsoft Power BI](#)

Identify your advanced analytics scenario

After you have answered the questions in the previous section, you are ready to determine which scenario best fits your case. The sample scenarios are outlined in [Scenarios for advanced analytics in Azure Machine Learning](#).

Next steps

[What is the Team Data Science Process \(TDSP\)?](#)

Load data into storage environments for analytics

1/30/2019 • 2 minutes to read

The Team Data Science Process requires that data be ingested or loaded into a variety of different storage environments to be processed or analyzed in the most appropriate way in each stage of the process. Data destinations commonly used for processing include Azure Blob Storage, SQL Azure databases, SQL Server on Azure VM, HDInsight (Hadoop), and Azure Machine Learning.

The following articles describe how to ingest data into various target environments where the data is stored and processed.

- To/From [Azure Blob Storage](#)
- To [SQL Server on Azure VM](#)
- To [Azure SQL database](#)
- To [Hive tables](#)
- To [SQL partitioned tables](#)
- From [On-premises SQL Server](#)

Technical and business needs, as well as the initial location, format, and size of your data will determine the target environments into which the data needs to be ingested to achieve the goals of your analysis. It is not uncommon for a scenario to require data to be moved between several environments to achieve the variety of tasks required to construct a predictive model. This sequence of tasks can include, for example, data exploration, pre-processing, cleaning, down-sampling, and model training.

Move data to and from Azure Blob storage

1/30/2019 • 2 minutes to read

The Team Data Science Process requires that data be ingested or loaded into a variety of different storage environments to be processed or analyzed in the most appropriate way in each stage of the process.

Different technologies for moving data

The following articles describe how to move data to and from Azure Blob storage using different technologies.

- [Azure Storage-Explorer](#)
- [AzCopy](#)
- [Python](#)
- [SSIS](#)

Which method is best for you depends on your scenario. The [Scenarios for advanced analytics in Azure Machine Learning](#) article helps you determine the resources you need for a variety of data science workflows used in the advanced analytics process.

NOTE

For a complete introduction to Azure blob storage, refer to [Azure Blob Basics](#) and to [Azure Blob Service](#).

Using Azure Data Factory

As an alternative, you can use [Azure Data Factory](#) to:

- create and schedule a pipeline that downloads data from Azure blob storage,
- pass it to a published Azure Machine Learning web service,
- receive the predictive analytics results, and
- upload the results to storage.

For more information, see [Create predictive pipelines using Azure Data Factory and Azure Machine Learning](#).

Prerequisites

This article assumes that you have an Azure subscription, a storage account, and the corresponding storage key for that account. Before uploading/downloading data, you must know your Azure storage account name and account key.

- To set up an Azure subscription, see [Free one-month trial](#).
- For instructions on creating a storage account and for getting account and key information, see [About Azure storage accounts](#).

Move data to and from Azure Blob Storage using Azure Storage Explorer

3/12/2019 • 2 minutes to read

Azure Storage Explorer is a free tool from Microsoft that allows you to work with Azure Storage data on Windows, macOS, and Linux. This topic describes how to use it to upload and download data from Azure blob storage. The tool can be downloaded from [Microsoft Azure Storage Explorer](#).

This menu links to technologies you can use to move data to and from Azure Blob storage:

NOTE

If you are using VM that was set up with the scripts provided by [Data Science Virtual machines in Azure](#), then Azure Storage Explorer is already installed on the VM.

NOTE

For a complete introduction to Azure blob storage, refer to [Azure Blob Basics](#) and [Azure Blob Service](#).

Prerequisites

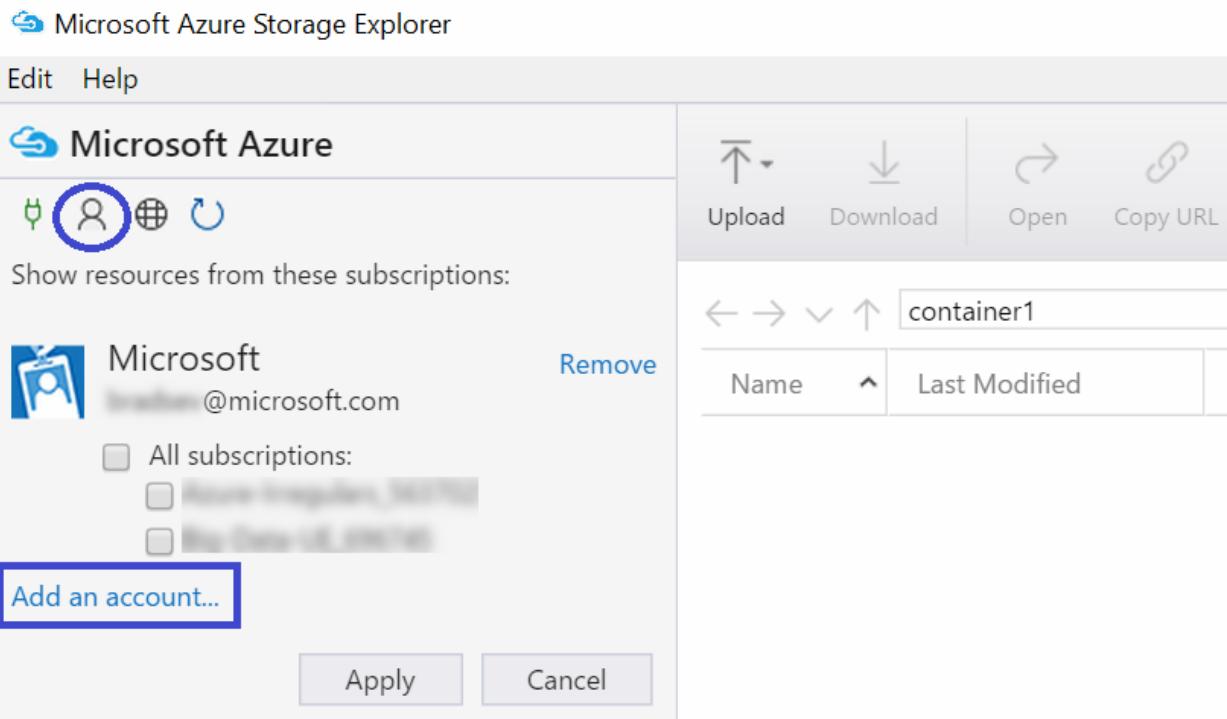
This document assumes that you have an Azure subscription, a storage account, and the corresponding storage key for that account. Before uploading/downloading data, you must know your Azure storage account name and account key.

- To set up an Azure subscription, see [Free one-month trial](#).
- For instructions on creating a storage account and for getting account and key information, see [About Azure storage accounts](#). Make a note the access key for your storage account as you need this key to connect to the account with the Azure Storage Explorer tool.
- The Azure Storage Explorer tool can be downloaded from [Microsoft Azure Storage Explorer](#). Accept the defaults during install.

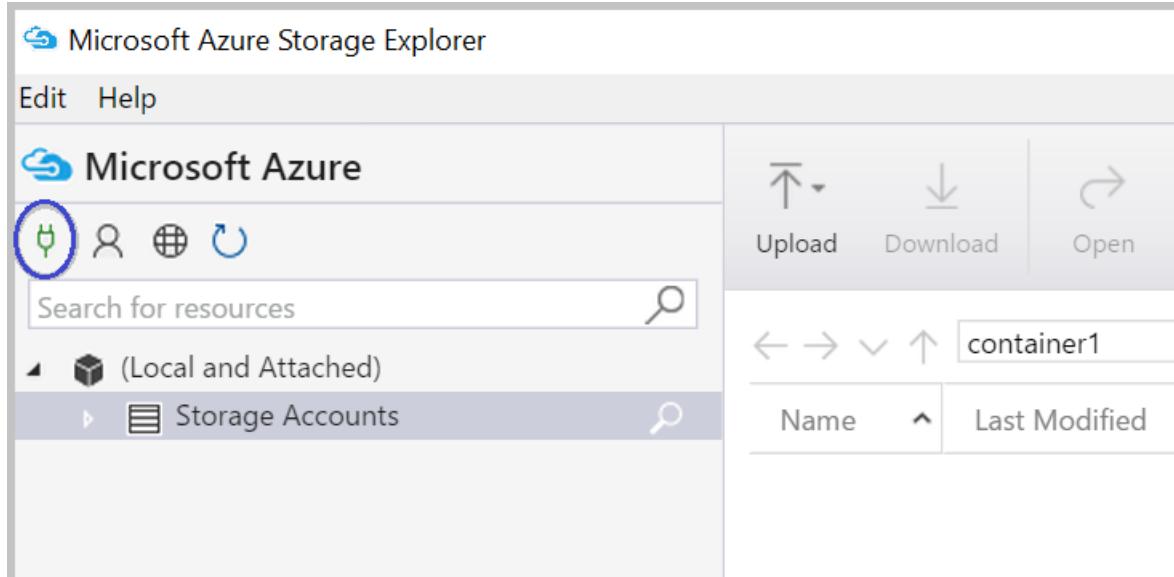
Use Azure Storage Explorer

The following steps document how to upload/download data using Azure Storage Explorer.

1. Launch Microsoft Azure Storage Explorer.
2. To bring up the **Sign in to your account...** wizard, select **Azure account settings** icon, then **Add an account** and enter you credentials.



3. To bring up the **Connect to Azure Storage** wizard, select the **Connect to Azure storage** icon.



4. Enter the access key from your Azure storage account on the **Connect to Azure Storage** wizard and then **Next**.

Connect to Azure Storage

Enter a connection string, Shared Access Signature (SAS) URI, or an account key.

Back

Next

Connect

Cancel

5. Enter storage account name in the **Account name** box and then select **Next**.

Attach External Storage

Enter information to connect to the Microsoft Azure storage account

Account name:

 Enter a storage account name

Account key:

Storage endpoints domain:

- Microsoft Azure Default
- Microsoft Azure China
- Other (specify below)

 core.windows.net

- Use HTTP (Not recommended)

[Online privacy statement](#)

Back

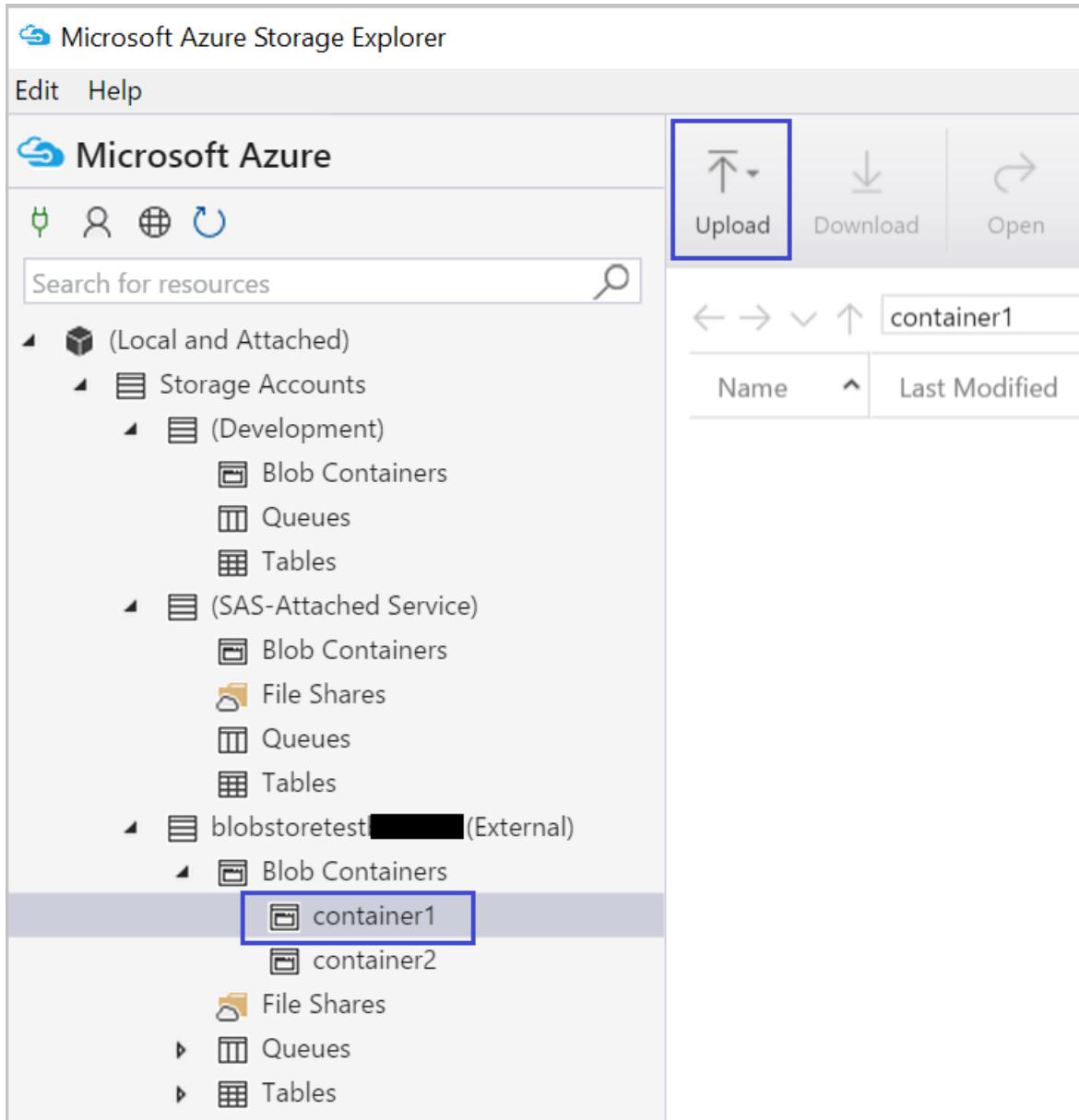
Next

Connect

Cancel

6. The storage account added should now be listed. To create a blob container in a storage account, right-click the **Blob Containers** node in that account, select **Create Blob Container**, and enter a name.

7. To upload data to a container, select the target container and click the **Upload** button.



8. Click on the ... to the right of the **Files** box, select one or multiple files to upload from the file system and click **Upload** to begin uploading the files.

Upload files

Files

No files selected



Blob type

Block Blob



Upload .vhdx files as page blobs (recommended)

Upload to folder (optional)

Upload

Cancel

9. To download data, selecting the blob in the corresponding container to download and click **Download**.

The screenshot shows the Microsoft Azure Storage Explorer interface. On the left, there is a navigation pane with sections for Local and Attached storage accounts, SAS-Attached Service, and External accounts. Under blobstoretestbradsev (External), the container1 is selected. In the main pane, the 'Download' button is highlighted with a blue box. Below it, a file named 'trip_data_1.csv.zip' is listed with its last modified date as 'Wed, 31 Aug 2016 14:31:13 GMT'. The file name is also highlighted with a blue box.

Copy data to and from Azure Blob Storage using AzCopy

2/12/2019 • 2 minutes to read

AzCopy is a command-line utility designed for uploading, downloading, and copying data to and from Microsoft Azure blob, file, and table storage.

For instructions on installing AzCopy and additional information on using it with the Azure platform, see [Getting Started with the AzCopy Command-Line Utility](#).

This menu links to technologies you can use to move data to and from Azure Blob storage:

NOTE

If you are using VM that was set up with the scripts provided by [Data Science Virtual machines in Azure](#), then AzCopy is already installed on the VM.

NOTE

For a complete introduction to Azure blob storage, refer to [Azure Blob Basics](#) and to [Azure Blob Service](#).

Prerequisites

This document assumes that you have an Azure subscription, a storage account and the corresponding storage key for that account. Before uploading/downloading data, you must know your Azure storage account name and account key.

- To set up an Azure subscription, see [Free one-month trial](#).
- For instructions on creating a storage account and for getting account and key information, see [About Azure storage accounts](#).

Run AzCopy commands

To run AzCopy commands, open a command window and navigate to the AzCopy installation directory on your computer, where the AzCopy.exe executable is located.

The basic syntax for AzCopy commands is:

```
AzCopy /Source:<source> /Dest:<destination> [Options]
```

NOTE

You can add the AzCopy installation location to your system path and then run the commands from any directory. By default, AzCopy is installed to %ProgramFiles(x86)%\Microsoft SDKs\Azure\AzCopy or %ProgramFiles%\Microsoft SDKs\Azure\AzCopy.

Upload files to an Azure blob

To upload a file, use the following command:

```
# Upload from local file system
AzCopy /Source:<your_local_directory> /Dest:
https://<your_account_name>.blob.core.windows.net/<your_container_name> /DestKey:<your_account_key> /S
```

Download files from an Azure blob

To download a file from an Azure blob, use the following command:

```
# Downloading blobs to local file system
AzCopy
/Source:https://<your_account_name>.blob.core.windows.net/<your_container_name>/<your_sub_directory_at_blob>
/Dest:<your_local_directory> /SourceKey:<your_account_key> /Pattern:<file_pattern> /S
```

Copy blobs between Azure containers

To copy blobs between Azure containers, use the following command:

```
# Copying blobs between Azure containers
AzCopy
/Source:https://<your_account_name1>.blob.core.windows.net/<your_container_name1>/<your_sub_directory_at_blob1>
/Dest:https://<your_account_name2>.blob.core.windows.net/<your_container_name2>/<your_sub_directory_at_blob2>
/SourceKey:<your_account_key1> /DestKey:<your_account_key2> /Pattern:<file_pattern> /S

<your_account_name>: your storage account name
<your_account_key>: your storage account key
<your_container_name>: your container name
<your_sub_directory_at_blob>: the sub directory in the container
<your_local_directory>: directory of local file system where files to be uploaded from or the directory of local file system files to be downloaded to
<file_pattern>: pattern of file names to be copied. The standard wildcards are supported
```

Tips for using AzCopy

TIP

1. When **uploading** files, **/S** uploads files recursively. Without this parameter, files in subdirectories are not uploaded.
2. When **downloading** file, **/S** searches the container recursively until all files in the specified directory and its subdirectories, or all files that match the specified pattern in the given directory and its subdirectories, are downloaded.
3. You cannot specify a **specific blob file** to download using the **/Source** parameter. To download a specific file, specify the blob file name to download using the **/Pattern** parameter. **/S** parameter can be used to have AzCopy look for a file name pattern recursively. Without the pattern parameter, AzCopy downloads all files in that directory.

2 minutes to read

Move data to or from Azure Blob Storage using SSIS connectors

3/12/2019 • 3 minutes to read

The [SQL Server Integration Services Feature Pack for Azure](#) provides components to connect to Azure, transfer data between Azure and on-premises data sources, and process data stored in Azure.

This menu links to technologies you can use to move data to and from Azure Blob storage:

Once customers have moved on-premises data into the cloud, they can access it from any Azure service to leverage the full power of the suite of Azure technologies. It may be used, for example, in Azure Machine Learning or on an HDInsight cluster.

This is typically be the first step for the [SQL](#) and [HDInsight](#) walkthroughs.

For a discussion of canonical scenarios that use SSIS to accomplish business needs common in hybrid data integration scenarios, see [Doing more with SQL Server Integration Services Feature Pack for Azure](#) blog.

NOTE

For a complete introduction to Azure blob storage, refer to [Azure Blob Basics](#) and to [Azure Blob Service](#).

Prerequisites

To perform the tasks described in this article, you must have an Azure subscription and an Azure storage account set up. You must know your Azure storage account name and account key to upload or download data.

- To set up an **Azure subscription**, see [Free one-month trial](#).
- For instructions on creating a **storage account** and for getting account and key information, see [About Azure storage accounts](#).

To use the **SSIS connectors**, you must download:

- **SQL Server 2014 or 2016 Standard (or above)**: Install includes SQL Server Integration Services.
- **Microsoft SQL Server 2014 or 2016 Integration Services Feature Pack for Azure**: These can be downloaded, respectively, from the [SQL Server 2014 Integration Services](#) and [SQL Server 2016 Integration Services](#) pages.

NOTE

SSIS is installed with SQL Server, but is not included in the Express version. For information on what applications are included in various editions of SQL Server, see [SQL Server Editions](#)

For training materials on SSIS, see [Hands On Training for SSIS](#)

For information on how to get up-and-running using SSIS to build simple extraction, transformation, and load (ETL) packages, see [SSIS Tutorial: Creating a Simple ETL Package](#).

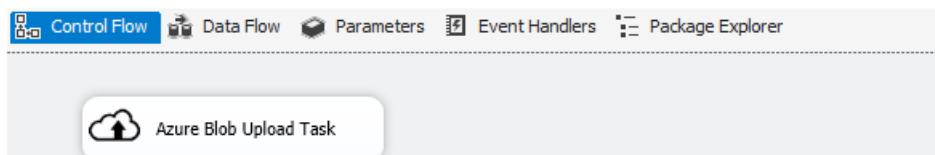
Download NYC Taxi dataset

The example described here use a publicly available dataset -- the [NYC Taxi Trips](#) dataset. The dataset consists of

about 173 million taxi rides in NYC in the year 2013. There are two types of data: trip details data and fare data. As there is a file for each month, we have 24 files in all, each of which is approximately 2GB uncompressed.

Upload data to Azure blob storage

To move data using the SSIS feature pack from on-premises to Azure blob storage, we use an instance of the [Azure Blob Upload Task](#), shown here:



The parameters that the task uses are described here:

FIELD	DESCRIPTION
AzureStorageConnection	Specifies an existing Azure Storage Connection Manager or creates a new one that refers to an Azure storage account that points to where the blob files are hosted.
BlobContainer	Specifies the name of the blob container that hold the uploaded files as blobs.
BlobDirectory	Specifies the blob directory where the uploaded file is stored as a block blob. The blob directory is a virtual hierarchical structure. If the blob already exists, it is replaced.
LocalDirectory	Specifies the local directory that contains the files to be uploaded.
FileName	Specifies a name filter to select files with the specified name pattern. For example, MySheet*.xls* includes files such as MySheet001.xls and MySheetABC.xlsx
TimeRangeFrom/TimeRangeTo	Specifies a time range filter. Files modified after <i>TimeRangeFrom</i> and before <i>TimeRangeTo</i> are included.

NOTE

The **AzureStorageConnection** credentials need to be correct and the **BlobContainer** must exist before the transfer is attempted.

Download data from Azure blob storage

To download data from Azure blob storage to on-premises storage with SSIS, use an instance of the [Azure Blob Download Task](#).

More advanced SSIS-Azure scenarios

The SSIS feature pack allows for more complex flows to be handled by packaging tasks together. For example, the blob data could feed directly into an HDInsight cluster, whose output could be downloaded back to a blob and then to on-premises storage. SSIS can run Hive and Pig jobs on an HDInsight cluster using additional SSIS connectors:

- To run a Hive script on an Azure HDInsight cluster with SSIS, use [Azure HDInsight Hive Task](#).
- To run a Pig script on an Azure HDInsight cluster with SSIS, use [Azure HDInsight Pig Task](#).

Move data to SQL Server on an Azure virtual machine

3/12/2019 • 8 minutes to read

This article outlines the options for moving data either from flat files (CSV or TSV formats) or from an on-premises SQL Server to SQL Server on an Azure virtual machine. These tasks for moving data to the cloud are part of the Team Data Science Process.

For a topic that outlines the options for moving data to an Azure SQL Database for Machine Learning, see [Move data to an Azure SQL Database for Azure Machine Learning](#).

The following table summarizes the options for moving data to SQL Server on an Azure virtual machine.

SOURCE	DESTINATION: SQL SERVER ON AZURE VM
Flat File	<ol style="list-style-type: none">1. Command-line bulk copy utility (BCP)2. Bulk Insert SQL Query3. Graphical Built-in Utilities in SQL Server
On-Premises SQL Server	<ol style="list-style-type: none">1. Deploy a SQL Server Database to a Microsoft Azure VM wizard2. Export to a flat File3. SQL Database Migration Wizard4. Database back up and restore

Note that this document assumes that SQL commands are executed from SQL Server Management Studio or Visual Studio Database Explorer.

TIP

As an alternative, you can use [Azure Data Factory](#) to create and schedule a pipeline that will move data to a SQL Server VM on Azure. For more information, see [Copy data with Azure Data Factory \(Copy Activity\)](#).

Prerequisites

This tutorial assumes you have:

- An **Azure subscription**. If you do not have a subscription, you can sign up for a [free trial](#).
- An **Azure storage account**. You will use an Azure storage account for storing the data in this tutorial. If you don't have an Azure storage account, see the [Create a storage account](#) article. After you have created the storage account, you will need to obtain the account key used to access the storage. See [Manage your storage access keys](#).
- Provisioned **SQL Server on an Azure VM**. For instructions, see [Set up an Azure SQL Server virtual machine as an IPython Notebook server for advanced analytics](#).
- Installed and configured **Azure PowerShell** locally. For instructions, see [How to install and configure Azure PowerShell](#).

Moving data from a flat file source to SQL Server on an Azure VM

If your data is in a flat file (arranged in a row/column format), it can be moved to SQL Server VM on Azure via the

following methods:

1. [Command-line bulk copy utility \(BCP\)](#)
2. [Bulk Insert SQL Query](#)
3. [Graphical Built-in Utilities in SQL Server \(Import/Export, SSIS\)](#)

Command-line bulk copy utility (BCP)

BCP is a command-line utility installed with SQL Server and is one of the quickest ways to move data. It works across all three SQL Server variants (On-premises SQL Server, SQL Azure and SQL Server VM on Azure).

NOTE

Where should my data be for BCP?

While it is not required, having files containing source data located on the same machine as the target SQL Server allows for faster transfers (network speed vs local disk IO speed). You can move the flat files containing data to the machine where SQL Server is installed using various file copying tools such as [AZCopy](#), [Azure Storage Explorer](#) or windows copy/paste via Remote Desktop Protocol (RDP).

1. Ensure that the database and the tables are created on the target SQL Server database. Here is an example of how to do that using the [Create Database](#) and [Create Table](#) commands:

```
CREATE DATABASE <database_name>

CREATE TABLE <tablename>
(
    <columnname1> <datatype> <constraint>,
    <columnname2> <datatype> <constraint>,
    <columnname3> <datatype> <constraint>
)
```

1. Generate the format file that describes the schema for the table by issuing the following command from the command-line of the machine where bcp is installed.

```
bcp dbname..tablename format nul -c -x -f exportformatfilename.xml -S servername\sqlinstance -T -t \t -r \n
```

2. Insert the data into the database using the bcp command as follows. This should work from the command-line assuming that the SQL Server is installed on same machine:

```
bcp dbname..tablename in datafilename.tsv -f exportformatfilename.xml -S servername\sqlinstancename -U
username -P password -b block_size_to_move_in_single_attempt -t \t -r \n
```

Optimizing BCP Inserts Please refer the following article '[Guidelines for Optimizing Bulk Import](#)' to optimize such inserts.

Parallelizing Inserts for Faster Data Movement

If the data you are moving is large, you can speed things up by simultaneously executing multiple BCP commands in parallel in a PowerShell Script.

NOTE

Big data Ingestion To optimize data loading for large and very large datasets, partition your logical and physical database tables using multiple file groups and partition tables. For more information about creating and loading data to partition tables, see [Parallel Load SQL Partition Tables](#).

The following sample PowerShell script demonstrates parallel inserts using bcp:

```

$NO_OF_PARALLEL_JOBS=2

Set-ExecutionPolicy RemoteSigned #set execution policy for the script to execute
# Define what each job does
$ScriptBlock = {
    param($partitionnumber)

    #Explicitly using SQL username password
    bcp database..tablename in datafile_path.csv -F 2 -f format_file_path.xml -U username@servername -S
    tcp:servername -P password -b block_size_to_move_in_single_attempt -t "," -r \n -o
    path_to_outputfile.$partitionnumber.txt

    #Trusted connection w/o username password (if you are using windows auth and are signed in with that
    #credentials)
    #bcp database..tablename in datafile_path.csv -o path_to_outputfile.$partitionnumber.txt -h "TABLOCK" -F 2
    #-f format_file_path.xml -T -b block_size_to_move_in_single_attempt -t "," -r \n
}

# Background processing of all partitions
for ($i=1; $i -le $NO_OF_PARALLEL_JOBS; $i++)
{
    Write-Debug "Submit loading partition # $i"
    Start-Job $ScriptBlock -Arg $i
}

# Wait for it all to complete
While (Get-Job -State "Running")
{
    Start-Sleep 10
    Get-Job
}

# Getting the information back from the jobs
Get-Job | Receive-Job
Set-ExecutionPolicy Restricted #reset the execution policy

```

Bulk Insert SQL Query

[Bulk Insert SQL Query](#) can be used to import data into the database from row/column based files (the supported types are covered in the[Prepare Data for Bulk Export or Import \(SQL Server\)](#)) topic.

Here are some sample commands for Bulk Insert are as below:

1. Analyze your data and set any custom options before importing to make sure that the SQL Server database assumes the same format for any special fields such as dates. Here is an example of how to set the date format as year-month-day (if your data contains the date in year-month-day format):

```
SET DATEFORMAT ymd;
```

1. Import data using bulk import statements:

```

BULK INSERT <tablename>
FROM
'<datafilename>'
WITH
(
    FirstRow = 2,
    FIELDTERMINATOR = ',', --this should be column separator in your data
    ROWTERMINATOR = '\n'   --this should be the row separator in your data
)

```

Built-in Utilities in SQL Server

You can use SQL Server Integrations Services (SSIS) to import data into SQL Server VM on Azure from a flat file. SSIS is available in two studio environments. For details, see [Integration Services \(SSIS\) and Studio Environments](#):

- For details on SQL Server Data Tools, see [Microsoft SQL Server Data Tools](#)
- For details on the Import/Export Wizard, see [SQL Server Import and Export Wizard](#)

Moving Data from on-premises SQL Server to SQL Server on an Azure VM

You can also use the following migration strategies:

1. [Deploy a SQL Server Database to a Microsoft Azure VM wizard](#)
2. [Export to Flat File](#)
3. [SQL Database Migration Wizard](#)
4. [Database back up and restore](#)

We describe each of these below:

Deploy a SQL Server Database to a Microsoft Azure VM wizard

The **Deploy a SQL Server Database to a Microsoft Azure VM wizard** is a simple and recommended way to move data from an on-premises SQL Server instance to SQL Server on an Azure VM. For detailed steps as well as a discussion of other alternatives, see [Migrate a database to SQL Server on an Azure VM](#).

Export to Flat File

Various methods can be used to bulk export data from an On-Premises SQL Server as documented in the [Bulk Import and Export of Data \(SQL Server\)](#) topic. This document will cover the Bulk Copy Program (BCP) as an example. Once data is exported into a flat file, it can be imported to another SQL server using bulk import.

1. Export the data from on-premises SQL Server to a File using the bcp utility as follows

```
bcp dbname..tablename out datafile.tsv -S servername\sqlinstancename -T -t \t -t \n -c
```

2. Create the database and the table on SQL Server VM on Azure using the `create database` and `create table` for the table schema exported in step 1.
3. Create a format file for describing the table schema of the data being exported/imported. Details of the format file are described in [Create a Format File \(SQL Server\)](#).

Format file generation when running BCP from the SQL Server machine

```
bcp dbname..tablename format nul -c -x -f exportformatfilename.xml -S servername\sqlinstance -T -t \t -r \n
```

Format file generation when running BCP remotely against a SQL Server

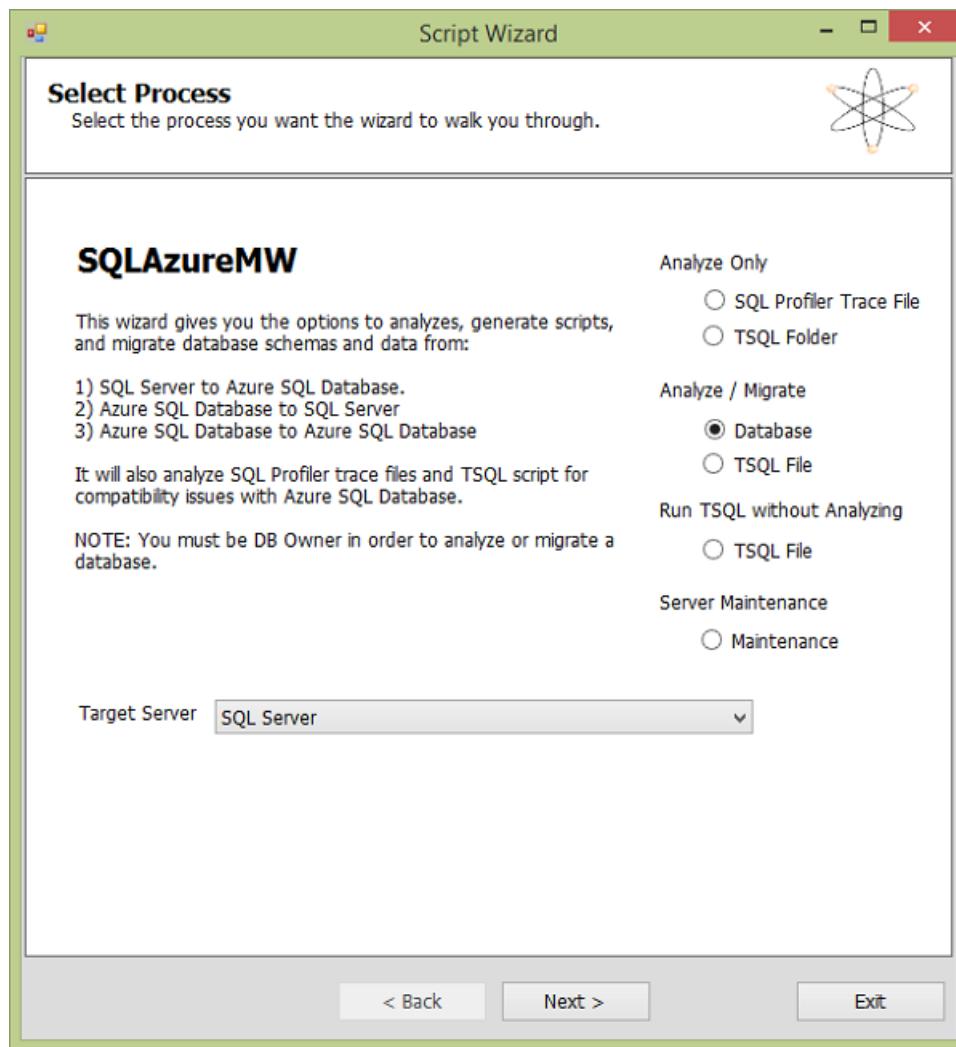
```
bcp dbname..tablename format nul -c -x -f exportformatfilename.xml -U username@servername.database.windows.net -S tcp:servername -P password --t \t -r \n
```

4. Use any of the methods described in section [Moving Data from File Source](#) to move the data in flat files to a SQL Server.

SQL Database Migration Wizard

[SQL Server Database Migration Wizard](#) provides a user-friendly way to move data between two SQL server

instances. It allows the user to map the data schema between sources and destination tables, choose column types and various other functionalities. It uses bulk copy (BCP) under the covers. A screenshot of the welcome screen for the SQL Database Migration wizard is shown below.



Database back up and restore

SQL Server supports:

1. [Database back up and restore functionality](#) (both to a local file or bacpac export to blob) and [Data Tier Applications](#) (using bacpac).
2. Ability to directly create SQL Server VMs on Azure with a copied database or copy to an existing SQL Azure database. For more details, see [Use the Copy Database Wizard](#).

A screenshot of the Database back up/restore options from SQL Server Management Studio is shown below.

Screenshot of Microsoft SQL Server Management Studio (SSMS) showing a database named 'aml_process_createstest' in Object Explorer. The 'Tasks' context menu is open over a table named 'tblProcessCreate'. The menu options include:

- Detach...
- Take Offline
- Bring Online
- Shrink
- Back Up...
- Restore...
- Mirror...
- Launch Database Mirroring Monitor...
- Ship Transaction Log(s)...
- Generate Scripts...
- Extract Data-tier Application...
- Deploy Database to Windows Azure SQL Database...
- Deploy Database to a Windows Azure VM...
- Export Data-tier Application...
- Register a Data-tier Application...
- Upgrade Data-tier Application...
- Delete Data-tier Application...
- Import Data...
- Export Data...
- Copy Database(s)...
- Manage Database Encryption...

The 'Generate Scripts...' option is highlighted. The table 'tblProcessCreate' has 10 rows displayed in the results grid.

id	arid_tanularity	arid_taniness	arid_id	arid_latitude	arid_location	arid_longitude	arid_mbld	arid_mbago
0639902515496	0.46131033741	ARINULV12298800C91	0	NULL	Eswatini-E91d4748cc-ba7d-5e8a5d320c54	0	NULL	NULL
067179330534	0.386806364066	ARINULV111878940578	37.77916	San Francisco, CA	-122.42005	0	0110e6b62384a8b9a336059e41a53	NULL
0581793765946	0.40190543384	ARINULV111878990FB1	0	California - LA	0	077e51a5-4f81-45b3-9347-2051811e366	NULL	NULL
0479340991297	0.307000168813	ARINULV111878990FE	23.62574	Mexico	-101.90625	0163c842-41b3-435d-9d62-9986e054344	NULL	NULL
0630030037599	0.417499644971	ARINULV111878994F3	35.14968	Memphis, TN	-90.04692	1c78a6c5-d334-4339-a0b7-7e8dd71849c2	classic pop and ro	classic pop and ro
00131579730599	0.379382374332	ARINULV1118789950B8	0	New Jersey	0	c51294b7-8954-441b-a9d5-e3c8d770881	NULL	NULL
0531334211069	0.33839951594	ARINULV111878990479	64.56563	Charleston, SC	12.66538	c795e5a7-7327-4375-b394-69614162071	NULL	NULL
0623827825331	0.387161655902	ARINULV1118789942BD	37.77916	San Francisco, CA	-122.42005	2d533be-25d4-4409-89e6-699e5038e01	NULL	NULL
04217993845	0.338655660667	ARINULV1118789900F	51.50632	London	-0.12714	c1632a11-027c-4a44-a465-ba5d8890055	uk.brightlight.p	uk.brightlight.p
0487396790338	0.343426978297	ARINULV1118789904D4	0	NULL	0	7e273984-e5d9-4451-9c4d-59b389856cd	NULL	NULL

Resources

Migrate a Database to SQL Server on an Azure VM

SQL Server on Azure Virtual Machines overview

Move data to an Azure SQL Database for Azure Machine Learning

3/12/2019 • 3 minutes to read

This article outlines the options for moving data either from flat files (CSV or TSV formats) or from data stored in an on-premises SQL Server to an Azure SQL database. These tasks for moving data to the cloud are part of the Team Data Science Process.

For a topic that outlines the options for moving data to an on-premises SQL Server for Machine Learning, see [Move data to SQL Server on an Azure virtual machine](#).

The following table summarizes the options for moving data to an Azure SQL Database.

SOURCE	DESTINATION: AZURE SQL DATABASE
Flat file (CSV or TSV formatted)	Bulk Insert SQL Query
On-premises SQL Server	1. Export to Flat File 2. SQL Database Migration Wizard 3. Database back up and restore 4. Azure Data Factory

Prerequisites

The procedures outlined here require that you have:

- An **Azure subscription**. If you do not have a subscription, you can sign up for a [free trial](#).
- An **Azure storage account**. You use an Azure storage account for storing the data in this tutorial. If you don't have an Azure storage account, see the [Create a storage account](#) article. After you have created the storage account, you need to obtain the account key used to access the storage. See [Manage your storage access keys](#).
- Access to an **Azure SQL Database**. If you must set up an Azure SQL Database, [Getting Started with Microsoft Azure SQL Database](#) provides information on how to provision a new instance of an Azure SQL Database.
- Installed and configured **Azure PowerShell** locally. For instructions, see [How to install and configure Azure PowerShell](#).

Data: The migration processes are demonstrated using the [NYC Taxi dataset](#). The NYC Taxi dataset contains information on trip data and fares and is available on Azure blob storage: [NYC Taxi Data](#). A sample and description of these files are provided in [NYC Taxi Trips Dataset Description](#).

You can either adapt the procedures described here to a set of your own data or follow the steps as described by using the NYC Taxi dataset. To upload the NYC Taxi dataset into your on-premises SQL Server database, follow the procedure outlined in [Bulk Import Data into SQL Server Database](#). These instructions are for a SQL Server on an Azure Virtual Machine, but the procedure for uploading to the on-premises SQL Server is the same.

Moving data from a flat file source to an Azure SQL database

Data in flat files (CSV or TSV formatted) can be moved to an Azure SQL database using a Bulk Insert SQL Query.

Bulk Insert SQL Query

The steps for the procedure using the Bulk Insert SQL Query are similar to those covered in the sections for

moving data from a flat file source to SQL Server on an Azure VM. For details, see [Bulk Insert SQL Query](#).

Moving Data from on-premises SQL Server to an Azure SQL database

If the source data is stored in an on-premises SQL Server, there are various possibilities for moving the data to an Azure SQL database:

1. [Export to Flat File](#)
2. [SQL Database Migration Wizard](#)
3. [Database back up and restore](#)
4. [Azure Data Factory](#)

The steps for the first three are very similar to those sections in [Move data to SQL Server on an Azure virtual machine](#) that cover these same procedures. Links to the appropriate sections in that topic are provided in the following instructions.

Export to Flat File

The steps for this exporting to a flat file are similar to those covered in [Export to Flat File](#).

SQL Database Migration Wizard

The steps for using the SQL Database Migration Wizard are similar to those covered in [SQL Database Migration Wizard](#).

Database back up and restore

The steps for using database back up and restore are similar to those covered in [Database back up and restore](#).

Azure Data Factory

The procedure for moving data to an Azure SQL database with Azure Data Factory (ADF) is provided in the topic [Move data from an on-premises SQL server to SQL Azure with Azure Data Factory](#). This topic shows how to move data from an on-premises SQL Server database to an Azure SQL database via Azure Blob Storage using ADF.

Consider using ADF when data needs to be continually migrated in a hybrid scenario that accesses both on-premises and cloud resources, and when the data is transacted or needs to be modified or have business logic added to it when being migrated. ADF allows for the scheduling and monitoring of jobs using simple JSON scripts that manage the movement of data on a periodic basis. ADF also has other capabilities such as support for complex operations.

Create Hive tables and load data from Azure Blob Storage

3/12/2019 • 10 minutes to read

This article presents generic Hive queries that create Hive tables and load data from Azure blob storage. Some guidance is also provided on partitioning Hive tables and on using the Optimized Row Columnar (ORC) formatting to improve query performance.

Prerequisites

This article assumes that you have:

- Created an Azure storage account. If you need instructions, see [About Azure storage accounts](#).
- Provisioned a customized Hadoop cluster with the HDInsight service. If you need instructions, see [Setup Clusters in HDInsight](#).
- Enabled remote access to the cluster, logged in, and opened the Hadoop Command-Line console. If you need instructions, see [Manage Apache Hadoop clusters](#).

Upload data to Azure blob storage

If you created an Azure virtual machine by following the instructions provided in [Set up an Azure virtual machine for advanced analytics](#), this script file should have been downloaded to the `C:\Users\<user name>\Documents\Data Science Scripts` directory on the virtual machine. These Hive queries only require that you plug in your own data schema and Azure blob storage configuration in the appropriate fields to be ready for submission.

We assume that the data for Hive tables is in an **uncompressed** tabular format, and that the data has been uploaded to the default (or to an additional) container of the storage account used by the Hadoop cluster.

If you want to practice on the **NYC Taxi Trip Data**, you need to:

- **download** the 24 [NYC Taxi Trip Data](#) files (12 Trip files and 12 Fare files),
- **unzip** all files into .csv files, and then
- **upload** them to the default (or appropriate container) of the Azure storage account; options for such an account appear at [Use Azure storage with Azure HDInsight clusters](#) topic. The process to upload the .csv files to the default container on the storage account can be found on this [page](#).

How to submit Hive queries

Hive queries can be submitted by using:

1. [Submit Hive queries through Hadoop Command Line in headnode of Hadoop cluster](#)
2. [Submit Hive queries with the Hive Editor](#)
3. [Submit Hive queries with Azure PowerShell Commands](#)

Hive queries are SQL-like. If you are familiar with SQL, you may find the [Hive for SQL Users Cheat Sheet](#) useful.

When submitting a Hive query, you can also control the destination of the output from Hive queries, whether it be on the screen or to a local file on the head node or to an Azure blob.

1. Submit Hive queries through Hadoop Command Line in headnode of Hadoop cluster

If the Hive query is complex, submitting it directly in the head node of the Hadoop cluster typically leads to faster turn around than submitting it with a Hive Editor or Azure PowerShell scripts.

Log in to the head node of the Hadoop cluster, open the Hadoop Command Line on the desktop of the head node, and enter command `cd %hive_home%\bin`.

You have three ways to submit Hive queries in the Hadoop Command Line:

- directly
- using .hql files
- with the Hive command console

Submit Hive queries directly in Hadoop Command Line.

You can run command like `hive -e "<your hive query>"` to submit simple Hive queries directly in Hadoop Command Line. Here is an example, where the red box outlines the command that submits the Hive query, and the green box outlines the output from the Hive query.

The screenshot shows a Windows command prompt window titled "Hadoop Command Line". The command entered is `hive -e "show tables;"`. The output shows the results of the query, which is empty in this case. The entire command line is highlighted with a red box, and the output is highlighted with a green box.

```
C:\apps\dist\hadoop-2.4.0.2.1.10.0-2290>hive -e "show tables;"  
'hive' is not recognized as an internal or external command,  
operable program or batch file.  
C:\apps\dist\hadoop-2.4.0.2.1.10.0-2290>cd %hive_home%\bin  
C:\apps\dist\hive-0.13.0.2.1.10.0-2290\bin>hive -e "show tables;"  
  
Logging initialized using configuration in file:/C:/apps/dist/hive-0.13.0.2.1.10.  
.0-2290/conf/hive-log4j.properties  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hadoop-2.4.0.2.1.10.0-2290/share/  
hadoop/common/lib/slf4j-log4j12-1.7.5.jar!org/slf4j/impl/StaticLoggerBinder.cl  
ass]  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hbase-0.98.0.2.1.10.0-2290-hadoo  
p2/lib/slf4j-log4j12-1.7.5.jar!org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
OK  
hivesamplitable  
Time taken: 1.582 seconds. Fetched: 1 row(s)  
C:\apps\dist\hive-0.13.0.2.1.10.0-2290\bin>
```

Submit Hive queries in .hql files

When the Hive query is more complicated and has multiple lines, editing queries in command line or Hive command console is not practical. An alternative is to use a text editor in the head node of the Hadoop cluster to save the Hive queries in a .hql file in a local directory of the head node. Then the Hive query in the .hql file can be submitted by using the `-f` argument as follows:



The screenshot shows a Windows command prompt window titled "Hadoop Command Line". The command entered is `hive -f /C:/apps/temp/hivequeryinfile.hql`. The output shows the results of the query, which is a list of mobile device models and their characteristics. The entire command line is highlighted with a red box, and the output is highlighted with a green box.

```
hive -f /C:/apps/temp/hivequeryinfile.hql  
Logging initialized using configuration in file:/C:/apps/dist/hive-0.13.0.2.1.10.  
.0-2290/conf/hive-log4j.properties  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hadoop-2.4.0.2.1.10.0-2290/share/  
hadoop/common/lib/slf4j-log4j12-1.7.5.jar!org/slf4j/impl/StaticLoggerBinder.cl  
ass]  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hbase-0.98.0.2.1.10.0-2290-hadoo  
p2/lib/slf4j-log4j12-1.7.5.jar!org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
OK  
select  
from hivesamplitable  
limit 50;  
  
+-----+-----+-----+-----+-----+-----+  
| _id | model | manufacturer | os | screen_size | screen_resolution |  
+-----+-----+-----+-----+-----+-----+  
| 1878428 | as-US | Android Samsung SGS-iSCM | California | 4.0 | 1024x600 |  
United States | 13.5208007 |  
| 20 | as-US | Android HTC Incredible | California | 4.0 | 1024x600 |  
United States | NILL |  
| 20 | as-US | Android HTC Incredible | California | 4.0 | 1024x600 |  
United States | 1.4757942 |  
| 20 | as-US | Android HTC Incredible | California | 4.0 | 1024x600 |  
United States | 0.265968 |  
| 20 | as-US | Android Motorola Droid R | Colorado | 4.0 | 1024x600 |  
United States | 28.3095378 |  
| 20 | as-US | Android Motorola Droid X | Colorado | 4.0 | 1024x600 |  
United States | 16.2981648 |  
| 20 | as-US | Android Motorola Droid X | Colorado | 4.0 | 1024x600 |  
United States | 1.7715228 |  
| 20 | as-US | Android Motorola Droid X Utah | United States | 4.0 | 1024x600 |  
United States | 16.046721 |  
| 20 | as-US | Android Motorola Droid X Utah | United States | 4.0 | 1024x600 |  
United States | 18.453161 |  
| 20 | as-US | Android Motorola Droid X Utah | United States | 4.0 | 1024x600 |  
United States | 0.177119 |  
| 20 | as-US | Android Motorola Droid X Colorado | Colorado | 4.0 | 1024x600 |  
United States | 28.3095378 |  
Time taken: 2.995 seconds. Fetched: 10 rows  
C:\apps\dist\hive-0.13.0.2.1.10.0-2290\bin>
```

SUPPRESS PROGRESS STATUS SCREEN PRINT OF HIVE QUERIES

By default, after Hive query is submitted in Hadoop Command Line, the progress of the Map/Reduce job is printed out on screen. To suppress the screen print of the Map/Reduce job progress, you can use an argument `-s`

("S" in upper case) in the command line as follows:

```
hive -S -f "<path to the .hql file>"  
hive -S -e "<Hive queries>"
```

Submit Hive queries in Hive command console.

You can also first enter the Hive command console by running command `hive` in Hadoop Command Line, and then submit Hive queries in Hive command console. Here is an example. In this example, the two red boxes highlight the commands used to enter the Hive command console, and the Hive query submitted in Hive command console, respectively. The green box highlights the output from the Hive query.

The screenshot shows a Windows command prompt window titled "Hadoop Command Line - hive". The command `hive` is run, followed by a query: `select * from hivesampletable limit 3;`. The output shows three rows of data from the sample table, each with columns: timestamp, location, device, provider, signal, and state. The entire output is highlighted with a green box.

```
hive> select * from hivesampletable limit 3;  
8 18:54:20 en-US Android Samsung SCH-i500 California  
United States 13.9204007 0 0  
23 19:19:44 en-US Android HTC Incredible Pennsylvania  
United States NULL 0 0  
23 19:19:46 en-US Android HTC Incredible Pennsylvania  
United States 1.4757422 0 1  
Time taken: 2.897 seconds. Fetched: 3 row(s)
```

The previous examples directly output the Hive query results on screen. You can also write the output to a local file on the head node, or to an Azure blob. Then, you can use other tools to further analyze the output of Hive queries.

Output Hive query results to a local file. To output Hive query results to a local directory on the head node, you have to submit the Hive query in the Hadoop Command Line as follows:

```
hive -e "<hive query>" > <local path in the head node>
```

In the following example, the output of Hive query is written into a file `hivequeryoutput.txt` in directory `C:\apps\temp`.

The screenshot shows the Hadoop Command Line window and a File Explorer window. The command `hive -e "select * from hivesampletable limit 10;" > output.txt` is run, and the resulting file `hivequeryoutput.txt` is shown in the File Explorer under the path `C:\apps\temp`. The file is a text document with the same data as the previous screenshot.

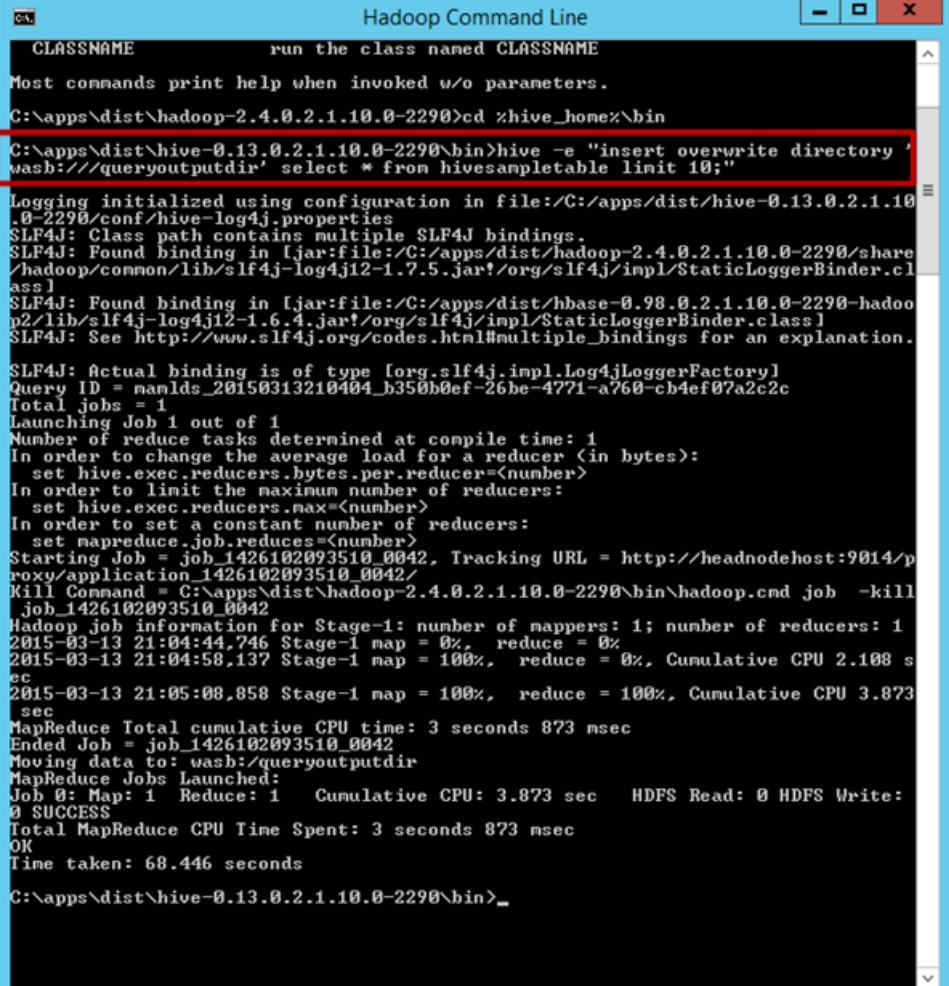
```
hive> select * from hivesampletable limit 10;  
8 18:54:20 en-US Android Samsung SCH-i500 California  
United States 13.9204007 0 0  
23 19:19:44 en-US Android HTC Incredible Pennsylvania  
United States NULL 0 0  
23 19:19:46 en-US Android HTC Incredible Pennsylvania  
United States 1.4757422 0 1  
Time taken: 2.897 seconds. Fetched: 3 row(s)  
hive> quit  
C:\apps\dist>hive -e "select * from hivesampletable limit 10;" > C:\apps\temp\hivequeryoutput.txt  
Logging initialized using configuration in file:/C:/apps/dist/hive-0.13.0.2.1.10-2290/conf/hive-log4j.properties  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hadoop-2.4.0.2.1.10-2290/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hbase-0.98.0.2.1.10-2290-hadoop2/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type org.slf4j.impl.Log4jLoggerFactory  
hive> select * from hivesampletable limit 3;  
8 18:54:20 en-US Android Samsung SCH-i500 California  
United States 13.9204007 0 0  
23 19:19:44 en-US Android HTC Incredible Pennsylvania  
United States NULL 0 0  
23 19:19:46 en-US Android HTC Incredible Pennsylvania  
United States 1.4757422 0 1  
Time taken: 2.897 seconds. Fetched: 3 row(s)  
hive> quit  
C:\apps\dist>hive -e "select * from hivesampletable limit 10;" > C:\apps\temp\hivequeryoutput.txt  
Logging initialized using configuration in file:/C:/apps/dist/hive-0.13.0.2.1.10-2290/conf/hive-log4j.properties  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hadoop-2.4.0.2.1.10-2290/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hbase-0.98.0.2.1.10-2290-hadoop2/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type org.slf4j.impl.Log4jLoggerFactory  
OK  
Time taken: 2.138 seconds. Fetched: 10 row(s)  
C:\apps\dist>hive -e "select * from hivesampletable limit 10;" > C:\apps\temp\hivequeryoutput.txt  
Logging initialized using configuration in file:/C:/apps/dist/hive-0.13.0.2.1.10-2290/conf/hive-log4j.properties  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hadoop-2.4.0.2.1.10-2290/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/C:/apps/dist/hbase-0.98.0.2.1.10-2290-hadoop2/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type org.slf4j.impl.Log4jLoggerFactory  
OK  
Time taken: 25.04 seconds. Fetched: 10 row(s)  
C:\apps\dist>hive -e "select * from hivesampletable limit 10;" > C:\apps\temp\hivequeryoutput.txt
```

Output Hive query results to an Azure blob

You can also output the Hive query results to an Azure blob, within the default container of the Hadoop cluster. The Hive query for this is as follows:

```
insert overwrite directory wasb:///<directory within the default container> <select clause from ...>
```

In the following example, the output of Hive query is written to a blob directory `queryoutputdir` within the default container of the Hadoop cluster. Here, you only need to provide the directory name, without the blob name. An error is thrown if you provide both directory and blob names, such as `wasb:///queryoutputdir/queryoutput.txt`.



The screenshot shows a Windows command-line interface window titled "Hadoop Command Line". The command entered is:

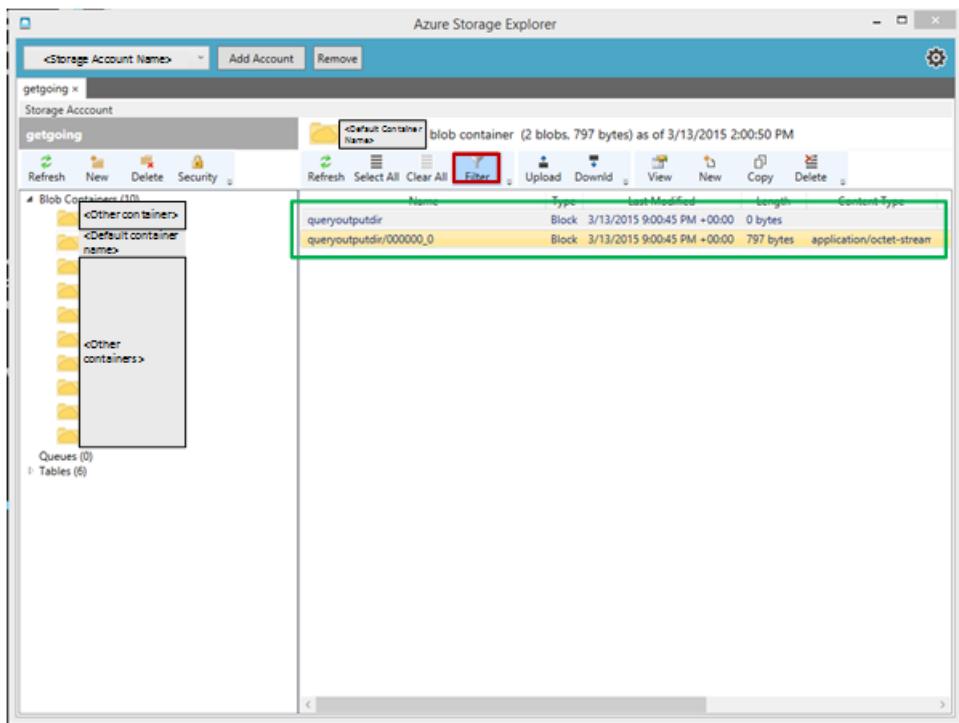
```
C:\apps\dist\hive-0.13.0.2.1.10.0-2290\bin>hive -e "insert overwrite directory 'wasb:///queryoutputdir' select * from hivesamptable limit 10;"
```

The output of the command is displayed below the command line. It shows the configuration of the SLF4J logger, the execution of the Hive job, and the completion of the job with a success message.

```
Logging initialized using configuration in file:/C:/apps/dist/hive-0.13.0.2.1.10.0-2290/conf/hive-log4j.properties
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/apps/dist/hadoop-2.4.0.2.1.10.0-2290/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/jimpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/apps/dist/hbase-0.98.0.2.1.10.0-2290-hadoop2/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/jimpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Query ID = mamlds_20150313210404_b350b0ef-26be-4771-a760-ch4ef07a2c2c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer <in bytes>
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1426102093510_0042, Tracking URL = http://headnodehost:9014/proxy/application_1426102093510_0042/
Kill Command = C:\apps\dist\hadoop-2.4.0.2.1.10.0-2290\bin\hadoop.cmd job -kill job_1426102093510_0042
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-03-13 21:04:44,746 Stage-1 map = 0%, reduce = 0%
2015-03-13 21:04:58,137 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.108 sec
2015-03-13 21:05:08,858 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.873 sec
MapReduce Total cumulative CPU time: 3 seconds 873 msec
Ended Job = job_1426102093510_0042
Moving data to: wasb:/queryoutputdir
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 1   Cumulative CPU: 3.873 sec   HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 873 msec
OK
Time taken: 68.446 seconds
C:\apps\dist\hive-0.13.0.2.1.10.0-2290\bin>_
```

If you open the default container of the Hadoop cluster using Azure Storage Explorer, you can see the output of the Hive query as shown in the following figure. You can apply the filter (highlighted by red box) to only retrieve the blob with specified letters in names.



2. Submit Hive queries with the Hive Editor

You can also use the Query Console (Hive Editor) by entering a URL of the form <https://azurehdinsight.net/Home/HiveEditor> into a web browser. You must be logged in to see this console and so you need your Hadoop cluster credentials here.

3. Submit Hive queries with Azure PowerShell Commands

You can also use PowerShell to submit Hive queries. For instructions, see [Submit Hive jobs using PowerShell](#).

Create Hive database and tables

The Hive queries are shared in the [GitHub repository](#) and can be downloaded from there.

Here is the Hive query that creates a Hive table.

```
create database if not exists <database name>;
CREATE EXTERNAL TABLE if not exists <database name>.<table name>
(
    field1 string,
    field2 int,
    field3 float,
    field4 double,
    ...,
    fieldN string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '<field separator>' lines terminated by '<line separator>'
STORED AS TEXTFILE LOCATION '<storage location>' TBLPROPERTIES("skip.header.line.count"="1");
```

Here are the descriptions of the fields that you need to plug in and other configurations:

- **<database name>**: the name of the database that you want to create. If you just want to use the default database, the query *create database...* can be omitted.
- **<table name>**: the name of the table that you want to create within the specified database. If you want to use the default database, the table can be directly referred by *<table name>* without *<database name>*.
- **<field separator>**: the separator that delimits fields in the data file to be uploaded to the Hive table.
- **<line separator>**: the separator that delimits lines in the data file.
- **<storage location>**: the Azure storage location to save the data of Hive tables. If you do not specify

LOCATION <storage location>, the database and the tables are stored in *hive/warehouse*/ directory in the default container of the Hive cluster by default. If you want to specify the storage location, the storage location has to be within the default container for the database and tables. This location has to be referred as location relative to the default container of the cluster in the format of '*wasb:///<directory 1>/*' or '*wasb:///<directory 1>/<directory 2>/*', etc. After the query is executed, the relative directories are created within the default container.

- **TBLPROPERTIES("skip.header.line.count" = "1")**: If the data file has a header line, you have to add this property **at the end** of the *create table* query. Otherwise, the header line is loaded as a record to the table. If the data file does not have a header line, this configuration can be omitted in the query.

Load data to Hive tables

Here is the Hive query that loads data into a Hive table.

```
LOAD DATA INPATH '<path to blob data>' INTO TABLE <database name>.<table name>;
```

- **<path to blob data>**: If the blob file to be uploaded to the Hive table is in the default container of the HDInsight Hadoop cluster, the *<path to blob data>* should be in the format '*wasb:///*'. The blob file can also be in an additional container of the HDInsight Hadoop cluster. In this case, *<path to blob data>* should be in the format '*wasb://blob.core.windows.net/*'.

NOTE

The blob data to be uploaded to Hive table has to be in the default or additional container of the storage account for the Hadoop cluster. Otherwise, the *LOAD DATA* query fails complaining that it cannot access the data.

Advanced topics: partitioned table and store Hive data in ORC format

If the data is large, partitioning the table is beneficial for queries that only need to scan a few partitions of the table. For instance, it is reasonable to partition the log data of a web site by dates.

In addition to partitioning Hive tables, it is also beneficial to store the Hive data in the Optimized Row Columnar (ORC) format. For more information on ORC formatting, see [Using ORC files improves performance when Hive is reading, writing, and processing data](#).

Partitioned table

Here is the Hive query that creates a partitioned table and loads data into it.

```
CREATE EXTERNAL TABLE IF NOT EXISTS <database name>.<table name>
  (field1 string,
  ...
  fieldN string
)
PARTITIONED BY (<partitionfieldname> vartype) ROW FORMAT DELIMITED FIELDS TERMINATED BY '<field separator>'
  lines terminated by '<line separator>' TBLPROPERTIES("skip.header.line.count"="1");
LOAD DATA INPATH '<path to the source file>' INTO TABLE <database name>.<partitioned table name>
  PARTITION (<partitionfieldname>=<partitionfieldvalue>);
```

When querying partitioned tables, it is recommended to add the partition condition in the **beginning** of the **where** clause as this improves the efficacy of searching significantly.

```
select
  field1, field2, ..., fieldN
from <database name>.<partitioned table name>
where <partitionfieldname>=<partitionfieldvalue> and ...;
```

Store Hive data in ORC format

You cannot directly load data from blob storage into Hive tables that is stored in the ORC format. Here are the steps that you need to take to load data from Azure blobs to Hive tables stored in ORC format.

Create an external table **STORED AS TEXTFILE** and load data from blob storage to the table.

```
CREATE EXTERNAL TABLE IF NOT EXISTS <database name>.<external textfile table name>
(
  field1 string,
  field2 int,
  ...
  fieldN date
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '<field separator>'
lines terminated by '<line separator>' STORED AS TEXTFILE
LOCATION 'wasb://<directory in Azure blob>' TBLPROPERTIES("skip.header.line.count"="1");

LOAD DATA INPATH '<path to the source file>' INTO TABLE <database name>.<table name>;
```

Create an internal table with the same schema as the external table in step 1, with the same field delimiter, and store the Hive data in the ORC format.

```
CREATE TABLE IF NOT EXISTS <database name>.<ORC table name>
(
  field1 string,
  field2 int,
  ...
  fieldN date
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '<field separator>' STORED AS ORC;
```

Select data from the external table in step 1 and insert into the ORC table

```
INSERT OVERWRITE TABLE <database name>.<ORC table name>
  SELECT * FROM <database name>.<external textfile table name>;
```

NOTE

If the TEXTFILE table <database name>.<external textfile table name> has partitions, in STEP 3, the

`SELECT * FROM <database name>.<external textfile table name>` command selects the partition variable as a field in the returned data set. Inserting it into the <database name>.<ORC table name> fails since <database name>.<ORC table name> does not have the partition variable as a field in the table schema. In this case, you need to specifically select the fields to be inserted to <database name>.<ORC table name> as follows:

```
INSERT OVERWRITE TABLE <database name>.<ORC table name> PARTITION (<partition variable>=<partition value>)
  SELECT field1, field2, ..., fieldN
  FROM <database name>.<external textfile table name>
  WHERE <partition variable>=<partition value>;
```

It is safe to drop the <external textfile table name> when using the following query after all data has been

inserted into <database name>.<ORC table name>:

```
DROP TABLE IF EXISTS <database name>.<external textfile table name>;
```

After following this procedure, you should have a table with data in the ORC format ready to use.

Build and optimize tables for fast parallel import of data into a SQL Server on an Azure VM

3/12/2019 • 5 minutes to read

This article describes how to build partitioned tables for fast parallel bulk importing of data to a SQL Server database. For big data loading/transfer to a SQL database, importing data to the SQL DB and subsequent queries can be improved by using *Partitioned Tables and Views*.

Create a new database and a set of filegroups

- [Create a new database](#), if it doesn't exist already.
- Add database filegroups to the database, which holds the partitioned physical files.
- This can be done with `CREATE DATABASE` if new or `ALTER DATABASE` if the database exists already.
- Add one or more files (as needed) to each database filegroup.

NOTE

Specify the target filegroup, which holds data for this partition and the physical database file name(s) where the filegroup data is stored.

The following example creates a new database with three filegroups other than the primary and log groups, containing one physical file in each. The database files are created in the default SQL Server Data folder, as configured in the SQL Server instance. For more information about the default file locations, see [File Locations for Default and Named Instances of SQL Server](#).

```
DECLARE @data_path nvarchar(256);
SET @data_path = (SELECT SUBSTRING(physical_name, 1, CHARINDEX(N'master.mdf', LOWER(physical_name)) - 1)
    FROM master.sys.master_files
    WHERE database_id = 1 AND file_id = 1);

EXECUTE (
    CREATE DATABASE <database_name>
        ON PRIMARY
            ( NAME = ''Primary'', FILENAME = ''' + @data_path + '<primary_file_name>.mdf'',
                SIZE = 4096KB , FILEGROWTH = 1024KB ),
        FILEGROUP [filegroup_1]
            ( NAME = ''FileGroup1'', FILENAME = ''' + @data_path + '<file_name_1>.ndf'',
                SIZE = 4096KB , FILEGROWTH = 1024KB ),
        FILEGROUP [filegroup_2]
            ( NAME = ''FileGroup2'', FILENAME = ''' + @data_path + '<file_name_2>.ndf'',
                SIZE = 4096KB , FILEGROWTH = 1024KB ),
        FILEGROUP [filegroup_3]
            ( NAME = ''FileGroup3'', FILENAME = ''' + @data_path + '<file_name_3>.ndf'',
                SIZE = 102400KB , FILEGROWTH = 10240KB )
        LOG ON
            ( NAME = ''LogFileGroup'', FILENAME = ''' + @data_path + '<log_file_name>.ldf'',
                SIZE = 1024KB , FILEGROWTH = 10% )
)
```

Create a partitioned table

To create partitioned table(s) according to the data schema, mapped to the database filegroups created in the

previous step, you must first create a partition function and scheme. When data is bulk imported to the partitioned table(s), records are distributed among the filegroups according to a partition scheme, as described below.

1. Create a partition function

[Create a partition function](#) This function defines the range of values/boundaries to be included in each individual partition table, for example, to limit partitions by month(some_datetime_field) in the year 2013:

```
CREATE PARTITION FUNCTION <DatetimeFieldPFN>(<datetime_field>)
AS RANGE RIGHT FOR VALUES (
    '20130201', '20130301', '20130401',
    '20130501', '20130601', '20130701', '20130801',
    '20130901', '20131001', '20131101', '20131201' )
```

2. Create a partition scheme

[Create a partition scheme](#). This scheme maps each partition range in the partition function to a physical filegroup, for example:

```
CREATE PARTITION SCHEME <DatetimeFieldPScheme> AS
PARTITION <DatetimeFieldPFN> TO (
<filegroup_1>, <filegroup_2>, <filegroup_3>, <filegroup_4>,
<filegroup_5>, <filegroup_6>, <filegroup_7>, <filegroup_8>,
<filegroup_9>, <filegroup_10>, <filegroup_11>, <filegroup_12> )
```

To verify the ranges in effect in each partition according to the function/scheme, run the following query:

```
SELECT psch.name as PartitionScheme,
       prng.value AS PartitionValue,
       prng.boundary_id AS BoundaryID
  FROM sys.partition_functions AS pfun
 INNER JOIN sys.partition_schemes psch ON pfun.function_id = psch.function_id
 INNER JOIN sys.partition_range_values prng ON prng.function_id=pfun.function_id
 WHERE pfun.name = <DatetimeFieldPFN>
```

3. Create a partition table

[Create partitioned table\(s\)](#) according to your data schema, and specify the partition scheme and constraint field used to partition the table, for example:

```
CREATE TABLE <table_name> ( [include schema definition here] )
ON <TablePScheme>(<partition_field>)
```

For more information, see [Create Partitioned Tables and Indexes](#).

Bulk import the data for each individual partition table

- You may use BCP, BULK INSERT, or other methods such as [SQL Server Migration Wizard](#). The example provided uses the BCP method.
- [Alter the database](#) to change transaction logging scheme to BULK_LOGGED to minimize overhead of logging, for example:

```
ALTER DATABASE <database_name> SET RECOVERY BULK_LOGGED
```

- To expedite data loading, launch the bulk import operations in parallel. For tips on expediting bulk importing of big data into SQL Server databases, see [Load 1TB in less than 1 hour](#).

The following PowerShell script is an example of parallel data loading using BCP.

```
# Set database name, input data directory, and output log directory
# This example loads comma-separated input data files
# The example assumes the partitioned data files are named as <base_file_name>_<partition_number>.csv
# Assumes the input data files include a header line. Loading starts at line number 2.

$dbname = "<database_name>"
$indir = "<path_to_data_files>"
$logdir = "<path_to_log_directory>

# Select authentication mode
$sqlauth = 0

# For SQL authentication, set the server and user credentials
$sqlusr = "<user@server>"
$server = "<tcp:serverdns>"
$pass = "<password>

# Set number of partitions per table - Should match the number of input data files per table
$numofparts = <number_of_partitions>

# Set table name to be loaded, basename of input data files, input format file, and number of partitions
$tbname = "<table_name>"
$basename = "<base_input_data_filename_no_extension>"
$fmtfile = "<full_path_to_format_file>

# Create log directory if it does not exist
New-Item -ErrorAction Ignore -ItemType directory -Path $logdir

# BCP example using Windows authentication
$ScriptBlock1 = {
    param($dbname, $tbname, $basename, $fmtfile, $indir, $logdir, $num)
    bcp ($dbname + ".." + $tbname) in ($indir + "\" + $basename + "_" + $num + ".csv") -o ($logdir + "\" + $tbname + "_" + $num + ".txt") -h "TABLOCK" -F 2 -C "RAW" -f ($fmtfile) -T -b 2500 -t "," -r \n
}

# BCP example using SQL authentication
$ScriptBlock2 = {
    param($dbname, $tbname, $basename, $fmtfile, $indir, $logdir, $num, $sqlusr, $server, $pass)
    bcp ($dbname + ".." + $tbname) in ($indir + "\" + $basename + "_" + $num + ".csv") -o ($logdir + "\" + $tbname + "_" + $num + ".txt") -h "TABLOCK" -F 2 -C "RAW" -f ($fmtfile) -U $sqlusr -S $server -P $pass -b 2500 -t "," -r \n
}

# Background processing of all partitions
for ($i=1; $i -le $numofparts; $i++)
{
    Write-Output "Submit loading trip and fare partitions # $i"
    if ($sqlauth -eq 0) {
        # Use Windows authentication
        Start-Job -ScriptBlock $ScriptBlock1 -Arg ($dbname, $tbname, $basename, $fmtfile, $indir, $logdir, $i)
    }
    else {
        # Use SQL authentication
        Start-Job -ScriptBlock $ScriptBlock2 -Arg ($dbname, $tbname, $basename, $fmtfile, $indir, $logdir, $i, $sqlusr, $server, $pass)
    }
}

Get-Job

# Optional - Wait till all jobs complete and report date and time
date
While (Get-Job -State "Running") { Start-Sleep 10 }
date
```

Create indexes to optimize joins and query performance

- If you extract data for modeling from multiple tables, create indexes on the join keys to improve the join performance.
- [Create indexes](#) (clustered or non-clustered) targeting the same filegroup for each partition, for example:

```
CREATE CLUSTERED INDEX <table_idx> ON <table_name>( [include index columns here] )
ON <TablePScheme>(<partition>field)
```

or,

```
CREATE INDEX <table_idx> ON <table_name>( [include index columns here] )
ON <TablePScheme>(<partition>field)
```

NOTE

You may choose to create the indexes before bulk importing the data. Index creation before bulk importing slows down the data loading.

Advanced Analytics Process and Technology in Action Example

For an end-to-end walkthrough example using the Team Data Science Process with a public dataset, see [Team Data Science Process in Action: using SQL Server](#).

Move data from an on-premises SQL server to SQL Azure with Azure Data Factory

3/12/2019 • 8 minutes to read

This article shows how to move data from an on-premises SQL Server Database to a SQL Azure Database via Azure Blob Storage using the Azure Data Factory (ADF).

For a table that summarizes various options for moving data to an Azure SQL Database, see [Move data to an Azure SQL Database for Azure Machine Learning](#).

Introduction: What is ADF and when should it be used to migrate data?

Azure Data Factory is a fully managed cloud-based data integration service that orchestrates and automates the movement and transformation of data. The key concept in the ADF model is pipeline. A pipeline is a logical grouping of Activities, each of which defines the actions to perform on the data contained in Datasets. Linked services are used to define the information needed for Data Factory to connect to the data resources.

With ADF, existing data processing services can be composed into data pipelines that are highly available and managed in the cloud. These data pipelines can be scheduled to ingest, prepare, transform, analyze, and publish data, and ADF manages and orchestrates the complex data and processing dependencies. Solutions can be quickly built and deployed in the cloud, connecting a growing number of on-premises and cloud data sources.

Consider using ADF:

- when data needs to be continually migrated in a hybrid scenario that accesses both on-premises and cloud resources
- when the data is transacted or needs to be modified or have business logic added to it when being migrated.

ADF allows for the scheduling and monitoring of jobs using simple JSON scripts that manage the movement of data on a periodic basis. ADF also has other capabilities such as support for complex operations. For more information on ADF, see the documentation at [Azure Data Factory \(ADF\)](#).

The Scenario

We set up an ADF pipeline that composes two data migration activities. Together they move data on a daily basis between an on-premises SQL database and an Azure SQL Database in the cloud. The two activities are:

- copy data from an on-premises SQL Server database to an Azure Blob Storage account
- copy data from the Azure Blob Storage account to an Azure SQL Database.

NOTE

The steps shown here have been adapted from the more detailed tutorial provided by the ADF team: [Copy data from an on-premises SQL Server database to Azure Blob storage](#). References to the relevant sections of that topic are provided when appropriate.

Prerequisites

This tutorial assumes you have:

- An **Azure subscription**. If you do not have a subscription, you can sign up for a [free trial](#).
- An **Azure storage account**. You use an Azure storage account for storing the data in this tutorial. If you don't have an Azure storage account, see the [Create a storage account](#) article. After you have created the storage account, you need to obtain the account key used to access the storage. See [Manage your storage access keys](#).
- Access to an **Azure SQL Database**. If you must set up an Azure SQL Database, the topic [Getting Started with Microsoft Azure SQL Database](#) provides information on how to provision a new instance of an Azure SQL Database.
- Installed and configured **Azure PowerShell** locally. For instructions, see [How to install and configure Azure PowerShell](#).

NOTE

This procedure uses the [Azure portal](#).

Upload the data to your on-premises SQL Server

We use the [NYC Taxi dataset](#) to demonstrate the migration process. The NYC Taxi dataset is available, as noted in that post, on Azure blob storage [NYC Taxi Data](#). The data has two files, the trip_data.csv file, which contains trip details, and the trip_fare.csv file, which contains details of the fare paid for each trip. A sample and description of these files are provided in [NYC Taxi Trips Dataset Description](#).

You can either adapt the procedure provided here to a set of your own data or follow the steps as described by using the NYC Taxi dataset. To upload the NYC Taxi dataset into your on-premises SQL Server database, follow the procedure outlined in [Bulk Import Data into SQL Server Database](#). These instructions are for a SQL Server on an Azure Virtual Machine, but the procedure for uploading to the on-premises SQL Server is the same.

Create an Azure Data Factory

The instructions for creating a new Azure Data Factory and a resource group in the [Azure portal](#) are provided [Create an Azure Data Factory](#). Name the new ADF instance *adfdsp* and name the resource group created *adfdsp*.

Install and configure Azure Data Factory Integration Runtime

The Integration Runtime is a customer managed data integration infrastructure used by Azure Data Factory to provide data integration capabilities across different network environments. This runtime was formerly called "Data Management Gateway".

To set up, [follow the instructions for creating a pipeline](#)

Create linked services to connect to the data resources

A linked service defines the information needed for Azure Data Factory to connect to a data resource. We have three resources in this scenario for which linked services are needed:

1. On-premises SQL Server
2. Azure Blob Storage
3. Azure SQL database

The step-by-step procedure for creating linked services is provided in [Create linked services](#).

Define and create tables to specify how to access the datasets

Create tables that specify the structure, location, and availability of the datasets with the following script-based procedures. JSON files are used to define the tables. For more information on the structure of these files, see [Datasets](#).

NOTE

You should execute the `Add-AzureAccount` cmdlet before executing the `New-AzureDataFactoryTable` cmdlet to confirm that the right Azure subscription is selected for the command execution. For documentation of this cmdlet, see [Add-AzureAccount](#).

The JSON-based definitions in the tables use the following names:

- the **table name** in the on-premises SQL server is *nyctaxi_data*
- the **container name** in the Azure Blob Storage account is *containername*

Three table definitions are needed for this ADF pipeline:

1. [SQL on-premises Table](#)
2. [Blob Table](#)
3. [SQL Azure Table](#)

NOTE

These procedures use Azure PowerShell to define and create the ADF activities. But these tasks can also be accomplished using the Azure portal. For details, see [Create datasets](#).

SQL on-premises Table

The table definition for the on-premises SQL Server is specified in the following JSON file:

```
{  
    "name": "OnPremSQLTable",  
    "properties":  
    {  
        "location":  
        {  
            "type": "OnPremisesSqlServerTableLocation",  
            "tableName": "nyctaxi_data",  
            "linkedServiceName": "adfออนพรีเมี่ยส"  
        },  
        "availability":  
        {  
            "frequency": "Day",  
            "interval": 1,  
            "waitOnExternal":  
            {  
                "retryInterval": "00:01:00",  
                "retryTimeout": "00:10:00",  
                "maximumRetry": 3  
            }  
        }  
    }  
}
```

The column names were not included here. You can sub-select on the column names by including them here (for details check the [ADF documentation](#) topic).

Copy the JSON definition of the table into a file called *onpremtabledef.json* file and save it to a known location (here assumed to be C:\temp\onpremtabledef.json). Create the table in ADF with the following Azure PowerShell

cmdlet:

```
New-AzureDataFactoryTable -ResourceGroupName ADFdsprg -DataFactoryName ADFdsp -File  
C:\temp\onpremtabledef.json
```

Blob Table

Definition for the table for the output blob location is in the following (this maps the ingested data from on-premises to Azure blob):

```
{
  "name": "OutputBlobTable",
  "properties":
  {
    "location":
    {
      "type": "AzureBlobLocation",
      "folderPath": "containername",
      "format":
      {
        "type": "TextFormat",
        "columnDelimiter": "\t"
      },
      "linkedServiceName": "adfds"
    },
    "availability":
    {
      "frequency": "Day",
      "interval": 1
    }
  }
}
```

Copy the JSON definition of the table into a file called *bloboutputtabledef.json* file and save it to a known location (here assumed to be C:\temp\bloboutputtabledef.json). Create the table in ADF with the following Azure PowerShell cmdlet:

```
New-AzureDataFactoryTable -ResourceGroupName adfdsp -DataFactoryName adfdsp -File  
C:\temp\bloboutputtabledef.json
```

SQL Azure Table

Definition for the table for the SQL Azure output is in the following (this schema maps the data coming from the blob):

```
{
  "name": "OutputSQLAzureTable",
  "properties":
  {
    "structure":
    [
      { "name": "column1", "type": "String"},  

      { "name": "column2", "type": "String"}
    ],
    "location":
    {
      "type": "AzureSqlTableLocation",
      "tableName": "your_db_name",
      "linkedServiceName": "adfsqlazure_linked_servicename"
    },
    "availability":
    {
      "frequency": "Day",
      "interval": 1
    }
  }
}
```

Copy the JSON definition of the table into a file called *AzureSqlTable.json* file and save it to a known location (here assumed to be *C:\temp\AzureSqlTable.json*). Create the table in ADF with the following Azure PowerShell cmdlet:

```
New-AzureDataFactoryTable -ResourceGroupName adfdsp -DataFactoryName adfdsp -File C:\temp\AzureSqlTable.json
```

Define and create the pipeline

Specify the activities that belong to the pipeline and create the pipeline with the following script-based procedures. A JSON file is used to define the pipeline properties.

- The script assumes that the **pipeline name** is *AMLDSPProcessPipeline*.
- Also note that we set the periodicity of the pipeline to be executed on daily basis and use the default execution time for the job (12 am UTC).

NOTE

The following procedures use Azure PowerShell to define and create the ADF pipeline. But this task can also be accomplished using the Azure portal. For details, see [Create pipeline](#).

Using the table definitions provided previously, the pipeline definition for the ADF is specified as follows:

```
{
  "name": "AMLDSPProcessPipeline",
  "properties":
  {
    "description" : "This pipeline has one Copy activity that copies data from an on-premises SQL to Azure blob",
    "activities":
    [
      {
        "name": "CopyFromSQLtoBlob",
        "description": "Copy data from on-premises SQL server to blob",
        "type": "CopyActivity",
        "inputs": [ {"name": "OnPremSQLTable"} ],
        "outputs": [ {"name": "OutputBlobTable"} ],
        "transformation":
        {
          "source":
          {
            "type": "SqlSource",
            "sqlReaderQuery": "select * from nytaxi_data"
          },
          "sink":
          {
            "type": "BlobSink"
          }
        },
        "Policy":
        {
          "concurrency": 3,
          "executionPriorityOrder": "NewestFirst",
          "style": "StartOfInterval",
          "retry": 0,
          "timeout": "01:00:00"
        }
      },
      {
        "name": "CopyFromBlobtoSQLAzure",
        "description": "Push data to Sql Azure",
        "type": "CopyActivity",
        "inputs": [ {"name": "OutputBlobTable"} ],
        "outputs": [ {"name": "OutputSQLAzureTable"} ],
        "transformation":
        {
          "source":
          {
            "type": "BlobSource"
          },
          "sink":
          {
            "type": "SqlSink",
            "WriteBatchTimeout": "00:5:00",
          }
        },
        "Policy":
        {
          "concurrency": 3,
          "executionPriorityOrder": "NewestFirst",
          "style": "StartOfInterval",
          "retry": 2,
          "timeout": "02:00:00"
        }
      }
    ]
  }
}
```

Copy this JSON definition of the pipeline into a file called *pipelinedef.json* file and save it to a known location (here assumed to be C:\temp\pipelinedef.json). Create the pipeline in ADF with the following Azure PowerShell cmdlet:

```
New-AzureDataFactoryPipeline -ResourceGroupName adfdsp -DataFactoryName adfdsp -File  
C:\temp\pipelinedef.json
```

Start the Pipeline

The pipeline can now be run using the following command:

```
Set-AzureDataFactoryPipelineActivePeriod -ResourceGroupName ADFdsprg -DataFactoryName ADFdsp -StartTime  
startdateZ -EndDateTime enddateZ -Name AMLDSPProcessPipeline
```

The *startdate* and *enddate* parameter values need to be replaced with the actual dates between which you want the pipeline to run.

Once the pipeline executes, you should be able to see the data show up in the container selected for the blob, one file per day.

Note that we have not leveraged the functionality provided by ADF to pipe data incrementally. For more information on how to do this and other capabilities provided by ADF, see the [ADF documentation](#).

Tasks to prepare data for enhanced machine learning

1/30/2019 • 6 minutes to read

Pre-processing and cleaning data are important tasks that typically must be conducted before dataset can be used effectively for machine learning. Raw data is often noisy and unreliable, and may be missing values. Using such data for modeling can produce misleading results. These tasks are part of the Team Data Science Process (TDSP) and typically follow an initial exploration of a dataset used to discover and plan the pre-processing required. For more detailed instructions on the TDSP process, see the steps outlined in the [Team Data Science Process](#).

Pre-processing and cleaning tasks, like the data exploration task, can be carried out in a wide variety of environments, such as SQL or Hive or Azure Machine Learning Studio, and with various tools and languages, such as R or Python, depending where your data is stored and how it is formatted. Since TDSP is iterative in nature, these tasks can take place at various steps in the workflow of the process.

This article introduces various data processing concepts and tasks that can be undertaken either before or after ingesting data into Azure Machine Learning.

For an example of data exploration and pre-processing done inside Azure Machine Learning studio, see the [Pre-processing data in Azure Machine Learning Studio](#) video.

Why pre-process and clean data?

Real world data is gathered from various sources and processes and it may contain irregularities or corrupt data compromising the quality of the dataset. The typical data quality issues that arise are:

- **Incomplete:** Data lacks attributes or containing missing values.
- **Noisy:** Data contains erroneous records or outliers.
- **Inconsistent:** Data contains conflicting records or discrepancies.

Quality data is a prerequisite for quality predictive models. To avoid "garbage in, garbage out" and improve data quality and therefore model performance, it is imperative to conduct a data health screen to spot data issues early and decide on the corresponding data processing and cleaning steps.

What are some typical data health screens that are employed?

We can check the general quality of data by checking:

- The number of **records**.
- The number of **attributes** (or **features**).
- The attribute **data types** (nominal, ordinal, or continuous).
- The number of **missing values**.
- **Well-formedness** of the data.
 - If the data is in TSV or CSV, check that the column separators and line separators always correctly separate columns and lines.
 - If the data is in HTML or XML format, check whether the data is well formed based on their respective standards.
 - Parsing may also be necessary in order to extract structured information from semi-structured or unstructured data.
- **Inconsistent data records.** Check the range of values are allowed. e.g. If the data contains student GPA, check if the GPA is in the designated range, say 0~4.

When you find issues with data, **processing steps** are necessary which often involves cleaning missing values, data normalization, discretization, text processing to remove and/or replace embedded characters which may affect data alignment, mixed data types in common fields, and others.

Azure Machine Learning consumes well-formed tabular data. If the data is already in tabular form, data pre-processing can be performed directly with Azure Machine Learning in the Machine Learning Studio. If data is not in tabular form, say it is in XML, parsing may be required in order to convert the data to tabular form.

What are some of the major tasks in data pre-processing?

- **Data cleaning:** Fill in or missing values, detect and remove noisy data and outliers.
- **Data transformation:** Normalize data to reduce dimensions and noise.
- **Data reduction:** Sample data records or attributes for easier data handling.
- **Data discretization:** Convert continuous attributes to categorical attributes for ease of use with certain machine learning methods.
- **Text cleaning:** remove embedded characters which may cause data misalignment, for e.g., embedded tabs in a tab-separated data file, embedded new lines which may break records, etc.

The sections below detail some of these data processing steps.

How to deal with missing values?

To deal with missing values, it is best to first identify the reason for the missing values to better handle the problem. Typical missing value handling methods are:

- **Deletion:** Remove records with missing values
- **Dummy substitution:** Replace missing values with a dummy value: e.g, *unknown* for categorical or 0 for numerical values.
- **Mean substitution:** If the missing data is numerical, replace the missing values with the mean.
- **Frequent substitution:** If the missing data is categorical, replace the missing values with the most frequent item
- **Regression substitution:** Use a regression method to replace missing values with regressed values.

How to normalize data?

Data normalization re-scales numerical values to a specified range. Popular data normalization methods include:

- **Min-Max Normalization:** Linearly transform the data to a range, say between 0 and 1, where the min value is scaled to 0 and max value to 1.
- **Z-score Normalization:** Scale data based on mean and standard deviation: divide the difference between the data and the mean by the standard deviation.
- **Decimal scaling:** Scale the data by moving the decimal point of the attribute value.

How to discretize data?

Data can be discretized by converting continuous values to nominal attributes or intervals. Some ways of doing this are:

- **Equal-Width Binning:** Divide the range of all possible values of an attribute into N groups of the same size, and assign the values that fall in a bin with the bin number.
- **Equal-Height Binning:** Divide the range of all possible values of an attribute into N groups, each containing the same number of instances, then assign the values that fall in a bin with the bin number.

How to reduce data?

There are various methods to reduce data size for easier data handling. Depending on data size and the domain, the following methods can be applied:

- **Record Sampling:** Sample the data records and only choose the representative subset from the data.
- **Attribute Sampling:** Select only a subset of the most important attributes from the data.
- **Aggregation:** Divide the data into groups and store the numbers for each group. For example, the daily revenue numbers of a restaurant chain over the past 20 years can be aggregated to monthly revenue to reduce the size of the data.

How to clean text data?

Text fields in tabular data may include characters which affect columns alignment and/or record boundaries. For e.g., embedded tabs in a tab-separated file cause column misalignment, and embedded new line characters break record lines. Improper text encoding handling while writing/reading text leads to information loss, inadvertent introduction of unreadable characters, e.g., nulls, and may also affect text parsing. Careful parsing and editing may be required in order to clean text fields for proper alignment and/or to extract structured data from unstructured or semi-structured text data.

Data exploration offers an early view into the data. A number of data issues can be uncovered during this step and corresponding methods can be applied to address those issues. It is important to ask questions such as what is the source of the issue and how the issue may have been introduced. This also helps you decide on the data processing steps that need to be taken to resolve them. The kind of insights one intends to derive from the data can also be used to prioritize the data processing effort.

References

Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann, 2011, Jiawei Han, Micheline Kamber, and Jian Pei

Explore data in the Team Data Science Process

3/12/2019 • 2 minutes to read

Exploring data is a step in the [Team Data Science Process](#).

The following articles describe how to explore data in three different storage environments that are typically used in the Data Science Process:

- Explore [Azure blob container](#) data using the [Pandas](#) Python package.
- Explore [SQL Server](#) data by using SQL and by using a programming language like Python.
- Explore [Hive table](#) data using Hive queries.

In addition, the video, [Preprocessing Data in Azure Machine Learning Studio](#), describes the commonly used modules for cleaning and transforming data in Azure Machine Learning Studio.

Explore data in Azure blob storage with pandas

3/12/2019 • 2 minutes to read

This article covers how to explore data that is stored in Azure blob container using [pandas](#) Python package.

This task is a step in the [Team Data Science Process](#).

Prerequisites

This article assumes that you have:

- Created an Azure storage account. If you need instructions, see [Create an Azure Storage account](#)
- Stored your data in an Azure blob storage account. If you need instructions, see [Moving data to and from Azure Storage](#)

Load the data into a pandas DataFrame

To explore and manipulate a dataset, it must first be downloaded from the blob source to a local file, which can then be loaded in a pandas DataFrame. Here are the steps to follow for this procedure:

1. Download the data from Azure blob with the following Python code sample using blob service. Replace the variable in the following code with your specific values:

```
from azure.storage.blob import BlobService
import tables

STORAGEACCOUNTNAME= <storage_account_name>
STORAGEACCOUNTKEY= <storage_account_key>
LOCALFILENAME= <local_file_name>
CONTAINERNAME= <container_name>
BLOBNAME= <blob_name>

#download from blob
t1=time.time()
blob_service=BlobService(account_name=STORAGEACCOUNTNAME,account_key=STORAGEACCOUNTKEY)
blob_service.get_blob_to_path(CONTAINERNAME,BLOBNAME,LOCALFILENAME)
t2=time.time()
print(("It takes %s seconds to download "+blobname) % (t2 - t1))
```

1. Read the data into a pandas DataFrame from the downloaded file.

```
#LOCALFILE is the file path
dataframe_blobdata = pd.read_csv(LOCALFILE)
```

Now you are ready to explore the data and generate features on this dataset.

Examples of data exploration using pandas

Here are a few examples of ways to explore data using pandas:

1. Inspect the **number of rows and columns**

```
print 'the size of the data is: %d rows and %d columns' % dataframe_blobdata.shape
```

1. **Inspect** the first or last few **rows** in the following dataset:

```
dataframe_blobdata.head(10)
```

```
dataframe_blobdata.tail(10)
```

1. Check the **data type** each column was imported as using the following sample code

```
for col in dataframe_blobdata.columns:  
    print dataframe_blobdata[col].name, ':', dataframe_blobdata[col].dtype
```

1. Check the **basic stats** for the columns in the data set as follows

```
dataframe_blobdata.describe()
```

1. Look at the number of entries for each column value as follows

```
dataframe_blobdata['<column_name>'].value_counts()
```

1. **Count missing values** versus the actual number of entries in each column using the following sample code

```
miss_num = dataframe_blobdata.shape[0] - dataframe_blobdata.count()  
print miss_num
```

1. If you have **missing values** for a specific column in the data, you can drop them as follows:

```
dataframe_blobdata_noNA = dataframe_blobdata.dropna()  
dataframe_blobdata_noNA.shape
```

Another way to replace missing values is with the mode function:

```
dataframe_blobdata_mode =  
dataframe_blobdata.fillna({'<column_name>':dataframe_blobdata['<column_name>'].mode()[0]})
```

1. Create a **histogram** plot using variable number of bins to plot the distribution of a variable

```
dataframe_blobdata['<column_name>'].value_counts().plot(kind='bar')  
  
np.log(dataframe_blobdata['<column_name>']+1).hist(bins=50)
```

1. Look at **correlations** between variables using a scatterplot or using the built-in correlation function

```
#relationship between column_a and column_b using scatter plot  
plt.scatter(dataframe_blobdata['<column_a>'], dataframe_blobdata['<column_b>'])  
  
#correlation between column_a and column_b  
dataframe_blobdata[['<column_a>', '<column_b>']].corr()
```

Explore data in SQL Server Virtual Machine on Azure

3/12/2019 • 2 minutes to read

This article covers how to explore data that is stored in a SQL Server VM on Azure. This can be done by data wrangling using SQL or by using a programming language like Python.

This task is a step in the [Team Data Science Process](#).

NOTE

The sample SQL statements in this document assume that data is in SQL Server. If it isn't, refer to the cloud data science process map to learn how to move your data to SQL Server.

Explore SQL data with SQL scripts

Here are a few sample SQL scripts that can be used to explore data stores in SQL Server.

1. Get the count of observations per day

```
SELECT CONVERT(date, <date_columnname>) as date, count(*) as c from <tablename> group by CONVERT(date, <date_columnname>)
```

2. Get the levels in a categorical column

```
select distinct <column_name> from <databasename>
```

3. Get the number of levels in combination of two categorical columns

```
select <column_a>, <column_b>, count(*) from <tablename> group by <column_a>, <column_b>
```

4. Get the distribution for numerical columns

```
select <column_name>, count(*) from <tablename> group by <column_name>
```

NOTE

For a practical example, you can use the [NYC Taxi dataset](#) and refer to the IPNB titled [NYC Data wrangling using IPython Notebook and SQL Server](#) for an end-to-end walk-through.

Explore SQL data with Python

Using Python to explore data and generate features when the data is in SQL Server is similar to processing data in Azure blob using Python, as documented in [Process Azure Blob data in your data science environment](#). The data needs to be loaded from the database into a pandas DataFrame and then can be processed further. We document the process of connecting to the database and loading the data into the DataFrame in this section.

The following connection string format can be used to connect to a SQL Server database from Python using pyodbc (replaceservername, dbname, username, and password with your specific values):

```
#Set up the SQL Azure connection
import pyodbc
conn = pyodbc.connect('DRIVER={SQL Server};SERVER=<servername>;DATABASE=<dbname>;UID=<username>;PWD=<password>')
```

The [Pandas library](#) in Python provides a rich set of data structures and data analysis tools for data manipulation for Python programming. The following code reads the results returned from a SQL Server database into a Pandas data frame:

```
# Query database and load the returned results in pandas data frame
data_frame = pd.read_sql('''select <columnname1>, <columnname2>... from <tablename>''', conn)
```

Now you can work with the Pandas DataFrame as covered in the topic [Process Azure Blob data in your data science environment](#).

The Team Data Science Process in action example

For an end-to-end walkthrough example of the Cortana Analytics Process using a public dataset, see [The Team Data Science Process in action: using SQL Server](#).

Explore data in Hive tables with Hive queries

3/12/2019 • 2 minutes to read

This article provides sample Hive scripts that are used to explore data in Hive tables in an HDInsight Hadoop cluster.

This task is a step in the [Team Data Science Process](#).

Prerequisites

This article assumes that you have:

- Created an Azure storage account. If you need instructions, see [Create an Azure Storage account](#)
- Provisioned a customized Hadoop cluster with the HDInsight service. If you need instructions, see [Customize Azure HDInsight Clusters for Advanced Analytics](#).
- The data has been uploaded to Hive tables in Azure HDInsight Hadoop clusters. If it has not, follow the instructions in [Create and load data to Hive tables](#) to upload data to Hive tables first.
- Enabled remote access to the cluster. If you need instructions, see [Access the Head Node of Hadoop Cluster](#).
- If you need instructions on how to submit Hive queries, see [How to Submit Hive Queries](#)

Example Hive query scripts for data exploration

1. Get the count of observations per partition

```
SELECT <partitionfieldname>, count(*) from <databasename>.<tablename> group by <partitionfieldname>;
```

2. Get the count of observations per day

```
SELECT to_date(<date_columnname>), count(*) from <databasename>.<tablename> group by  
to_date(<date_columnname>);
```

3. Get the levels in a categorical column

```
SELECT distinct <column_name> from <databasename>.<tablename>
```

4. Get the number of levels in combination of two categorical columns

```
SELECT <column_a>, <column_b>, count(*) from <databasename>.<tablename> group by <column_a>, <column_b>
```

5. Get the distribution for numerical columns

```
SELECT <column_name>, count(*) from <databasename>.<tablename> group by <column_name>
```

6. Extract records from joining two tables

```
SELECT
    a.<common_columnname1> as <new_name1>,
    a.<common_columnname2> as <new_name2>,
    a.<a_column_name1> as <new_name3>,
    a.<a_column_name2> as <new_name4>,
    b.<b_column_name1> as <new_name5>,
    b.<b_column_name2> as <new_name6>
FROM
(
    SELECT <common_columnname1>,
        <common_columnname2>,
        <a_column_name1>,
        <a_column_name2>,
    FROM <database>.<table>
) a
join
(
    SELECT <common_columnname1>,
        <common_columnname2>,
        <b_column_name1>,
        <b_column_name2>,
    FROM <database>.<table>
) b
ON a.<common_columnname1>=b.<common_columnname1> and a.<common_columnname2>=b.<common_columnname2>
```

Additional query scripts for taxi trip data scenarios

Examples of queries that are specific to [NYC Taxi Trip Data](#) scenarios are also provided in [GitHub repository](#). These queries already have data schema specified and are ready to be submitted to run.

Sample data in Azure blob containers, SQL Server, and Hive tables

1/30/2019 • 2 minutes to read

The following articles describe how to sample data that is stored in one of three different Azure locations:

- [Azure blob container data](#) is sampled by downloading it programmatically and then sampling it with sample Python code.
- [SQL Server data](#) is sampled using both SQL and the Python Programming Language.
- [Hive table data](#) is sampled using Hive queries.

This sampling task is a step in the [Team Data Science Process \(TDSP\)](#).

Why sample data?

If the dataset you plan to analyze is large, it's usually a good idea to down-sample the data to reduce it to a smaller but representative and more manageable size. This facilitates data understanding, exploration, and feature engineering. Its role in the Cortana Analytics Process is to enable fast prototyping of the data processing functions and machine learning models.

Sample data in Azure blob storage

1/30/2019 • 2 minutes to read

This article covers sampling data stored in Azure blob storage by downloading it programmatically and then sampling it using procedures written in Python.

Why sample your data? If the dataset you plan to analyze is large, it's usually a good idea to down-sample the data to reduce it to a smaller but representative and more manageable size. This facilitates data understanding, exploration, and feature engineering. Its role in the Cortana Analytics Process is to enable fast prototyping of the data processing functions and machine learning models.

This sampling task is a step in the [Team Data Science Process \(TDSP\)](#).

Download and down-sample data

1. Download the data from Azure blob storage using the blob service from the following sample Python code:

```
from azure.storage.blob import BlobService
import tables

STORAGEACCOUNTNAME= <storage_account_name>
STORAGEACCOUNTKEY= <storage_account_key>
LOCALFILENAME= <local_file_name>
CONTAINERNAME= <container_name>
BLOBNAME= <blob_name>

#download from blob
t1=time.time()
blob_service=BlobService(account_name=STORAGEACCOUNTNAME,account_key=STORAGEACCOUNTKEY)
blob_service.get_blob_to_path(CONTAINERNAME,BLOBNAME,LOCALFILENAME)
t2=time.time()
print(("It takes %s seconds to download "+blobname) % (t2 - t1))
```

2. Read data into a Pandas data-frame from the file downloaded above.

```
import pandas as pd

#directly ready from file on disk
dataframe_blobdata = pd.read_csv(LOCALFILE)
```

3. Down-sample the data using the `numpy`'s `random.choice` as follows:

```
# A 1 percent sample
sample_ratio = 0.01
sample_size = np.round(dataframe_blobdata.shape[0] * sample_ratio)
sample_rows = np.random.choice(dataframe_blobdata.index.values, sample_size)
dataframe_blobdata_sample = dataframe_blobdata.ix[sample_rows]
```

Now you can work with the above data frame with the 1 Percent sample for further exploration and feature generation.

Upload data and read it into Azure Machine Learning

You can use the following sample code to down-sample the data and use it directly in Azure Machine Learning:

1. Write the data frame to a local file

```
dataframe.to_csv(os.path.join(os.getcwd(), LOCALFILENAME), sep='\t', encoding='utf-8', index=False)
```

2. Upload the local file to an Azure blob using the following sample code:

```
from azure.storage.blob import BlobService
import tables

STORAGEACCOUNTNAME= <storage_account_name>
LOCALFILENAME= <local_file_name>
STORAGEACCOUNTKEY= <storage_account_key>
CONTAINERNAME= <container_name>
BLOBNAME= <blob_name>

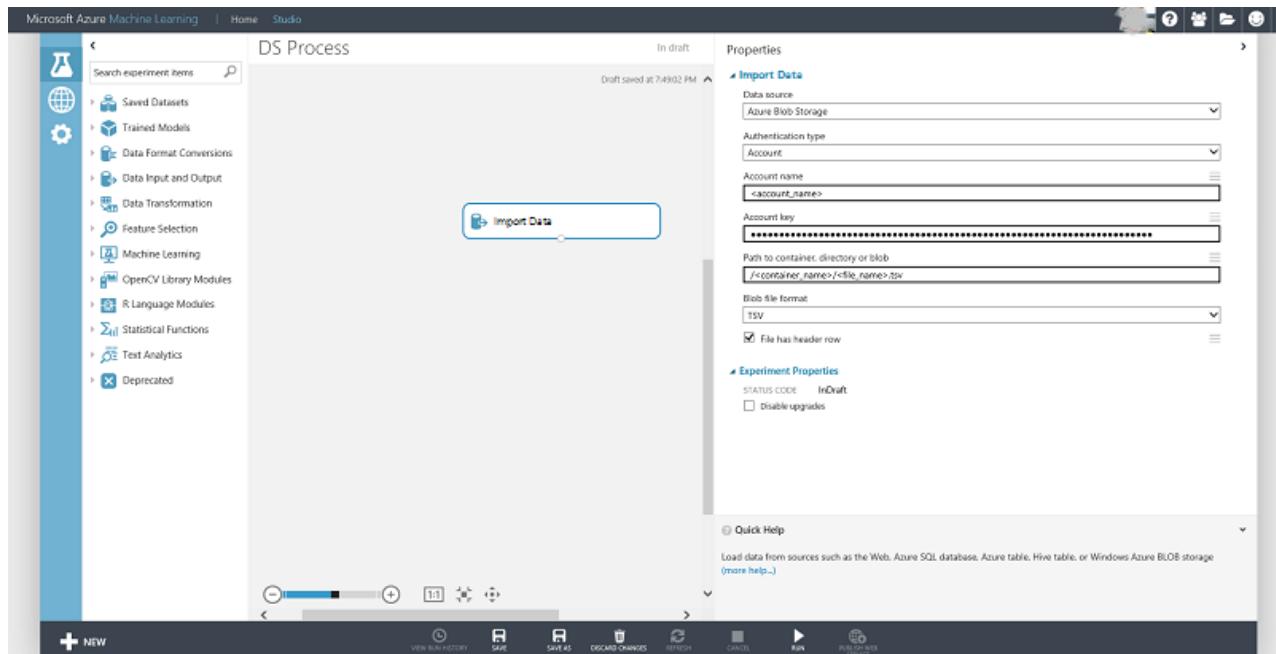
output_blob_service=BlobService(account_name=STORAGEACCOUNTNAME,account_key=STORAGEACCOUNTKEY)
localfileprocessed = os.path.join(os.getcwd(),LOCALFILENAME) #assuming file is in current working directory

try:

    #perform upload
    output_blob_service.put_block_blob_from_path(CONTAINERNAME,BLOBNAME,localfileprocessed)

except:
    print ("Something went wrong with uploading to the blob:"+ BLOBNAME)
```

3. Read the data from the Azure blob using Azure Machine Learning [Import Data](#) as shown in the image below:



Sample data in SQL Server on Azure

3/12/2019 • 3 minutes to read

This article shows how to sample data stored in SQL Server on Azure using either SQL or the Python programming language. It also shows how to move sampled data into Azure Machine Learning by saving it to a file, uploading it to an Azure blob, and then reading it into Azure Machine Learning Studio.

The Python sampling uses the [pyodbc](#) ODBC library to connect to SQL Server on Azure and the [Pandas](#) library to do the sampling.

NOTE

The sample SQL code in this document assumes that the data is in a SQL Server on Azure. If it is not, refer to [Move data to SQL Server on Azure](#) article for instructions on how to move your data to SQL Server on Azure.

Why sample your data? If the dataset you plan to analyze is large, it's usually a good idea to down-sample the data to reduce it to a smaller but representative and more manageable size. This facilitates data understanding, exploration, and feature engineering. Its role in the [Team Data Science Process \(TDSP\)](#) is to enable fast prototyping of the data processing functions and machine learning models.

This sampling task is a step in the [Team Data Science Process \(TDSP\)](#).

Using SQL

This section describes several methods using SQL to perform simple random sampling against the data in the database. Choose a method based on your data size and its distribution.

The following two items show how to use `newid` in SQL Server to perform the sampling. The method you choose depends on how random you want the sample to be (pk_id in the following sample code is assumed to be an auto-generated primary key).

1. Less strict random sample

```
select * from <table_name> where <primary_key> in  
(select top 10 percent <primary_key> from <table_name> order by newid())
```

2. More random sample

```
SELECT * FROM <table_name>  
WHERE 0.1 >= CAST(CHECKSUM(NEWID(), <primary_key>) & 0xffffffff AS float)/ CAST (0xffffffff AS int)
```

Tablesample can be used for sampling the data as well. This may be a better approach if your data size is large (assuming that data on different pages is not correlated) and for the query to complete in a reasonable time.

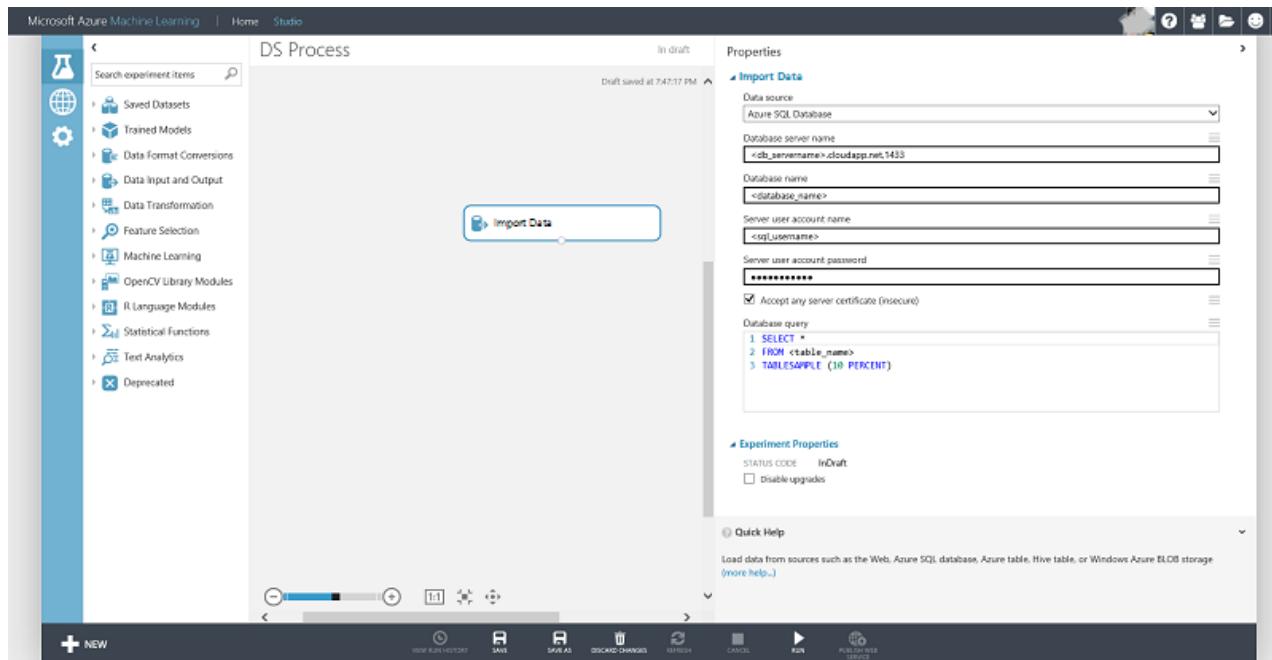
```
SELECT *  
FROM <table_name>  
TABLESAMPLE (10 PERCENT)
```

NOTE

You can explore and generate features from this sampled data by storing it in a new table

Connecting to Azure Machine Learning

You can directly use the sample queries above in the Azure Machine Learning [Import Data](#) module to down-sample the data on the fly and bring it into an Azure Machine Learning experiment. A screenshot of using the reader module to read the sampled data is shown here:



Using the Python programming language

This section demonstrates using the [pyodbc library](#) to establish an ODBC connect to a SQL server database in Python. The database connection string is as follows: (replace servername, dbname, username and password with your configuration):

```
#Set up the SQL Azure connection
import pyodbc
conn = pyodbc.connect('DRIVER={SQL Server};SERVER=<servername>;DATABASE=<dbname>;UID=<username>;PWD=<password>')
```

The [Pandas](#) library in Python provides a rich set of data structures and data analysis tools for data manipulation for Python programming. The following code reads a 0.1% sample of the data from a table in Azure SQL database into a Pandas data frame:

```
import pandas as pd

# Query database and load the returned results in pandas data frame
data_frame = pd.read_sql('''select column1, column2... from <table_name> tablesample (0.1 percent)''', conn)
```

You can now work with the sampled data in the Pandas data frame.

Connecting to Azure Machine Learning

You can use the following sample code to save the down-sampled data to a file and upload it to an Azure blob. The data in the blob can be directly read into an Azure Machine Learning Experiment using the [Import Data](#) module. The steps are as follows:

1. Write the pandas data frame to a local file

```
dataframe.to_csv(os.path.join(os.getcwd(),LOCALFILENAME), sep='\t', encoding='utf-8', index=False)
```

2. Upload local file to Azure blob

```
from azure.storage import BlobService
import tables

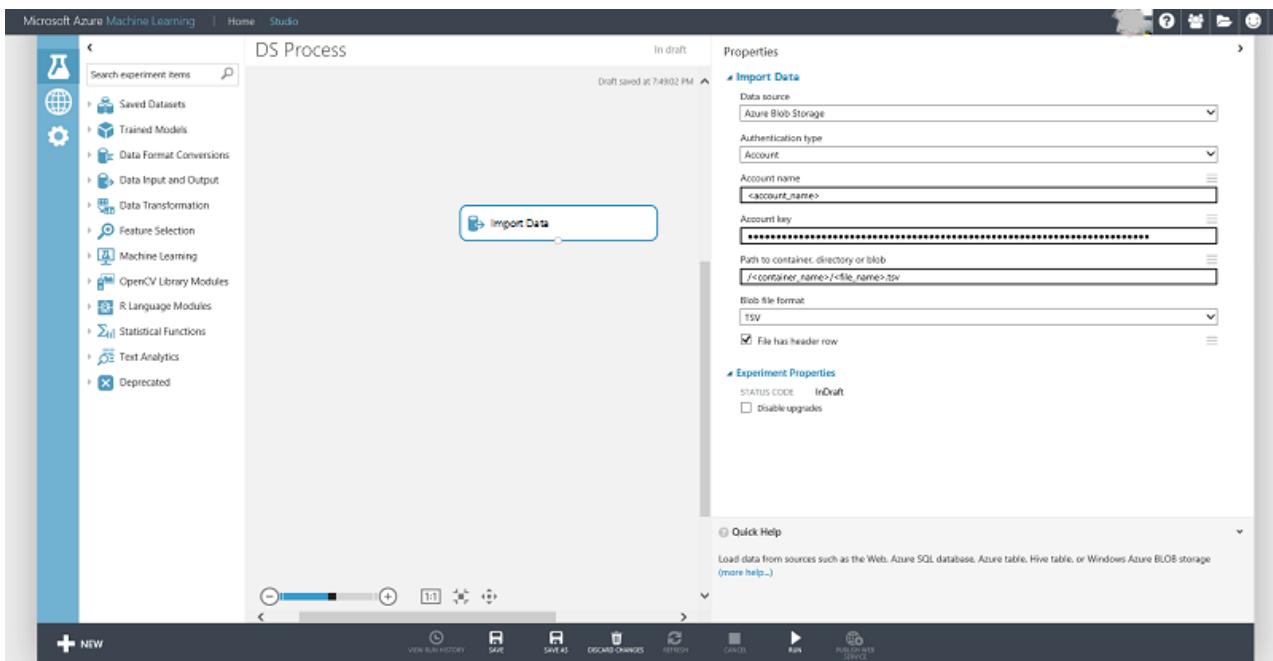
STORAGEACCOUNTNAME= <storage_account_name>
LOCALFILENAME= <local_file_name>
STORAGEACCOUNTKEY= <storage_account_key>
CONTAINERNAME= <container_name>
BLOBNAME= <blob_name>

output_blob_service=BlobService(account_name=STORAGEACCOUNTNAME,account_key=STORAGEACCOUNTKEY)
localfileprocessed = os.path.join(os.getcwd(),LOCALFILENAME) #assuming file is in current working directory

try:
    #perform upload
    output_blob_service.put_block_blob_from_path(CONTAINERNAME,BLOBNAME,localfileprocessed)

except:
    print ("Something went wrong with uploading blob:"+BLOBNAME)
```

3. Read data from Azure blob using Azure Machine Learning Import Data module as shown in the following screen grab:



The Team Data Science Process in Action example

To walkthrough an example of the Team Data Science Process a using a public dataset, see [Team Data Science Process in Action: using SQL Server](#).

Sample data in Azure HDInsight Hive tables

1/30/2019 • 2 minutes to read

This article describes how to down-sample data stored in Azure HDInsight Hive tables using Hive queries to reduce it to a size more manageable for analysis. It covers three popularly used sampling methods:

- Uniform random sampling
- Random sampling by groups
- Stratified sampling

Why sample your data? If the dataset you plan to analyze is large, it's usually a good idea to down-sample the data to reduce it to a smaller but representative and more manageable size. Down-sampling facilitates data understanding, exploration, and feature engineering. Its role in the Team Data Science Process is to enable fast prototyping of the data processing functions and machine learning models.

This sampling task is a step in the [Team Data Science Process \(TDSP\)](#).

How to submit Hive queries

Hive queries can be submitted from the Hadoop Command-Line console on the head node of the Hadoop cluster. To do this, log into the head node of the Hadoop cluster, open the Hadoop Command-Line console, and submit the Hive queries from there. For instructions on submitting Hive queries in the Hadoop Command-Line console, see [How to Submit Hive Queries](#).

Uniform random sampling

Uniform random sampling means that each row in the data set has an equal chance of being sampled. It can be implemented by adding an extra field rand() to the data set in the inner "select" query, and in the outer "select" query condition on that random field.

Here is an example query:

```
SET sampleRate=<sample rate, 0-1>;
select
    field1, field2, ..., fieldN
from
(
    select
        field1, field2, ..., fieldN, rand() as samplekey
    from <hive table name>
) a
where samplekey<='${hiveconf:sampleRate}'
```

Here, `<sample rate, 0-1>` specifies the proportion of records that the users want to sample.

Random sampling by groups

When sampling categorical data, you may want to either include or exclude all of the instances for some value of the categorical variable. This sort of sampling is called "sampling by group". For example, if you have a categorical variable "State", which has values such as NY, MA, CA, NJ, and PA, you want records from each state to be together, whether they are sampled or not.

Here is an example query that samples by group:

```

SET sampleRate=<sample rate, 0-1>;
select
    b.field1, b.field2, ..., b.catfield, ..., b.fieldN
from
(
    select
        field1, field2, ..., catfield, ..., fieldN
    from <table name>
) b
join
(
    select
        catfield
    from
    (
        select
            catfield, rand() as samplekey
        from <table name>
        group by catfield
    ) a
    where samplekey<='${hiveconf:sampleRate}'
) c
on b.catfield=c.catfield

```

Stratified sampling

Random sampling is stratified with respect to a categorical variable when the samples obtained have categorical values that are present in the same ratio as they were in the parent population. Using the same example as above, suppose your data has the following observations by states: NJ has 100 observations, NY has 60 observations, and WA has 300 observations. If you specify the rate of stratified sampling to be 0.5, then the sample obtained should have approximately 50, 30, and 150 observations of NJ, NY, and WA respectively.

Here is an example query:

```

SET sampleRate=<sample rate, 0-1>;
select
    field1, field2, field3, ..., fieldN, state
from
(
    select
        field1, field2, field3, ..., fieldN, state,
        count(*) over (partition by state) as state_cnt,
        rank() over (partition by state order by rand()) as state_rank
    from <table name>
) a
where state_rank <= state_cnt* '${hiveconf:sampleRate}'

```

For information on more advanced sampling methods that are available in Hive, see [LanguageManual Sampling](#).

Access datasets with Python using the Azure Machine Learning Python client library

1/30/2019 • 8 minutes to read

The preview of Microsoft Azure Machine Learning Python client library can enable secure access to your Azure Machine Learning datasets from a local Python environment and enables the creation and management of datasets in a workspace.

This topic provides instructions on how to:

- install the Machine Learning Python client library
- access and upload datasets, including instructions on how to get authorization to access Azure Machine Learning datasets from your local Python environment
- access intermediate datasets from experiments
- use the Python client library to enumerate datasets, access metadata, read the contents of a dataset, create new datasets and update existing datasets

Try [Azure Machine Learning Studio](#), available in paid or free options.

Prerequisites

The Python client library has been tested under the following environments:

- Windows, Mac and Linux
- Python 2.7, 3.3 and 3.4

It has a dependency on the following packages:

- requests
- python-dateutil
- pandas

We recommend using a Python distribution such as [Anaconda](#) or [Canopy](#), which come with Python, IPython and the three packages listed above installed. Although IPython is not strictly required, it is a great environment for manipulating and visualizing data interactively.

How to install the Azure Machine Learning Python client library

The Azure Machine Learning Python client library must also be installed to complete the tasks outlined in this topic. It is available from the [Python Package Index](#). To install it in your Python environment, run the following command from your local Python environment:

```
pip install azureml
```

Alternatively, you can download and install from the sources on [GitHub](#).

```
python setup.py install
```

If you have git installed on your machine, you can use pip to install directly from the git repository:

```
pip install git+https://github.com/Azure/Azure-MachineLearning-ClientLibrary-Python.git
```

Use Studio Code snippets to access datasets

The Python client library gives you programmatic access to your existing datasets from experiments that have been run.

From the Studio web interface, you can generate code snippets that include all the necessary information to download and deserialize datasets as pandas DataFrame objects on your local machine.

Security for data access

The code snippets provided by Studio for use with the Python client library includes your workspace id and authorization token. These provide full access to your workspace and must be protected, like a password.

For security reasons, the code snippet functionality is only available to users that have their role set as **Owner** for the workspace. Your role is displayed in Azure Machine Learning Studio on the **USERS** page under **Settings**.

The screenshot shows the 'settings' page in the Azure Machine Learning Studio. On the left sidebar, 'SETTINGS' is selected. The main area displays a table titled 'USERS' with columns: NAME, EMAIL, ROLE, and STATUS. One row is visible, showing a user with the role 'Owner' and status 'Active'. The 'NAME' and 'EMAIL' columns contain redacted text.

If your role is not set as **Owner**, you can either request to be reinvited as an owner, or ask the owner of the workspace to provide you with the code snippet.

To obtain the authorization token, you can do one of the following:

- Ask for a token from an owner. Owners can access their authorization tokens from the Settings page of their workspace in Studio. Select **Settings** from the left pane and click **AUTHORIZATION TOKENS** to see the primary and secondary tokens. Although either the primary or the secondary authorization tokens can be used in the code snippet, it is recommended that owners only share the secondary authorization tokens.

The screenshot shows the 'settings' page in the Azure Machine Learning Studio. On the left sidebar, 'SETTINGS' is selected. The main area displays a table with tabs: 'NAME', 'AUTHORIZATION TOKENS', and 'USERS'. The 'AUTHORIZATION TOKENS' tab is active, showing two fields: 'PRIMARY AUTHORIZATION TOKEN' and 'SECONDARY AUTHORIZATION TOKEN', each with a redacted value and a 'Regenerate' button.

- Ask to be promoted to role of owner. To do this, a current owner of the workspace needs to first remove you from the workspace then re-invite you to it as an owner.

Once developers have obtained the workspace id and authorization token, they are able to access the workspace using the code snippet regardless of their role.

Authorization tokens are managed on the **AUTHORIZATION TOKENS** page under **SETTINGS**. You can regenerate them, but this procedure revokes access to the previous token.

Access datasets from a local Python application

1. In Machine Learning Studio, click **DATASETS** in the navigation bar on the left.
2. Select the dataset you would like to access. You can select any of the datasets from the **MY DATASETS** list or from the **SAMPLES** list.
3. From the bottom toolbar, click **Generate Data Access Code**. If the data is in a format incompatible with the Python client library, this button is disabled.

The screenshot shows the Machine Learning Studio interface. On the left, there is a sidebar with icons for EXPERIMENTS, WEB SERVICES, MODULES, DATASETS (which is selected and highlighted in blue), TRAINED MODELS, and SETTINGS. The main area is titled "datasets" and shows a table of datasets. The table has columns: NAME, SUBMITTED BY, DESCRIPTION, DATA TYPE, and CREATION DATE. One dataset is listed: "My Data.tsv" submitted by "ptvsazure" with description "My Test Data", type "GenericTSV", and creation date "1/20/2015 12:3...". At the bottom of the table is a search icon. Below the table is a toolbar with "NEW", "DOWNLOAD", "DELETE", and a "GENERATE DATA ACCESS CODE..." button, which is highlighted with a red box.

4. Select the code snippet from the window that appears and copy it to your clipboard.

The screenshot shows a modal dialog box titled "GENERATE DATA ACCESS CODE". It contains the text "Use this code to access your data" and "To programmatically access this dataset, copy the code snippet into your favorite development environment. [Learn More](#)". A note below says "Note: this code includes your workspace access token, which provides full access to your workspace. It should be treated like a password." The "CODE SNIPPET" section is titled "Python" and contains the following code:

```
from azureml import Workspace
ws = Workspace(
    workspace_id='[REDACTED]',
    authorization_token='[REDACTED]',
)
ds = ws.datasets['My Data.tsv']
frame = ds.to_dataframe()
```

Below the code snippet is a checkbox labeled "USE SECONDARY TOKEN" and a checkmark icon.

5. Paste the code into the notebook of your local Python application.

The screenshot shows an IPython Notebook interface. In cell [3], Python code is run to import the Azure ML library and load a dataset named 'My Data.tsv' into a DataFrame. In cell [4], the resulting DataFrame is displayed as a table.

```
In [3]: from azureml import Workspace
ws = Workspace(
    workspace_id='[REDACTED]',
    authorization_token='[REDACTED]')
ds = ws.datasets['My Data.tsv']
frame = ds.to_dataframe()

In [4]: frame
```

Out[4]:

	fLength	fWidth	fSize	fConc	fConcl	fAsym	fM3Long	fM3Trans	fAlpha	fDist	Class
0	29.4491	12.7271	2.6637	0.3536	0.1876	-19.4070	-18.1295	7.1258	8.1501	252.1500	g
1	51.5830	10.7969	2.6222	0.5227	0.2733	-64.0583	-30.1280	4.3855	15.0428	269.4810	g
2	36.3558	10.3843	2.8531	0.5309	0.3599	15.4179	51.9328	16.2848	9.1263	208.7935	h
3	37.2577	12.0793	2.4354	0.3560	0.2037	5.1882	-17.8545	6.0370	30.0150	61.1727	h
4	34.8906	15.7072	2.7147	0.3587	0.1938	-8.5682	-12.6514	-14.3676	89.7920	222.4690	h
5	60.4957	17.6753	2.9380	0.1753	0.0894	31.3257	-15.9801	14.4117	15.6390	113.1820	g
6	18.5731	16.2365	2.6758	0.4895	0.3091	1.1158	8.1104	-11.7988	50.9791	224.8780	h
7	21.5929	12.4539	2.3512	0.4900	0.3096	0.1172	-4.3583	2.5525	58.3290	52.0772	g
8	45.5590	9.9957	2.5809	0.3885	0.1955	17.0284	34.9608	-7.8856	14.4173	177.6820	g
9	62.3597	21.6946	3.1739	0.3087	0.2041	-45.3412	52.1023	22.5905	36.9876	274.7409	h

Access intermediate datasets from Machine Learning experiments

After an experiment is run in the Machine Learning Studio, it is possible to access the intermediate datasets from the output nodes of modules. Intermediate datasets are data that has been created and used for intermediate steps when a model tool has been run.

Intermediate datasets can be accessed as long as the data format is compatible with the Python client library.

The following formats are supported (constants for these are in the `azureml.DataTypeIds` class):

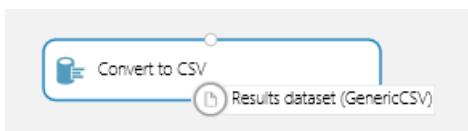
- PlainText
- GenericCSV
- GenericTSV
- GenericCSVNoHeader
- GenericTSVNoHeader

You can determine the format by hovering over a module output node. It is displayed along with the node name, in a tooltip.

Some of the modules, such as the [Split](#) module, output to a format named `Dataset`, which is not supported by the Python client library.

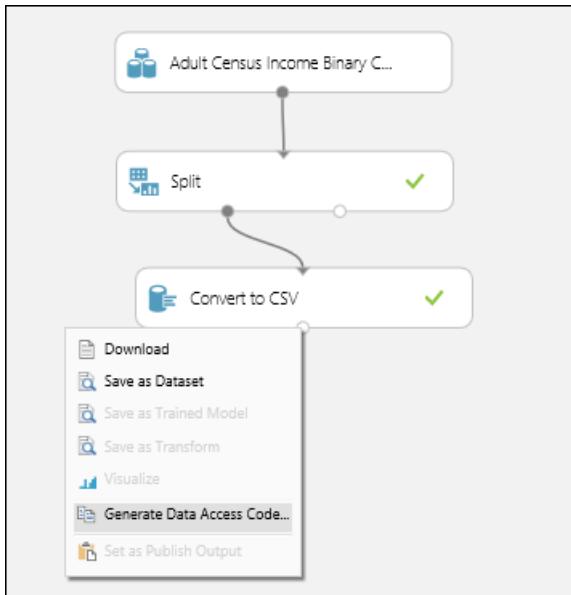


You need to use a conversion module, such as [Convert to CSV](#), to get an output into a supported format.

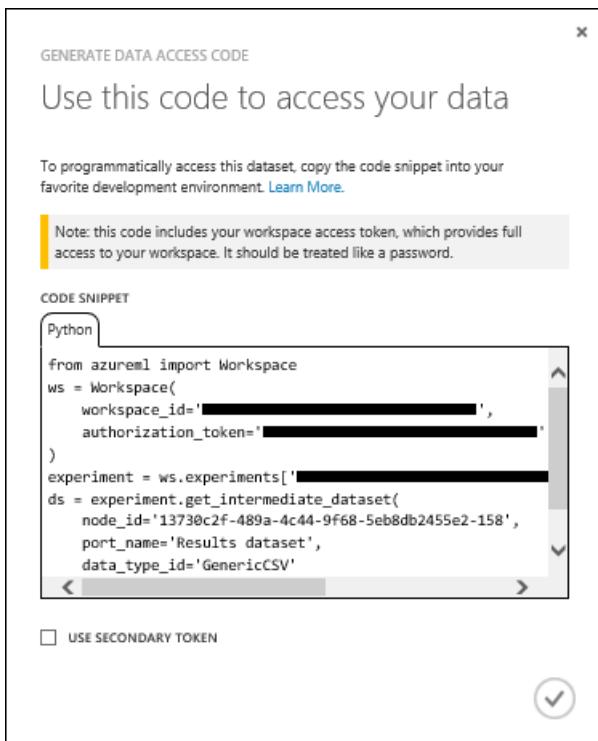


The following steps show an example that creates an experiment, runs it and accesses the intermediate dataset.

1. Create a new experiment.
2. Insert an **Adult Census Income Binary Classification** dataset module.
3. Insert a **Split** module, and connect its input to the dataset module output.
4. Insert a **Convert to CSV** module and connect its input to one of the **Split** module outputs.
5. Save the experiment, run it, and wait for it to finish running.
6. Click the output node on the **Convert to CSV** module.
7. When the context menu appears, select **Generate Data Access Code**.



8. Select the code snippet and copy it to your clipboard from the window that appears.



9. Paste the code in your notebook.

IP[y]: Notebook My Test Notebook Last Checkpoint: Jan 20 12:58 (unsaved changes)

File Edit View Insert Cell Kernel Help

Cell Toolbar: None

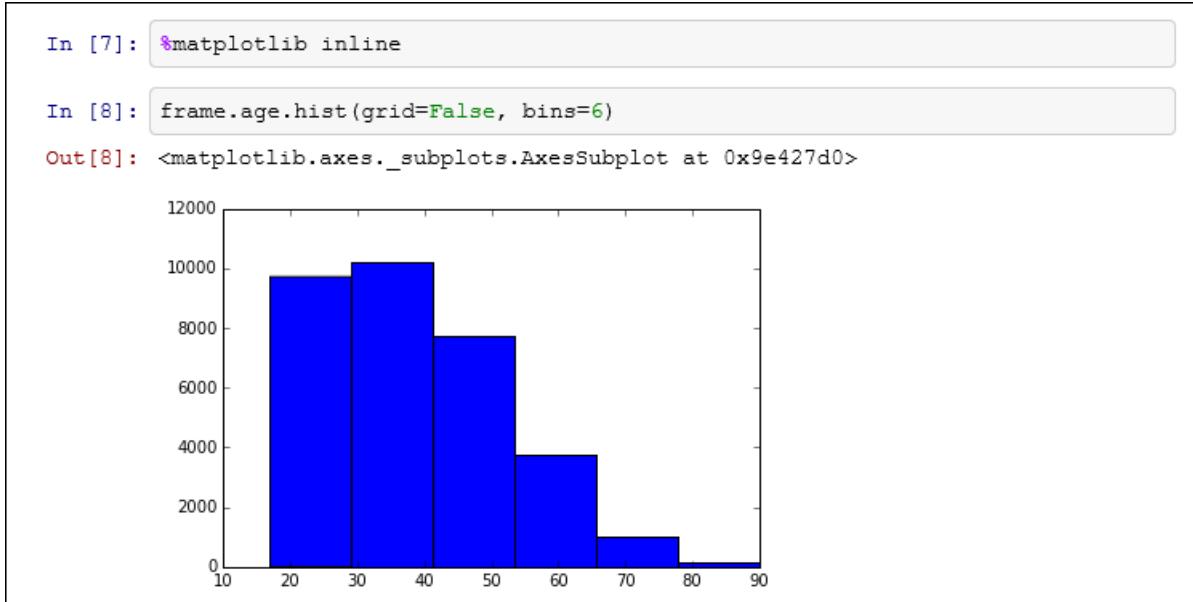
```
In [6]: from azureml import Workspace
ws = Workspace(
    workspace_id='[REDACTED]',
    authorization_token='[REDACTED]')
experiment = ws.experiments['[REDACTED]']
ds = experiment.get_intermediate_dataset(
    node_id='13730c2f-489a-4c44-9f68-5eb8db2455e2-158',
    port_name='Results dataset',
    data_type_id='GenericCSV')
frame = ds.to_dataframe()
```

```
In [7]: frame
```

```
Out[7]:
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss
0	52	Private	225317	5th-6th	3	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0
1	29	Private	154017	HS-grad	9	Never-married	Sales	Not-in-family	White	Female	0	0
2	25	Private	203570	HS-grad	9	Separated	Other-service	Unmarried	Black	Male	0	0

10. You can visualize the data using matplotlib. This displays in a histogram for the age column:



Use the Machine Learning Python client library to access, read, create, and manage datasets

Workspace

The workspace is the entry point for the Python client library. Provide the `Workspace` class with your workspace id and authorization token to create an instance:

```
ws = Workspace(workspace_id='4c29e1adeba2e5a7cbeb0e4f4adfb4df',
               authorization_token='f4f3ade2c6aefdb1afb043cd8bcf3daf')
```

Enumerate datasets

To enumerate all datasets in a given workspace:

```
for ds in ws.datasets:  
    print(ds.name)
```

To enumerate just the user-created datasets:

```
for ds in ws.user_datasets:  
    print(ds.name)
```

To enumerate just the example datasets:

```
for ds in ws.example_datasets:  
    print(ds.name)
```

You can access a dataset by name (which is case-sensitive):

```
ds = ws.datasets['my dataset name']
```

Or you can access it by index:

```
ds = ws.datasets[0]
```

Metadata

Datasets have metadata, in addition to content. (Intermediate datasets are an exception to this rule and do not have any metadata.)

Some metadata values are assigned by the user at creation time:

```
print(ds.name)  
print(ds.description)  
print(ds.family_id)  
print(ds.data_type_id)
```

Others are values assigned by Azure ML:

```
print(ds.id)  
print(ds.created_date)  
print(ds.size)
```

See the `SourceDataset` class for more on the available metadata.

Read contents

The code snippets provided by Machine Learning Studio automatically download and deserialize the dataset to a pandas DataFrame object. This is done with the `to_dataframe` method:

```
frame = ds.to_dataframe()
```

If you prefer to download the raw data, and perform the deserialization yourself, that is an option. At the moment, this is the only option for formats such as 'ARFF', which the Python client library cannot deserialize.

To read the contents as text:

```
text_data = ds.read_as_text()
```

To read the contents as binary:

```
binary_data = ds.read_as_binary()
```

You can also just open a stream to the contents:

```
with ds.open() as file:  
    binary_data_chunk = file.read(1000)
```

Create a new dataset

The Python client library allows you to upload datasets from your Python program. These datasets are then available for use in your workspace.

If you have your data in a pandas DataFrame, use the following code:

```
from azureml import DataTypeIds  
  
dataset = ws.datasets.add_from_dataframe(  
    dataframe=frame,  
    data_type_id=DataTypeIds.GenericCSV,  
    name='my new dataset',  
    description='my description'  
)
```

If your data is already serialized, you can use:

```
from azureml import DataTypeIds  
  
dataset = ws.datasets.add_from_raw_data(  
    raw_data=raw_data,  
    data_type_id=DataTypeIds.GenericCSV,  
    name='my new dataset',  
    description='my description'  
)
```

The Python client library is able to serialize a pandas DataFrame to the following formats (constants for these are in the `azureml.DataTypeIds` class):

- PlainText
- GenericCSV
- GenericTSV
- GenericCSVNoHeader
- GenericTSVNoHeader

Update an existing dataset

If you try to upload a new dataset with a name that matches an existing dataset, you should get a conflict error.

To update an existing dataset, you first need to get a reference to the existing dataset:

```
dataset = ws.datasets['existing dataset']

print(dataset.data_type_id) # 'GenericCSV'
print(dataset.name)        # 'existing dataset'
print(dataset.description) # 'data up to jan 2015'
```

Then use `update_from_dataframe` to serialize and replace the contents of the dataset on Azure:

```
dataset = ws.datasets['existing dataset']

dataset.update_from_dataframe(frame2)

print(dataset.data_type_id) # 'GenericCSV'
print(dataset.name)        # 'existing dataset'
print(dataset.description) # 'data up to jan 2015'
```

If you want to serialize the data to a different format, specify a value for the optional `data_type_id` parameter.

```
from azureml import DataTypeIds

dataset = ws.datasets['existing dataset']

dataset.update_from_dataframe(
    dataframe=frame2,
    data_type_id=DataTypeIds.GenericTSV,
)

print(dataset.data_type_id) # 'GenericTSV'
print(dataset.name)        # 'existing dataset'
print(dataset.description) # 'data up to jan 2015'
```

You can optionally set a new description by specifying a value for the `description` parameter.

```
dataset = ws.datasets['existing dataset']

dataset.update_from_dataframe(
    dataframe=frame2,
    description='data up to feb 2015',
)

print(dataset.data_type_id) # 'GenericCSV'
print(dataset.name)        # 'existing dataset'
print(dataset.description) # 'data up to feb 2015'
```

You can optionally set a new name by specifying a value for the `name` parameter. From now on, you'll retrieve the dataset using the new name only. The following code updates the data, name and description.

```
dataset = ws.datasets['existing dataset']

dataset.update_from_dataframe(
    dataframe=frame2,
    name='existing dataset v2',
    description='data up to feb 2015',
)

print(dataset.data_type_id)                      # 'GenericCSV'
print(dataset.name)                            # 'existing dataset v2'
print(dataset.description)                     # 'data up to feb 2015'

print(ws.datasets['existing dataset v2'].name) # 'existing dataset v2'
print(ws.datasets['existing dataset'].name)   # IndexError
```

The `data_type_id`, `name` and `description` parameters are optional and default to their previous value. The `dataframe` parameter is always required.

If your data is already serialized, use `update_from_raw_data` instead of `update_from_dataframe`. If you just pass in `raw_data` instead of `dataframe`, it works in a similar way.

Process Azure blob data with advanced analytics

1/30/2019 • 3 minutes to read

This document covers exploring data and generating features from data stored in Azure Blob storage.

Load the data into a Pandas data frame

In order to explore and manipulate a dataset, it must be downloaded from the blob source to a local file which can then be loaded in a Pandas data frame. Here are the steps to follow for this procedure:

1. Download the data from Azure blob with the following sample Python code using blob service. Replace the variable in the code below with your specific values:

```
from azure.storage.blob import BlobService
import tables

STORAGEACCOUNTNAME= <storage_account_name>
STORAGEACCOUNTKEY= <storage_account_key>
LOCALFILENAME= <local_file_name>
CONTAINERNAME= <container_name>
BLOBNAME= <blob_name>

#download from blob
t1=time.time()
blob_service=BlobService(account_name=STORAGEACCOUNTNAME,account_key=STORAGEACCOUNTKEY)
blob_service.get_blob_to_path(CONTAINERNAME,BLOBNAME,LOCALFILENAME)
t2=time.time()
print(("It takes %s seconds to download "+blobname) % (t2 - t1))
```

2. Read the data into a Pandas data-frame from the downloaded file.

```
#LOCALFILE is the file path
dataframe_blobdata = pd.read_csv(LOCALFILE)
```

Now you are ready to explore the data and generate features on this dataset.

Data Exploration

Here are a few examples of ways to explore data using Pandas:

1. Inspect the number of rows and columns

```
print 'the size of the data is: %d rows and %d columns' % dataframe_blobdata.shape
```

2. Inspect the first or last few rows in the dataset as below:

```
dataframe_blobdata.head(10)

dataframe_blobdata.tail(10)
```

3. Check the data type each column was imported as using the following sample code

```
for col in dataframe_blobdata.columns:  
    print dataframe_blobdata[col].name, ':\t', dataframe_blobdata[col].dtype
```

4. Check the basic stats for the columns in the data set as follows

```
dataframe_blobdata.describe()
```

5. Look at the number of entries for each column value as follows

```
dataframe_blobdata['<column_name>'].value_counts()
```

6. Count missing values versus the actual number of entries in each column using the following sample code

```
miss_num = dataframe_blobdata.shape[0] - dataframe_blobdata.count()  
print miss_num
```

7. If you have missing values for a specific column in the data, you can drop them as follows:

```
dataframe_blobdata_noNA = dataframe_blobdata.dropna() dataframe_blobdata_noNA.shape
```

Another way to replace missing values is with the mode function:

```
dataframe_blobdata_mode =  
dataframe_blobdata.fillna({'<column_name>':dataframe_blobdata['<column_name>'].mode()[0]})
```

8. Create a histogram plot using variable number of bins to plot the distribution of a variable

```
dataframe_blobdata['<column_name>'].value_counts().plot(kind='bar')  
  
np.log(dataframe_blobdata['<column_name>']+1).hist(bins=50)
```

9. Look at correlations between variables using a scatterplot or using the built-in correlation function

```
#relationship between column_a and column_b using scatter plot  
plt.scatter(dataframe_blobdata['<column_a>'], dataframe_blobdata['<column_b>'])  
  
#correlation between column_a and column_b  
dataframe_blobdata[['<column_a>', '<column_b>']].corr()
```

Feature Generation

We can generate features using Python as follows:

Indicator value based Feature Generation

Categorical features can be created as follows:

1. Inspect the distribution of the categorical column:

```
dataframe_blobdata['<categorical_column>'].value_counts()
```

2. Generate indicator values for each of the column values

```
#generate the indicator column
dataframe_blobdata_identity = pd.get_dummies(dataframe_blobdata['<categorical_column>'],
prefix='<categorical_column>_identity')
```

3. Join the indicator column with the original data frame

```
#Join the dummy variables back to the original data frame
dataframe_blobdata_with_identity = dataframe_blobdata.join(dataframe_blobdata_identity)
```

4. Remove the original variable itself:

```
#Remove the original column rate_code in df1_with_dummy
dataframe_blobdata_with_identity.drop('<categorical_column>', axis=1, inplace=True)
```

Binning Feature Generation

For generating binned features, we proceed as follows:

1. Add a sequence of columns to bin a numeric column

```
bins = [0, 1, 2, 4, 10, 40]
dataframe_blobdata_bin_id = pd.cut(dataframe_blobdata['<numeric_column>'], bins)
```

2. Convert binning to a sequence of boolean variables

```
dataframe_blobdata_bin_bool = pd.get_dummies(dataframe_blobdata_bin_id, prefix='<numeric_column>')
```

3. Finally, Join the dummy variables back to the original data frame

```
dataframe_blobdata_with_bin_bool = dataframe_blobdata.join(dataframe_blobdata_bin_bool)
```

Writing data back to Azure blob and consuming in Azure Machine Learning

After you have explored the data and created the necessary features, you can upload the data (sampled or featurized) to an Azure blob and consume it in Azure Machine Learning using the following steps: Note that additional features can be created in the Azure Machine Learning Studio as well.

1. Write the data frame to local file

```
dataframe.to_csv(os.path.join(os.getcwd(), LOCALFILENAME), sep='\t', encoding='utf-8', index=False)
```

2. Upload the data to Azure blob as follows:

```

from azure.storage.blob import BlobService
import tables

STORAGEACCOUNTNAME= <storage_account_name>
LOCALFILENAME= <local_file_name>
STORAGEACCOUNTKEY= <storage_account_key>
CONTAINERNAME= <container_name>
BLOBNAME= <blob_name>

output_blob_service=BlobService(account_name=STORAGEACCOUNTNAME,account_key=STORAGEACCOUNTKEY)
localfileprocessed = os.path.join(os.getcwd(),LOCALFILENAME) #assuming file is in current working directory

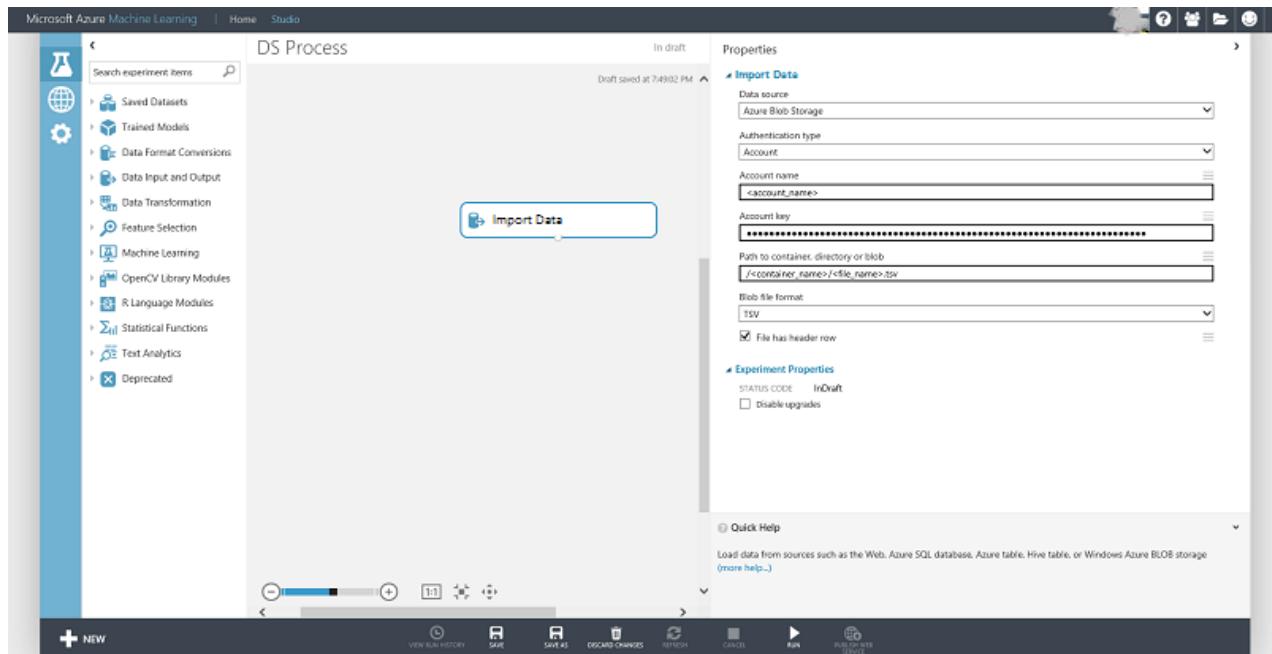
try:

#perform upload
output_blob_service.put_block_blob_from_path(CONTAINERNAME,BLOBNAME,localfileprocessed)

except:
    print ("Something went wrong with uploading blob:"+BLOBNAME)

```

3. Now the data can be read from the blob using the Azure Machine Learning [Import Data](#) module as shown in the screen below:



Scalable Data Science with Azure Data Lake: An end-to-end Walkthrough

3/14/2019 • 19 minutes to read

This walkthrough shows how to use Azure Data Lake to do data exploration and binary classification tasks on a sample of the NYC taxi trip and fare dataset to predict whether or not a tip is paid by a fare. It walks you through the steps of the [Team Data Science Process](#), end-to-end, from data acquisition to model training, and then to the deployment of a web service that publishes the model.

Azure Data Lake Analytics

The [Microsoft Azure Data Lake](#) has all the capabilities required to make it easy for data scientists to store data of any size, shape and speed, and to conduct data processing, advanced analytics, and machine learning modeling with high scalability in a cost-effective way. You pay on a per-job basis, only when data is actually being processed. Azure Data Lake Analytics includes U-SQL, a language that blends the declarative nature of SQL with the expressive power of C# to provide scalable distributed query capability. It enables you to process unstructured data by applying schema on read, insert custom logic and user-defined functions (UDFs), and includes extensibility to enable fine grained control over how to execute at scale. To learn more about the design philosophy behind U-SQL, see [Visual Studio blog post](#).

Data Lake Analytics is also a key part of Cortana Analytics Suite and works with Azure SQL Data Warehouse, Power BI, and Data Factory. This gives you a complete cloud big data and advanced analytics platform.

This walkthrough begins by describing how to install the prerequisites and resources that are needed to complete data science process tasks. Then it outlines the data processing steps using U-SQL and concludes by showing how to use Python and Hive with Azure Machine Learning Studio to build and deploy the predictive models.

U-SQL and Visual Studio

This walkthrough recommends using Visual Studio to edit U-SQL scripts to process the dataset. The U-SQL scripts are described here and provided in a separate file. The process includes ingesting, exploring, and sampling the data. It also shows how to run a U-SQL scripted job from the Azure portal. Hive tables are created for the data in an associated HDInsight cluster to facilitate the building and deployment of a binary classification model in Azure Machine Learning Studio.

Python

This walkthrough also contains a section that shows how to build and deploy a predictive model using Python with Azure Machine Learning Studio. It provides a Jupyter notebook with the Python scripts for the steps in this process. The notebook includes code for some additional feature engineering steps and models construction such as multiclass classification and regression modeling in addition to the binary classification model outlined here. The regression task is to predict the amount of the tip based on other tip features.

Azure Machine Learning

Azure Machine Learning Studio is used to build and deploy the predictive models. This is done using two approaches: first with Python scripts and then with Hive tables on an HDInsight (Hadoop) cluster.

Scripts

Only the principal steps are outlined in this walkthrough. You can download the full **U-SQL script** and **Jupyter Notebook** from [GitHub](#).

Prerequisites

Before you begin these topics, you must have the following:

- An Azure subscription. If you do not already have one, see [Get Azure free trial](#).
- [Recommended] Visual Studio 2013 or later. If you do not already have one of these versions installed, you can download a free Community version from [Visual Studio Community](#).

NOTE

Instead of Visual Studio, you can also use the Azure portal to submit Azure Data Lake queries. Instructions are provided on how to do so both with Visual Studio and on the portal in the section titled **Process data with U-SQL**.

Prepare data science environment for Azure Data Lake

To prepare the data science environment for this walkthrough, create the following resources:

- Azure Data Lake Store (ADLS)
- Azure Data Lake Analytics (ADLA)
- Azure Blob storage account
- Azure Machine Learning Studio account
- Azure Data Lake Tools for Visual Studio (Recommended)

This section provides instructions on how to create each of these resources. If you choose to use Hive tables with Azure Machine Learning, instead of Python, to build a model, you also need to provision an HDInsight (Hadoop) cluster. This alternative procedure is described in the Option 2 section.

NOTE

The **Azure Data Lake Store** can be created either separately or when you create the **Azure Data Lake Analytics** as the default storage. Instructions are referenced for creating each of these resources separately, but the Data Lake storage account need not be created separately.

Create an Azure Data Lake Store

Create an ADLS from the [Azure portal](#). For details, see [Create an HDInsight cluster with Data Lake Store using Azure portal](#). Be sure to set up the Cluster AAD Identity in the **DataSource** blade of the **Optional Configuration** blade described there.

The screenshot shows the Microsoft Azure portal interface. In the top left, there's a 'Preview UI' button, followed by the 'Microsoft Azure' logo and a dropdown menu. Below that is a navigation bar with 'New', 'Data + Storage', 'Report bug', and a search bar. On the left, a sidebar has a 'New' button highlighted with a red box. The main area is titled 'New' and 'Data + Storage'. Under 'MARKETPLACE', 'Data + Storage' is selected and highlighted with a red box. In the 'FEATURED APPS' section, 'Data Lake Store' is listed with its icon and description: 'Hyper-scale repository for big data analytic workloads'. A red box highlights this entry.

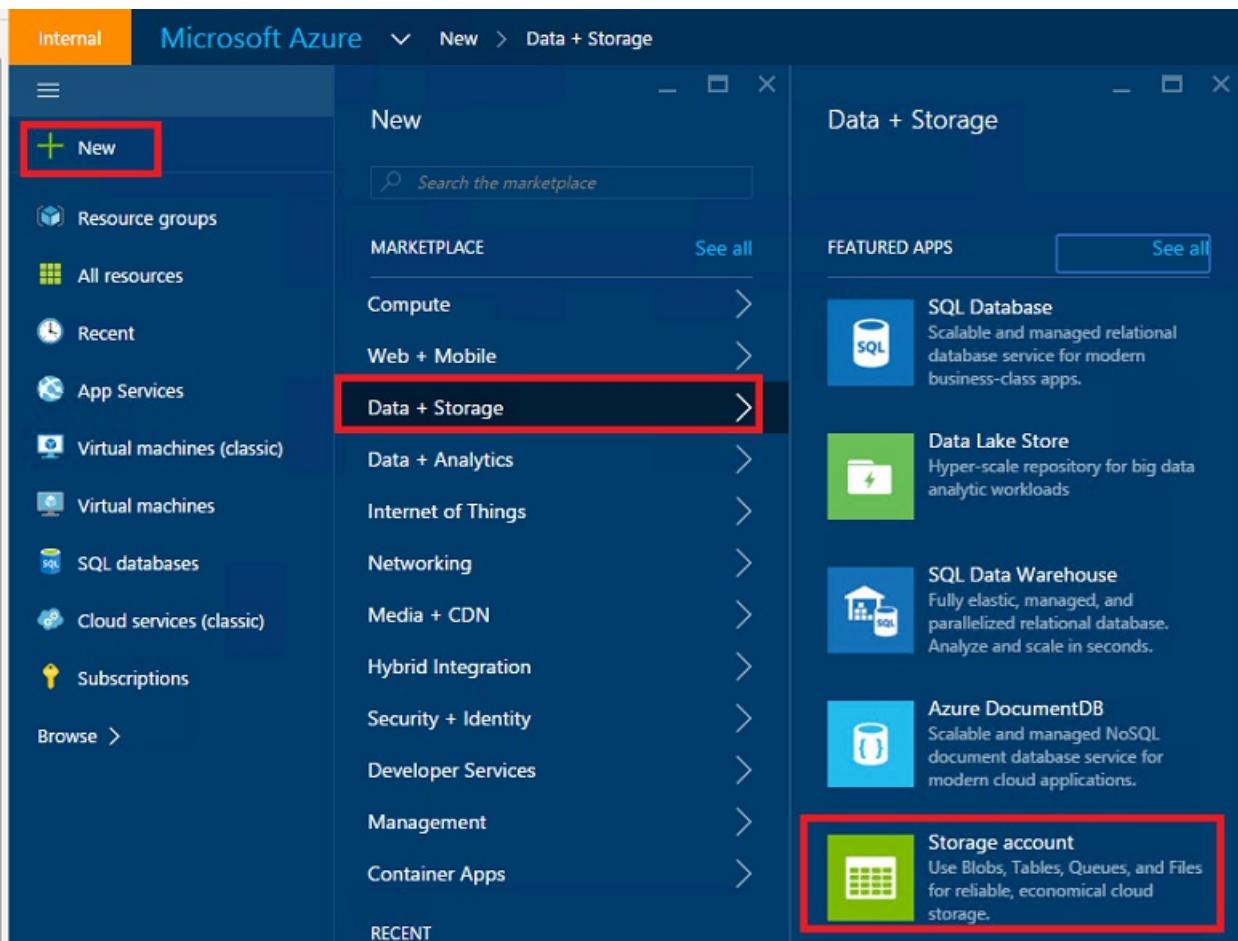
Create an Azure Data Lake Analytics account

Create an ADLA account from the [Azure portal](#). For details, see [Tutorial: get started with Azure Data Lake Analytics using Azure portal](#).

This screenshot is similar to the previous one but focuses on 'Data + Analytics'. The 'Data + Storage' item in the 'Data + Analytics' section of the marketplace is highlighted with a red box. In the 'Data + Analytics' featured apps section, 'Data Lake Analytics' is highlighted with a red box. Its description reads: 'Big data analytics made easy'.

Create an Azure Blob storage account

Create an Azure Blob storage account from the [Azure portal](#). For details, see the Create a storage account section in [About Azure storage accounts](#).



Set up an Azure Machine Learning Studio account

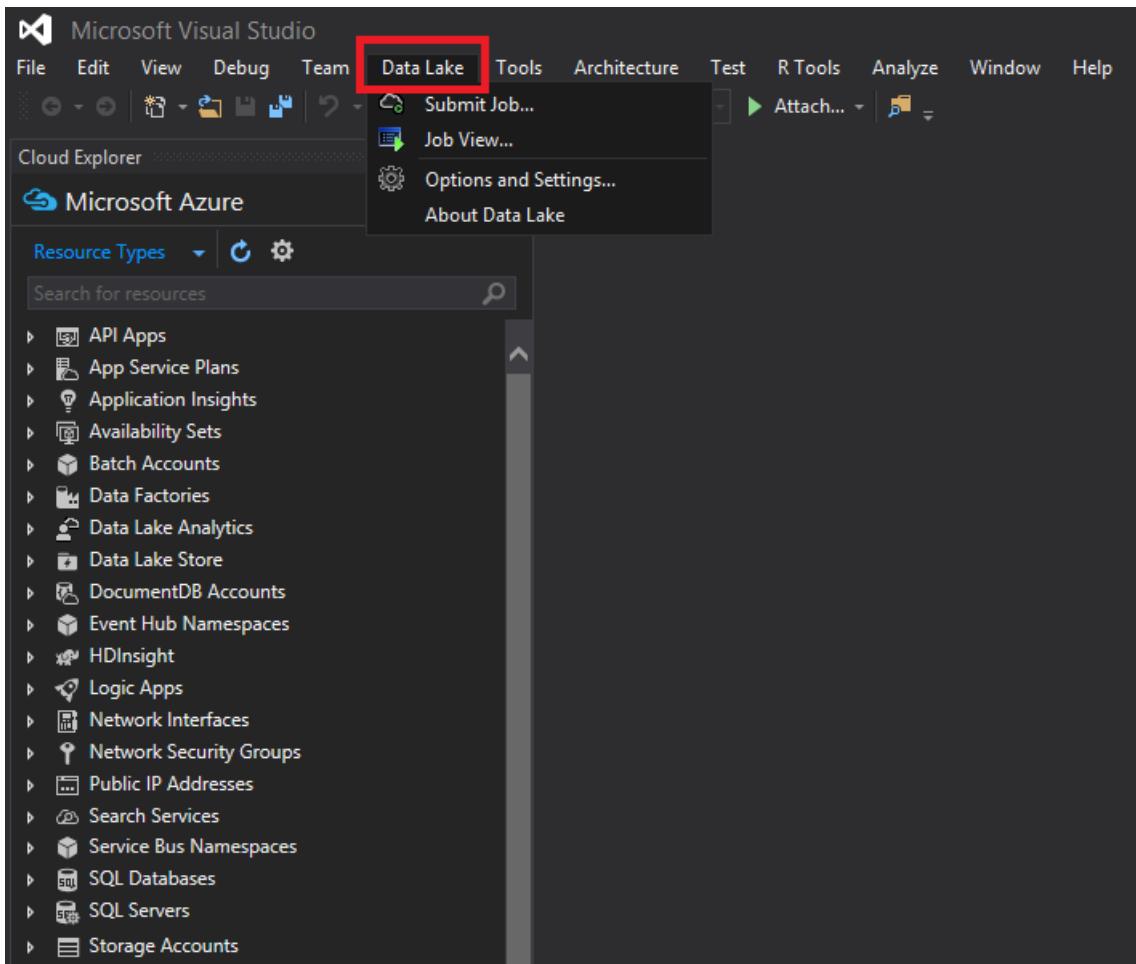
Sign up/into Azure Machine Learning Studio from the [Azure Machine Learning studio](#) page. Click on the **Get started now** button and then choose a "Free Workspace" or "Standard Workspace". Now you are ready to create experiments in Azure Machine Learning studio.

Install Azure Data Lake Tools [Recommended]

Install Azure Data Lake Tools for your version of Visual Studio from [Azure Data Lake Tools for Visual Studio](#).

This screenshot shows the download page for the Azure Data Lake Tools for Visual Studio. At the top, it says 'Azure Data Lake Tools for Visual Studio'. Below that, there's a language selection 'Language: English' and a large red 'Download' button, which is circled in blue. The page also includes a brief description: 'Plug-in for Azure Data Lake development using Visual Studio' and three expandable sections: 'Details', 'System Requirements', and 'Install Instructions'.

After the installation finishes successfully, open up Visual Studio. You should see the Data Lake tab the menu at the top. Your Azure resources should appear in the left panel when you sign into your Azure account.



The NYC Taxi Trips dataset

The data set used here is a publicly available dataset -- the [NYC Taxi Trips dataset](#). The NYC Taxi Trip data consists of about 20 GB of compressed CSV files (~48 GB uncompressed), recording more than 173 million individual trips and the fares paid for each trip. Each trip record includes the pickup and dropoff locations and times, anonymized hack (driver's) license number, and the medallion (taxi's unique ID) number. The data covers all trips in the year 2013 and is provided in the following two datasets for each month:

The 'trip_data' CSV contains trip details, such as number of passengers, pickup and dropoff points, trip duration, and trip length. Here are a few sample records:

```
medallion,hack_license,vendor_id,rate_code,store_and_fwd_flag,pickup_datetime,dropoff_datetime,passenger_count
,trip_time_in_secs,trip_distance,pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
89D227B655E5C82AECE13C3F540D4CF4,BA96DE419E711691B9445D6A6307C170,CMT,1,N,2013-01-01 15:11:48,2013-01-01
15:18:10,4,382,1.00,-73.978165,40.757977,-73.989838,40.751171
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,1,N,2013-01-06 00:18:35,2013-01-06
00:22:54,1,259,1.50,-74.006683,40.731781,-73.994499,40.75066
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,1,N,2013-01-05 18:49:41,2013-01-05
18:54:23,1,282,1.10,-74.004707,40.73777,-74.009834,40.726002
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:54:15,2013-01-07
23:58:20,2,244,.70,-73.974602,40.759945,-73.984734,40.759388
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:25:03,2013-01-07
23:34:24,1,560,2.10,-73.97625,40.748528,-74.002586,40.747868
```

The 'trip_fare' CSV contains details of the fare paid for each trip, such as payment type, fare amount, surcharge and taxes, tips and tolls, and the total amount paid. Here are a few sample records:

```

medallion, hack_license, vendor_id, pickup_datetime, payment_type, fare_amount, surcharge, mta_tax,
tip_amount, tolls_amount, total_amount
89D227B655E5C82AECF13C3F540D4CF4,BA96DE419E711691B9445D6A6307C170,CMT,2013-01-01
15:11:48,CSH,6.5,0,0.5,0,0,7
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,2013-01-06
00:18:35,CSH,6,0.5,0.5,0,0,7
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,2013-01-05
18:49:41,CSH,5.5,1,0.5,0,0,7
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,2013-01-07
23:54:15,CSH,5,0.5,0.5,0,0,6
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,2013-01-07
23:25:03,CSH,9.5,0.5,0.5,0,0,10.5

```

The unique key to join trip_data and trip_fare is composed of the following three fields: medallion, hack_license and pickup_datetime. The raw CSV files can be accessed from a public Azure storage blob. The U-SQL script for this join is in the [Join trip and fare tables](#) section.

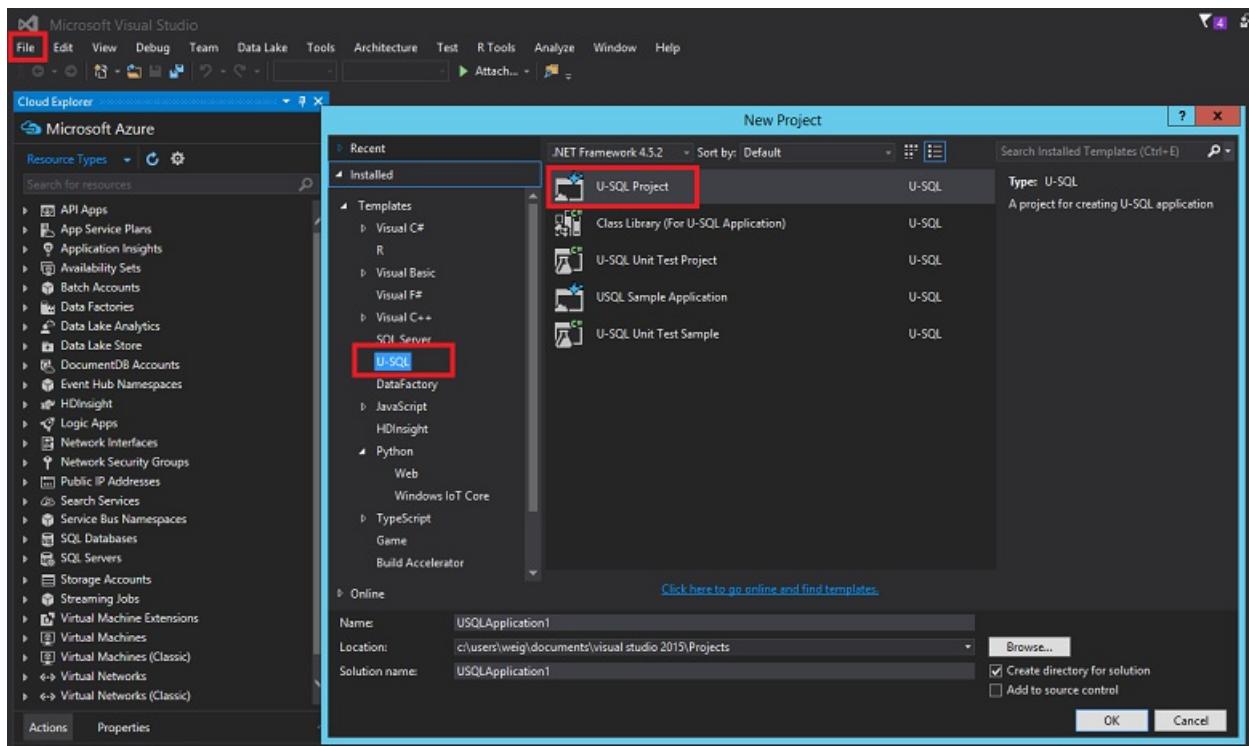
Process data with U-SQL

The data processing tasks illustrated in this section include ingesting, checking quality, exploring, and sampling the data. How to join trip and fare tables is also shown. The final section shows run a U-SQL scripted job from the Azure portal. Here are links to each subsection:

- [Data ingestion: read in data from public blob](#)
- [Data quality checks](#)
- [Data exploration](#)
- [Join trip and fare tables](#)
- [Data sampling](#)
- [Run U-SQL jobs](#)

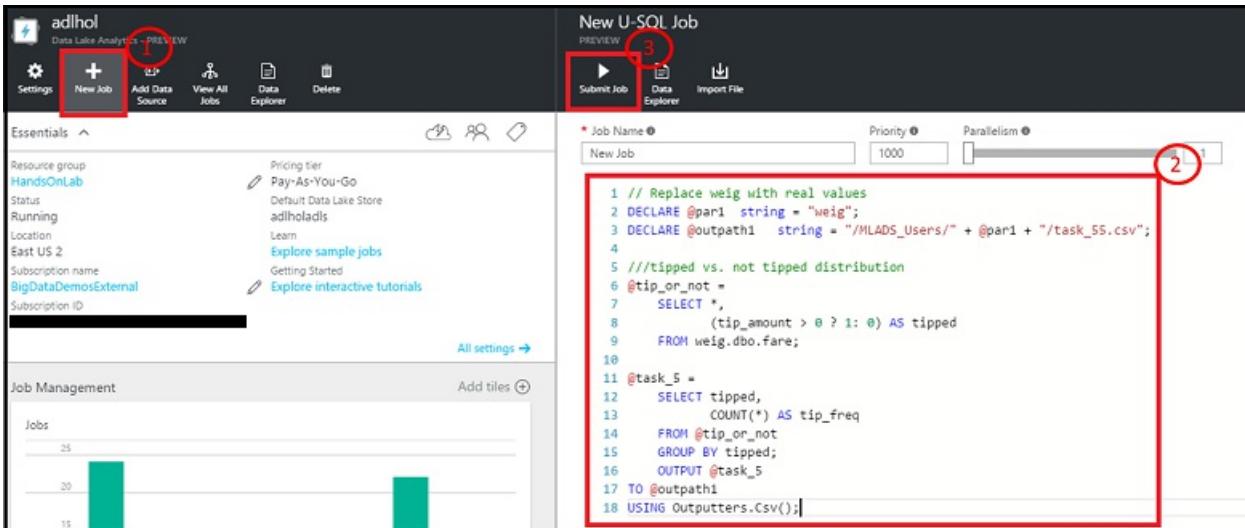
The U-SQL scripts are described here and provided in a separate file. You can download the full **U-SQL scripts** from [GitHub](#).

To execute U-SQL, Open Visual Studio, click **File --> New --> Project**, choose **U-SQL Project**, name and save it to a folder.



NOTE

It is possible to use the Azure Portal to execute U-SQL instead of Visual Studio. You can navigate to the Azure Data Lake Analytics resource on the portal and submit queries directly as illustrated in the following figure:



Data Ingestion: Read in data from public blob

The location of the data in the Azure blob is referenced as

wasb://container_name@blob_storage_account_name.blob.core.windows.net/blob_name and can be extracted using **Extractors.Csv()**. Substitute your own container name and storage account name in following scripts for *containername@blob_storage_account_name* in the wasb address. Since the file names are in same format, it is possible to use ****trip_data{*}.csv**** to read in all 12 trip files.

```
///Read in Trip data
@trip0 =
    EXTRACT
        medallion string,
        hack_license string,
        vendor_id string,
        rate_code string,
        store_and_fwd_flag string,
        pickup_datetime string,
        dropoff_datetime string,
        passenger_count string,
        trip_time_in_secs string,
        trip_distance string,
        pickup_longitude string,
        pickup_latitude string,
        dropoff_longitude string,
        dropoff_latitude string
    // This is reading 12 trip data from blob
    FROM "wasb://container_name@blob_storage_account_name.blob.core.windows.net/nyctaxitrip/trip_data_{*}.csv"
    USING Extractors.Csv();
```

Since there are headers in the first row, you need to remove the headers and change column types into appropriate ones. You can either save the processed data to Azure Data Lake Storage using **swebhdfs://data_lake_storage_name.azuredatastorage.net/folder_name/file_name** or to Azure Blob storage account using **wasb://container_name@blob_storage_account_name.blob.core.windows.net/blob_name**.

```

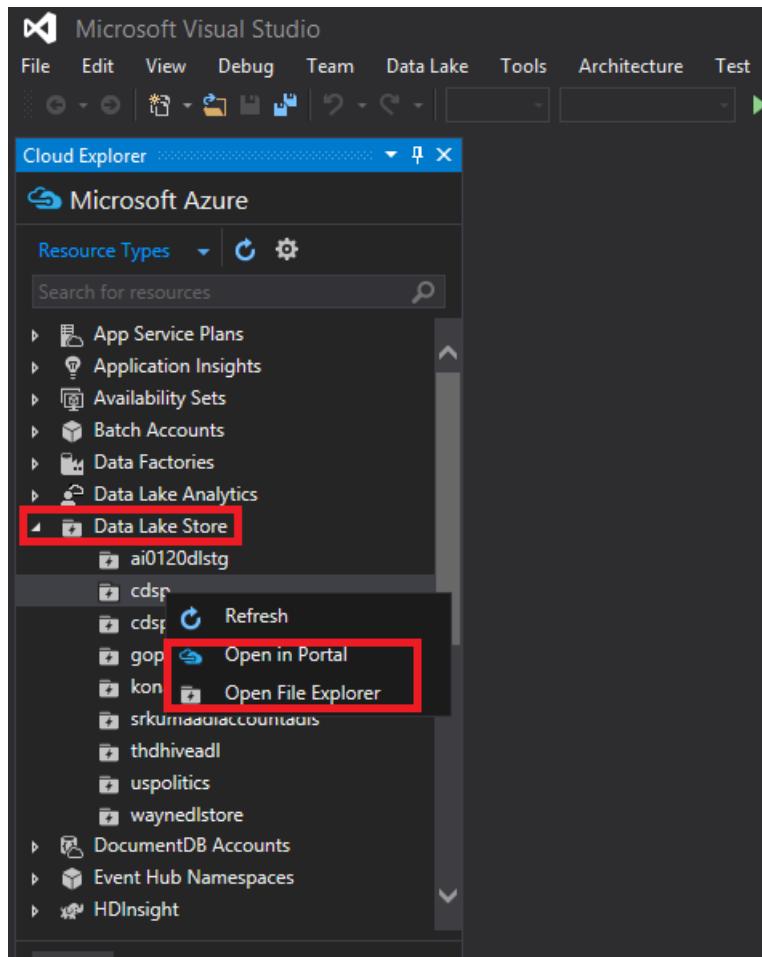
// change data types
@trip =
    SELECT
        medallion,
        hack_license,
        vendor_id,
        rate_code,
        store_and_fwd_flag,
        DateTime.Parse(pickup_datetime) AS pickup_datetime,
        DateTime.Parse(dropoff_datetime) AS dropoff_datetime,
        Int32.Parse(passenger_count) AS passenger_count,
        Double.Parse(trip_time_in_secs) AS trip_time_in_secs,
        Double.Parse(trip_distance) AS trip_distance,
        (pickup_longitude==string.Empty ? 0: float.Parse(pickup_longitude)) AS pickup_longitude,
        (pickup_latitude==string.Empty ? 0: float.Parse(pickup_latitude)) AS pickup_latitude,
        (dropoff_longitude==string.Empty ? 0: float.Parse(dropoff_longitude)) AS dropoff_longitude,
        (dropoff_latitude==string.Empty ? 0: float.Parse(dropoff_latitude)) AS dropoff_latitude
    FROM @trip0
    WHERE medallion != "medallion";

////output data to ADL
OUTPUT @trip
TO "swebhdfs://data_lake_storage_name.azuredatalakestore.net/nytaxi_folder/demo_trip.csv"
USING Outputters.Csv();

////Output data to blob
OUTPUT @trip
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_trip.csv"
USING Outputters.Csv();

```

Similarly you can read in the fare data sets. Right-click Azure Data Lake Store, you can choose to look at your data in **Azure portal --> Data Explorer** or **File Explorer** within Visual Studio.



NAME	SIZE	LAST MODIFIED
trip0test.csv		4/12/2016, 2:57:42 PM
demo_ex_7_full_data.csv	37.6 GB	4/13/2016, 7:35:22 AM
demo_ex_9_stratified_1_1000_copy.csv	39.1 MB	4/13/2016, 3:39:07 AM
demo_ex_9_stratified_1_1000.csv	39.1 MB	4/13/2016, 7:35:25 AM
demo_fare.csv	22.5 GB	4/13/2016, 7:35:17 AM
demo_trip.csv	33.2 GB	4/13/2016, 7:35:15 AM
trip0_test.csv	2.87 GB	4/12/2016, 2:51:48 PM

Data quality checks

After trip and fare tables have been read in, data quality checks can be done in the following way. The resulting CSV files can be output to Azure Blob storage or Azure Data Lake Store.

Find the number of medallions and unique number of medallions:

```
//check the number of medallions and unique number of medallions
@trip2 =
    SELECT
        medallion,
        vendor_id,
        pickup_datetime.Month AS pickup_month
    FROM @trip;

@ex_1 =
    SELECT
        pickup_month,
        COUNT(medallion) AS cnt_medallion,
        COUNT(DISTINCT(medallion)) AS unique_medallion
    FROM @trip2
    GROUP BY pickup_month;
    OUTPUT @ex_1
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_1.csv"
USING Outputters.Csv();
```

Find those medallions that had more than 100 trips:

```
//find those medallions that had more than 100 trips
@ex_2 =
    SELECT medallion,
        COUNT(medallion) AS cnt_medallion
    FROM @trip2
    //where pickup_datetime >= "2013-01-01t00:00:00.0000000" and pickup_datetime <= "2013-04-
01t00:00:00.0000000"
    GROUP BY medallion
    HAVING COUNT(medallion) > 100;
    OUTPUT @ex_2
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_2.csv"
USING Outputters.Csv();
```

Find those invalid records in terms of pickup_longitude:

```

///find those invalid records in terms of pickup_longitude
@ex_3 =
    SELECT COUNT(medallion) AS cnt_invalid_pickup_longitude
    FROM @trip
    WHERE
        pickup_longitude < -90 OR pickup_longitude > 90;
    OUTPUT @ex_3
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_3.csv"
USING Outputters.Csv();

```

Find missing values for some variables:

```

//check missing values
@res =
    SELECT *,
        (medallion == null? 1 : 0) AS missing_medallion
    FROM @trip;

@trip_summary6 =
    SELECT
        vendor_id,
        SUM(missing_medallion) AS medallion_empty,
        COUNT(medallion) AS medallion_total,
        COUNT(DISTINCT(medallion)) AS medallion_total_unique
    FROM @res
    GROUP BY vendor_id;
OUTPUT @trip_summary6
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_16.csv"
USING Outputters.Csv();

```

Data exploration

Do some data exploration with the following scripts to get a better understanding of the data.

Find the distribution of tipped and non-tipped trips:

```

///tipped vs. not tipped distribution
@tip_or_not =
    SELECT *,
        (tip_amount > 0 ? 1: 0) AS tipped
    FROM @fare;

@ex_4 =
    SELECT tipped,
        COUNT(*) AS tip_freq
    FROM @tip_or_not
    GROUP BY tipped;
    OUTPUT @ex_4
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_4.csv"
USING Outputters.Csv();

```

Find the distribution of tip amount with cut-off values: 0, 5, 10, and 20 dollars.

```

//tip class/range distribution
@tip_class =
    SELECT *,
        (tip_amount >20? 4: (tip_amount >10? 3:(tip_amount >5 ? 2:(tip_amount > 0 ? 1: 0)))) AS tip_class
    FROM @fare;
@ex_5 =
    SELECT tip_class,
        COUNT(*) AS tip_freq
    FROM @tip_class
    GROUP BY tip_class;
    OUTPUT @ex_5
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_5.csv"
USING Outputters.Csv();

```

Find basic statistics of trip distance:

```

// find basic statistics for trip_distance
@trip_summary4 =
    SELECT
        vendor_id,
        COUNT(*) AS cnt_row,
        MIN(trip_distance) AS min_trip_distance,
        MAX(trip_distance) AS max_trip_distance,
        AVG(trip_distance) AS avg_trip_distance
    FROM @trip
    GROUP BY vendor_id;
OUTPUT @trip_summary4
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_14.csv"
USING Outputters.Csv();

```

Find the percentiles of trip distance:

```

// find percentiles of trip_distance
@trip_summary3 =
    SELECT DISTINCT vendor_id AS vendor,
        PERCENTILE_DISC(0.25) WITHIN GROUP(ORDER BY trip_distance) OVER(PARTITION BY vendor_id) AS
median_trip_distance_disc,
        PERCENTILE_DISC(0.5) WITHIN GROUP(ORDER BY trip_distance) OVER(PARTITION BY vendor_id) AS
median_trip_distance_disc,
        PERCENTILE_DISC(0.75) WITHIN GROUP(ORDER BY trip_distance) OVER(PARTITION BY vendor_id) AS
median_trip_distance_disc
    FROM @trip;
    // group by vendor_id;
OUTPUT @trip_summary3
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_13.csv"
USING Outputters.Csv();

```

Join trip and fare tables

Trip and fare tables can be joined by medallion, hack_license, and pickup_time.

```

//join trip and fare table

@model_data_full =
SELECT t.*,
f.payment_type, f.fare_amount, f.surcharge, f.mta_tax, f.tolls_amount, f.total_amount, f.tip_amount,
(f.tip_amount > 0 ? 1: 0) AS tipped,
(f.tip_amount >20? 4: (f.tip_amount >10? 3:(f.tip_amount >5 ? 2:(f.tip_amount > 0 ? 1: 0)))) AS tip_class
FROM @trip AS t JOIN  @fare AS f
ON  (t.medallion == f.medallion AND t.hack_license == f.hack_license AND t.pickup_datetime ==
f.pickup_datetime)
WHERE  (pickup_longitude != 0 AND dropoff_longitude != 0 );

//// output to blob
OUTPUT @model_data_full
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_7_full_data.csv"
USING Outputters.Csv();

////output data to ADL
OUTPUT @model_data_full
TO "swebhdfs://data_lake_storage_name.azuredatalakestore.net/nyctaxi_folder/demo_ex_7_full_data.csv"
USING Outputters.Csv();

```

For each level of passenger count, calculate the number of records, average tip amount, variance of tip amount, percentage of tipped trips.

```

// contingency table
@trip_summary8 =
SELECT passenger_count,
COUNT(*) AS cnt,
AVG(tip_amount) AS avg_tip_amount,
VAR(tip_amount) AS var_tip_amount,
SUM(tipped) AS cnt_tipped,
(float)SUM(tipped)/COUNT(*) AS pct_tipped
FROM @model_data_full
GROUP BY passenger_count;
OUTPUT @trip_summary8
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_17.csv"
USING Outputters.Csv();

```

Data sampling

First, randomly select 0.1% of the data from the joined table:

```

//random select 1/1000 data for modeling purpose
@addrownumberres_randomsample =
SELECT *,
ROW_NUMBER() OVER() AS rounum
FROM @model_data_full;

@model_data_random_sample_1_1000 =
SELECT *
FROM @addrownumberres_randomsample
WHERE rounum % 1000 == 0;

OUTPUT @model_data_random_sample_1_1000
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_7_random_1_1000.csv"
USING Outputters.Csv();

```

Then do stratified sampling by binary variable tip_class:

```

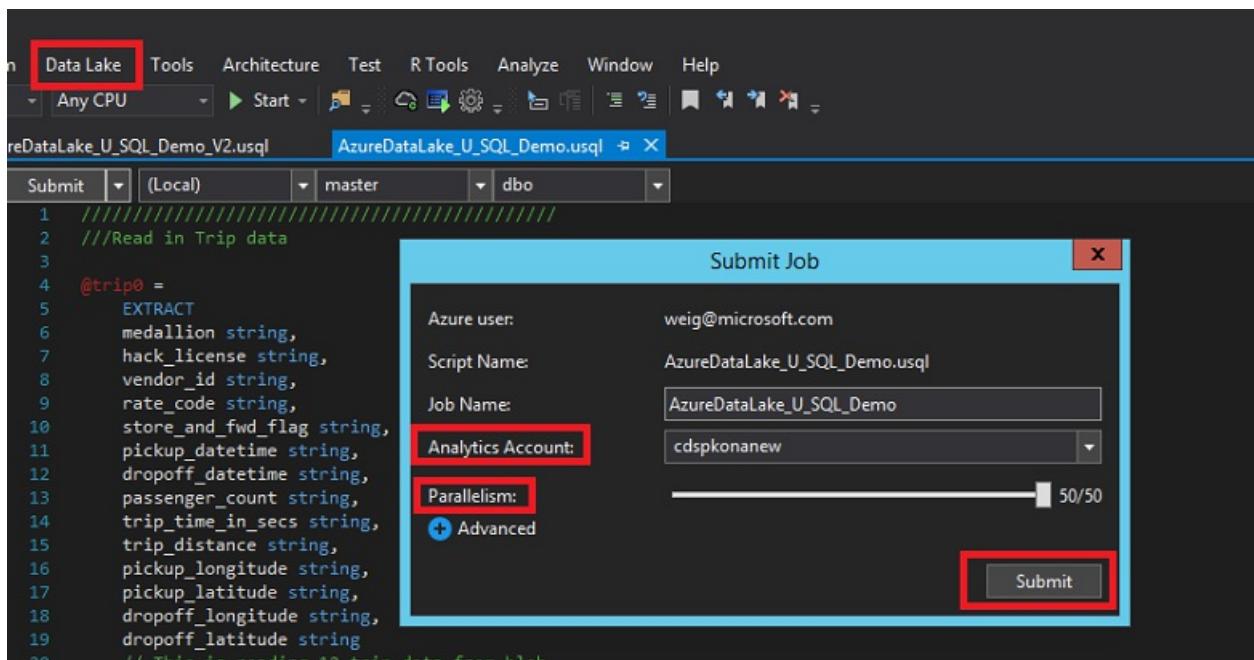
//stratified random select 1/1000 data for modeling purpose
@addrownumberres_stratifiedsample =
SELECT *,
       ROW_NUMBER() OVER(PARTITION BY tip_class) AS rounum
FROM @model_data_full;

@model_data_stratified_sample_1_1000 =
SELECT *
FROM @addrownumberres_stratifiedsample
WHERE rounum % 1000 == 0;
//// output to blob
OUTPUT @model_data_stratified_sample_1_1000
TO "wasb://container_name@blob_storage_account_name.blob.core.windows.net/demo_ex_9_stratified_1_1000.csv"
USING Outputters.Csv();
////output data to ADL
OUTPUT @model_data_stratified_sample_1_1000
TO "swebhdfs://data_lake_storage_name.azuredatalakestore.net/nytaxi_folder/demo_ex_9_stratified_1_1000.csv"
USING Outputters.Csv();

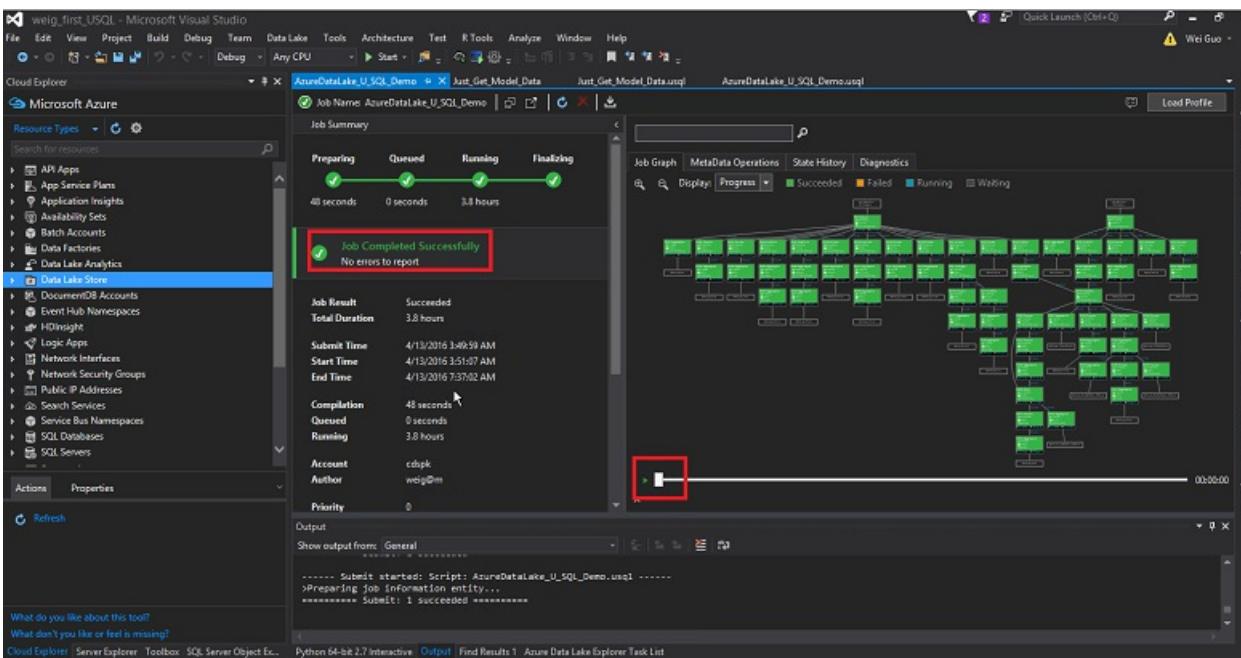
```

Run U-SQL jobs

When you finish editing U-SQL scripts, you can submit them to the server using your Azure Data Lake Analytics account. Click **Data Lake**, **Submit Job**, select your **Analytics Account**, choose **Parallelism**, and click **Submit** button.



When the job is complied successfully, the status of your job is displayed in Visual Studio for monitoring. After the job finishes running, you can even replay the job execution process and find out the bottleneck steps to improve your job efficiency. You can also go to Azure portal to check the status of your U-SQL jobs.



This screenshot shows the Azure Data Lake Analytics portal. On the left, the 'cdspk' resource group details are displayed, including the 'View All Jobs' button which is highlighted with a red box. On the right, the 'All Jobs' list shows a total of 238 jobs. One job, 'AzureDataLake_U_SQL_Demo', is highlighted with a red box and marked as 'Succeeded'. The table includes columns for Status, Job Name, and Language (U-SQL).

Status	Job Name	Language
Succeeded	AzureDataLake_U_SQL_Demo	U-SQL
		U-SQL
		U-SQL
		U-SQL

Now you can check the output files in either Azure Blob storage or Azure portal. Use the stratified sample data for our modeling in the next step.

This screenshot shows the Azure Storage Explorer interface. The left sidebar lists 'Storage Account' (weigstoragefordsvm) and 'Blob Containers' (7), with 'test1' selected. The main pane displays the contents of the 'test1' blob container, which contains 98 blobs (188.94G) as of 4/27/2016 12:20:15 AM. A large table lists the blobs, all of which are CSV files named 'demo_ex_1.csv' through 'demo_trip.csv'. The table includes columns for Name, Type, Last Modified, Length, and Content Type.

Name	Type	Last Modified	Length	Content Type
demo_ex_1.csv	Block	4/13/2016 2:35:23 PM +0:00	219 bytes	text/plain; charset=utf-8
demo_ex_10.csv	Block	4/13/2016 2:35:26 PM +0:00	40 bytes	text/plain; charset=utf-8
demo_ex_11.csv	Block	4/13/2016 2:35:18 PM +0:00	32 bytes	text/plain; charset=utf-8
demo_ex_12.csv	Block	4/13/2016 2:35:18 PM +0:00	118 bytes	text/plain; charset=utf-8
demo_ex_13.csv	Block	4/13/2016 2:35:20 PM +0:00	30 bytes	text/plain; charset=utf-8
demo_ex_14.csv	Block	4/13/2016 2:35:19 PM +0:00	89 bytes	text/plain; charset=utf-8
demo_ex_15.csv	Block	4/13/2016 2:35:20 PM +0:00	22 bytes	text/plain; charset=utf-8
demo_ex_16.csv	Block	4/13/2016 2:35:20 PM +0:00	46 bytes	text/plain; charset=utf-8
demo_ex_17.csv	Block	4/13/2016 2:35:24 PM +0:00	695 bytes	text/plain; charset=utf-8
demo_ex_2.csv	Block	4/13/2016 2:35:21 PM +0:00	550.96K	text/plain; charset=utf-8
demo_ex_3.csv	Block	4/13/2016 2:35:19 PM +0:00	6 bytes	text/plain; charset=utf-8
demo_ex_4.csv	Block	4/13/2016 2:35:19 PM +0:00	24 bytes	text/plain; charset=utf-8
demo_ex_5.csv	Block	4/13/2016 2:35:19 PM +0:00	55 bytes	text/plain; charset=utf-8
demo_ex_6.csv	Block	4/13/2016 2:35:19 PM +0:00	187.02K	text/plain; charset=utf-8
demo_ex_7_full_data.csv	Block	4/13/2016 2:35:22 PM +0:00	35.05G	text/plain; charset=utf-8
demo_ex_7_random_1_1000.csv	Block	4/13/2016 2:35:26 PM +0:00	37.41M	text/plain; charset=utf-8
demo_ex_8.csv	Block	4/13/2016 2:35:27 PM +0:00	40 bytes	text/plain; charset=utf-8
demo_ex_9_stratified_1_1000.csv	Block	4/13/2016 2:35:26 PM +0:00	37.32M	text/plain; charset=utf-8
demo_ex_9_stratified_1_1000_copy.csv	Block	4/13/2016 10:39:08 AM +0:00	37.32M	text/plain; charset=utf-8
demo_fare.csv	Block	4/13/2016 2:35:18 PM +0:00	20.97G	text/plain; charset=utf-8
demo_trip.csv	Block	4/13/2016 2:35:16 PM +0:00	30.88G	text/plain; charset=utf-8

The screenshot shows the Azure Data Lake Store - PREVIEW interface. At the top, there's a navigation bar with icons for Filter, New Folder, Upload, Access, Rename Folder, Folder Properties, Delete Folder, and Refresh. Below the navigation bar, the path 'cdsp > nyctaxi_weig' is displayed. A pencil icon is on the right of the path. The main area is a table listing files:

NAME	SIZE	LAST MODIFIED
trip0test.csv		4/12/2016, 2:57:42 PM
demo_ex_7_full_data.csv	37.6 GB	4/13/2016, 7:35:22 AM
demo_ex_9_stratified_1_1000_copy.csv	39.1 MB	4/13/2016, 3:39:07 AM
demo_ex_9_stratified_1_1000.csv	39.1 MB	4/13/2016, 7:35:25 AM
demo_fare.csv	22.5 GB	4/13/2016, 7:35:17 AM
demo_trip.csv	33.2 GB	4/13/2016, 7:35:15 AM
trip0_test.csv	2.87 GB	4/12/2016, 2:51:48 PM

Build and deploy models in Azure Machine Learning

Two options are available for you to pull data into Azure Machine Learning to build and

- In the first option, you use the sampled data that has been written to an Azure Blob (in the **Data sampling** step above) and use Python to build and deploy models from Azure Machine Learning.
- In the second option, you query the data in Azure Data Lake directly using a Hive query. This option requires that you create a new HDInsight cluster or use an existing HDInsight cluster where the Hive tables point to the NY Taxi data in Azure Data Lake Storage. Both these options are discussed in the following sections.

Option 1: Use Python to build and deploy machine learning models

To build and deploy machine learning models using Python, create a Jupyter Notebook on your local machine or in Azure Machine Learning Studio. The Jupyter Notebook provided on [GitHub](#) contains the full code to explore, visualize data, feature engineering, modeling and deployment. In this article, just the modeling and deployment are covered.

Import Python libraries

In order to run the sample Jupyter Notebook or the Python script file, the following Python packages are needed. If you are using the Azure Machine Learning Notebook service, these packages have been pre-installed.

```

import pandas as pd
from pandas import Series, DataFrame
import numpy as np
import matplotlib.pyplot as plt
from time import time
import pyodbc
import os
from azure.storage.blob import BlobService
import tables
import time
import zipfile
import random
import sklearn
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import train_test_split
from sklearn import metrics
from __future__ import division
from sklearn import linear_model
from azureml import services

```

Read in the data from blob

- Connection String

```

CONTAINERNAME = 'test1'
STORAGEACCOUNTNAME = 'XXXXXXXXXX'
STORAGEACCOUNTKEY = 'YYYYYYYYYYYYYYYYYYYYYYYYYYYY'
BLOBNAME = 'demo_ex_9_stratified_1_1000_copy.csv'
blob_service = BlobService(account_name=STORAGEACCOUNTNAME,account_key=STORAGEACCOUNTKEY)

```

- Read in as text

```

t1 = time.time()
data = blob_service.get_blob_to_text(CONTAINERNAME,BLOBNAME).split("\n")
t2 = time.time()
print(("It takes %s seconds to read in "+BLOBNAME) % (t2 - t1))

```

It takes 1.61118912697 seconds to read in demo_ex_9_stratified_1_1000_copy.csv

- Add column names and separate columns

```

colnames =
['medallion','hack_license','vendor_id','rate_code','store_and_fwd_flag','pickup_datetime','dropoff_datetime',
'passenger_count','trip_time_in_secs','trip_distance','pickup_longitude','pickup_latitude','dropoff_longitude',
'dropoff_latitude',
'payment_type', 'fare_amount', 'surcharge', 'mta_tax', 'tolls_amount', 'total_amount', 'tip_amount',
'tipped', 'tip_class', 'rownum']
df1 = pd.DataFrame([sub.split(",") for sub in data], columns = colnames)

```

- Change some columns to numeric

```

cols_2_float =
['trip_time_in_secs','pickup_longitude','pickup_latitude','dropoff_longitude','dropoff_latitude',
'fare_amount', 'surcharge','mta_tax','tolls_amount','total_amount','tip_amount',
'passenger_count','trip_distance'
,'tipped','tip_class','rownum']
for col in cols_2_float:
    df1[col] = df1[col].astype(float)

```

Build machine learning models

Here you build a binary classification model to predict whether a trip is tipped or not. In the Jupyter Notebook you can find other two models: multiclass classification, and regression models.

- First you need to create dummy variables that can be used in scikit-learn models

```

df1_payment_type_dummy = pd.get_dummies(df1['payment_type'], prefix='payment_type_dummy')
df1_vendor_id_dummy = pd.get_dummies(df1['vendor_id'], prefix='vendor_id_dummy')

```

- Create data frame for the modeling

```

cols_to_keep = ['tipped', 'trip_distance', 'passenger_count']
data = df1[cols_to_keep].join([df1_payment_type_dummy,df1_vendor_id_dummy])

X = data.iloc[:,1:]
Y = data.tipped

```

- Training and testing 60-40 split

```

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.4, random_state=0)

```

- Logistic Regression in training set

```

model = LogisticRegression()
logit_fit = model.fit(X_train, Y_train)
print ('Coefficients: \n', logit_fit.coef_)
Y_train_pred = logit_fit.predict(X_train)

![c1](./media/data-lake-walkthrough/c1-py-logit-coefficient.PNG)

```

- Score testing data set

```

Y_test_pred = logit_fit.predict(X_test)

```

- Calculate Evaluation metrics

```

fpr_train, tpr_train, thresholds_train = metrics.roc_curve(Y_train, Y_train_pred)
print fpr_train, tpr_train, thresholds_train

fpr_test, tpr_test, thresholds_test = metrics.roc_curve(Y_test, Y_test_pred)
print fpr_test, tpr_test, thresholds_test

#AUC
print metrics.auc(fpr_train,tpr_train)
print metrics.auc(fpr_test,tpr_test)

#Confusion Matrix
print metrics.confusion_matrix(Y_train,Y_train_pred)
print metrics.confusion_matrix(Y_test,Y_test_pred)

![c2](./media/data-lake-walkthrough/c2-py-logit-evaluation.PNG)

```

Build Web Service API and consume it in Python

You want to operationalize the machine learning model after it has been built. The binary logistic model is used here as an example. Make sure the scikit-learn version in your local machine is 0.15.1. You don't have to worry about this if you use Azure Machine Learning studio.

- Find your workspace credentials from Azure Machine Learning studio settings. In Azure Machine Learning Studio, click **Settings** --> **Name** --> **Authorization Tokens**.

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a sidebar with icons for Projects, Experiments, Web Services, Notebooks, Datasets, Trained Models, and Settings. The Settings icon is highlighted with a red box. The main area is titled 'settings' and shows workspace details. The 'NAME' tab is selected, showing 'WORKSPACE NAME' as 'Wei-WorkSpace-Azure-ML'. The 'AUTHORIZATION TOKENS' tab is also visible. The 'SETTINGS' tab is highlighted with a red box. Other visible fields include 'WORKSPACE DESCRIPTION' (empty), 'WORKSPACE TYPE' (Standard), 'WORKSPACE ID' (highlighted with a red box), 'CREATION TIME' (4/3/2015, 1:33:43 PM), 'OWNER'S EMAIL' (outlook.com), 'SUBSCRIPTION ID' (e8:), and 'STORAGE ACCOUNT' (account).

```

workspaceid = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxx'
auth_token = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxx'

```

- Create Web Service

```

@services.publish(workspaceid, auth_token)
@services.types(trip_distance = float, passenger_count = float, payment_type_dummy_CRD = float,
payment_type_dummy_CSH=float, payment_type_dummy_DIS = float, payment_type_dummy_NOC = float,
payment_type_dummy_UNK = float, vendor_id_dummy_CMT = float, vendor_id_dummy_VTS = float)
@services.returns(int) #0, or 1
def predictNYCTAXI(trip_distance, passenger_count, payment_type_dummy_CRD,
payment_type_dummy_CSH,payment_type_dummy_DIS, payment_type_dummy_NOC, payment_type_dummy_UNK,
vendor_id_dummy_CMT, vendor_id_dummy_VTS ):
    inputArray = [trip_distance, passenger_count, payment_type_dummy_CRD, payment_type_dummy_CSH,
payment_type_dummy_DIS, payment_type_dummy_NOC, payment_type_dummy_UNK, vendor_id_dummy_CMT,
vendor_id_dummy_VTS]
    return logit_fit.predict(inputArray)

```

- Get web service credentials

```

url = predictNYCTAXI.service.url
api_key = predictNYCTAXI.service.api_key

print url
print api_key

@services.service(url, api_key)
@services.types(trip_distance = float, passenger_count = float, payment_type_dummy_CRD = float,
payment_type_dummy_CSH=float,payment_type_dummy_DIS = float, payment_type_dummy_NOC = float,
payment_type_dummy_UNK = float, vendor_id_dummy_CMT = float, vendor_id_dummy_VTS = float)
@services.returns(float)
def NYCTAXIPredictor(trip_distance, passenger_count, payment_type_dummy_CRD,
payment_type_dummy_CSH,payment_type_dummy_DIS, payment_type_dummy_NOC, payment_type_dummy_UNK,
vendor_id_dummy_CMT, vendor_id_dummy_VTS ):
    pass

```

- Call Web service API. You have to wait 5-10 seconds after the previous step.

```

NYCTAXIPredictor(1,2,1,0,0,0,0,0,1)

![c4](./media/data-lake-walkthrough/c4-call-API.PNG)

```

Option 2: Create and deploy models directly in Azure Machine Learning

Azure Machine Learning Studio can read data directly from Azure Data Lake Store and then be used to create and deploy models. This approach uses a Hive table that points at the Azure Data Lake Store. This requires that a separate Azure HDInsight cluster be provisioned, on which the Hive table is created. The following sections show how to do this.

Create an HDInsight Linux Cluster

Create an HDInsight Cluster (Linux) from the [Azure portal](#). For details, see the **Create an HDInsight cluster with access to Azure Data Lake Store** section in [Create an HDInsight cluster with Data Lake Store using Azure portal](#).

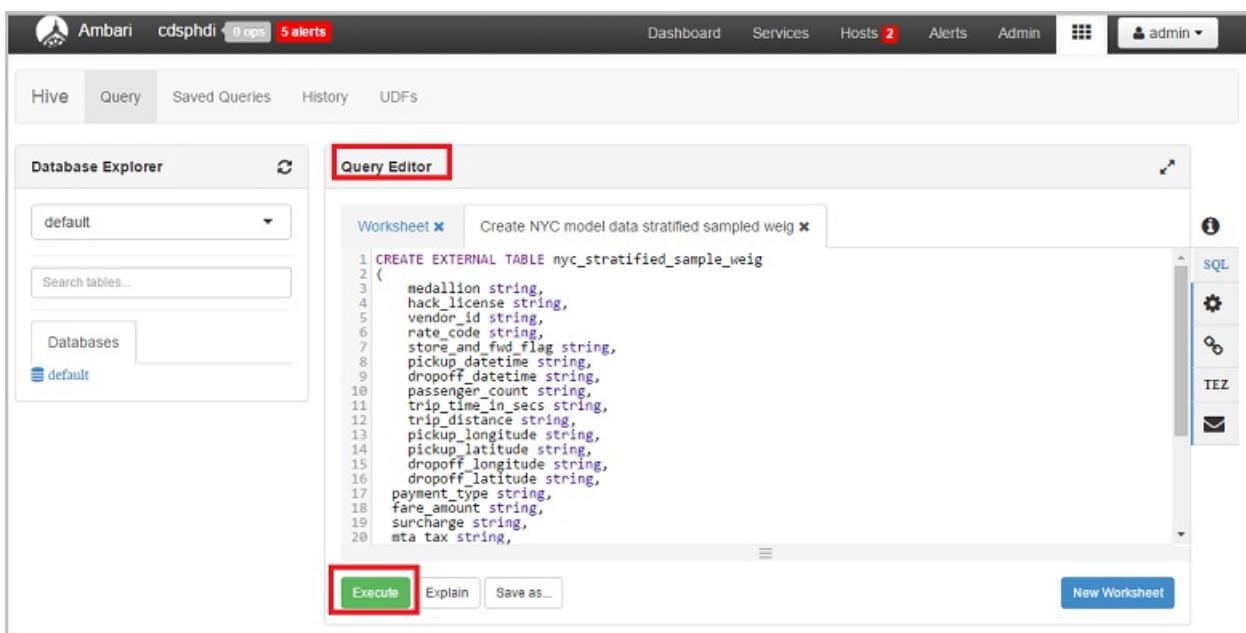
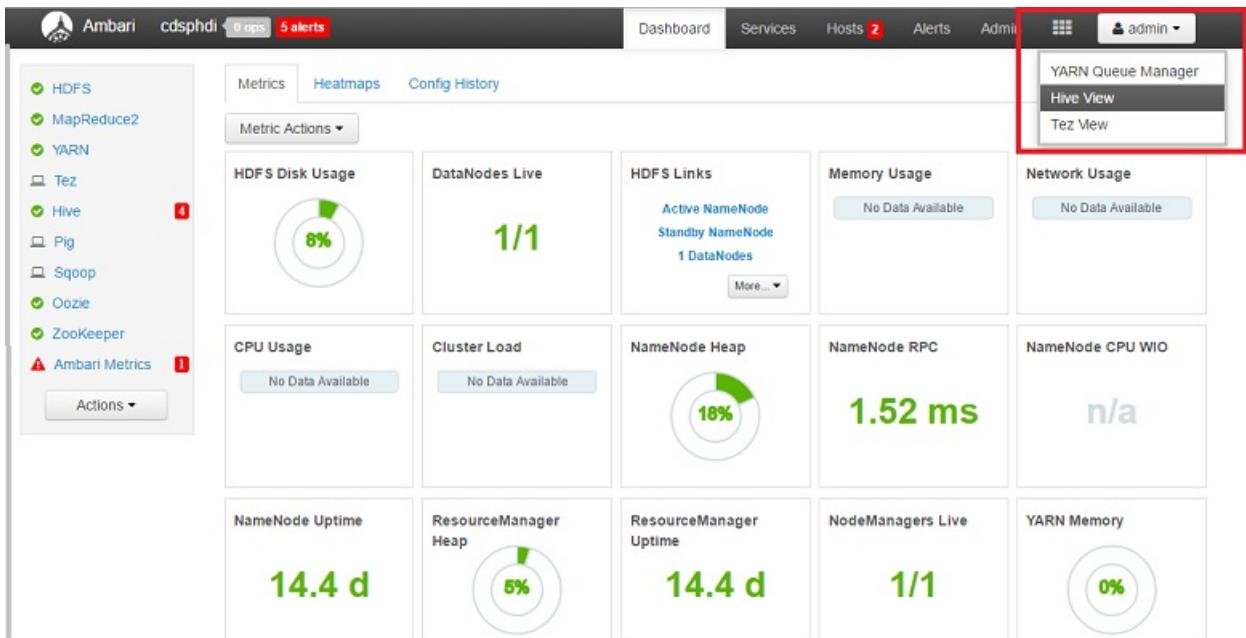
The screenshot shows the Microsoft Azure portal's 'New' blade. The 'Data + Analytics' category is selected and highlighted with a red box. On the right, the 'Data + Analytics' section is displayed with several service cards. The 'HDInsight' service card is also highlighted with a red box.

Create Hive table in HDInsight

Now you create Hive tables to be used in Azure Machine Learning Studio in the HDInsight cluster using the data stored in Azure Data Lake Store in the previous step. Go to the HDInsight cluster created. Click **Settings** --> **Properties** --> **Cluster AAD Identity** --> **ADLS Access**, make sure your Azure Data Lake Store account is added in the list with read, write and execute rights.

The screenshot shows the Azure HDInsight cluster dashboard. The 'Settings' blade is open, specifically the 'Cluster AAD Identity' section. The 'ADLS Access' tab is selected and highlighted with a red box. In the 'Data Lake Store Root Folder Access' section, a list of ADLS accounts is shown with their respective permission levels (READ, WRITE, EXECUTE). One account, 'adls', is highlighted with a red box and has all three checkboxes checked.

Then click **Dashboard** next to the **Settings** button and a window pops up. Click **Hive View** in the upper right corner of the page and you should see the **Query Editor**.



Paste in the following Hive scripts to create a table. The location of data source is in Azure Data Lake Store reference in this way: **adl://data_lake_store_name.azuredatalakestore.net:443/folder_name/file_name**.

```

CREATE EXTERNAL TABLE nyc_stratified_sample
(
    medallion string,
    hack_license string,
    vendor_id string,
    rate_code string,
    store_and_fwd_flag string,
    pickup_datetime string,
    dropoff_datetime string,
    passenger_count string,
    trip_time_in_secs string,
    trip_distance string,
    pickup_longitude string,
    pickup_latitude string,
    dropoff_longitude string,
    dropoff_latitude string,
    payment_type string,
    fare_amount string,
    surcharge string,
    mta_tax string,
    tolls_amount string,
    total_amount string,
    tip_amount string,
    tipped string,
    tip_class string,
    rownum string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' lines terminated by '\n'
LOCATION
'adl://data_lake_storage_name.azuredatalakestore.net:443/nyctaxi_folder/demo_ex_9_stratified_1_1000_copy.csv';

```

When the query finishes running, you should see the results like this:

Query Process Results (Status: Succeeded)			Save results... ▾
Logs	Results		
Filter columns...		previous	next
nyc_stratified_sample_weig.medallion	nyc_stratified_sample_weig.hack_license	nyc_stratified_sample_weig.vendor_id	nyc_stratified_sample_weig.rate_code
"0053334C798EC6C8E637657962030F99"	"9EF690115D60940E7A8039A67542642E"	"VTS"	"1"
"00B99071EE4DC8266384113B91E6AC13"	"081B28EDA7F73E5E2C97573F0DBAC25D"	"CMT"	"0"
"011E4CBA1987A8553F8EA5681FFBF7F1"	"F53B26859F6C46C24E39EEAEE8096268"	"CMT"	"1"
"1109955CCAABC BCE1A22BCED5F1DBFF5"	"EAA69F43239E5C6609CD8FF4D6725836"	"CMT"	"0"
"0A415B814A6479EE706972D2FD6DA08A"	"12A32F655B8C9CE3AC7872477A641974"	"VTS"	"1"
"02B29197FB7470B583ED12167D19E998"	"C1EEBC8298A619F86637D7E97F6BDD5C"	"CMT"	"0"
"03055B956C21B1F915DCDB118AA79F21"	"E0C7A67293DE535DB04F8AFD8BF28F73"	"VTS"	"1"
"037673EEAE0DCB912D06BED04E89D89D"	"3B247BD1230E95A0D9FED3E47FECB8D9"	"CMT"	"0"
"03BF54085C92C385889B957D804780D1"	"776D5633A2041CEBDE03092E401D61DB"	"CMT"	"1"
"0437660BC3704C2F185301D539434A64"	"AE9AB8C79A2A0EB6DDA76B7024FF6029"	"CMT"	"0"

Build and deploy models in Azure Machine Learning Studio

You are now ready to build and deploy a model that predicts whether or not a tip is paid with Azure Machine Learning. The stratified sample data is ready to be used in this binary classification (tip or not) problem. The

predictive models using multiclass classification (tip_class) and regression (tip_amount) can also be built and deployed with Azure Machine Learning Studio, but here it is only shown how to handle the case using the binary classification model.

1. Get the data into Azure Machine Learning studio using the **Import Data** module, available in the **Data Input and Output** section. For more information, see the [Import Data module](#) reference page.
2. Select **Hive Query** as the **Data source** in the **Properties** panel.
3. Paste the following Hive script in the **Hive database query** editor

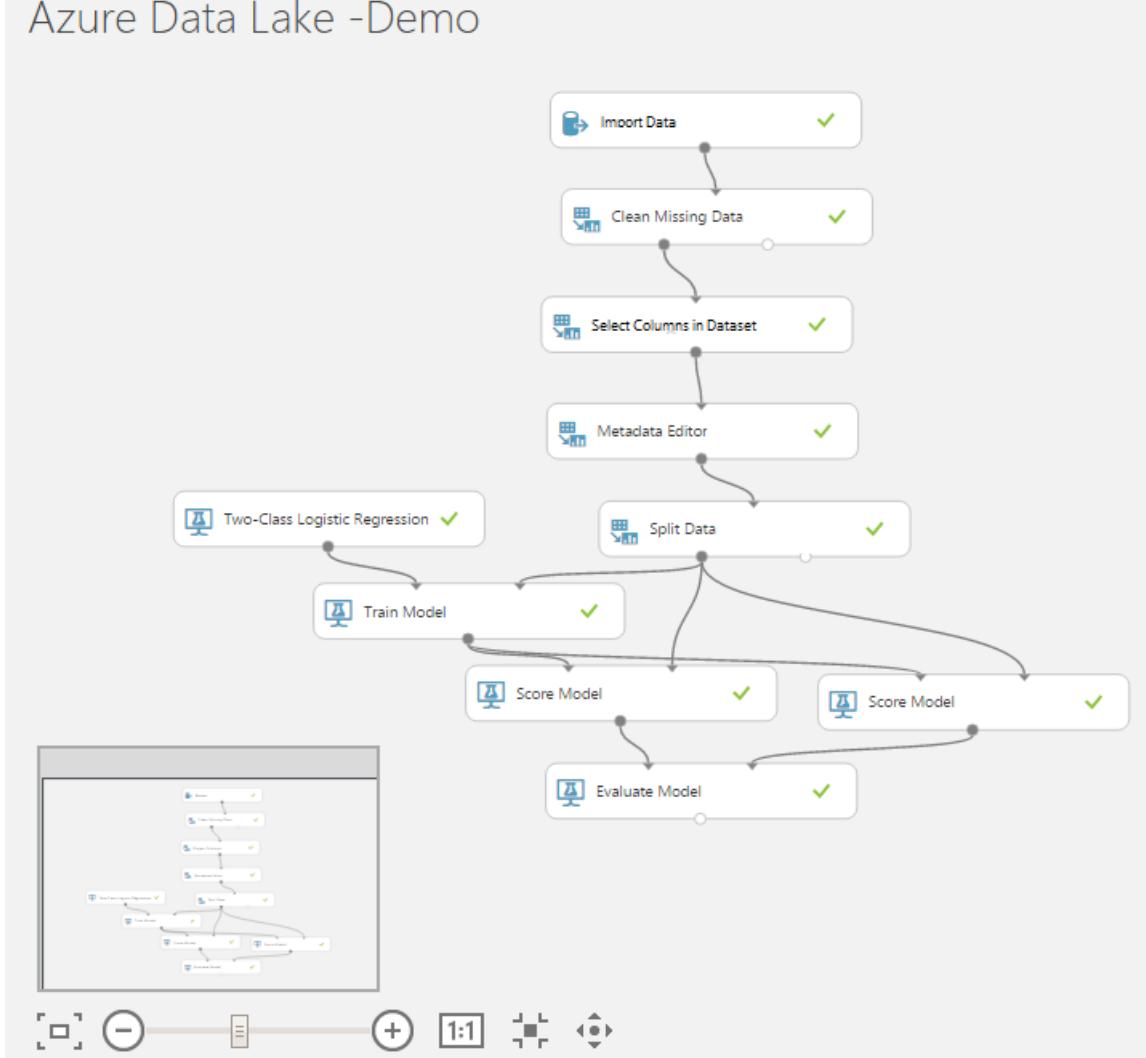
```
select * from nyc_stratified_sample;
```

4. Enter the URI of HDInsight cluster (this can be found in Azure portal), Hadoop credentials, location of output data, and Azure storage account name/key/container name.



An example of a binary classification experiment reading data from Hive table is shown in the following figure:

Azure Data Lake -Demo



After the experiment is created, click **Set Up Web Service --> Predictive Web Service**

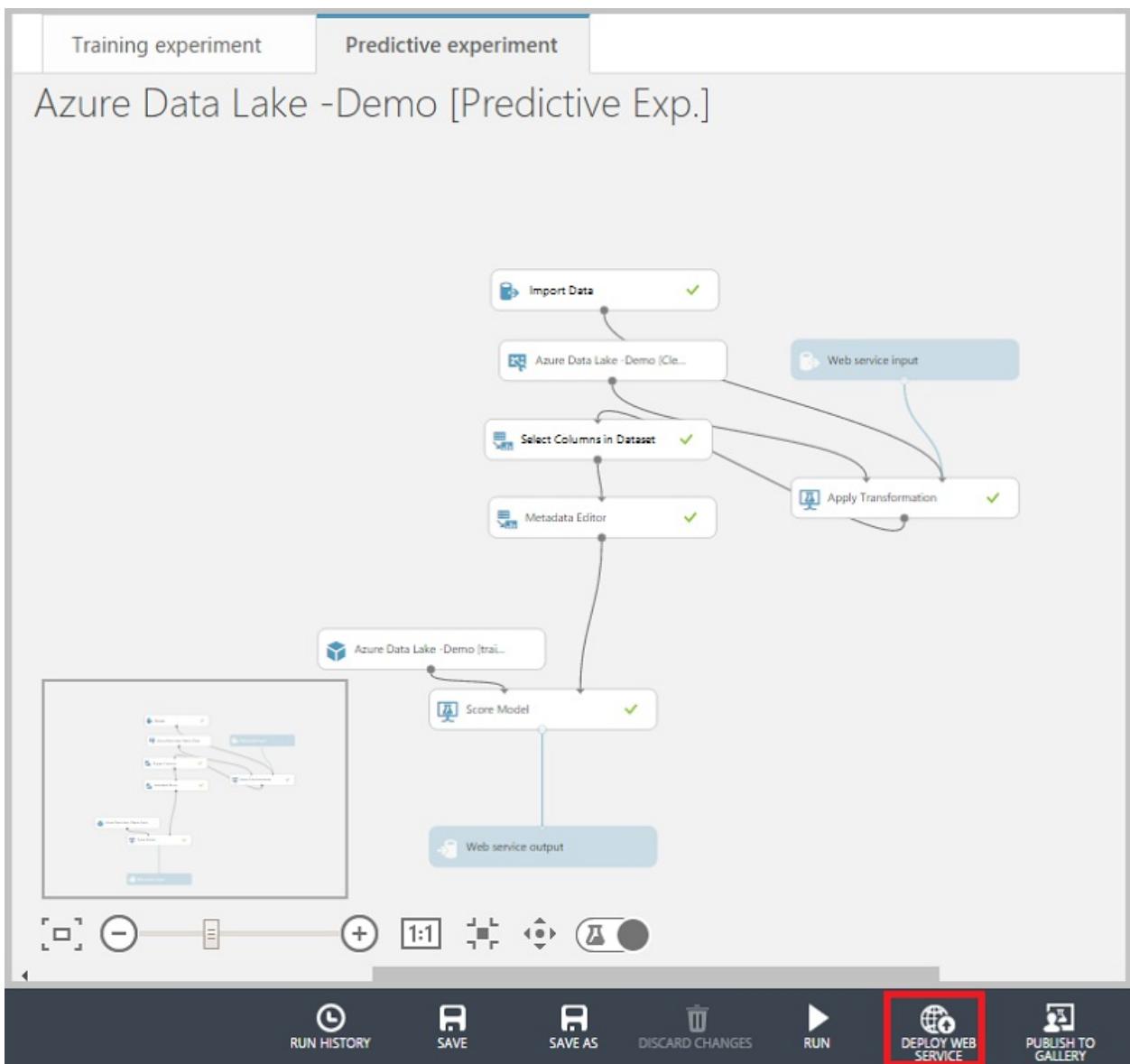
Azure Data Lake -Demo

Finished running ✓

Draft saved at 3:54:48 AM



Run the automatically created scoring experiment, when it finishes, click **Deploy Web Service**



The web service dashboard displays shortly:

Microsoft Azure Machine Learning

azure data lake -demo [predictive exp.]

DASHBOARD **CONFIGURATION**

General

Published experiment
[View snapshot](#) [View latest](#)

Description
 No description provided for this web service.

API key
 h86t ... 51fw==

Default Endpoint

API HELP PAGE	TEST	APPS
REQUEST/RESPONSE	Test	<small>Excel 2013 or later Excel 2010 or earlier workbook</small>
BATCH EXECUTION		<small>Excel 2013 or later workbook</small>

Additional endpoints
 Number of additional endpoints created for this web service: 0
[Manage endpoints in Azure management portal](#)

Summary

By completing this walkthrough you have created a data science environment for building scalable end-to-end

solutions in Azure Data Lake. This environment was used to analyze a large public dataset, taking it through the canonical steps of the Data Science Process, from data acquisition through model training, and then to the deployment of the model as a web service. U-SQL was used to process, explore and sample the data. Python and Hive were used with Azure Machine Learning Studio to build and deploy predictive models.

What's next?

The learning path for the [Team Data Science Process \(TDSP\)](#) provides links to topics describing each step in the advanced analytics process. There are a series of walkthroughs itemized on the [Team Data Science Process walkthroughs](#) page that showcase how to use resources and services in various predictive analytics scenarios:

- [The Team Data Science Process in action: using SQL Data Warehouse](#)
- [The Team Data Science Process in action: using HDInsight Hadoop clusters](#)
- [The Team Data Science Process: using SQL Server](#)
- [Overview of the Data Science Process using Spark on Azure HDInsight](#)

Process Data in SQL Server Virtual Machine on Azure

3/12/2019 • 6 minutes to read

This document covers how to explore data and generate features for data stored in a SQL Server VM on Azure. This can be done by data wrangling using SQL or by using a programming language like Python.

NOTE

The sample SQL statements in this document assume that data is in SQL Server. If it isn't, refer to the cloud data science process map to learn how to move your data to SQL Server.

Using SQL

We describe the following data wrangling tasks in this section using SQL:

1. [Data Exploration](#)
2. [Feature Generation](#)

Data Exploration

Here are a few sample SQL scripts that can be used to explore data stores in SQL Server.

NOTE

For a practical example, you can use the [NYC Taxi dataset](#) and refer to the IPNB titled [NYC Data wrangling using IPython Notebook and SQL Server](#) for an end-to-end walk-through.

1. Get the count of observations per day

```
SELECT CONVERT(date, <date_columnname>) as date, count(*) as c from <tablename> group by CONVERT(date, <date_columnname>)
```

2. Get the levels in a categorical column

```
select distinct <column_name> from <databasename>
```

3. Get the number of levels in combination of two categorical columns

```
select <column_a>, <column_b>, count(*) from <tablename> group by <column_a>, <column_b>
```

4. Get the distribution for numerical columns

```
select <column_name>, count(*) from <tablename> group by <column_name>
```

Feature Generation

In this section, we describe ways of generating features using SQL:

1. [Count based Feature Generation](#)
2. [Binning Feature Generation](#)
3. [Rolling out the features from a single column](#)

NOTE

Once you generate additional features, you can either add them as columns to the existing table or create a new table with the additional features and primary key, that can be joined with the original table.

Count based Feature Generation

The following examples demonstrate two ways of generating count features. The first method uses conditional sum and the second method uses the 'where' clause. These can then be joined with the original table (using primary key columns) to have count features alongside the original data.

```
select <column_name1>,<column_name2>,<column_name3>, COUNT(*) as Count_Features from <tablename> group by  
<column_name1>,<column_name2>,<column_name3>  
  
select <column_name1>,<column_name2> , sum(1) as Count_Features from <tablename>  
where <column_name3> = '<some_value>' group by <column_name1>,<column_name2>
```

Binning Feature Generation

The following example shows how to generate binned features by binning (using five bins) a numerical column that can be used as a feature instead:

```
`SELECT <column_name>, NTILE(5) OVER (ORDER BY <column_name>) AS BinNumber from <tablename>`
```

Rolling out the features from a single column

In this section, we demonstrate how to roll out a single column in a table to generate additional features. The example assumes that there is a latitude or longitude column in the table from which you are trying to generate features.

Here is a brief primer on latitude/longitude location data (resourced from stackoverflow [How to measure the accuracy of latitude and longitude?](#)). This is useful to understand before featurizing the location field:

- The sign tells us whether we are north or south, east or west on the globe.
- A nonzero hundreds digit tells us that we're using longitude, not latitude!
- The tens digit gives a position to about 1,000 kilometers. It gives us useful information about what continent or ocean we are on.
- The units digit (one decimal degree) gives a position up to 111 kilometers (60 nautical miles, about 69 miles). It can tell us roughly what large state or country we are in.
- The first decimal place is worth up to 11.1 km: it can distinguish the position of one large city from a neighboring large city.
- The second decimal place is worth up to 1.1 km: it can separate one village from the next.
- The third decimal place is worth up to 110 m: it can identify a large agricultural field or institutional campus.
- The fourth decimal place is worth up to 11 m: it can identify a parcel of land. It is comparable to the typical accuracy of an uncorrected GPS unit with no interference.
- The fifth decimal place is worth up to 1.1 m: it distinguishes trees from each other. Accuracy to this level with commercial GPS units can only be achieved with differential correction.
- The sixth decimal place is worth up to 0.11 m: you can use this for laying out structures in detail, for designing landscapes, building roads. It should be more than good enough for tracking movements of glaciers and rivers. This can be achieved by taking painstaking measures with GPS, such as differentially corrected GPS.

The location information can be featurized as follows, separating out region, location, and city information. Note that you can also call a REST end point such as Bing Maps API available at [Find a Location by Point](#) to get the region/district information.

```

select
    <location_columnname>
    ,round(<location_columnname>,0) as l1
    ,12=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1)) >=
1 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,1) else
'0' end
    ,13=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1)) >=
2 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,2,1) else
'0' end
    ,14=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1)) >=
3 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,3,1) else
'0' end
    ,15=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1)) >=
4 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,4,1) else
'0' end
    ,16=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1)) >=
5 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,5,1) else
'0' end
    ,17=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1)) >=
6 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,6,1) else
'0' end
from <tablename>

```

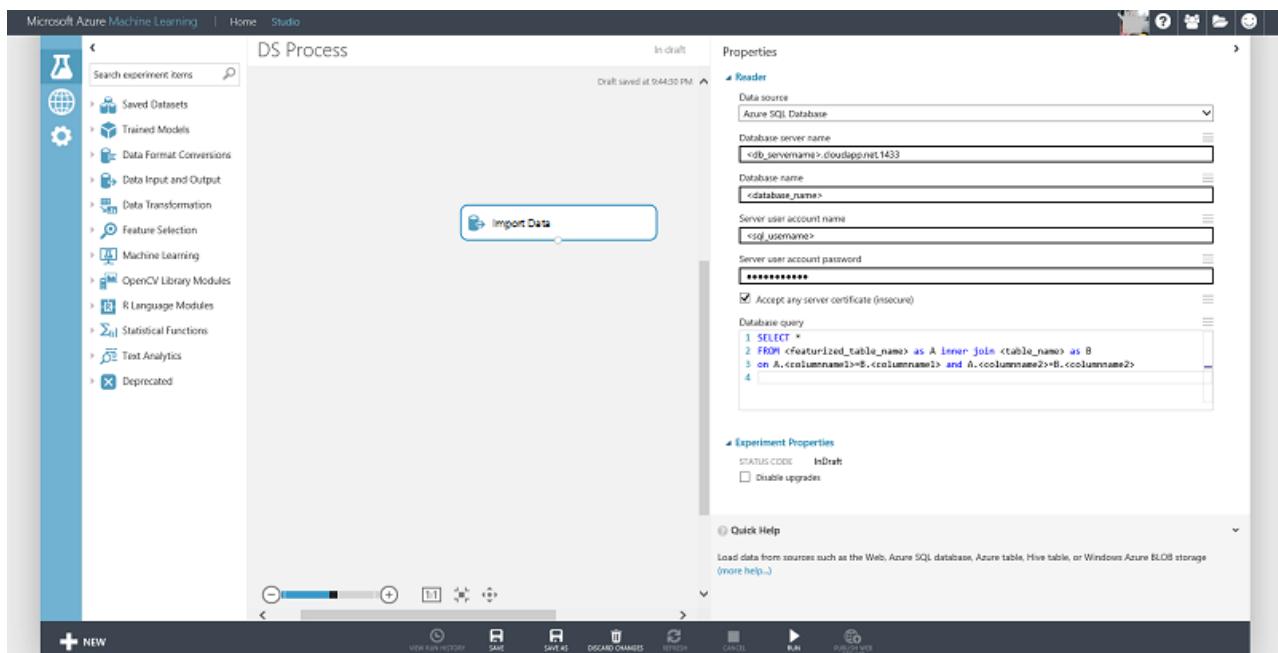
These location-based features can be further used to generate additional count features as described earlier.

TIP

You can programmatically insert the records using your language of choice. You may need to insert the data in chunks to improve write efficiency (for an example of how to do this using pyodbc, see [A HelloWorld sample to access SQLServer with python](#)). Another alternative is to insert data in the database using the [BCP utility](#).

Connecting to Azure Machine Learning

The newly generated feature can be added as a column to an existing table or stored in a new table and joined with the original table for machine learning. Features can be generated or accessed if already created, using the [Import Data](#) module in Azure Machine Learning as shown below:



Using a programming language like Python

Using Python to explore data and generate features when the data is in SQL Server is similar to processing data in

Azure blob using Python as documented in [Process Azure Blob data in your data science environment](#). The data needs to be loaded from the database into a pandas data frame and then can be processed further. We document the process of connecting to the database and loading the data into the data frame in this section.

The following connection string format can be used to connect to a SQL Server database from Python using pyodbc (replace servername, dbname, username, and password with your specific values):

```
#Set up the SQL Azure connection
import pyodbc
conn = pyodbc.connect('DRIVER={SQL Server};SERVER=<servername>;DATABASE=<dbname>;UID=<username>;PWD=<password>')
```

The [Pandas library](#) in Python provides a rich set of data structures and data analysis tools for data manipulation for Python programming. The code below reads the results returned from a SQL Server database into a Pandas data frame:

```
# Query database and load the returned results in pandas data frame
data_frame = pd.read_sql('''select <columnname1>, <columnname2>... from <tablename>''', conn)
```

Now you can work with the Pandas data frame as covered in the article [Process Azure Blob data in your data science environment](#).

Azure Data Science in Action Example

For an end-to-end walkthrough example of the Azure Data Science Process using a public dataset, see [Azure Data Science Process in Action](#).

Cheat sheet for an automated data pipeline for Azure Machine Learning predictions

1/30/2019 • 2 minutes to read

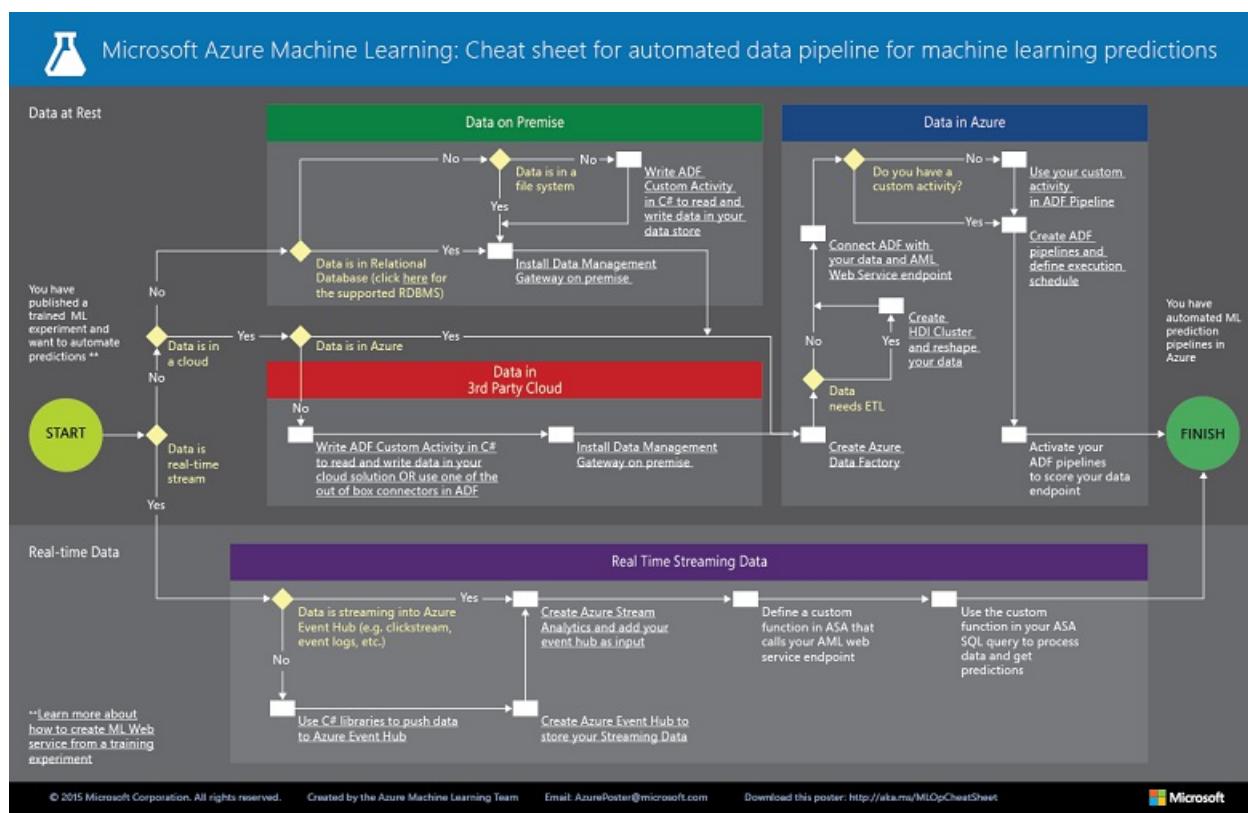
The **Microsoft Azure Machine Learning automated data pipeline cheat sheet** helps you navigate through the technology you can use to get your data to your Machine Learning web service where it can be scored by your predictive analytics model.

Depending on whether your data is on-premises, in the cloud, or streaming real-time, there are different mechanisms available to move the data to your web service endpoint for scoring. This cheat sheet walks you through the decisions you need to make, and it offers links to articles that can help you develop your solution.

Download the Machine Learning automated data pipeline cheat sheet

Once you download the cheat sheet, you can print it in tabloid size (11 x 17 in.).

Download the cheat sheet here: [Microsoft Azure Machine Learning automated data pipeline cheat sheet](#)



More help with Machine Learning Studio

- For an overview of Microsoft Azure Machine Learning, see [Introduction to machine learning on Microsoft Azure](#).
- For an explanation of how to deploy a scoring web service, see [Deploy an Azure Machine Learning web service](#).
- For a discussion of how to consume a scoring web service, see [How to consume an Azure Machine Learning Web service](#).

Try [Azure Machine Learning Studio](#), available in paid or free options.

Overview of data science using Spark on Azure HDInsight

3/12/2019 • 9 minutes to read

This suite of topics shows how to use HDInsight Spark to complete common data science tasks such as data ingestion, feature engineering, modeling, and model evaluation. The data used is a sample of the 2013 NYC taxi trip and fare dataset. The models built include logistic and linear regression, random forests, and gradient boosted trees. The topics also show how to store these models in Azure blob storage (WASB) and how to score and evaluate their predictive performance. More advanced topics cover how models can be trained using cross-validation and hyper-parameter sweeping. This overview topic also references the topics that describe how to set up the Spark cluster that you need to complete the steps in the walkthroughs provided.

Spark and MLlib

[Spark](#) is an open-source parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications. The Spark processing engine is built for speed, ease of use, and sophisticated analytics. Spark's in-memory distributed computation capabilities make it a good choice for the iterative algorithms used in machine learning and graph computations. [MLlib](#) is Spark's scalable machine learning library that brings the algorithmic modeling capabilities to this distributed environment.

HDInsight Spark

[HDInsight Spark](#) is the Azure hosted offering of open-source Spark. It also includes support for [Jupyter](#)

PySpark notebooks on the Spark cluster that can run Spark SQL interactive queries for transforming, filtering, and visualizing data stored in Azure Blobs (WASB). PySpark is the Python API for Spark. The code snippets that provide the solutions and show the relevant plots to visualize the data here run in Jupyter notebooks installed on the Spark clusters. The modeling steps in these topics contain code that shows how to train, evaluate, save, and consume each type of model.

Setup: Spark clusters and Jupyter notebooks

Setup steps and code are provided in this walkthrough for using an HDInsight Spark 1.6. But Jupyter notebooks are provided for both HDInsight Spark 1.6 and Spark 2.0 clusters. A description of the notebooks and links to them are provided in the [Readme.md](#) for the GitHub repository containing them. Moreover, the code here and in the linked notebooks is generic and should work on any Spark cluster. If you are not using HDInsight Spark, the cluster setup and management steps may be slightly different from what is shown here. For convenience, here are the links to the Jupyter notebooks for Spark 1.6 (to be run in the pySpark kernel of the Jupyter Notebook server) and Spark 2.0 (to be run in the pySpark3 kernel of the Jupyter Notebook server):

Spark 1.6 notebooks

These notebooks are to be run in the pySpark kernel of Jupyter notebook server.

- [pySpark-machine-learning-data-science-spark-data-exploration-modeling.ipynb](#): Provides information on how to perform data exploration, modeling, and scoring with several different algorithms.
- [pySpark-machine-learning-data-science-spark-advanced-data-exploration-modeling.ipynb](#): Includes topics in notebook #1, and model development using hyperparameter tuning and cross-validation.
- [pySpark-machine-learning-data-science-spark-model-consumption.ipynb](#): Shows how to operationalize a saved model using Python on HDInsight clusters.

Spark 2.0 notebooks

These notebooks are to be run in the pySpark3 kernel of Jupyter notebook server.

- [Spark2.0-pySpark3-machine-learning-data-science-spark-advanced-data-exploration-modeling.ipynb](#): This file provides information on how to perform data exploration, modeling, and scoring in Spark 2.0 clusters using the NYC Taxi trip and fare data-set described [here](#). This notebook may be a good starting point for quickly exploring the code we have provided for Spark 2.0. For a more detailed notebook analyzes the NYC Taxi data, see the next notebook in this list. See the notes following this list that compare these notebooks.
- [Spark2.0-pySpark3_NYC_Taxi_Tip_Regression.ipynb](#): This file shows how to perform data wrangling (Spark SQL and dataframe operations), exploration, modeling and scoring using the NYC Taxi trip and fare data-set described [here](#).
- [Spark2.0-pySpark3_Airline_Departure_Delay_Classification.ipynb](#): This file shows how to perform data wrangling (Spark SQL and dataframe operations), exploration, modeling and scoring using the well-known Airline On-time departure dataset from 2011 and 2012. We integrated the airline dataset with the airport weather data (e.g. windspeed, temperature, altitude etc.) prior to modeling, so these weather features can be included in the model.

NOTE

The airline dataset was added to the Spark 2.0 notebooks to better illustrate the use of classification algorithms. See the following links for information about airline on-time departure dataset and weather dataset:

- Airline on-time departure data: <https://www.transtats.bts.gov/ONTIME/>
- Airport weather data: <https://www.ncdc.noaa.gov/>

NOTE

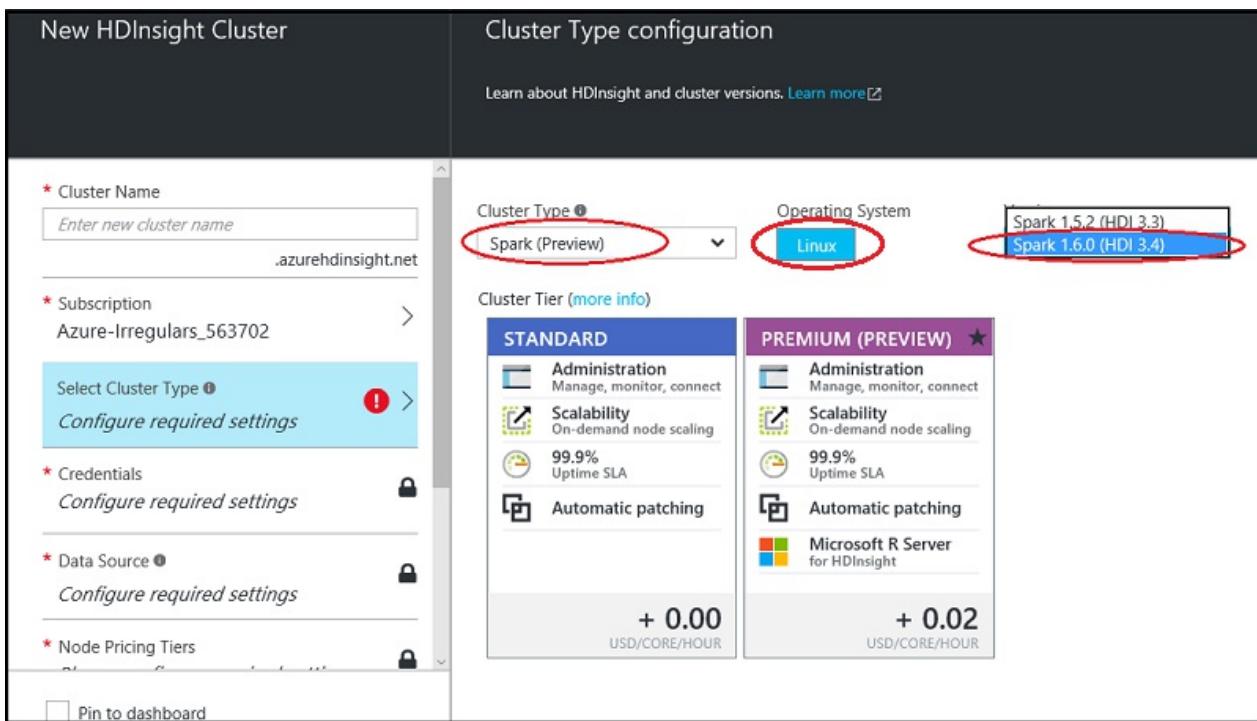
The Spark 2.0 notebooks on the NYC taxi and airline flight delay data-sets can take 10 mins or more to run (depending on the size of your HDI cluster). The first notebook in the above list shows many aspects of the data exploration, visualization and ML model training in a notebook that takes less time to run with down-sampled NYC data set, in which the taxi and fare files have been pre-joined: [Spark2.0-pySpark3-machine-learning-data-science-spark-advanced-data-exploration-modeling.ipynb](#). This notebook takes a much shorter time to finish (2-3 mins) and may be a good starting point for quickly exploring the code we have provided for Spark 2.0.

For guidance on the operationalization of a Spark 2.0 model and model consumption for scoring, see the [Spark 1.6 document on consumption](#) for an example outlining the steps required. To use this on Spark 2.0, replace the Python code file with [this file](#).

Prerequisites

The following procedures are related to Spark 1.6. For the Spark 2.0 version, use the notebooks described and linked to previously.

1. You must have an Azure subscription. If you do not already have one, see [Get Azure free trial](#).
2. You need a Spark 1.6 cluster to complete this walkthrough. To create one, see the instructions provided in [Get started: create Apache Spark on Azure HDInsight](#). The cluster type and version is specified from the **Select Cluster Type** menu.



NOTE

For a topic that shows how to use Scala rather than Python to complete tasks for an end-to-end data science process, see the [Data Science using Scala with Spark on Azure](#).

WARNING

Billing for HDInsight clusters is prorated per minute, whether you use them or not. Be sure to delete your cluster after you finish using it. See [how to delete an HDInsight cluster](#).

The NYC 2013 Taxi data

The NYC Taxi Trip data is about 20 GB of compressed comma-separated values (CSV) files (~48 GB uncompressed), comprising more than 173 million individual trips and the fares paid for each trip. Each trip record includes the pick up and dropoff location and time, anonymized hack (driver's) license number and medallion (taxi's unique id) number. The data covers all trips in the year 2013 and is provided in the following two datasets for each month:

1. The 'trip_data' CSV files contain trip details, such as number of passengers, pick up and dropoff points, trip duration, and trip length. Here are a few sample records:

```
medallion,hack_license,vendor_id,rate_code,store_and_fwd_flag,pickup_datetime,dropoff_datetime,passenger_count,trip_time_in_secs,trip_distance,pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
89D227B655E5C82AEFC13C3F540D4CF4,BA96DE419E711691B9445D6A6307C170,CMT,1,N,2013-01-01 15:11:48,2013-01-01 15:18:10,4,382,1.00,-73.978165,40.757977,-73.989838,40.751171
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,1,N,2013-01-06 00:18:35,2013-01-06 00:22:54,1,259,1.50,-74.006683,40.731781,-73.994499,40.75066
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,1,N,2013-01-05 18:49:41,2013-01-05 18:54:23,1,282,1.10,-74.004707,40.73777,-74.009834,40.726002
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:54:15,2013-01-07 23:58:20,2,244,.70,-73.974602,40.759945,-73.984734,40.759388
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,1,N,2013-01-07 23:25:03,2013-01-07 23:34:24,1,560,2.10,-73.97625,40.748528,-74.002586,40.747868
```

2. The 'trip_fare' CSV files contain details of the fare paid for each trip, such as payment type, fare amount, surcharge and taxes, tips and tolls, and the total amount paid. Here are a few sample records:

```

medallion, hack_license, vendor_id, pickup_datetime, payment_type, fare_amount, surcharge, mta_tax,
tip_amount, tolls_amount, total_amount
89D227B655E5C82AECE13C3F540D4CF4,BA96DE419E711691B9445D6A6307C170,CMT,2013-01-01
15:11:48,CSH,6.5,0,0.5,0,0,7
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,2013-01-06
00:18:35,CSH,6,0.5,0.5,0,0,7
0BD7C8F5BA12B88E0B67BED28BEA73D8,9FD8F69F0804BDB5549F40E9DA1BE472,CMT,2013-01-05
18:49:41,CSH,5.5,1,0.5,0,0,7
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,2013-01-07
23:54:15,CSH,5,0.5,0.5,0,0,6
DFD2202EE08F7A8DC9A57B02ACB81FE2,51EE87E3205C985EF8431D850C786310,CMT,2013-01-07
23:25:03,CSH,9.5,0.5,0.5,0,0,10.5

```

We have taken a 0.1% sample of these files and joined the trip_data and trip_fare CVS files into a single dataset to use as the input dataset for this walkthrough. The unique key to join trip_data and trip_fare is composed of the fields: medallion, hack_licence and pickup_datetime. Each record of the dataset contains the following attributes representing a NYC Taxi trip:

FIELD	BRIEF DESCRIPTION
medallion	Anonymized taxi medallion (unique taxi id)
hack_license	Anonymized Hackney Carriage License number
vendor_id	Taxi vendor id
rate_code	NYC taxi rate of fare
store_and_fwd_flag	Store and forward flag
pickup_datetime	Pick up date & time
dropoff_datetime	Dropoff date & time
pickup_hour	Pick up hour
pickup_week	Pick up week of the year
weekday	Weekday (range 1-7)
passenger_count	Number of passengers in a taxi trip
trip_time_in_secs	Trip time in seconds
trip_distance	Trip distance traveled in miles
pickup_longitude	Pick up longitude
pickup_latitude	Pick up latitude
dropoff_longitude	Dropoff longitude

FIELD	BRIEF DESCRIPTION
dropoff_latitude	Dropoff latitude
direct_distance	Direct distance between pick up and dropoff locations
payment_type	Payment type (cash, credit-card etc.)
fare_amount	Fare amount in
surcharge	Surcharge
mta_tax	Mta tax
tip_amount	Tip amount
tolls_amount	Tolls amount
total_amount	Total amount
tipped	Tipped (0/1 for no or yes)
tip_class	Tip class (0: \$0, 1: \$0-5, 2: \$6-10, 3: \$11-20, 4: > \$20)

Execute code from a Jupyter notebook on the Spark cluster

You can launch the Jupyter Notebook from the Azure portal. Find your Spark cluster on your dashboard and click it to enter management page for your cluster. To open the notebook associated with the Spark cluster, click **Cluster Dashboards -> Jupyter Notebook**.

The screenshot shows the Azure HDInsight Cluster dashboard. On the left, under 'Essentials', there's a 'Resource group' section with a redacted URL, a 'Status' section showing 'Running' in 'South Central US', and a 'Subscription name' and 'Subscription ID' section. Below this is a 'Quick Links' section with three items: 'Cluster Dashboards' (circled in red), 'Ambari Views', and 'Scale Cluster'. Under 'Usage', there's a donut chart titled 'Cores in South Central US for subscription' showing 2,100 cores. To the right, there's a 'Cluster Dashboards' section with tiles for 'HDInsight Cluster Dashboard', 'Jupyter Notebook' (circled in red), 'Spark History Server', and 'Yarn'. An 'Add a section' button is at the bottom.

You can also browse to <https://CLUSTERNAME.azurehdinsight.net/jupyter> to access the Jupyter Notebooks. Replace the CLUSTERNAME part of this URL with the name of your own cluster. You need the password for your admin account to access the notebooks.

The screenshot shows the Jupyter notebook interface. At the top, there are tabs for 'Files', 'Running', and 'Clusters'. Below that is a list of files: 'PySpark' (circled in red) and 'Scala'. In the top right corner, there are buttons for 'Upload', 'New', and a refresh icon. A blue border surrounds the entire interface.

Select PySpark to see a directory that contains a few examples of pre-packaged notebooks that use the PySpark API. The notebooks that contain the code samples for this suite of Spark topic are available at [GitHub](#)

You can upload the notebooks directly from [GitHub](#) to the Jupyter notebook server on your Spark cluster. On the home page of your Jupyter, click the **Upload** button on the right part of the screen. It opens a file explorer. Here you can paste the GitHub (raw content) URL of the Notebook and click **Open**.

You see the file name on your Jupyter file list with an **Upload** button again. Click this **Upload** button. Now you have imported the notebook. Repeat these steps to upload the other notebooks from this walkthrough.

TIP

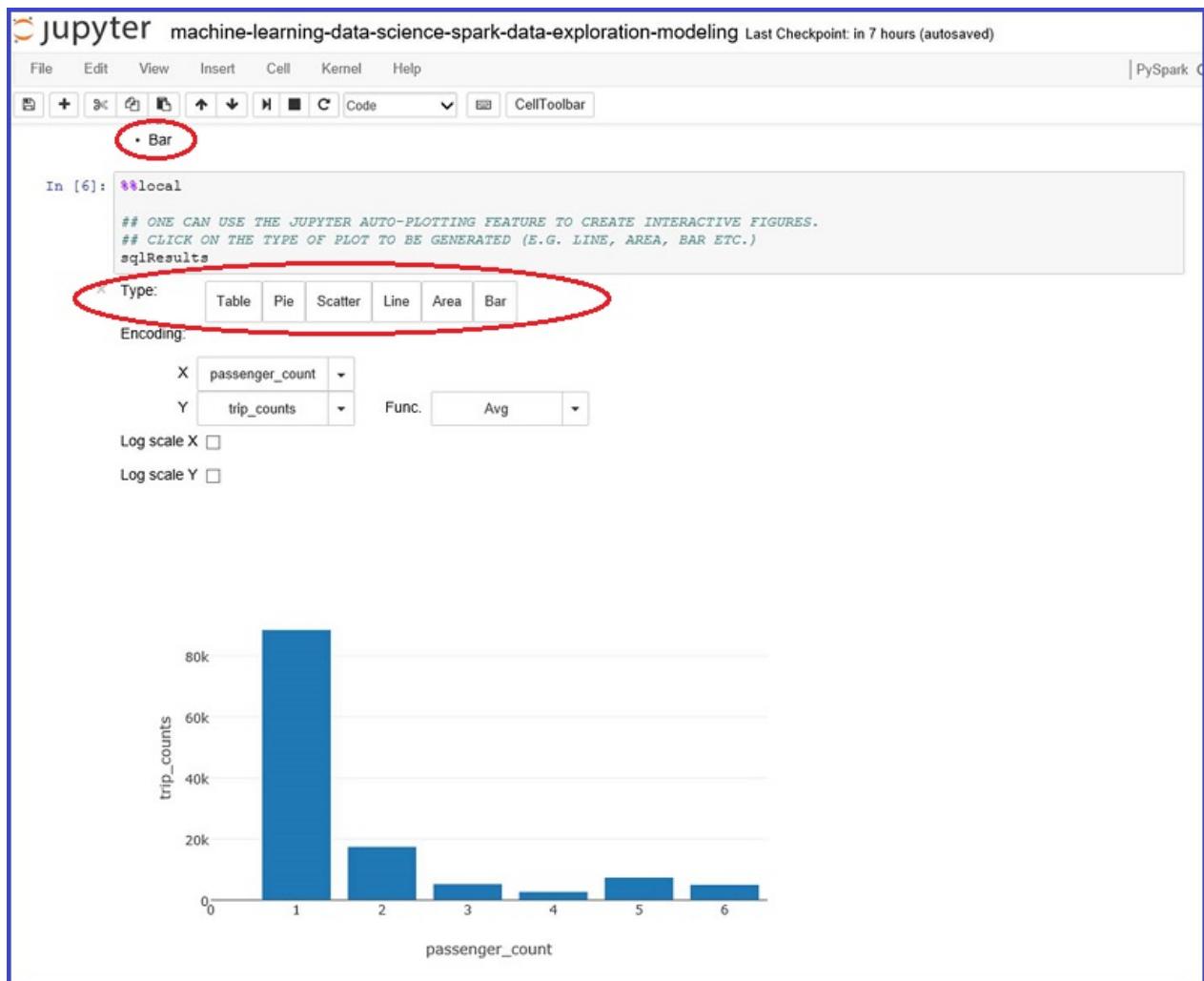
You can right-click the links on your browser and select **Copy Link** to get the GitHub raw content URL. You can paste this URL into the Jupyter Upload file explorer dialog box.

Now you can:

- See the code by clicking the notebook.
- Execute each cell by pressing **SHIFT-ENTER**.
- Run the entire notebook by clicking on **Cell -> Run**.
- Use the automatic visualization of queries.

TIP

The PySpark kernel automatically visualizes the output of SQL (HiveQL) queries. You are given the option to select among several different types of visualizations (Table, Pie, Line, Area, or Bar) by using the **Type** menu buttons in the notebook:



What's next?

Now that you are set up with an HDInsight Spark cluster and have uploaded the Jupyter notebooks, you are ready to work through the topics that correspond to the three PySpark notebooks. They show how to explore your data and then how to create and consume models. The advanced data exploration and modeling notebook shows how to include cross-validation, hyper-parameter sweeping, and model evaluation.

Data Exploration and modeling with Spark: Explore the dataset and create, score, and evaluate the machine

learning models by working through the [Create binary classification and regression models for data with the Spark MLlib toolkit](#) topic.

Model consumption: To learn how to score the classification and regression models created in this topic, see [Score and evaluate Spark-built machine learning models](#).

Cross-validation and hyperparameter sweeping: See [Advanced data exploration and modeling with Spark](#) on how models can be trained using cross-validation and hyper-parameter sweeping

Data Science using Scala and Spark on Azure

3/12/2019 • 34 minutes to read

This article shows you how to use Scala for supervised machine learning tasks with the Spark scalable MLlib and Spark ML packages on an Azure HDInsight Spark cluster. It walks you through the tasks that constitute the [Data Science process](#): data ingestion and exploration, visualization, feature engineering, modeling, and model consumption. The models in the article include logistic and linear regression, random forests, and gradient-boosted trees (GBTs), in addition to two common supervised machine learning tasks:

- Regression problem: Prediction of the tip amount (\$) for a taxi trip
- Binary classification: Prediction of tip or no tip (1/0) for a taxi trip

The modeling process requires training and evaluation on a test data set and relevant accuracy metrics. In this article, you can learn how to store these models in Azure Blob storage and how to score and evaluate their predictive performance. This article also covers the more advanced topics of how to optimize models by using cross-validation and hyper-parameter sweeping. The data used is a sample of the 2013 NYC taxi trip and fare data set available on GitHub.

[Scala](#), a language based on the Java virtual machine, integrates object-oriented and functional language concepts. It's a scalable language that is well suited to distributed processing in the cloud, and runs on Azure Spark clusters.

[Spark](#) is an open-source parallel-processing framework that supports in-memory processing to boost the performance of big data analytics applications. The Spark processing engine is built for speed, ease of use, and sophisticated analytics. Spark's in-memory distributed computation capabilities make it a good choice for iterative algorithms in machine learning and graph computations. The [spark.ml](#) package provides a uniform set of high-level APIs built on top of data frames that can help you create and tune practical machine learning pipelines. [MLlib](#) is Spark's scalable machine learning library, which brings modeling capabilities to this distributed environment.

[HDInsight Spark](#) is the Azure-hosted offering of open-source Spark. It also includes support for Jupyter Scala notebooks on the Spark cluster, and can run Spark SQL interactive queries to transform, filter, and visualize data stored in Azure Blob storage. The Scala code snippets in this article that provide the solutions and show the relevant plots to visualize the data run in Jupyter notebooks installed on the Spark clusters. The modeling steps in these topics have code that shows you how to train, evaluate, save, and consume each type of model.

The setup steps and code in this article are for Azure HDInsight 3.4 Spark 1.6. However, the code in this article and in the [Scala Jupyter Notebook](#) are generic and should work on any Spark cluster. The cluster setup and management steps might be slightly different from what is shown in this article if you are not using HDInsight Spark.

NOTE

For a topic that shows you how to use Python rather than Scala to complete tasks for an end-to-end Data Science process, see [Data Science using Spark on Azure HDInsight](#).

Prerequisites

- You must have an Azure subscription. If you do not already have one, [get an Azure free trial](#).
- You need an Azure HDInsight 3.4 Spark 1.6 cluster to complete the following procedures. To create a cluster, see the instructions in [Get started: Create Apache Spark on Azure HDInsight](#). Set the cluster type and version on the **Select Cluster Type** menu.

New HDInsight Cluster

Cluster Type configuration

Learn about HDInsight and cluster versions. [Learn more](#)

* Cluster Name
Enter new cluster name .azurehdinsight.net

* Subscription
Azure-Irregulars_563702

Select Cluster Type ⓘ Configure required settings ! >

* Credentials
Configure required settings

* Data Source ⓘ
Configure required settings

* Node Pricing Tiers
Configure required settings

Pin to dashboard

Cluster Type ⓘ Spark (Preview)

Operating System Linux

Spark 1.5.2 (HDI 3.3)
Spark 1.6.0 (HDI 3.4)

Cluster Tier ([more info](#))

STANDARD

- Administration Manage, monitor, connect
- Scalability On-demand node scaling
- 99.9% Uptime SLA
- Automatic patching

+ 0.00 USD/CORE/HOUR

PREMIUM (PREVIEW) ★

- Administration Manage, monitor, connect
- Scalability On-demand node scaling
- 99.9% Uptime SLA
- Automatic patching
- Microsoft R Server for HDInsight

+ 0.02 USD/CORE/HOUR

WARNING

Billing for HDInsight clusters is prorated per minute, whether you use them or not. Be sure to delete your cluster after you finish using it. See [how to delete an HDInsight cluster](#).

For a description of the NYC taxi trip data and instructions on how to execute code from a Jupyter notebook on the Spark cluster, see the relevant sections in [Overview of Data Science using Spark on Azure HDInsight](#).

Execute Scala code from a Jupyter notebook on the Spark cluster

You can launch a Jupyter notebook from the Azure portal. Find the Spark cluster on your dashboard, and then click it to enter the management page for your cluster. Next, click **Cluster Dashboards**, and then click **Jupyter Notebook** to open the notebook associated with the Spark cluster.

The screenshot shows the Azure HDInsight Cluster blade on the left and the Cluster Dashboards page on the right. The Cluster Dashboards page contains four tiles:

- HDInsight Cluster Dashboard**: Shows a yellow elephant icon.
- Jupyter Notebook**: Shows an orange and white circular icon. This tile is circled in red.
- Spark History Server**: Shows a blue square icon.
- Yarn**: Shows a blue square icon.

You also can access Jupyter notebooks at <https://<clustername>.azurehdinsight.net/jupyter>. Replace *clustername* with the name of your cluster. You need the password for your administrator account to access the Jupyter notebooks.

The screenshot shows the Jupyter notebook interface with a file explorer on the left. The Scala directory is circled in red. The top right corner features an **Upload** button, which is also circled in red.

Select **Scala** to see a directory that has a few examples of prepackaged notebooks that use the PySpark API. The Exploration Modeling and Scoring using Scala.ipynb notebook that contains the code samples for this suite of Spark topics is available on [GitHub](#).

You can upload the notebook directly from GitHub to the Jupyter Notebook server on your Spark cluster. On your Jupyter home page, click the **Upload** button. In the file explorer, paste the GitHub (raw content) URL of the Scala notebook, and then click **Open**. The Scala notebook is available at the following URL:

[Exploration-Modeling-and-Scoring-using-Scala.ipynb](#)

Setup: Preset Spark and Hive contexts, Spark magics, and Spark libraries

Preset Spark and Hive contexts

```
# SET THE START TIME
import java.util.Calendar
val beginningTime = Calendar.getInstance().getTime()
```

The Spark kernels that are provided with Jupyter notebooks have preset contexts. You don't need to explicitly set the Spark or Hive contexts before you start working with the application you are developing. The preset contexts are:

- `sc` for `SparkContext`
- `sqlContext` for `HiveContext`

Spark magics

The Spark kernel provides some predefined "magics," which are special commands that you can call with `%%`. Two of these commands are used in the following code samples.

- `%%local` specifies that the code in subsequent lines will be executed locally. The code must be valid Scala code.
- `%%sql -o <variable name>` executes a Hive query against `sqlContext`. If the `-o` parameter is passed, the result of the query is persisted in the `%%local` Scala context as a Spark data frame.

For more information about the kernels for Jupyter notebooks and their predefined "magics" that you call with `%%` (for example, `%%local`), see [Kernels available for Jupyter notebooks with HDInsight Spark Linux clusters on HDInsight](#).

Import libraries

Import the Spark, MLlib, and other libraries you'll need by using the following code.

```

# IMPORT SPARK AND JAVA LIBRARIES
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.functions._
import java.text.SimpleDateFormat
import java.util.Calendar
import sqlContext.implicits._
import org.apache.spark.sql.Row

# IMPORT SPARK SQL FUNCTIONS
import org.apache.spark.sql.types.{StructType, StructField, StringType, IntegerType, FloatType, DoubleType}
import org.apache.spark.sql.functions.rand

# IMPORT SPARK ML FUNCTIONS
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.feature.{StringIndexer, VectorAssembler, OneHotEncoder, VectorIndexer, Binarizer}
import org.apache.spark.ml.tuning.{ParamGridBuilder, TrainValidationSplit, CrossValidator}
import org.apache.spark.ml.regression.{LinearRegression, LinearRegressionModel, RandomForestRegressor,
RandomForestRegressionModel, GBTRRegressor, GBTRRegressionModel}
import org.apache.spark.ml.classification.{LogisticRegression, LogisticRegressionModel,
RandomForestClassifier, RandomForestClassificationModel, GBTCClassifier, GBTCClassificationModel}
import org.apache.spark.ml.evaluation.{BinaryClassificationEvaluator, RegressionEvaluator,
MulticlassClassificationEvaluator}

# IMPORT SPARK MLLIB FUNCTIONS
import org.apache.spark.mllib.linalg.{Vector, Vectors}
import org.apache.spark.mllib.util.MLUtils
import org.apache.spark.mllib.classification.{LogisticRegressionWithLBFGS, LogisticRegressionModel}
import org.apache.spark.mllib.regression.{LabeledPoint, LinearRegressionWithSGD, LinearRegressionModel}
import org.apache.spark.mllib.tree.{GradientBoostedTrees, RandomForest}
import org.apache.spark.mllib.tree.configuration.BoostingStrategy
import org.apache.spark.mllib.tree.model.{GradientBoostedTreesModel, RandomForestModel, Predict}
import org.apache.spark.mllib.evaluation.{BinaryClassificationMetrics, MulticlassMetrics, RegressionMetrics}

# SPECIFY SQLCONTEXT
val sqlContext = new SQLContext(sc)

```

Data ingestion

The first step in the Data Science process is to ingest the data that you want to analyze. You bring the data from external sources or systems where it resides into your data exploration and modeling environment. In this article, the data you ingest is a joined 0.1% sample of the taxi trip and fare file (stored as a .tsv file). The data exploration and modeling environment is Spark. This section contains the code to complete the following series of tasks:

1. Set directory paths for data and model storage.
2. Read in the input data set (stored as a .tsv file).
3. Define a schema for the data and clean the data.
4. Create a cleaned data frame and cache it in memory.
5. Register the data as a temporary table in SQLContext.
6. Query the table and import the results into a data frame.

Set directory paths for storage locations in Azure Blob storage

Spark can read and write to Azure Blob storage. You can use Spark to process any of your existing data, and then store the results again in Blob storage.

To save models or files in Blob storage, you need to properly specify the path. Reference the default container attached to the Spark cluster by using a path that begins with `wasb:///`. Reference other locations by using `wasb://`.

The following code sample specifies the location of the input data to be read and the path to Blob storage that is attached to the Spark cluster where the model will be saved.

```

# SET PATHS TO DATA AND MODEL FILE LOCATIONS
# INGEST DATA AND SPECIFY HEADERS FOR COLUMNS
val taxi_train_file =
sc.textFile("wasb://mllibwalkthroughs@cdspsparkssamples.blob.core.windows.net/Data/NYCTaxi/JoinedTaxiTripFare.P
oint1Pct.Train.tsv")
val header = taxi_train_file.first;

# SET THE MODEL STORAGE DIRECTORY PATH
# NOTE THAT THE FINAL BACKSLASH IN THE PATH IS REQUIRED.
val modelDir = "wasb:///user/remoteuser/NYCTaxi/Models/";

```

Import data, create an RDD, and define a data frame according to the schema

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# DEFINE THE SCHEMA BASED ON THE HEADER OF THE FILE
val sqlContext = new SQLContext(sc)
val taxi_schema = StructType(
  Array(
    StructField("medallion", StringType, true),
    StructField("hack_license", StringType, true),
    StructField("vendor_id", StringType, true),
    StructField("rate_code", DoubleType, true),
    StructField("store_and_fwd_flag", StringType, true),
    StructField("pickup_datetime", StringType, true),
    StructField("dropoff_datetime", StringType, true),
    StructField("pickup_hour", DoubleType, true),
    StructField("pickup_week", DoubleType, true),
    StructField("weekday", DoubleType, true),
    StructField("passenger_count", DoubleType, true),
    StructField("trip_time_in_secs", DoubleType, true),
    StructField("trip_distance", DoubleType, true),
    StructField("pickup_longitude", DoubleType, true),
    StructField("pickup_latitude", DoubleType, true),
    StructField("dropoff_longitude", DoubleType, true),
    StructField("dropoff_latitude", DoubleType, true),
    StructField("direct_distance", StringType, true),
    StructField("payment_type", StringType, true),
    StructField("fare_amount", DoubleType, true),
    StructField("surcharge", DoubleType, true),
    StructField("mta_tax", DoubleType, true),
    StructField("tip_amount", DoubleType, true),
    StructField("tolls_amount", DoubleType, true),
    StructField("total_amount", DoubleType, true),
    StructField("tipped", DoubleType, true),
    StructField("tip_class", DoubleType, true)
  )
)

# CAST VARIABLES ACCORDING TO THE SCHEMA
val taxi_temp = (taxi_train_file.map(_.split("\t"))
  .filter((r) => r(0) != "medallion")
  .map(p => Row(p(0), p(1), p(2),
    p(3).toDouble, p(4), p(5), p(6), p(7).toDouble, p(8).toDouble, p(9).toDouble,
    p(10).toDouble,
    p(11).toDouble, p(12).toDouble, p(13).toDouble, p(14).toDouble, p(15).toDouble,
    p(16).toDouble,
    p(17), p(18), p(19).toDouble, p(20).toDouble, p(21).toDouble, p(22).toDouble,
    p(23).toDouble, p(24).toDouble, p(25).toDouble, p(26).toDouble)))
)

# CREATE AN INITIAL DATA FRAME AND DROP COLUMNS, AND THEN CREATE A CLEANED DATA FRAME BY FILTERING FOR
UNWANTED VALUES OR OUTLIERS
val taxi_train_df = sqlContext.createDataFrame(taxi_temp, taxi_schema)

```

```

val taxi_df_train_cleaned = (taxi_train_df.drop(taxi_train_df.col("medallion"))
    .drop(taxi_train_df.col("hack_license")).drop(taxi_train_df.col("store_and_fwd_flag"))
    .drop(taxi_train_df.col("pickup_datetime")).drop(taxi_train_df.col("dropoff_datetime"))
    .drop(taxi_train_df.col("pickup_longitude")).drop(taxi_train_df.col("pickup_latitude"))
    .drop(taxi_train_df.col("dropoff_longitude")).drop(taxi_train_df.col("dropoff_latitude"))
    .drop(taxi_train_df.col("surcharge")).drop(taxi_train_df.col("mta_tax"))
    .drop(taxi_train_df.col("direct_distance")).drop(taxi_train_df.col("tolls_amount"))
    .drop(taxi_train_df.col("total_amount")).drop(taxi_train_df.col("tip_class"))
    .filter("passenger_count > 0 AND passenger_count < 8 AND payment_type in ('CSH', 'CRD') AND tip_amount
    >= 0 AND tip_amount < 30 AND fare_amount >= 1 AND fare_amount < 150 AND trip_distance > 0 AND trip_distance <
    100 AND trip_time_in_secs > 30 AND trip_time_in_secs < 7200"));

# CACHE AND MATERIALIZE THE CLEANED DATA FRAME IN MEMORY
taxi_df_train_cleaned.cache()
taxi_df_train_cleaned.count()

# REGISTER THE DATA FRAME AS A TEMPORARY TABLE IN SQLCONTEXT
taxi_df_train_cleaned.registerTempTable("taxi_train")

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime()) / 1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

```

Output:

Time to run the cell: 8 seconds.

Query the table and import results in a data frame

Next, query the table for fare, passenger, and tip data; filter out corrupt and outlying data; and print several rows.

```

# QUERY THE DATA
val sqlStatement = """
    SELECT fare_amount, passenger_count, tip_amount, tipped
    FROM taxi_train
    WHERE passenger_count > 0 AND passenger_count < 7
    AND fare_amount > 0 AND fare_amount < 200
    AND payment_type in ('CSH', 'CRD')
    AND tip_amount > 0 AND tip_amount < 25
"""
val sqlResultsDF = sqlContext.sql(sqlStatement)

# SHOW ONLY THE TOP THREE ROWS
sqlResultsDF.show(3)

```

Output:

FARE_AMOUNT	PASSENGER_COUNT	TIP_AMOUNT	TIPPED
13.5	1.0	2.9	1.0
16.0	2.0	3.4	1.0
10.5	2.0	1.0	1.0

Data exploration and visualization

After you bring the data into Spark, the next step in the Data Science process is to gain a deeper understanding of the data through exploration and visualization. In this section, you examine the taxi data by using SQL queries. Then, import the results into a data frame to plot the target variables and prospective features for visual inspection

by using the auto-visualization feature of Jupyter.

Use local and SQL magic to plot data

By default, the output of any code snippet that you run from a Jupyter notebook is available within the context of the session that is persisted on the worker nodes. If you want to save a trip to the worker nodes for every computation, and if all the data that you need for your computation is available locally on the Jupyter server node (which is the head node), you can use the `%%local` magic to run the code snippet on the Jupyter server.

- **SQL magic** (`%%sql`). The HDInsight Spark kernel supports easy inline HiveQL queries against `SQLContext`. The `(-o VARIABLE_NAME)` argument persists the output of the SQL query as a Pandas data frame on the Jupyter server. This means it'll be available in the local mode.
- **%%local magic**. The `%%local` magic runs the code locally on the Jupyter server, which is the head node of the HDInsight cluster. Typically, you use `%%local` magic in conjunction with the `%%sql` magic with the `-o` parameter. The `-o` parameter would persist the output of the SQL query locally, and then `%%local` magic would trigger the next set of code snippet to run locally against the output of the SQL queries that is persisted locally.

Query the data by using SQL

This query retrieves the taxi trips by fare amount, passenger count, and tip amount.

```
# RUN THE SQL QUERY
%%sql -q -o sqlResults
SELECT fare_amount, passenger_count, tip_amount, tipped FROM taxi_train WHERE passenger_count > 0 AND
passenger_count < 7 AND fare_amount > 0 AND fare_amount < 200 AND payment_type in ('CSH', 'CRD') AND
tip_amount > 0 AND tip_amount < 25
```

In the following code, the `%%local` magic creates a local data frame, `sqlResults`. You can use `sqlResults` to plot by using `matplotlib`.

TIP

Local magic is used multiple times in this article. If your data set is large, please sample to create a data frame that can fit in local memory.

Plot the data

You can plot by using Python code after the data frame is in local context as a Pandas data frame.

```
# RUN THE CODE LOCALLY ON THE JUPYTER SERVER
%%local

# USE THE JUPYTER AUTO-PLOTTING FEATURE TO CREATE INTERACTIVE FIGURES.
# CLICK THE TYPE OF PLOT TO GENERATE (LINE, AREA, BAR, ETC.)
sqlResults
```

The Spark kernel automatically visualizes the output of SQL (HiveQL) queries after you run the code. You can choose between several types of visualizations:

- Table
- Pie
- Line
- Area
- Bar

Here's the code to plot the data:

```

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
import matplotlib.pyplot as plt
%matplotlib inline

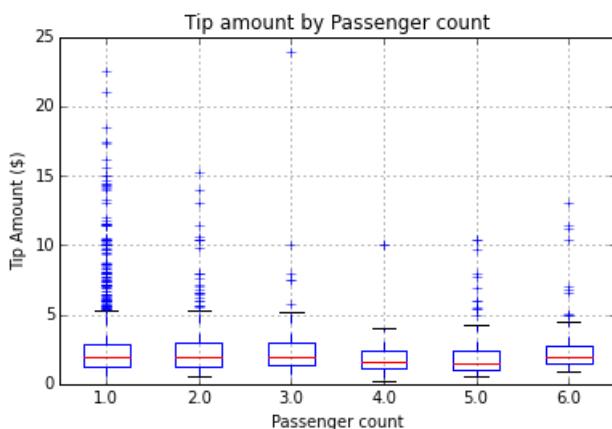
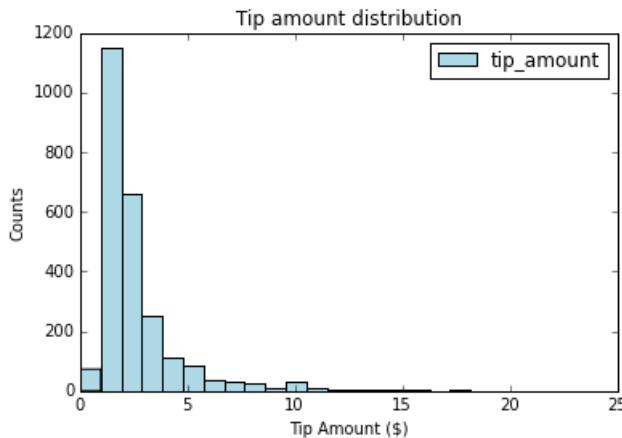
# PLOT TIP BY PAYMENT TYPE AND PASSENGER COUNT
ax1 = sqlResults[['tip_amount']].plot(kind='hist', bins=25, facecolor='lightblue')
ax1.set_title('Tip amount distribution')
ax1.set_xlabel('Tip Amount ($)')
ax1.set_ylabel('Counts')
plt.suptitle('')
plt.show()

# PLOT TIP BY PASSENGER COUNT
ax2 = sqlResults.boxplot(column=['tip_amount'], by=['passenger_count'])
ax2.set_title('Tip amount by Passenger count')
ax2.set_xlabel('Passenger count')
ax2.set_ylabel('Tip Amount ($)')
plt.suptitle('')
plt.show()

# PLOT TIP AMOUNT BY FARE AMOUNT; SCALE POINTS BY PASSENGER COUNT
ax = sqlResults.plot(kind='scatter', x= 'fare_amount', y = 'tip_amount', c='blue', alpha = 0.10, s=5*(sqlResults.passenger_count))
ax.set_title('Tip amount by Fare amount')
ax.set_xlabel('Fare Amount ($)')
ax.set_ylabel('Tip Amount ($)')
plt.axis([-2, 80, -2, 20])
plt.show()

```

Output:





Create features and transform features, and then prep data for input into modeling functions

For tree-based modeling functions from Spark ML and MLLib, you have to prepare target and features by using a variety of techniques, such as binning, indexing, one-hot encoding, and vectorization. Here are the procedures to follow in this section:

1. Create a new feature by **binning** hours into traffic time buckets.
2. Apply **indexing and one-hot encoding** to categorical features.
3. **Sample and split the data set** into training and test fractions.
4. **Specify training variable and features**, and then create indexed or one-hot encoded training and testing input labeled point resilient distributed datasets (RDDs) or data frames.
5. Automatically **categorize and vectorize features and targets** to use as inputs for machine learning models.

Create a new feature by binning hours into traffic time buckets

This code shows you how to create a new feature by binning hours into traffic time buckets and how to cache the resulting data frame in memory. Where RDDs and data frames are used repeatedly, caching leads to improved execution times. Accordingly, you'll cache RDDs and data frames at several stages in the following procedures.

```
# CREATE FOUR BUCKETS FOR TRAFFIC TIMES
val sqlStatement = """
    SELECT *,
    CASE
        WHEN (pickup_hour <= 6 OR pickup_hour >= 20) THEN "Night"
        WHEN (pickup_hour >= 7 AND pickup_hour <= 10) THEN "AMRush"
        WHEN (pickup_hour >= 11 AND pickup_hour <= 15) THEN "Afternoon"
        WHEN (pickup_hour >= 16 AND pickup_hour <= 19) THEN "PMRush"
    END as TrafficTimeBins
    FROM taxi_train
"""
val taxi_df_train_with_newFeatures = sqlContext.sql(sqlStatement)

# CACHE THE DATA FRAME IN MEMORY AND MATERIALIZE THE DATA FRAME IN MEMORY
taxi_df_train_with_newFeatures.cache()
taxi_df_train_with_newFeatures.count()
```

Indexing and one-hot encoding of categorical features

The modeling and predict functions of MLLib require features with categorical input data to be indexed or encoded prior to use. This section shows you how to index or encode categorical features for input into the modeling functions.

You need to index or encode your models in different ways, depending on the model. For example, logistic and linear regression models require one-hot encoding. For example, a feature with three categories can be expanded

into three feature columns. Each column would contain 0 or 1 depending on the category of an observation. MLlib provides the [OneHotEncoder](#) function for one-hot encoding. This encoder maps a column of label indices to a column of binary vectors with at most a single one-value. With this encoding, algorithms that expect numerical valued features, such as logistic regression, can be applied to categorical features.

Here you transform only four variables to show examples, which are character strings. You also can index other variables, such as weekday, represented by numerical values, as categorical variables.

For indexing, use `StringIndexer()`, and for one-hot encoding, use `OneHotEncoder()` functions from MLlib. Here is the code to index and encode categorical features:

```
# CREATE INDEXES AND ONE-HOT ENCODED VECTORS FOR SEVERAL CATEGORICAL FEATURES

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# INDEX AND ENCODE VENDOR_ID
val stringIndexer = new StringIndexer().setInputCol("vendor_id").setOutputCol("vendorIndex").fit(taxi_df_train_with_newFeatures)
val indexed = stringIndexer.transform(taxi_df_train_with_newFeatures)
val encoder = new OneHotEncoder().setInputCol("vendorIndex").setOutputCol("vendorVec")
val encoded1 = encoder.transform(indexed)

# INDEX AND ENCODE RATE_CODE
val stringIndexer = new StringIndexer().setInputCol("rate_code").setOutputCol("rateIndex").fit(encoded1)
val indexed = stringIndexer.transform(encoded1)
val encoder = new OneHotEncoder().setInputCol("rateIndex").setOutputCol("rateVec")
val encoded2 = encoder.transform(indexed)

# INDEX AND ENCODE PAYMENT_TYPE
val stringIndexer = new StringIndexer().setInputCol("payment_type").setOutputCol("paymentIndex").fit(encoded2)
val indexed = stringIndexer.transform(encoded2)
val encoder = new OneHotEncoder().setInputCol("paymentIndex").setOutputCol("paymentVec")
val encoded3 = encoder.transform(indexed)

# INDEX AND TRAFFIC TIME BINS
val stringIndexer = new StringIndexer().setInputCol("TrafficTimeBins").setOutputCol("TrafficTimeBinsIndex").fit(encoded3)
val indexed = stringIndexer.transform(encoded3)
val encoder = new OneHotEncoder().setInputCol("TrafficTimeBinsIndex").setOutputCol("TrafficTimeBinsVec")
val encodedFinal = encoder.transform(indexed)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");
```

Output:

Time to run the cell: 4 seconds.

Sample and split the data set into training and test fractions

This code creates a random sampling of the data (25%, in this example). Although sampling is not required for this example due to the size of the data set, the article shows you how you can sample so that you know how to use it for your own problems when needed. When samples are large, this can save significant time while you train models. Next, split the sample into a training part (75%, in this example) and a testing part (25%, in this example) to use in classification and regression modeling.

Add a random number (between 0 and 1) to each row (in a "rand" column) that can be used to select cross-validation folds during training.

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# SPECIFY SAMPLING AND SPLITTING FRACTIONS
val samplingFraction = 0.25;
val trainingFraction = 0.75;
val testingFraction = (1-trainingFraction);
val seed = 1234;
val encodedFinalSampledTmp = encodedFinal.sample(withReplacement = false, fraction = samplingFraction, seed = seed)
val sampledDFcount = encodedFinalSampledTmp.count().toInt

val generateRandomDouble = udf(() => {
    scala.util.Random.nextDouble
})

# ADD A RANDOM NUMBER FOR CROSS-VALIDATION
val encodedFinalSampled = encodedFinalSampledTmp.withColumn("rand", generateRandomDouble());

# SPLIT THE SAMPLED DATA FRAME INTO TRAIN AND TEST, WITH A RANDOM COLUMN ADDED FOR DOING CROSS-VALIDATION
# (SHOWN LATER)
# INCLUDE A RANDOM COLUMN FOR CREATING CROSS-VALIDATION FOLDS
val splits = encodedFinalSampled.randomSplit(Array(trainingFraction, testingFraction), seed = seed)
val trainData = splits(0)
val testData = splits(1)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

```

Output:

Time to run the cell: 2 seconds.

Specify training variable and features, and then create indexed or one-hot encoded training and testing input labeled point RDDs or data frames

This section contains code that shows you how to index categorical text data as a labeled point data type, and encode it so you can use it to train and test MLlib logistic regression and other classification models. Labeled point objects are RDDs that are formatted in a way that is needed as input data by most of machine learning algorithms in MLlib. A [labeled point](#) is a local vector, either dense or sparse, associated with a label/response.

In this code, you specify the target (dependent) variable and the features to use to train models. Then, you create indexed or one-hot encoded training and testing input labeled point RDDs or data frames.

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# MAP NAMES OF FEATURES AND TARGETS FOR CLASSIFICATION AND REGRESSION PROBLEMS
val featuresIndOneHot = List("paymentVec", "vendorVec", "rateVec", "TrafficTimeBinsVec", "pickup_hour",
"weekday", "passenger_count", "trip_time_in_secs", "trip_distance",
"fare_amount").map(encodedFinalSampled.columns.indexOf(_))
val featuresIndIndex = List("paymentIndex", "vendorIndex", "rateIndex", "TrafficTimeBinsIndex", "pickup_hour",
"weekday", "passenger_count", "trip_time_in_secs", "trip_distance",
"fare_amount").map(encodedFinalSampled.columns.indexOf(_))

# SPECIFY THE TARGET FOR CLASSIFICATION ('tipped') AND REGRESSION ('tip_amount') PROBLEMS
val targetIndBinary = List("tipped").map(encodedFinalSampled.columns.indexOf(_))
val targetIndRegression = List("tip_amount").map(encodedFinalSampled.columns.indexOf(_))

# CREATE INDEXED LABELED POINT RDD OBJECTS
val indexedTRAINbinary = trainData.rdd.map(r => LabeledPoint(r.getDouble(targetIndBinary(0).toInt),
Vectors.dense(featuresIndIndex.map(r.getDouble(_).toArray))))
val indexedTESTbinary = testData.rdd.map(r => LabeledPoint(r.getDouble(targetIndBinary(0).toInt),
Vectors.dense(featuresIndIndex.map(r.getDouble(_).toArray))))
val indexedTRAINreg = trainData.rdd.map(r => LabeledPoint(r.getDouble(targetIndRegression(0).toInt),
Vectors.dense(featuresIndIndex.map(r.getDouble(_).toArray))))
val indexedTESTreg = testData.rdd.map(r => LabeledPoint(r.getDouble(targetIndRegression(0).toInt),
Vectors.dense(featuresIndIndex.map(r.getDouble(_).toArray))))

# CREATE INDEXED DATA FRAMES THAT YOU CAN USE TO TRAIN BY USING SPARK ML FUNCTIONS
val indexedTRAINbinaryDF = indexedTRAINbinary.toDF()
val indexedTESTbinaryDF = indexedTESTbinary.toDF()
val indexedTRAINregDF = indexedTRAINreg.toDF()
val indexedTESTregDF = indexedTESTreg.toDF()

# CREATE ONE-HOT ENCODED (VECTORIZED) DATA FRAMES THAT YOU CAN USE TO TRAIN BY USING SPARK ML FUNCTIONS
val assemblerOneHot = new VectorAssembler().setInputCols(Array("paymentVec", "vendorVec", "rateVec",
"TrafficTimeBinsVec", "pickup_hour", "weekday", "passenger_count", "trip_time_in_secs", "trip_distance",
"fare_amount")).setOutputCol("features")
val OneHotTRAIN = assemblerOneHot.transform(trainData)
val OneHotTEST = assemblerOneHot.transform(testData)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime()) / 1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

```

Output:

Time to run the cell: 4 seconds.

Automatically categorize and vectorize features and targets to use as inputs for machine learning models

Use Spark ML to categorize the target and features to use in tree-based modeling functions. The code completes two tasks:

- Creates a binary target for classification by assigning a value of 0 or 1 to each data point between 0 and 1 by using a threshold value of 0.5.
- Automatically categorizes features. If the number of distinct numerical values for any feature is less than 32, that feature is categorized.

Here's the code for these two tasks.

```

# CATEGORIZE FEATURES AND BINARIZE THE TARGET FOR THE BINARY CLASSIFICATION PROBLEM

# TRAIN DATA
val indexer = new VectorIndexer().setInputCol("features").setOutputCol("featuresCat").setMaxCategories(32)
val indexerModel = indexer.fit(indexedTRAINbinaryDF)
val indexedTrainwithCatFeat = indexerModel.transform(indexedTRAINbinaryDF)
val binarizer: Binarizer = new Binarizer().setInputCol("label").setOutputCol("labelBin").setThreshold(0.5)
val indexedTRAINwithCatFeatBinTarget = binarizer.transform(indexedTrainwithCatFeat)

# TEST DATA
val indexerModel = indexer.fit(indexedTESTbinaryDF)
val indexedTrainwithCatFeat = indexerModel.transform(indexedTESTbinaryDF)
val binarizer: Binarizer = new Binarizer().setInputCol("label").setOutputCol("labelBin").setThreshold(0.5)
val indexedTESTwithCatFeatBinTarget = binarizer.transform(indexedTrainwithCatFeat)

# CATEGORIZE FEATURES FOR THE REGRESSION PROBLEM
# CREATE PROPERLY INDEXED AND CATEGORIZED DATA FRAMES FOR TREE-BASED MODELS

# TRAIN DATA
val indexer = new VectorIndexer().setInputCol("features").setOutputCol("featuresCat").setMaxCategories(32)
val indexerModel = indexer.fit(indexedTRAINregDF)
val indexedTRAINwithCatFeat = indexerModel.transform(indexedTRAINregDF)

# TEST DATA
val indexerModel = indexer.fit(indexedTESTbinaryDF)
val indexedTESTwithCatFeat = indexerModel.transform(indexedTESTregDF)

```

Binary classification model: Predict whether a tip should be paid

In this section, you create three types of binary classification models to predict whether or not a tip should be paid:

- A **logistic regression model** by using the Spark ML `LogisticRegression()` function
- A **random forest classification model** by using the Spark ML `RandomForestClassifier()` function
- A **gradient boosting tree classification model** by using the MLLib `GradientBoostedTrees()` function

Create a logistic regression model

Next, create a logistic regression model by using the Spark ML `LogisticRegression()` function. You create the model building code in a series of steps:

1. **Train the model** data with one parameter set.
2. **Evaluate the model** on a test data set with metrics.
3. **Save the model** in Blob storage for future consumption.
4. **Score the model** against test data.
5. **Plot the results** with receiver operating characteristic (ROC) curves.

Here's the code for these procedures:

```

# CREATE A LOGISTIC REGRESSION MODEL
val lr = new
LogisticRegression().setLabelCol("tipped").setFeaturesCol("features").setMaxIter(10).setRegParam(0.3).setElast
icNetParam(0.8)
val lrModel = lr.fit(OneHotTRAIN)

# PREDICT ON THE TEST DATA SET
val predictions = lrModel.transform(OneHotTEST)

# SELECT `BinaryClassificationEvaluator()` TO COMPUTE THE TEST ERROR
val evaluator = new
BinaryClassificationEvaluator().setLabelCol("tipped").setRawPredictionCol("probability").setMetricName("areaUn
derROC")
val ROC = evaluator.evaluate(predictions)
println("ROC on test data = " + ROC)

# SAVE THE MODEL
val timestamp = Calendar.getInstance().getTime().toString.replaceAll(" ", ".").replaceAll(":", "_");
val modelName = "LogisticRegression_"
val filename = modelDir.concat(modelName).concat(timestamp)
lrModel.save(filename);

```

Load, score, and save the results.

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# LOAD THE SAVED MODEL AND SCORE THE TEST DATA SET
val savedModel = org.apache.spark.ml.classification.LogisticRegressionModel.load(filename)
println(s"Coefficients: ${savedModel.coefficients} Intercept: ${savedModel.intercept}")

# SCORE THE MODEL ON THE TEST DATA
val predictions = savedModel.transform(OneHotTEST).select("tipped","probability","rawPrediction")
predictions.registerTempTable("testResults")

# SELECT `BinaryClassificationEvaluator()` TO COMPUTE THE TEST ERROR
val evaluator = new
BinaryClassificationEvaluator().setLabelCol("tipped").setRawPredictionCol("probability").setMetricName("areaUn
derROC")
val ROC = evaluator.evaluate(predictions)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.")

# PRINT THE ROC RESULTS
println("ROC on test data = " + ROC)

```

Output:

ROC on test data = 0.9827381497557599

Use Python on local Pandas data frames to plot the ROC curve.

```

# QUERY THE RESULTS
%%sql -q -o sqlResults
SELECT tipped, probability from testResults

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
%matplotlib inline
from sklearn.metrics import roc_curve,auc

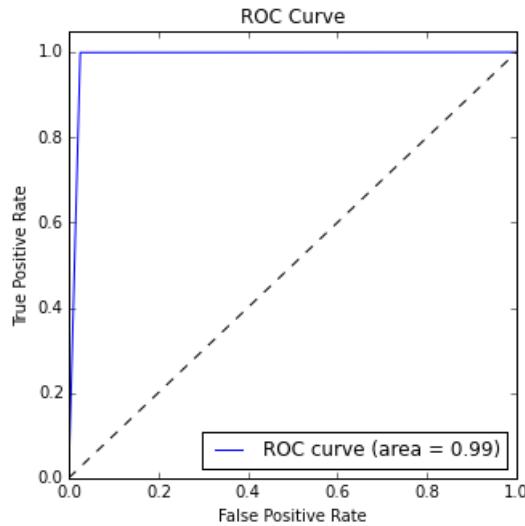
sqlResults['probFloat'] = sqlResults.apply(lambda row: row['probability'].values()[0][1], axis=1)
predictions_pddf = sqlResults[["tipped","probFloat"]]

# PREDICT THE ROC CURVE
# predictions_pddf = sqlResults.rename(columns={'_1': 'probability', 'tipped': 'label'})
prob = predictions_pddf["probFloat"]
fpr, tpr, thresholds = roc_curve(predictions_pddf['tipped'], prob, pos_label=1);
roc_auc = auc(fpr, tpr)

# PLOT THE ROC CURVE
plt.figure(figsize=(5,5))
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()

```

Output:



Create a random forest classification model

Next, create a random forest classification model by using the Spark ML `RandomForestClassifier()` function, and then evaluate the model on test data.

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# CREATE THE RANDOM FOREST CLASSIFIER MODEL
val rf = new
RandomForestClassifier().setLabelCol("labelBin").setFeaturesCol("featuresCat").setNumTrees(10).setSeed(1234)

# FIT THE MODEL
val rfModel = rf.fit(indexedTRAINwithCatFeatBinTarget)
val predictions = rfModel.transform(indexedTESTwithCatFeatBinTarget)

# EVALUATE THE MODEL
val evaluator = new
MulticlassClassificationEvaluator().setLabelCol("label").setPredictionCol("prediction").setMetricName("f1")
val Test_f1Score = evaluator.evaluate(predictions)
println("F1 score on test data: " + Test_f1Score);

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

# CALCULATE BINARY CLASSIFICATION EVALUATION METRICS
val evaluator = new
BinaryClassificationEvaluator().setLabelCol("label").setRawPredictionCol("probability").setMetricName("areaUnderROC")
val ROC = evaluator.evaluate(predictions)
println("ROC on test data = " + ROC)

```

Output:

ROC on test data = 0.9847103571552683

Create a GBT classification model

Next, create a GBT classification model by using MLlib's `GradientBoostedTrees()` function, and then evaluate the model on test data.

```

# TRAIN A GBT CLASSIFICATION MODEL BY USING MLLIB AND A LABELED POINT

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# DEFINE THE GBT CLASSIFICATION MODEL
val boostingStrategy = BoostingStrategy.defaultParams("Classification")
boostingStrategy.numIterations = 20
boostingStrategy.treeStrategy.numClasses = 2
boostingStrategy.treeStrategy.maxDepth = 5
boostingStrategy.treeStrategy.categoricalFeaturesInfo = Map[Int, Int]((0,2),(1,2),(2,6),(3,4))

# TRAIN THE MODEL
val gbtModel = GradientBoostedTrees.train(indexedTRAINbinary, boostingStrategy)

# SAVE THE MODEL IN BLOB STORAGE
val datestamp = Calendar.getInstance().getTime().toString.replaceAll(" ", ".").replaceAll(":", "_");
val modelName = "GBT_Classification_"
val filename = modelDir.concat(modelName).concat(datestamp)
gbtModel.save(sc, filename);

# EVALUATE THE MODEL ON TEST INSTANCES AND THE COMPUTE TEST ERROR
val labelAndPreds = indexedTESTbinary.map { point =>
    val prediction = gbtModel.predict(point.features)
    (point.label, prediction)
}
val testErr = labelAndPreds.filter(r => r._1 != r._2).count.toDouble / indexedTRAINbinary.count()
//println("Learned classification GBT model:\n" + gbtModel.toDebugString)
println("Test Error = " + testErr)

# USE BINARY AND MULTICLASS METRICS TO EVALUATE THE MODEL ON THE TEST DATA
val metrics = new MulticlassMetrics(labelAndPreds)
println(s"Precision: ${metrics.precision}")
println(s"Recall: ${metrics.recall}")
println(s"F1 Score: ${metrics.fMeasure}")

val metrics = new BinaryClassificationMetrics(labelAndPreds)
println(s"Area under PR curve: ${metrics.areaUnderPR}")
println(s"Area under ROC curve: ${metrics.areaUnderROC}")

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

# PRINT THE ROC METRIC
println(s"Area under ROC curve: ${metrics.areaUnderROC}")

```

Output:

Area under ROC curve: 0.9846895479241554

Regression model: Predict tip amount

In this section, you create two types of regression models to predict the tip amount:

- A **regularized linear regression model** by using the Spark ML `LinearRegression()` function. You'll save the model and evaluate the model on test data.
- A **gradient-boosting tree regression model** by using the Spark ML `GBTRegressor()` function.

Create a regularized linear regression model

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# CREATE A REGULARIZED LINEAR REGRESSION MODEL BY USING THE SPARK ML FUNCTION AND DATA FRAMES
val lr = new
LinearRegression().setLabelCol("tip_amount").setFeaturesCol("features").setMaxIter(10).setRegParam(0.3).setEla
sticNetParam(0.8)

# FIT THE MODEL BY USING DATA FRAMES
val lrModel = lr.fit(OneHotTRAIN)
println(s"Coefficients: ${lrModel.coefficients} Intercept: ${lrModel.intercept}")

# SUMMARIZE THE MODEL OVER THE TRAINING SET AND PRINT METRICS
val trainingSummary = lrModel.summary
println(s"numIterations: ${trainingSummary.totalIterations}")
println(s"objectiveHistory: ${trainingSummary.objectiveHistory.toList}")
trainingSummary.residuals.show()
println(s"RMSE: ${trainingSummary.rootMeanSquaredError}")
println(s"r2: ${trainingSummary.r2}")

# SAVE THE MODEL IN AZURE BLOB STORAGE
val timestamp = Calendar.getInstance().getTime().toString.replaceAll(" ", ".").replaceAll(":", "_");
val modelName = "LinearRegression_"
val filename = modelDir.concat(modelName).concat(timestamp)
lrModel.save(filename);

# PRINT THE COEFFICIENTS
println(s"Coefficients: ${lrModel.coefficients} Intercept: ${lrModel.intercept}")

# SCORE THE MODEL ON TEST DATA
val predictions = lrModel.transform(OneHotTEST)

# EVALUATE THE MODEL ON TEST DATA
val evaluator = new
RegressionEvaluator().setLabelCol("tip_amount").setPredictionCol("prediction").setMetricName("r2")
val r2 = evaluator.evaluate(predictions)
println("R-sqr on test data = " + r2)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

```

Output:

Time to run the cell: 13 seconds.

```

# LOAD A SAVED LINEAR REGRESSION MODEL FROM BLOB STORAGE AND SCORE A TEST DATA SET

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# LOAD A SAVED LINEAR REGRESSION MODEL FROM AZURE BLOB STORAGE
val savedModel = org.apache.spark.ml.regression.LinearRegressionModel.load(filename)
println(s"Coefficients: ${savedModel.coefficients} Intercept: ${savedModel.intercept}")

# SCORE THE MODEL ON TEST DATA
val predictions = savedModel.transform(OneHotTEST).select("tip_amount","prediction")
predictions.registerTempTable("testResults")

# EVALUATE THE MODEL ON TEST DATA
val evaluator = new
RegressionEvaluator().setLabelCol("tip_amount").setPredictionCol("prediction").setMetricName("r2")
val r2 = evaluator.evaluate(predictions)
println("R-sqr on test data = " + r2)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.")

# PRINT THE RESULTS
println("R-sqr on test data = " + r2)

```

Output:

R-sqr on test data = 0.5960320470835743

Next, query the test results as a data frame and use AutoVizWidget and matplotlib to visualize it.

```

# RUN A SQL QUERY
%%sql -q -o sqlResults
select * from testResults

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER
%%local

# USE THE JUPYTER AUTO-PLOTTING FEATURE TO CREATE INTERACTIVE FIGURES
# CLICK THE TYPE OF PLOT TO GENERATE (LINE, AREA, BAR, AND SO ON)
sqlResults

```

The code creates a local data frame from the query output and plots the data. The `%%local` magic creates a local data frame, `sqlResults`, which you can use to plot with matplotlib.

NOTE

This Spark magic is used multiple times in this article. If the amount of data is large, you should sample to create a data frame that can fit in local memory.

Create plots by using Python matplotlib.

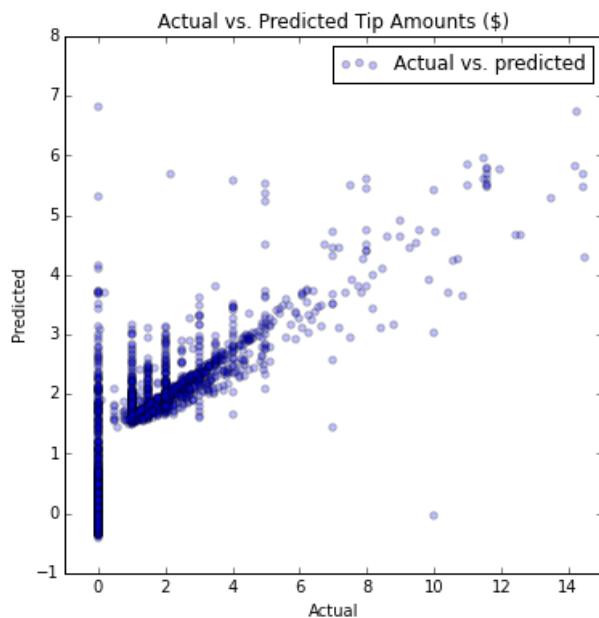
```

# RUN THE CODE LOCALLY ON THE JUPYTER SERVER AND IMPORT LIBRARIES
%%local
sqlResults
%matplotlib inline
import numpy as np

# PLOT THE RESULTS
ax = sqlResults.plot(kind='scatter', figsize = (6,6), x='tip_amount', y='prediction', color='blue', alpha = 0.25, label='Actual vs. predicted');
fit = np.polyfit(sqlResults['tip_amount'], sqlResults['prediction'], deg=1)
ax.set_title('Actual vs. Predicted Tip Amounts ($)')
ax.set_xlabel("Actual")
ax.set_ylabel("Predicted")
#ax.plot(sqlResults['tip_amount'], fit[0] * sqlResults['prediction'] + fit[1], color='magenta')
plt.axis([-1, 15, -1, 8])
plt.show(ax)

```

Output:



Create a GBT regression model

Create a GBT regression model by using the Spark ML `GBTRegressor()` function, and then evaluate the model on test data.

[Gradient-boosted trees](#) (GBTs) are ensembles of decision trees. GBTs train decision trees iteratively to minimize a loss function. You can use GBTs for regression and classification. They can handle categorical features, do not require feature scaling, and can capture nonlinearities and feature interactions. You also can use them in a multiclass-classification setting.

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# TRAIN A GBT REGRESSION MODEL
val gbt = new GBTRegressor().setLabelCol("label").setFeaturesCol("featuresCat").setMaxIter(10)
val gbtModel = gbt.fit(indexedTRAINwithCatFeat)

# MAKE PREDICTIONS
val predictions = gbtModel.transform(indexedTESTwithCatFeat)

# COMPUTE TEST SET R2
val evaluator = new
RegressionEvaluator().setLabelCol("label").setPredictionCol("prediction").setMetricName("r2")
val Test_R2 = evaluator.evaluate(predictions)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.")

# PRINT THE RESULTS
println("Test R-sqr is: " + Test_R2);

```

Output:

Test R-sqr is: 0.7655383534596654

Advanced modeling utilities for optimization

In this section, you use machine learning utilities that developers frequently use for model optimization. Specifically, you can optimize machine learning models three different ways by using parameter sweeping and cross-validation:

- Split the data into train and validation sets, optimize the model by using hyper-parameter sweeping on a training set, and evaluate on a validation set (linear regression)
- Optimize the model by using cross-validation and hyper-parameter sweeping by using Spark ML's CrossValidator function (binary classification)
- Optimize the model by using custom cross-validation and parameter-sweeping code to use any machine learning function and parameter set (linear regression)

Cross-validation is a technique that assesses how well a model trained on a known set of data will generalize to predict the features of data sets on which it has not been trained. The general idea behind this technique is that a model is trained on a data set of known data, and then the accuracy of its predictions is tested against an independent data set. A common implementation is to divide a data set into k -folds, and then train the model in a round-robin fashion on all but one of the folds.

Hyper-parameter optimization is the problem of choosing a set of hyper-parameters for a learning algorithm, usually with the goal of optimizing a measure of the algorithm's performance on an independent data set. A hyper-parameter is a value that you must specify outside the model training procedure. Assumptions about hyper-parameter values can affect the flexibility and accuracy of the model. Decision trees have hyper-parameters, for example, such as the desired depth and number of leaves in the tree. You must set a misclassification penalty term for a support vector machine (SVM).

A common way to perform hyper-parameter optimization is to use a grid search, also called a **parameter sweep**. In a grid search, an exhaustive search is performed through the values of a specified subset of the hyper-parameter space for a learning algorithm. Cross-validation can supply a performance metric to sort out the optimal results produced by the grid search algorithm. If you use cross-validation hyper-parameter sweeping, you

can help limit problems like overfitting a model to training data. This way, the model retains the capacity to apply to the general set of data from which the training data was extracted.

Optimize a linear regression model with hyper-parameter sweeping

Next, split data into train and validation sets, use hyper-parameter sweeping on a training set to optimize the model, and evaluate on a validation set (linear regression).

```
# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# RENAME `tip_amount` AS A LABEL
val OneHotTRAINLabeled =
OneHotTRAIN.select("tip_amount","features").withColumnRenamed(existingName="tip_amount",newName="label")
val OneHotTESTLabeled =
OneHotTEST.select("tip_amount","features") .withColumnRenamed(existingName="tip_amount",newName="label")
OneHotTRAINLabeled.cache()
OneHotTESTLabeled.cache()

# DEFINE THE ESTIMATOR FUNCTION: `THE LinearRegression()` FUNCTION
val lr = new LinearRegression().setLabelCol("label").setFeaturesCol("features").setMaxIter(10)

# DEFINE THE PARAMETER GRID
val paramGrid = new ParamGridBuilder().addGrid(lr.regParam, Array(0.1, 0.01,
0.001)).addGrid(lr.fitIntercept).addGrid(lr.elasticNetParam, Array(0.1, 0.5, 0.9)).build()

# DEFINE THE PIPELINE WITH A TRAIN/TEST VALIDATION SPLIT (75% IN THE TRAINING SET), AND THEN THE SPECIFY
ESTIMATOR, EVALUATOR, AND PARAMETER GRID
val trainPct = 0.75
val trainValidationSplit = new TrainValidationSplit().setEstimator(lr).setEvaluator(new
RegressionEvaluator).setEstimatorParamMaps(paramGrid).setTrainRatio(trainPct)

# RUN THE TRAIN VALIDATION SPLIT AND CHOOSE THE BEST SET OF PARAMETERS
val model = trainValidationSplit.fit(OneHotTRAINLabeled)

# MAKE PREDICTIONS ON THE TEST DATA BY USING THE MODEL WITH THE COMBINATION OF PARAMETERS THAT PERFORMS THE
BEST
val testResults = model.transform(OneHotTESTLabeled).select("label", "prediction")

# COMPUTE TEST SET R2
val evaluator = new
RegressionEvaluator().setLabelCol("label").setPredictionCol("prediction").setMetricName("r2")
val Test_R2 = evaluator.evaluate(testResults)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.")

println("Test R-sqr is: " + Test_R2);
```

Output:

Test R-sqr is: 0.6226484708501209

Optimize the binary classification model by using cross-validation and hyper-parameter sweeping

This section shows you how to optimize a binary classification model by using cross-validation and hyper-parameter sweeping. This uses the Spark ML `CrossValidator` function.

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# CREATE DATA FRAMES WITH PROPERLY LABELED COLUMNS TO USE WITH THE TRAIN AND TEST SPLIT
val indexedTRAINwithCatFeatBinTargetRF =
indexedTRAINwithCatFeatBinTarget.select("labelBin","featuresCat").withColumnRenamed(existingName="labelBin",new
wName="label").withColumnRenamed(existingName="featuresCat",newName="features")
val indexedTESTwithCatFeatBinTargetRF =
indexedTESTwithCatFeatBinTarget.select("labelBin","featuresCat").withColumnRenamed(existingName="labelBin",new
Name="label").withColumnRenamed(existingName="featuresCat",newName="features")
indexedTRAINwithCatFeatBinTargetRF.cache()
indexedTESTwithCatFeatBinTargetRF.cache()

# DEFINE THE ESTIMATOR FUNCTION
val rf = new
RandomForestClassifier().setLabelCol("label").setFeaturesCol("features").setImpurity("gini").setSeed(1234).set
FeatureSubsetStrategy("auto").setMaxBins(32)

# DEFINE THE PARAMETER GRID
val paramGrid = new ParamGridBuilder().addGrid(rf.maxDepth, Array(4,8)).addGrid(rf.numTrees,
Array(5,10)).addGrid(rf.minInstancesPerNode, Array(100,300)).build()

# SPECIFY THE NUMBER OF FOLDS
val numFolds = 3

# DEFINE THE TRAIN/TEST VALIDATION SPLIT (75% IN THE TRAINING SET)
val CrossValidator = new CrossValidator().setEstimator(rf).setEvaluator(new
BinaryClassificationEvaluator).setEstimatorParamMaps(paramGrid).setNumFolds(numFolds)

# RUN THE TRAIN VALIDATION SPLIT AND CHOOSE THE BEST SET OF PARAMETERS
val model = CrossValidator.fit(indexedTRAINwithCatFeatBinTargetRF)

# MAKE PREDICTIONS ON THE TEST DATA BY USING THE MODEL WITH THE COMBINATION OF PARAMETERS THAT PERFORMS THE
BEST
val testResults = model.transform(indexedTESTwithCatFeatBinTargetRF).select("label", "prediction")

# COMPUTE THE TEST F1 SCORE
val evaluator = new
MulticlassClassificationEvaluator().setLabelCol("label").setPredictionCol("prediction").setMetricName("f1")
val Test_f1Score = evaluator.evaluate(testResults)

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

```

Output:

Time to run the cell: 33 seconds.

Optimize the linear regression model by using custom cross-validation and parameter-sweeping code

Next, optimize the model by using custom code, and identify the best model parameters by using the criterion of highest accuracy. Then, create the final model, evaluate the model on test data, and save the model in Blob storage. Finally, load the model, score test data, and evaluate accuracy.

```

# RECORD THE START TIME
val starttime = Calendar.getInstance().getTime()

# DEFINE THE PARAMETER GRID AND THE NUMBER OF FOLDS
val paramGrid = new ParamGridBuilder().addGrid(rf.maxDepth, Array(5,10)).addGrid(rf.numTrees,
Array(10,25,50)).build()

val nFolds = 3
val numModels = paramGrid.size
val numParamsinGrid = 2

```

```

# SPECIFY THE NUMBER OF CATEGORIES FOR CATEGORICAL VARIABLES
val categoricalFeaturesInfo = Map[Int, Int]((0,2),(1,2),(2,6),(3,4))

var maxDepth = -1
var numTrees = -1
var param = ""
var paramval = -1
var validateLB = -1.0
var validateUB = -1.0
val h = 1.0 / nFolds;
val RMSE = Array.fill(numModels)(0.0)

# CREATE K-FOLDS
val splits = MLUtils.kFold(indexedTRAINbinary, numFolds = nFolds, seed=1234)

# LOOP THROUGH K-FOLDS AND THE PARAMETER GRID TO GET AND IDENTIFY THE BEST PARAMETER SET BY LEVEL OF ACCURACY
for (i <- 0 to (nFolds-1)) {
    validateLB = i * h
    validateUB = (i + 1) * h
    val validationCV = trainData.filter($"rand" >= validateLB && $"rand" < validateUB)
    val trainCV = trainData.filter($"rand" < validateLB || $"rand" >= validateUB)
    val validationLabPt = validationCV.rdd.map(r => LabeledPoint(r.getDouble(targetIndRegression(0).toInt),
    Vectors.dense(featuresIndIndex.map(r.getDouble(_)).toArray)));
    val trainCVLabPt = trainCV.rdd.map(r => LabeledPoint(r.getDouble(targetIndRegression(0).toInt),
    Vectors.dense(featuresIndIndex.map(r.getDouble(_)).toArray)));
    validationLabPt.cache()
    trainCVLabPt.cache()

    for (nParamSets <- 0 to (numModels-1)) {
        for (nParams <- 0 to (numParamsinGrid-1)) {
            param = paramGrid(nParamSets).toSeq(nParams).param.toString.split("__")(1)
            paramval = paramGrid(nParamSets).toSeq(nParams).value.toString.toInt
            if (param == "maxDepth") {maxDepth = paramval}
            if (param == "numTrees") {numTrees = paramval}
        }
        val rfModel = RandomForest.trainRegressor(trainCVLabPt,
        categoricalFeaturesInfo=categoricalFeaturesInfo,
                                            numTrees=numTrees, maxDepth=maxDepth,
                                            featureSubsetStrategy="auto",impurity="variance",
        maxBins=32)
        val labelAndPreds = validationLabPt.map { point =>
            val prediction = rfModel.predict(point.features)
            ( prediction, point.label )
        }
        val validMetrics = new RegressionMetrics(labelAndPreds)
        val rmse = validMetrics.rootMeanSquaredError
        RMSE(nParamSets) += rmse
    }
    validationLabPt.unpersist();
    trainCVLabPt.unpersist();
}
val minRMSEindex = RMSE.indexOf(RMSE.min)

# GET THE BEST PARAMETERS FROM A CROSS-VALIDATION AND PARAMETER SWEEP
var best_maxDepth = -1
var best_numTrees = -1
for (nParams <- 0 to (numParamsinGrid-1)) {
    param = paramGrid(minRMSEindex).toSeq(nParams).param.toString.split("__")(1)
    paramval = paramGrid(minRMSEindex).toSeq(nParams).value.toString.toInt
    if (param == "maxDepth") {best_maxDepth = paramval}
    if (param == "numTrees") {best_numTrees = paramval}
}

# CREATE THE BEST MODEL WITH THE BEST PARAMETERS AND A FULL TRAINING DATA SET
val best_rfModel = RandomForest.trainRegressor(indexedTRAINreg,
categoricalFeaturesInfo=categoricalFeaturesInfo,
                                            numTrees=best_numTrees, maxDepth=best_maxDepth

```

```

    featureSubsetStrategy="auto",impurity="variance",
maxBins=32)

# SAVE THE BEST RANDOM FOREST MODEL IN BLOB STORAGE
val datestamp = Calendar.getInstance().getTime().toString.replaceAll(" ", ".").replaceAll(":", "_");
val modelName = "BestCV_RF_Regression_"
val filename = modelDir.concat(modelName).concat(datestamp)
best_rfModel.save(sc, filename);

# PREDICT ON THE TRAINING SET WITH THE BEST MODEL AND THEN EVALUATE
val labelAndPreds = indexedTESTreg.map { point =>
    val prediction = best_rfModel.predict(point.features)
    ( prediction, point.label )
}

val test_rmse = new RegressionMetrics(labelAndPreds).rootMeanSquaredError
val test_rsqr = new RegressionMetrics(labelAndPreds).r2

# GET THE TIME TO RUN THE CELL
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000).toString;
println("Time taken to run the above cell: " + elapsedtime + " seconds.");

# LOAD THE MODEL
val savedRFModel = RandomForestModel.load(sc, filename)

val labelAndPreds = indexedTESTreg.map { point =>
    val prediction = savedRFModel.predict(point.features)
    ( prediction, point.label )
}

# TEST THE MODEL
val test_rmse = new RegressionMetrics(labelAndPreds).rootMeanSquaredError
val test_rsqr = new RegressionMetrics(labelAndPreds).r2

```

Output:

Time to run the cell: 61 seconds.

Consume Spark-built machine learning models automatically with Scala

For an overview of topics that walk you through the tasks that comprise the Data Science process in Azure, see [Team Data Science Process](#).

[Team Data Science Process walkthroughs](#) describes other end-to-end walkthroughs that demonstrate the steps in the Team Data Science Process for specific scenarios. The walkthroughs also illustrate how to combine cloud and on-premises tools and services into a workflow or pipeline to create an intelligent application.

[Score Spark-built machine learning models](#) shows you how to use Scala code to automatically load and score new data sets with machine learning models built in Spark and saved in Azure Blob storage. You can follow the instructions provided there, and simply replace the Python code with Scala code in this article for automated consumption.

Feature engineering in data science

3/12/2019 • 7 minutes to read

This article explains the purposes of feature engineering and provides examples of its role in the data enhancement process of machine learning. The examples used to illustrate this process are drawn from Azure Machine Learning Studio.

This task is a step in the [Team Data Science Process \(TDSP\)](#).

Feature engineering attempts to increase the predictive power of learning algorithms by creating features from raw data that help facilitate the learning process. The engineering and selection of features is one part of the TDSP outlined in the [What is the Team Data Science Process lifecycle?](#) Feature engineering and selection are parts of the **Develop features** step of the TDSP.

- **feature engineering:** This process attempts to create additional relevant features from the existing raw features in the data, and to increase the predictive power of the learning algorithm.
- **feature selection:** This process selects the key subset of original data features in an attempt to reduce the dimensionality of the training problem.

Normally **feature engineering** is applied first to generate additional features, and then the **feature selection** step is performed to eliminate irrelevant, redundant, or highly correlated features.

The training data used in machine learning can often be enhanced by extraction of features from the raw data collected. An example of an engineered feature in the context of learning how to classify the images of handwritten characters is creation of a bit density map constructed from the raw bit distribution data. This map can help locate the edges of the characters more efficiently than simply using the raw distribution directly.

To create features for data in specific environments, see the following articles:

- [Create features for data in SQL Server](#)
- [Create features for data in a Hadoop cluster using Hive queries](#)

Try [Azure Machine Learning Studio](#), available in paid or free options.

Create features from your data - feature engineering

The training data consists of a matrix composed of examples (records or observations stored in rows), each of which has a set of features (variables or fields stored in columns). The features specified in the experimental design are expected to characterize the patterns in the data. Although many of the raw data fields can be directly included in the selected feature set used to train a model, it is often the case that additional (engineered) features need to be constructed from the features in the raw data to generate an enhanced training dataset.

What kind of features should be created to enhance the dataset when training a model? Engineered features that enhance the training provide information that better differentiates the patterns in the data. The new features are expected to provide additional information that is not clearly captured or easily apparent in the original or existing feature set. But this process is something of an art. Sound and productive decisions often require some domain expertise.

When starting with Azure Machine Learning, it is easiest to grasp this process concretely using samples provided in the Studio. Two examples are presented here:

- A regression example [Prediction of the number of bike rentals](#) in a supervised experiment where the target values are known

- A text mining classification example using [Feature Hashing](#)

Example 1: Add temporal features for a regression model

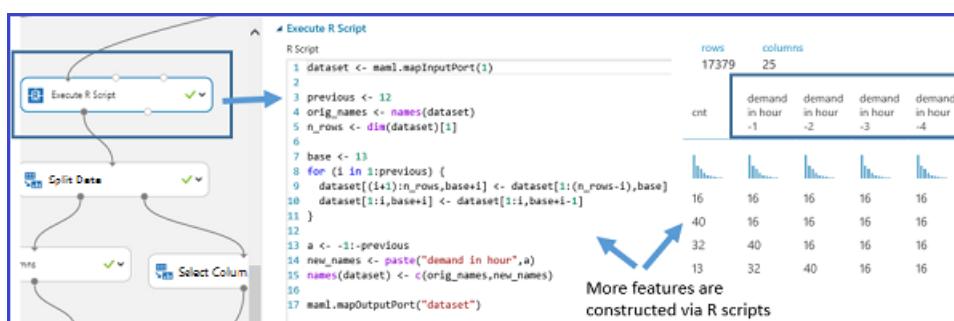
Let's use the experiment "Demand forecasting of bikes" in Azure Machine Learning Studio to demonstrate how to engineer features for a regression task. The objective of this experiment is to predict the demand for the bikes, that is, the number of bike rentals within a specific month/day/hour. The dataset "Bike Rental UCI dataset" is used as the raw input data. This dataset is based on real data from the Capital Bikeshare company that maintains a bike rental network in Washington DC in the United States. The dataset represents the number of bike rentals within a specific hour of a day in the years 2011 and year 2012 and contains 17379 rows and 17 columns. The raw feature set contains weather conditions (temperature/humidity/wind speed) and the type of the day (holiday/weekday). The field to predict is the "cnt" count, which represents the bike rentals within a specific hour and which ranges from 1 to 977.

With the goal of constructing effective features in the training data, four regression models are built using the same algorithm but with four different training datasets. The four datasets represent the same raw input data, but with an increasing number of features set. These features are grouped into four categories:

1. A = weather + holiday + weekday + weekend features for the predicted day
2. B = number of bikes that were rented in each of the previous 12 hours
3. C = number of bikes that were rented in each of the previous 12 days at the same hour
4. D = number of bikes that were rented in each of the previous 12 weeks at the same hour and the same day

Besides feature set A, which already exists in the original raw data, the other three sets of features are created through the feature engineering process. Feature set B captures very recent demand for the bikes. Feature set C captures the demand for bikes at a particular hour. Feature set D captures demand for bikes at particular hour and particular day of the week. The four training datasets each includes feature set A, A+B, A+B+C, and A+B+C+D, respectively.

In the Azure Machine Learning experiment, these four training datasets are formed via four branches from the pre-processed input dataset. Except the leftmost branch, each of these branches contains an [Execute R Script](#) module, in which the derived features (feature set B, C, and D) are respectively constructed and appended to the imported dataset. The following figure demonstrates the R script used to create feature set B in the second left branch.



A comparison of the performance results of the four models is summarized in the following table:

Features	Mean Absolute Error	Root Mean Square Error
A	89.7	124.9
A + B	51.7	88.3
A + B + C	47.6	81.1
A + B + C + D	48.3	82.1

The best results are shown by features A+B+C. Note that the error rate decreases when additional feature set are included in the training data. It verifies the presumption that the feature set B, C provide additional relevant information for the regression task. But adding the D feature does not seem to provide any additional reduction in

the error rate.

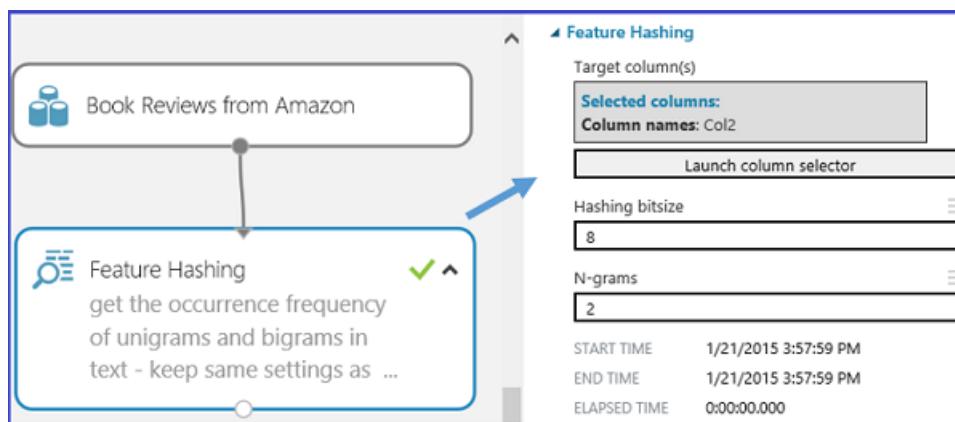
Example 2: Creating features in text mining

Feature engineering is widely applied in tasks related to text mining, such as document classification and sentiment analysis. For example, when you want to classify documents into several categories, a typical assumption is that the word/phrases included in one doc category are less likely to occur in another doc category. In other words, the frequency of the words/phrases distribution is able to characterize different document categories. In text mining applications, because individual pieces of text-contents usually serve as the input data, the feature engineering process is needed to create the features involving word/phrase frequencies.

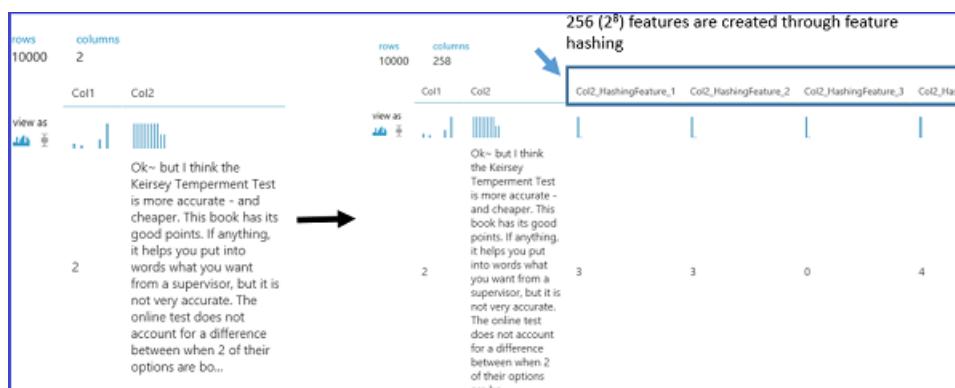
To achieve this task, a technique called **feature hashing** is applied to efficiently turn arbitrary text features into indices. Instead of associating each text feature (words/phrases) to a particular index, this method functions by applying a hash function to the features and using their hash values as indices directly.

In Azure Machine Learning, there is a [Feature Hashing](#) module that creates these word/phrase features conveniently. Following figure shows an example of using this module. The input dataset contains two columns: the book rating ranging from 1 to 5, and the actual review content. The goal of this [Feature Hashing](#) module is to retrieve a bunch of new features that show the occurrence frequency of the corresponding word(s)/phrase(s) within the particular book review. To use this module, complete the following steps:

- First, select the column that contains the input text ("Col2" in this example).
- Second, set the "Hashing bitsize" to 8, which means $2^8=256$ features will be created. The word/phase in all the text will be hashed to 256 indices. The parameter "Hashing bitsize" ranges from 1 to 31. The word(s)/phrase(s) are less likely to be hashed into the same index if setting it to be a larger number.
- Third, set the parameter "N-grams" to 2. This value gets the occurrence frequency of unigrams (a feature for every single word) and bigrams (a feature for every pair of adjacent words) from the input text. The parameter "N-grams" ranges from 0 to 10, which indicates the maximum number of sequential words to be included in a feature.



The following figure shows what these new feature look like.



Conclusion

Engineered and selected features increase the efficiency of the training process, which attempts to extract the key information contained in the data. They also improve the power of these models to classify the input data accurately and to predict outcomes of interest more robustly. Feature engineering and selection can also combine to make the learning more computationally tractable. It does so by enhancing and then reducing the number of features needed to calibrate or train a model. Mathematically speaking, the features selected to train the model are a minimal set of independent variables that explain the patterns in the data and then predict outcomes successfully.

It is not always necessarily to perform feature engineering or feature selection. Whether it is needed or not depends on the data to hand or collected, the algorithm selected, and the objective of the experiment.

Create features for data in SQL Server using SQL and Python

3/12/2019 • 5 minutes to read

This document shows how to generate features for data stored in a SQL Server VM on Azure that help algorithms learn more efficiently from the data. You can use SQL or a programming language like Python to accomplish this task. Both approaches are demonstrated here.

This task is a step in the [Team Data Science Process \(TDSP\)](#).

NOTE

For a practical example, you can consult the [NYC Taxi dataset](#) and refer to the IPNB titled [NYC Data wrangling using IPython Notebook and SQL Server](#) for an end-to-end walk-through.

Prerequisites

This article assumes that you have:

- Created an Azure storage account. If you need instructions, see [Create an Azure Storage account](#)
- Stored your data in SQL Server. If you have not, see [Move data to an Azure SQL Database for Azure Machine Learning](#) for instructions on how to move the data there.

Feature generation with SQL

In this section, we describe ways of generating features using SQL:

1. [Count based Feature Generation](#)
2. [Binning Feature Generation](#)
3. [Rolling out the features from a single column](#)

NOTE

Once you generate additional features, you can either add them as columns to the existing table or create a new table with the additional features and primary key, that can be joined with the original table.

Count based feature generation

This document demonstrates two ways of generating count features. The first method uses conditional sum and the second method uses the 'where` clause. These can then be joined with the original table (using primary key columns) to have count features alongside the original data.

```
select <column_name1>,<column_name2>,<column_name3>, COUNT(*) as Count_Features from <tablename> group by  
<column_name1>,<column_name2>,<column_name3>  
  
select <column_name1>,<column_name2> , sum(1) as Count_Features from <tablename>  
where <column_name3> = '<some_value>' group by <column_name1>,<column_name2>
```

Binning Feature Generation

The following example shows how to generate binned features by binning (using 5 bins) a numerical column that

can be used as a feature instead:

```
`SELECT <column_name>, NTILE(5) OVER (ORDER BY <column_name>) AS BinNumber from <tablename>`
```

Rolling out the features from a single column

In this section, we demonstrate how to roll out a single column in a table to generate additional features. The example assumes that there is a latitude or longitude column in the table from which you are trying to generate features.

Here is a brief primer on latitude/longitude location data (resourced from stackoverflow

<https://gis.stackexchange.com/questions/8650/how-to-measure-the-accuracy-of-latitude-and-longitude>). Here are some useful things to understand about location data before creating features from the field:

- The sign indicates whether we are north or south, east or west on the globe.
- A nonzero hundreds digit indicates longitude, not latitude is being used.
- The tens digit gives a position to about 1,000 kilometers. It gives useful information about what continent or ocean we are on.
- The units digit (one decimal degree) gives a position up to 111 kilometers (60 nautical miles, about 69 miles). It indicates, roughly, what large state or country we are in.
- The first decimal place is worth up to 11.1 km: it can distinguish the position of one large city from a neighboring large city.
- The second decimal place is worth up to 1.1 km: it can separate one village from the next.
- The third decimal place is worth up to 110 m: it can identify a large agricultural field or institutional campus.
- The fourth decimal place is worth up to 11 m: it can identify a parcel of land. It is comparable to the typical accuracy of an uncorrected GPS unit with no interference.
- The fifth decimal place is worth up to 1.1 m: it distinguishes trees from each other. Accuracy to this level with commercial GPS units can only be achieved with differential correction.
- The sixth decimal place is worth up to 0.11 m: you can use this for laying out structures in detail, for designing landscapes, building roads. It should be more than good enough for tracking movements of glaciers and rivers. This can be achieved by taking painstaking measures with GPS, such as differentially corrected GPS.

The location information can be featurized by separating out region, location, and city information. Note that once can also call a REST end point such as Bing Maps API available at

<https://msdn.microsoft.com/library/ff701710.aspx> to get the region/district information.

```

select
    <location_columnname>
    ,round(<location_columnname>,0) as l1
    ,l2=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1))
=> 1 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,1)
else '0' end
    ,l3=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1))
=> 2 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,2,1)
else '0' end
    ,l4=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1))
=> 3 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,3,1)
else '0' end
    ,l5=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1))
=> 4 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,4,1)
else '0' end
    ,l6=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1))
=> 5 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,5,1)
else '0' end
    ,l7=case when LEN (PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1))
=> 6 then substring(PARSENAME(round(ABS(<location_columnname>)) - FLOOR(ABS(<location_columnname>)),6),1,6,1)
else '0' end
from <tablename>

```

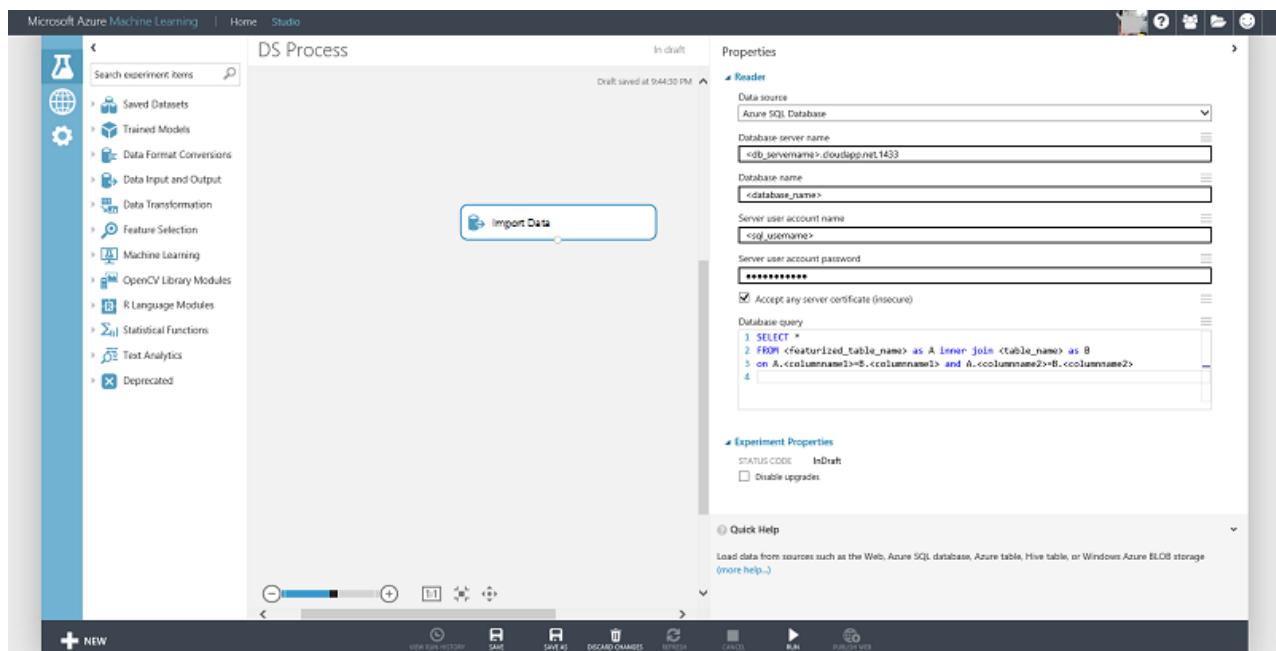
These location based features can be further used to generate additional count features as described earlier.

TIP

You can programmatically insert the records using your language of choice. You may need to insert the data in chunks to improve write efficiency. [Here is an example of how to do this using pyodbc](#). Another alternative is to insert data in the database using [BCP utility](#)

Connecting to Azure Machine Learning

The newly generated feature can be added as a column to an existing table or stored in a new table and joined with the original table for machine learning. Features can be generated or accessed if already created, using the [Import Data](#) module in Azure ML as shown below:



Using a programming language like Python

Using Python to generate features when the data is in SQL Server is similar to processing data in Azure blob using

Python. For comparison, see [Process Azure Blob data in your data science environment](#). Load the data from the database into a pandas data frame to process it further. The process of connecting to the database and loading the data into the data frame is documented in this section.

The following connection string format can be used to connect to a SQL Server database from Python using pyodbc (replace servername, dbname, username, and password with your specific values):

```
#Set up the SQL Azure connection
import pyodbc
conn = pyodbc.connect('DRIVER={SQL Server};SERVER=<servername>;DATABASE=<dbname>;UID=<username>;PWD=<password>')
```

The [Pandas library](#) in Python provides a rich set of data structures and data analysis tools for data manipulation for Python programming. The following code reads the results returned from a SQL Server database into a Pandas data frame:

```
# Query database and load the returned results in pandas data frame
data_frame = pd.read_sql('''select <columnname1>, <columnname2>... from <tablename>''', conn)
```

Now you can work with the Pandas data frame as covered in topics [Create features for Azure blob storage data using Panda](#).

Create features for data in a Hadoop cluster using Hive queries

3/14/2019 • 7 minutes to read

This document shows how to create features for data stored in an Azure HDInsight Hadoop cluster using Hive queries. These Hive queries use embedded Hive User-Defined Functions (UDFs), the scripts for which are provided.

The operations needed to create features can be memory intensive. The performance of Hive queries becomes more critical in such cases and can be improved by tuning certain parameters. The tuning of these parameters is discussed in the final section.

Examples of the queries that are presented are specific to the [NYC Taxi Trip Data](#) scenarios are also provided in [GitHub repository](#). These queries already have data schema specified and are ready to be submitted to run. In the final section, parameters that users can tune so that the performance of Hive queries can be improved are also discussed.

This task is a step in the [Team Data Science Process \(TDSP\)](#).

Prerequisites

This article assumes that you have:

- Created an Azure storage account. If you need instructions, see [Create an Azure Storage account](#)
- Provisioned a customized Hadoop cluster with the HDInsight service. If you need instructions, see [Customize Azure HDInsight Hadoop Clusters for Advanced Analytics](#).
- The data has been uploaded to Hive tables in Azure HDInsight Hadoop clusters. If it has not, follow [Create and load data to Hive tables](#) to upload data to Hive tables first.
- Enabled remote access to the cluster. If you need instructions, see [Access the Head Node of Hadoop Cluster](#).

Feature generation

In this section, several examples of the ways in which features can be generating using Hive queries are described. Once you have generated additional features, you can either add them as columns to the existing table or create a new table with the additional features and primary key, which can then be joined with the original table. Here are the examples presented:

1. [Frequency-based Feature Generation](#)
2. [Risks of Categorical Variables in Binary Classification](#)
3. [Extract features from Datetime Field](#)
4. [Extract features from Text Field](#)
5. [Calculate distance between GPS coordinates](#)

Frequency-based feature generation

It is often useful to calculate the frequencies of the levels of a categorical variable, or the frequencies of certain combinations of levels from multiple categorical variables. Users can use the following script to calculate these frequencies:

```

select
    a.<column_name1>, a.<column_name2>, a.sub_count/sum(a.sub_count) over () as frequency
from
(
    select
        <column_name1>,<column_name2>, count(*) as sub_count
    from <databasename>.<tablename> group by <column_name1>, <column_name2>
)a
order by frequency desc;

```

Risks of categorical variables in binary classification

In binary classification, non-numeric categorical variables must be converted into numeric features when the models being used only take numeric features. This conversion is done by replacing each non-numeric level with a numeric risk. This section shows some generic Hive queries that calculate the risk values (log odds) of a categorical variable.

```

set smooth_param1=1;
set smooth_param2=20;
select
    <column_name1>,<column_name2>,
    ln((sum_target+${hiveconf:smooth_param1})/(record_count-sum_target+${hiveconf:smooth_param2}-
${hiveconf:smooth_param1})) as risk
from
(
    select
        <column_name1>, <column_name2>, sum(binary_target) as sum_target, sum(1) as record_count
    from
        (
            select
                <column_name1>, <column_name2>, if(target_column>0,1,0) as binary_target
            from <databasename>.<tablename>
        )a
    group by <column_name1>, <column_name2>
)b

```

In this example, variables `smooth_param1` and `smooth_param2` are set to smooth the risk values calculated from the data. Risks have a range between -Inf and Inf. A risk > 0 indicates that the probability that the target is equal to 1 is greater than 0.5.

After the risk table is calculated, users can assign risk values to a table by joining it with the risk table. The Hive joining query was provided in previous section.

Extract features from datetime fields

Hive comes with a set of UDFs for processing datetime fields. In Hive, the default datetime format is 'yyyy-MM-dd 00:00:00' ('1970-01-01 12:21:32' for example). This section shows examples that extract the day of a month, the month from a datetime field, and other examples that convert a datetime string in a format other than the default format to a datetime string in default format.

```

select day(<datetime field>), month(<datetime field>)
from <databasename>.<tablename>;

```

This Hive query assumes that the is in the default datetime format.

If a datetime field is not in the default format, you need to convert the datetime field into Unix time stamp first, and then convert the Unix time stamp to a datetime string that is in the default format. When the datetime is in default format, users can apply the embedded datetime UDFs to extract features.

```
select from_unixtime(unix_timestamp(<datetime field>,'<pattern of the datetime field>'))
from <databasename>.<tablename>;
```

In this query, if the has the pattern like *03/26/2015 12:04:39*, the ' should be `'MM/dd/yyyy HH:mm:ss'`. To test it, users can run

```
select from_unixtime(unix_timestamp('05/15/2015 09:32:10','MM/dd/yyyy HH:mm:ss'))
from hivesampletable limit 1;
```

The *hivesampletable* in this query comes preinstalled on all Azure HDInsight Hadoop clusters by default when the clusters are provisioned.

Extract features from text fields

When the Hive table has a text field that contains a string of words that are delimited by spaces, the following query extracts the length of the string, and the number of words in the string.

```
select length(<text field>) as str_len, size(split(<text field>,' ')) as word_num
from <databasename>.<tablename>;
```

Calculate distances between sets of GPS coordinates

The query given in this section can be directly applied to the NYC Taxi Trip Data. The purpose of this query is to show how to apply an embedded mathematical function in Hive to generate features.

The fields that are used in this query are the GPS coordinates of pickup and dropoff locations, named *pickup_longitude*, *pickup_latitude*, *dropoff_longitude*, and *dropoff_latitude*. The queries that calculate the direct distance between the pickup and dropoff coordinates are:

```
set R=3959;
set pi=radians(180);
select pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude,
    ${hiveconf:R}*2*2*atan((1-sqrt(1-pow(sin((dropoff_latitude-pickup_latitude)
        *${hiveconf:pi}/180/2),2)-cos(pickup_latitude*${hiveconf:pi}/180)
        *cos(dropoff_latitude*${hiveconf:pi}/180)*pow(sin((dropoff_longitude-
pickup_longitude)*${hiveconf:pi}/180/2),2)))
        /sqrt(pow(sin((dropoff_latitude-pickup_latitude)*${hiveconf:pi}/180/2),2)
        +cos(pickup_latitude*${hiveconf:pi}/180)*cos(dropoff_latitude*${hiveconf:pi}/180)*
        pow(sin((dropoff_longitude-pickup_longitude)*${hiveconf:pi}/180/2),2))) as direct_distance
from nytaxi.trip
where pickup_longitude between -90 and 0
and pickup_latitude between 30 and 90
and dropoff_longitude between -90 and 0
and dropoff_latitude between 30 and 90
limit 10;
```

The mathematical equations that calculate the distance between two GPS coordinates can be found on the [Movable Type Scripts](#) site, authored by Peter Lapisu. In this Javascript, the function `toRad()` is just `lat_or_lon*pi/180`, which converts degrees to radians. Here, *lat_or_lon* is the latitude or longitude. Since Hive does not provide the function `atan2`, but provides the function `atan`, the `atan2` function is implemented by `atan` function in the above Hive query using the definition provided in [Wikipedia](#).

$$\text{atan2}(y, x) = 2 \arctan \frac{\sqrt{x^2 + y^2} - x}{y}.$$

A full list of Hive embedded UDFs can be found in the **Built-in Functions** section on the [Apache Hive wiki](#).

Advanced topics: Tune Hive parameters to improve query speed

The default parameter settings of Hive cluster might not be suitable for the Hive queries and the data that the queries are processing. This section discusses some parameters that users can tune to improve the performance of Hive queries. Users need to add the parameter tuning queries before the queries of processing data.

1. **Java heap space:** For queries involving joining large datasets, or processing long records, **running out of heap space** is one of the common errors. This error can be avoided by setting parameters *mapreduce.map.java.opts* and *mapreduce.task.io.sort.mb* to desired values. Here is an example:

```
set mapreduce.map.java.opts=-Xmx4096m;
set mapreduce.task.io.sort.mb=-Xmx1024m;
```

This parameter allocates 4GB memory to Java heap space and also makes sorting more efficient by allocating more memory for it. It is a good idea to play with these allocations if there are any job failure errors related to heap space.

2. **DFS block size:** This parameter sets the smallest unit of data that the file system stores. As an example, if the DFS block size is 128 MB, then any data of size less than and up to 128 MB is stored in a single block. Data that is larger than 128 MB is allotted extra blocks.
3. Choosing a small block size causes large overheads in Hadoop since the name node has to process many more requests to find the relevant block pertaining to the file. A recommended setting when dealing with gigabytes (or larger) data is:

```
set dfs.block.size=128m;
```

4. **Optimizing join operation in Hive:** While join operations in the map/reduce framework typically take place in the reduce phase, sometimes, enormous gains can be achieved by scheduling joins in the map phase (also called "mapjoins"). To direct Hive to do this whenever possible, set:

```
set hive.auto.convert.join=true;
```

5. **Specifying the number of mappers to Hive:** While Hadoop allows the user to set the number of reducers, the number of mappers is typically not be set by the user. A trick that allows some degree of control on this number is to choose the Hadoop variables *mapred.min.split.size* and *mapred.max.split.size* as the size of each map task is determined by:

```
num_maps = max(mapred.min.split.size, min(mapred.max.split.size, dfs.block.size))
```

Typically, the default value of:

- *mapred.min.split.size* is 0, that of
- *mapred.max.split.size* is **Long.MAX** and that of
- *dfs.block.size* is 64 MB.

As we can see, given the data size, tuning these parameters by "setting" them allows us to tune the number of mappers used.

6. Here are a few other more **advanced options** for optimizing Hive performance. These allow you to set the memory allocated to map and reduce tasks, and can be useful in tweaking performance. Keep in mind that the *mapreduce.reduce.memory.mb* cannot be greater than the physical memory size of each worker node in the Hadoop cluster.

```
set mapreduce.map.memory.mb = 2048;
set mapreduce.reduce.memory.mb=6144;
set mapreduce.reduce.java.opts=-Xmx8192m;
set mapred.reduce.tasks=128;
set mapred.tasktracker.reduce.tasks.maximum=128;
```

Feature selection in the Team Data Science Process (TDSP)

1/30/2019 • 4 minutes to read

This article explains the purposes of feature selection and provides examples of its role in the data enhancement process of machine learning. These examples are drawn from Azure Machine Learning Studio.

Try [Azure Machine Learning Studio](#), available in paid or free options.

The engineering and selection of features is one part of the Team Data Science Process (TDSP) outlined in the article [What is the Team Data Science Process?](#). Feature engineering and selection are parts of the **Develop features** step of the TDSP.

- **feature engineering:** This process attempts to create additional relevant features from the existing raw features in the data, and to increase predictive power to the learning algorithm.
- **feature selection:** This process selects the key subset of original data features in an attempt to reduce the dimensionality of the training problem.

Normally **feature engineering** is applied first to generate additional features, and then the **feature selection** step is performed to eliminate irrelevant, redundant, or highly correlated features.

Filter features from your data - feature selection

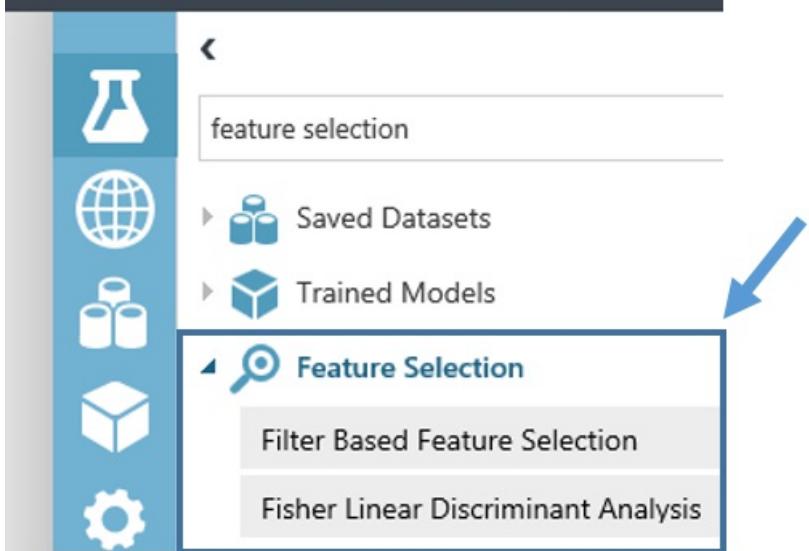
Feature selection is a process that is commonly applied for the construction of training datasets for predictive modeling tasks such as classification or regression tasks. The goal is to select a subset of the features from the original dataset that reduce its dimensions by using a minimal set of features to represent the maximum amount of variance in the data. This subset of features is used to train the model. Feature selection serves two main purposes.

- First, feature selection often increases classification accuracy by eliminating irrelevant, redundant, or highly correlated features.
- Second, it decreases the number of features, which makes the model training process more efficient. Efficiency is particularly important for learners that are expensive to train such as support vector machines.

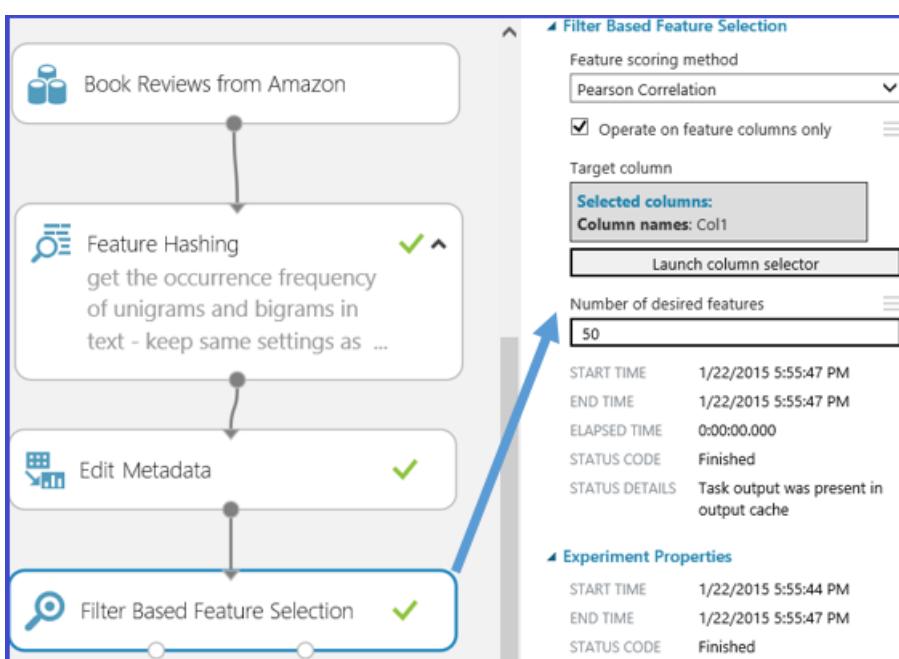
Although feature selection does seek to reduce the number of features in the dataset used to train the model, it is not referred to by the term "dimensionality reduction". Feature selection methods extract a subset of original features in the data without changing them. Dimensionality reduction methods employ engineered features that can transform the original features and thus modify them. Examples of dimensionality reduction methods include Principal Component Analysis, canonical correlation analysis, and Singular Value Decomposition.

Among others, one widely applied category of feature selection methods in a supervised context is called "filter-based feature selection". By evaluating the correlation between each feature and the target attribute, these methods apply a statistical measure to assign a score to each feature. The features are then ranked by the score, which may be used to help set the threshold for keeping or eliminating a specific feature. Examples of the statistical measures used in these methods include Person correlation, mutual information, and the Chi squared test.

In Azure Machine Learning Studio, there are modules provided for feature selection. As shown in the following figure, these modules include [Filter-Based Feature Selection](#) and [Fisher Linear Discriminant Analysis](#).



Consider, for example, the use of the [Filter-Based Feature Selection](#) module. For convenience, continue using the text mining example. Assume that you want to build a regression model after a set of 256 features are created through the [Feature Hashing](#) module, and that the response variable is the "Col1" that contains book review ratings ranging from 1 to 5. By setting "Feature scoring method" to be "Pearson Correlation", the "Target column" to be "Col1", and the "Number of desired features" to 50. Then the module [Filter-Based Feature Selection](#) produces a dataset containing 50 features together with the target attribute "Col1". The following figure shows the flow of this experiment and the input parameters:



The following figure shows the resulting datasets:

ROWS: 10000 COLUMNS: 51

view as:

Filter Based Feature Selection

Right click the 1st port

Right click the 2nd port

Col1 Col2_Hash Col2_HashingFeature_203 Col2_HashingFeature_146 Col2_HashingFeature_122 Col2_Hash

2	6	1	2	6
2	6	4	7	5
1	9	2	1	3
2	5	2	2	3
2	3	1	1	0
2	11	6	5	5
1	2	2	3	1
2	4	3	2	1
2	2	2	6	3
2	2	0	1	0
1	11	4	3	5
2	1	0	5	0
2	2	1	2	2
1	1	3	3	3

Each feature is scored based on the Pearson Correlation between itself and the target attribute "Col1". The features with top scores are kept.

The corresponding scores of the selected features are shown in the following figure:

ROWS: 1 COLUMNS: 51

view as:

Filter Based Feature Selection

Right click the 1st port

Right click the 2nd port

Col1 Col2_Hash Col2_HashingFeature_203 Col2_HashingFeature_146 Col2_HashingFeature_122 Col2_Hash

1	0.083607	0.060681	0.05716	0.056381	
---	----------	----------	---------	----------	--

By applying this **Filter-Based Feature Selection** module, 50 out of 256 features are selected because they have the most correlated features with the target variable "Col1", based on the scoring method "Pearson Correlation".

Conclusion

Feature engineering and feature selection are two commonly Engineered and selected features increase the efficiency of the training process which attempts to extract the key information contained in the data. They also improve the power of these models to classify the input data accurately and to predict outcomes of interest more robustly. Feature engineering and selection can also combine to make the learning more computationally tractable. It does so by enhancing and then reducing the number of features needed to calibrate or train a model. Mathematically speaking, the features selected to train the model are a minimal set of independent variables that explain the patterns in the data and then predict outcomes successfully.

It is not always necessarily to perform feature engineering or feature selection. Whether it is needed or not depends on the data collected, the algorithm selected, and the objective of the experiment.

How to choose algorithms for Azure Machine Learning Studio

3/12/2019 • 15 minutes to read

The answer to the question "What machine learning algorithm should I use?" is always "It depends." It depends on the size, quality, and nature of the data. It depends on what you want to do with the answer. It depends on how the math of the algorithm was translated into instructions for the computer you are using. And it depends on how much time you have. Even the most experienced data scientists can't tell which algorithm will perform best before trying them.

Machine Learning Studio provides state-of-the-art algorithms, such as Scalable Boosted Decision trees, Bayesian Recommendation systems, Deep Neural Networks, and Decision Jungles developed at Microsoft Research. Scalable open-source machine learning packages, like Vowpal Wabbit, are also included. Machine Learning Studio supports machine learning algorithms for multiclass and binary classification, regression, and clustering. See the complete list of [Machine Learning Modules](#). The documentation provides some information about each algorithm and how to tune parameters to optimize the algorithm for your use.

The Machine Learning Algorithm Cheat Sheet

The [Microsoft Azure Machine Learning Studio Algorithm Cheat Sheet](#) helps you choose the right machine learning algorithm for your predictive analytics solutions from the Azure Machine Learning Studio library of algorithms. This article walks you through how to use this cheat sheet.

NOTE

To download the cheat sheet and follow along with this article, go to [Machine learning algorithm cheat sheet for Microsoft Azure Machine Learning Studio](#).

This cheat sheet has a very specific audience in mind: a beginning data scientist with undergraduate-level machine learning, trying to choose an algorithm to start with in Azure Machine Learning Studio. That means that it makes some generalizations and oversimplifications, but it points you in a safe direction. It also means that there are lots of algorithms not listed here.

These recommendations are compiled feedback and tips from many data scientists and machine learning experts. We didn't agree on everything, but we've tried to harmonize our opinions into a rough consensus. Most of the statements of disagreement begin with "It depends..."

How to use the cheat sheet

Read the path and algorithm labels on the chart as "For *<path label>*, use *<algorithm>*." For example, "For *speed*, use *two class logistic regression*." Sometimes more than one branch applies. Sometimes none of them are a perfect fit. They're intended to be rule-of-thumb recommendations, so don't worry about it being exact. Several data scientists we talked with said that the only sure way to find the very best algorithm is to try all of them.

Here's an example from the [Azure AI Gallery](#) of an experiment that tries several algorithms against the same data and compares the results: [Compare Multi-class Classifiers: Letter recognition](#).

TIP

To download an easy-to-understand infographic overview of machine learning basics to learn about popular algorithms used to answer common machine learning questions, see [Machine learning basics with algorithm examples](#).

Flavors of machine learning

Supervised

Supervised learning algorithms make predictions based on a set of examples. For instance, historical stock prices can be used to make guesses about future prices. Each example used for training is labeled with the value of interest—in this case the stock price. A supervised learning algorithm looks for patterns in those value labels. It can use any information that might be relevant—the day of the week, the season, the company's financial data, the type of industry, the presence of disruptive geopolitical events—and each algorithm looks for different types of patterns. After the algorithm has found the best pattern it can, it uses that pattern to make predictions for unlabeled testing data—tomorrow's prices.

Supervised learning is a popular and useful type of machine learning. With one exception, all the modules in Azure Machine Learning Studio are supervised learning algorithms. There are several specific types of supervised learning that are represented within Azure Machine Learning Studio: classification, regression, and anomaly detection.

- **Classification.** When the data are being used to predict a category, supervised learning is also called classification. This is the case when assigning an image as a picture of either a 'cat' or a 'dog'. When there are only two choices, it's called **two-class** or **binomial classification**. When there are more categories, as when predicting the winner of the NCAA March Madness tournament, this problem is known as **multi-class classification**.
- **Regression.** When a value is being predicted, as with stock prices, supervised learning is called regression.
- **Anomaly detection.** Sometimes the goal is to identify data points that are simply unusual. In fraud detection, for example, any highly unusual credit card spending patterns are suspect. The possible variations are so numerous and the training examples so few, that it's not feasible to learn what fraudulent activity looks like. The approach that anomaly detection takes is to simply learn what normal activity looks like (using a history of non-fraudulent transactions) and identify anything that is significantly different.

Unsupervised

In unsupervised learning, data points have no labels associated with them. Instead, the goal of an unsupervised learning algorithm is to organize the data in some way or to describe its structure. This can mean grouping it into clusters or finding different ways of looking at complex data so that it appears simpler or more organized.

Reinforcement learning

In reinforcement learning, the algorithm gets to choose an action in response to each data point. The learning algorithm also receives a reward signal a short time later, indicating how good the decision was. Based on this, the algorithm modifies its strategy in order to achieve the highest reward. Currently there are no reinforcement learning algorithm modules in Azure Machine Learning Studio. Reinforcement learning is common in robotics, where the set of sensor readings at one point in time is a data point, and the algorithm must choose the robot's next action. It is also a natural fit for Internet of Things applications.

Considerations when choosing an algorithm

Accuracy

Getting the most accurate answer possible isn't always necessary. Sometimes an approximation is adequate, depending on what you want to use it for. If that's the case, you may be able to cut your processing time dramatically by sticking with more approximate methods. Another advantage of more approximate methods is

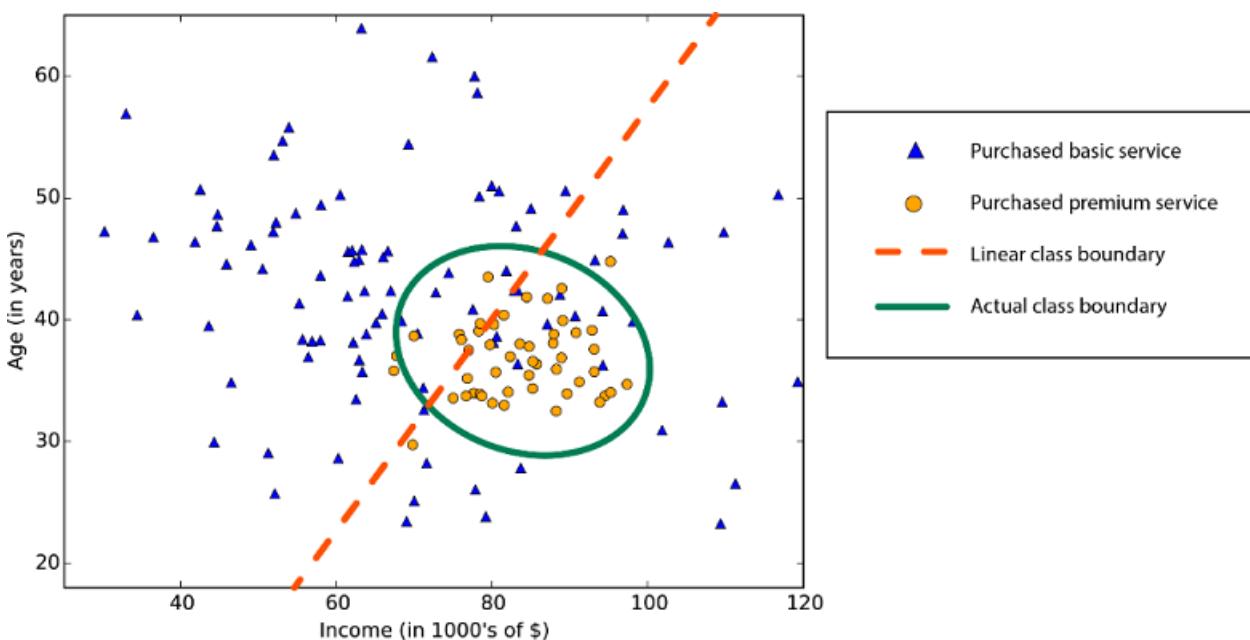
that they naturally tend to avoid overfitting.

Training time

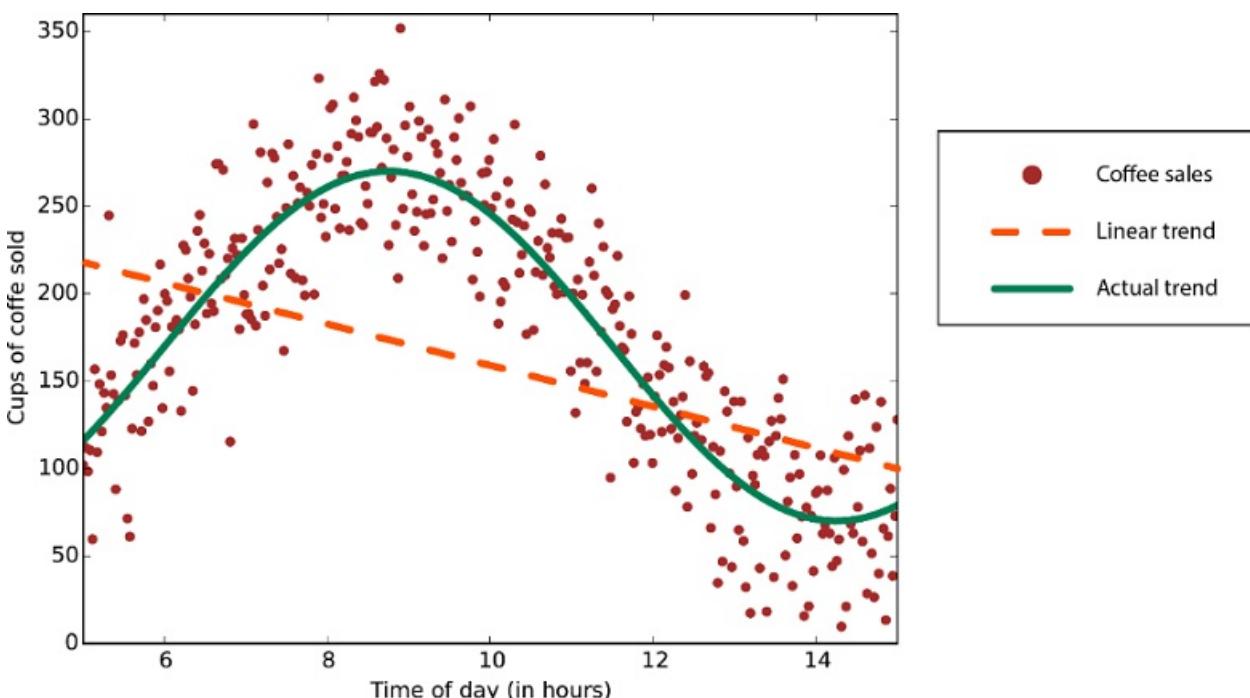
The number of minutes or hours necessary to train a model varies a great deal between algorithms. Training time is often closely tied to accuracy—one typically accompanies the other. In addition, some algorithms are more sensitive to the number of data points than others. When time is limited it can drive the choice of algorithm, especially when the data set is large.

Linearity

Lots of machine learning algorithms make use of linearity. Linear classification algorithms assume that classes can be separated by a straight line (or its higher-dimensional analog). These include logistic regression and support vector machines (as implemented in Azure Machine Learning Studio). Linear regression algorithms assume that data trends follow a straight line. These assumptions aren't bad for some problems, but on others they bring accuracy down.



Non-linear class boundary - relying on a linear classification algorithm would result in low accuracy



Data with a nonlinear trend - using a linear regression method would generate much larger errors than necessary

Despite their dangers, linear algorithms are very popular as a first line of attack. They tend to be algorithmically simple and fast to train.

Number of parameters

Parameters are the knobs a data scientist gets to turn when setting up an algorithm. They are numbers that affect the algorithm's behavior, such as error tolerance or number of iterations, or options between variants of how the algorithm behaves. The training time and accuracy of the algorithm can sometimes be quite sensitive to getting just the right settings. Typically, algorithms with large numbers of parameters require the most trial and error to find a good combination.

Alternatively, there is a [parameter sweeping](#) module block in Azure Machine Learning Studio that automatically tries all parameter combinations at whatever granularity you choose. While this is a great way to make sure you've spanned the parameter space, the time required to train a model increases exponentially with the number of parameters.

The upside is that having many parameters typically indicates that an algorithm has greater flexibility. It can often achieve very good accuracy, provided you can find the right combination of parameter settings.

Number of features

For certain types of data, the number of features can be very large compared to the number of data points. This is often the case with genetics or textual data. The large number of features can bog down some learning algorithms, making training time unfeasibly long. Support Vector Machines are particularly well suited to this case (see below).

Special cases

Some learning algorithms make particular assumptions about the structure of the data or the desired results. If you can find one that fits your needs, it can give you more useful results, more accurate predictions, or faster training times.

ALGORITHM	ACCURACY	TRAINING TIME	LINEARITY	PARAMETERS	NOTES
Two-class classification					
logistic regression		●	●	5	
decision forest	●	○		6	
decision jungle	●	○		6	Low memory footprint
boosted decision tree	●	○		6	Large memory footprint
neural network	●			9	Additional customization is possible
averaged perceptron	○	○	●	4	
support vector machine		○	●	5	Good for large feature sets

ALGORITHM	ACCURACY	TRAINING TIME	LINEARITY	PARAMETERS	NOTES
locally deep support vector machine	○			8	Good for large feature sets
Bayes' point machine		○	●	3	
Multi-class classification					
logistic regression		●	●	5	
decision forest	●	○		6	
decision jungle	●	○		6	Low memory footprint
neural network	●			9	Additional customization is possible
one-v-all	-	-	-	-	See properties of the two-class method selected
Regression					
linear		●	●	4	
Bayesian linear		○	●	2	
decision forest	●	○		6	
boosted decision tree	●	○		5	Large memory footprint
fast forest quantile	●	○		9	Distributions rather than point predictions
neural network	●			9	Additional customization is possible
Poisson			●	5	Technically log-linear. For predicting counts
ordinal				0	For predicting rank-ordering
Anomaly detection					

ALGORITHM	ACCURACY	TRAINING TIME	LINEARITY	PARAMETERS	NOTES
support vector machine	○	○		2	Especially good for large feature sets
PCA-based anomaly detection		○	●	3	
K-means		○	●	4	A clustering algorithm

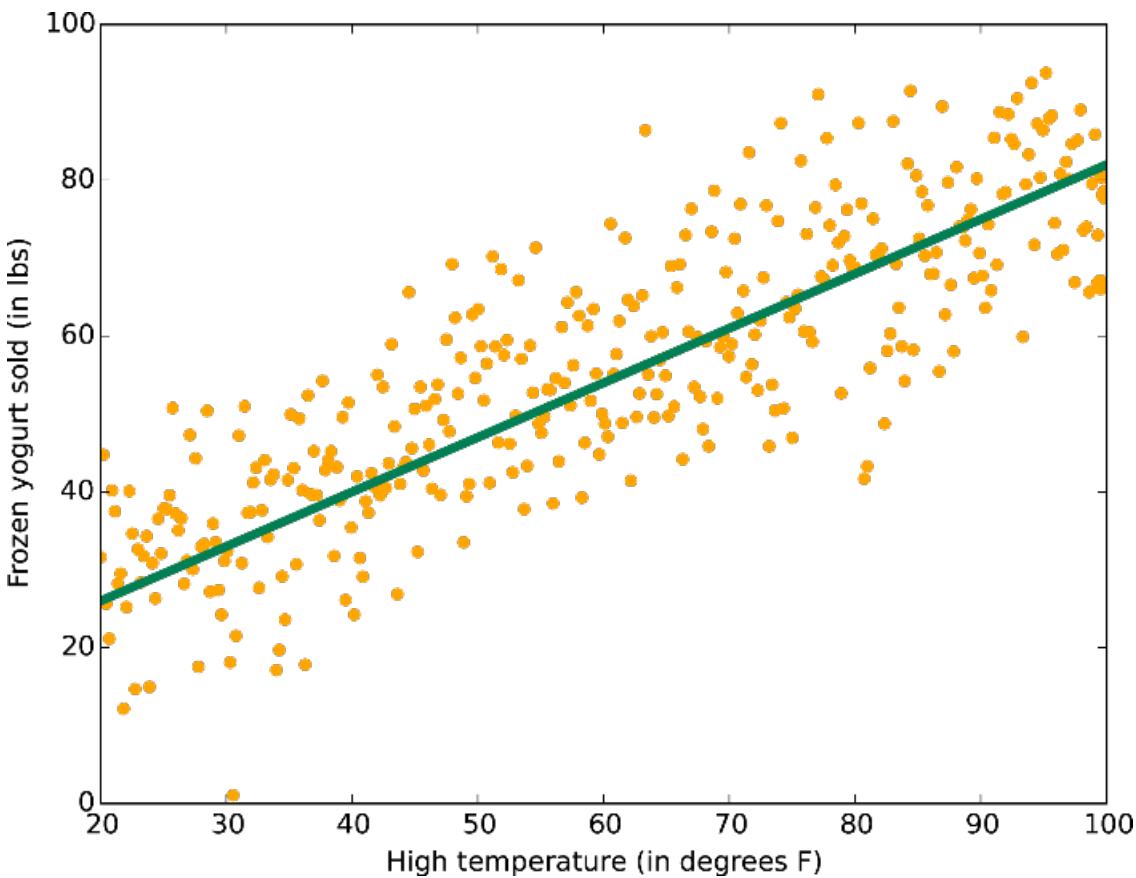
Algorithm properties:

- - shows excellent accuracy, fast training times, and the use of linearity
- - shows good accuracy and moderate training times

Algorithm notes

Linear regression

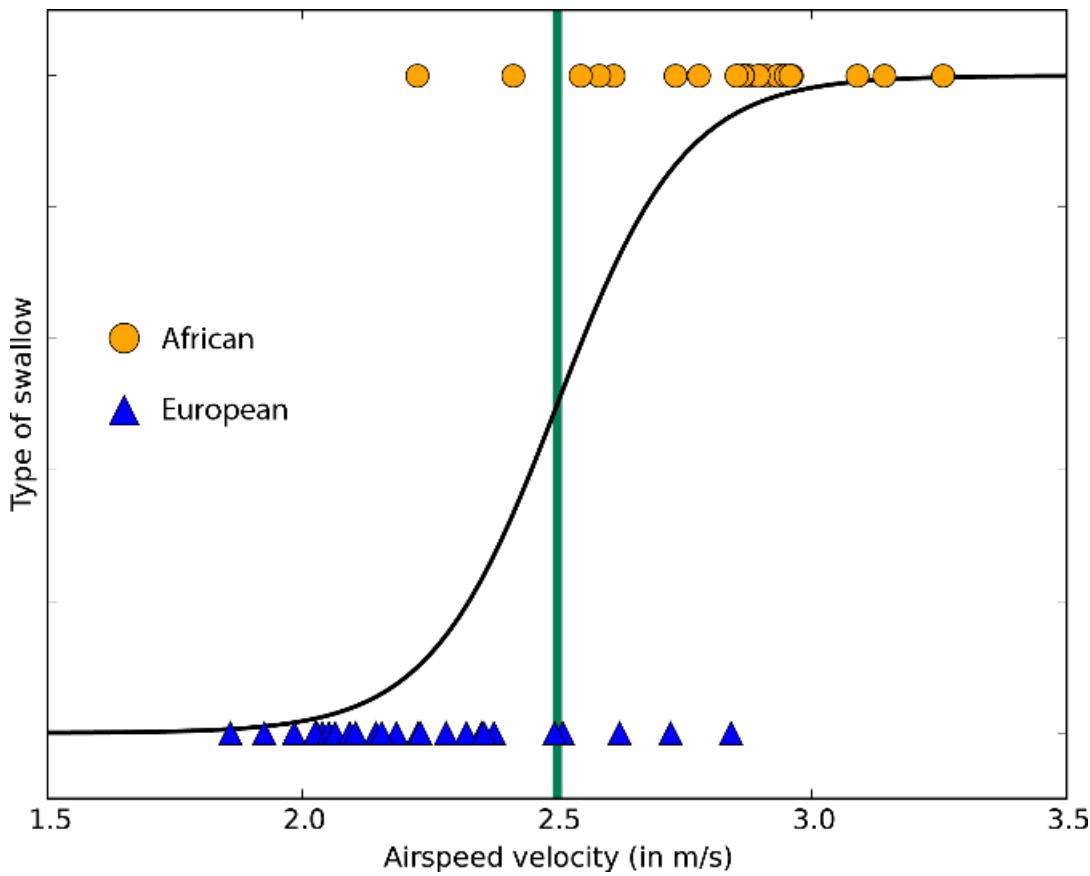
As mentioned previously, [linear regression](#) fits a line (or plane, or hyperplane) to the data set. It's a workhorse, simple and fast, but it may be overly simplistic for some problems.



Data with a linear trend

Logistic regression

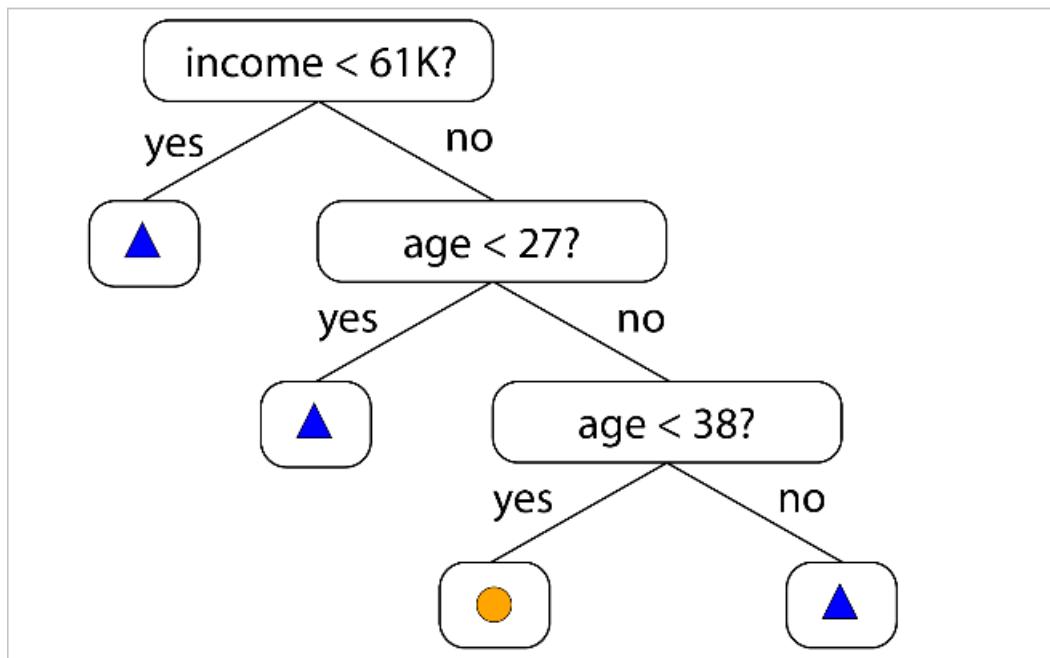
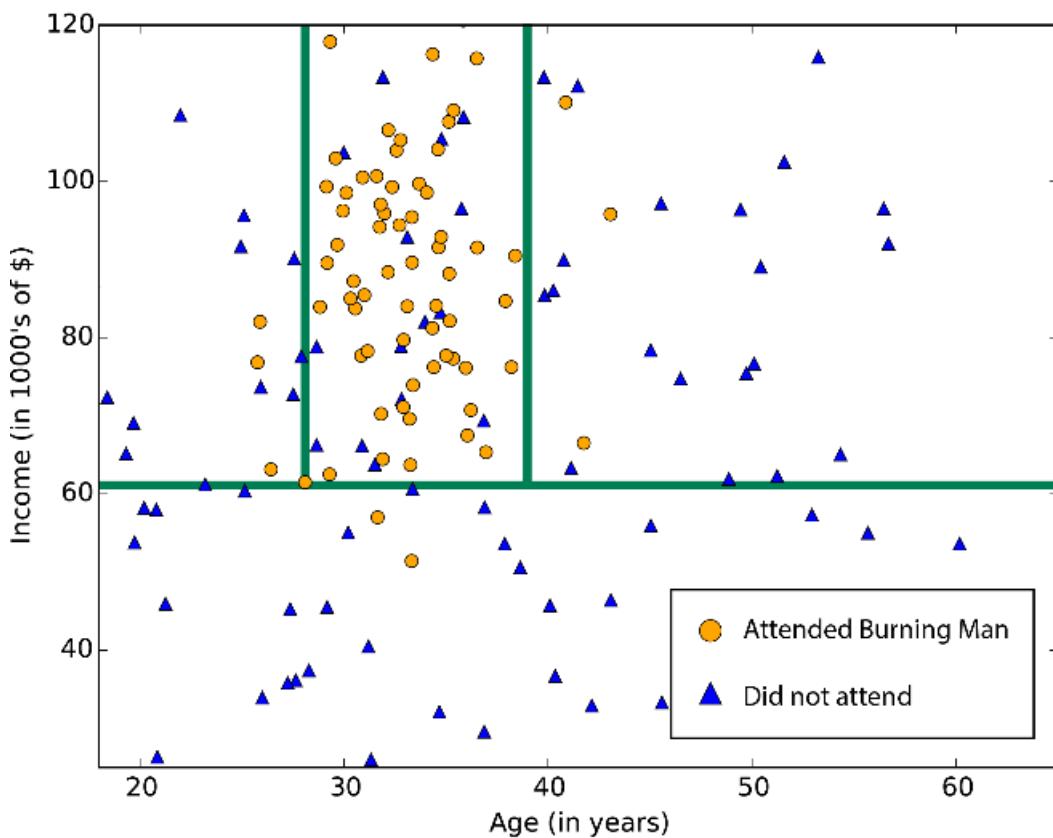
Although it includes 'regression' in the name, logistic regression is actually a powerful tool for [two-class](#) and [multiclass](#) classification. It's fast and simple. The fact that it uses an 'S'-shaped curve instead of a straight line makes it a natural fit for dividing data into groups. Logistic regression gives linear class boundaries, so when you use it, make sure a linear approximation is something you can live with.



A logistic regression to two-class data with just one feature - the class boundary is the point at which the logistic curve is just as close to both classes

Trees, forests, and jungles

Decision forests ([regression](#), [two-class](#), and [multiclass](#)), decision jungles ([two-class](#) and [multiclass](#)), and boosted decision trees ([regression](#) and [two-class](#)) are all based on decision trees, a foundational machine learning concept. There are many variants of decision trees, but they all do the same thing—subdivide the feature space into regions with mostly the same label. These can be regions of consistent category or of constant value, depending on whether you are doing classification or regression.



A decision tree subdivides a feature space into regions of roughly uniform values

Because a feature space can be subdivided into arbitrarily small regions, it's easy to imagine dividing it finely enough to have one data point per region. This is an extreme example of overfitting. In order to avoid this, a large set of trees are constructed with special mathematical care taken to ensure the trees are not correlated. The average of this "decision forest" is a tree that avoids overfitting. Decision forests can use a lot of memory. Decision jungles are a variant that consumes less memory at the expense of a slightly longer training time.

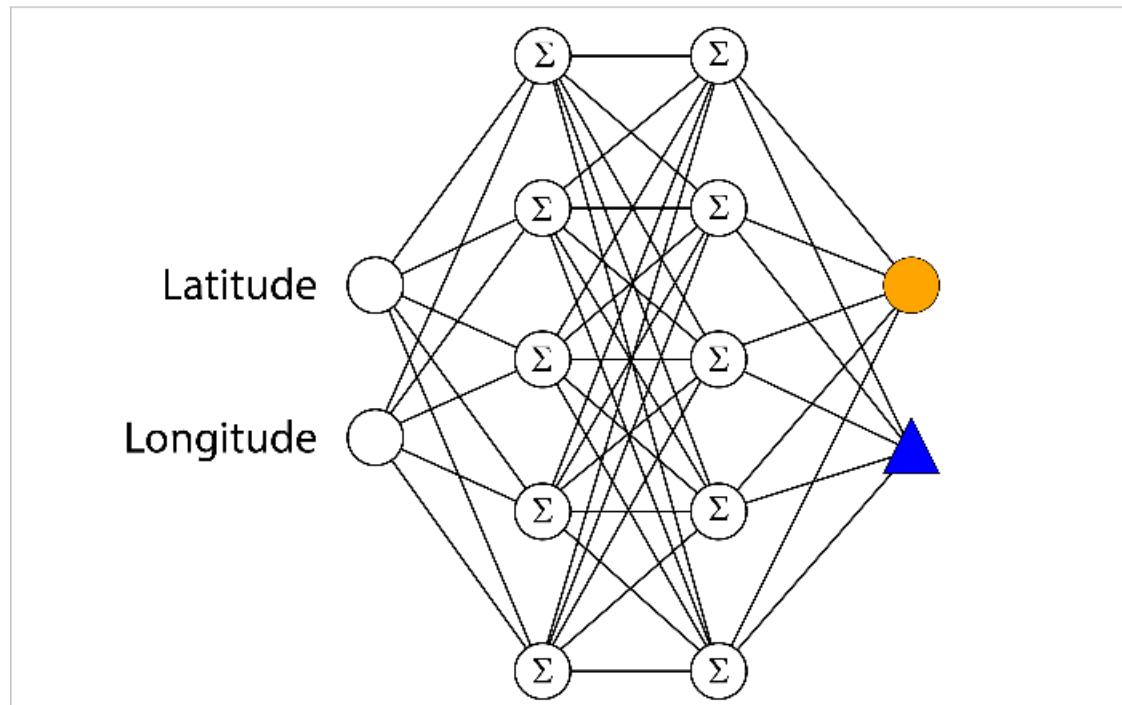
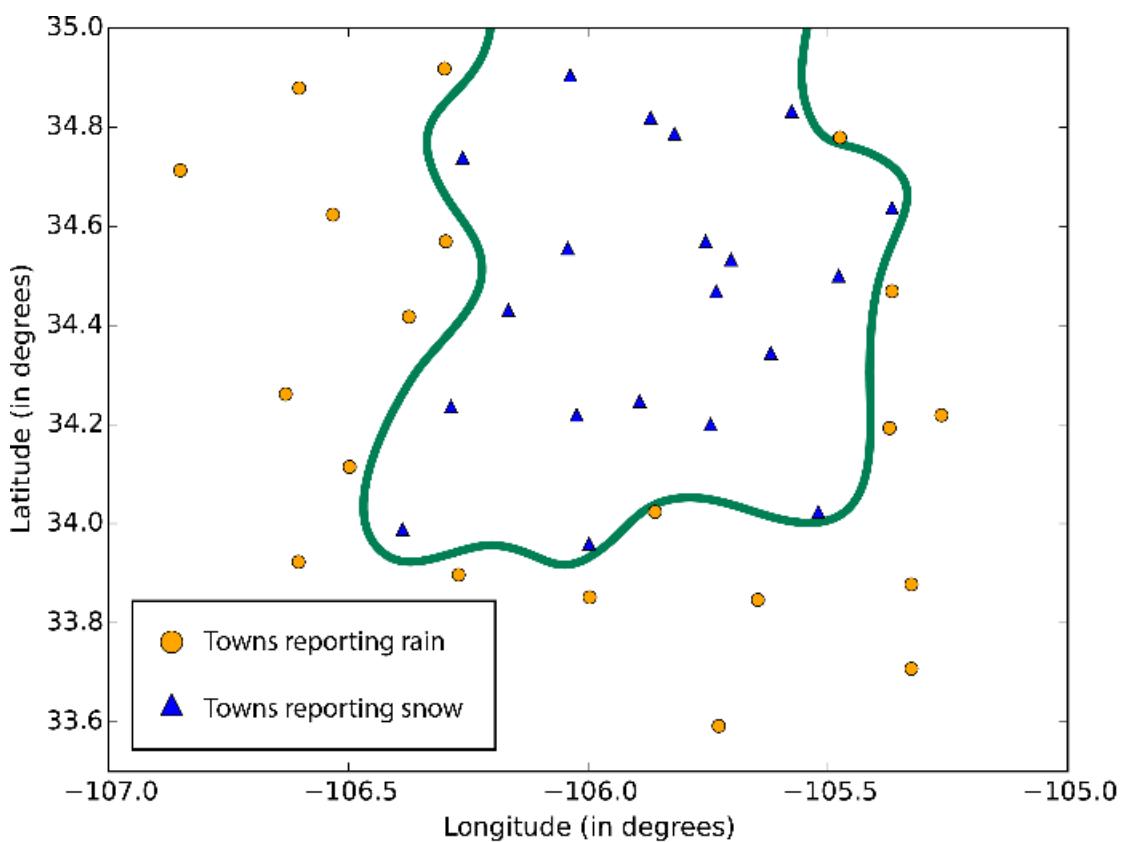
Boosted decision trees avoid overfitting by limiting how many times they can subdivide and how few data points are allowed in each region. The algorithm constructs a sequence of trees, each of which learns to compensate for the error left by the tree before. The result is a very accurate learner that tends to use a lot of memory. For the full technical description, check out [Friedman's original paper](#).

[Fast forest quantile regression](#) is a variation of decision trees for the special case where you want to know not only the typical (median) value of the data within a region, but also its distribution in the form of quantiles.

Neural networks and perceptrons

Neural networks are brain-inspired learning algorithms covering [multiclass](#), [two-class](#), and [regression](#) problems. They come in an infinite variety, but the neural networks within Azure Machine Learning Studio are all of the form of directed acyclic graphs. That means that input features are passed forward (never backward) through a sequence of layers before being turned into outputs. In each layer, inputs are weighted in various combinations, summed, and passed on to the next layer. This combination of simple calculations results in the ability to learn sophisticated class boundaries and data trends, seemingly by magic. Many-layered networks of this sort perform the "deep learning" that fuels so much tech reporting and science fiction.

This high performance doesn't come for free, though. Neural networks can take a long time to train, particularly for large data sets with lots of features. They also have more parameters than most algorithms, which means that parameter sweeping expands the training time a great deal. And for those overachievers who wish to [specify their own network structure](#), the possibilities are inexhaustible.

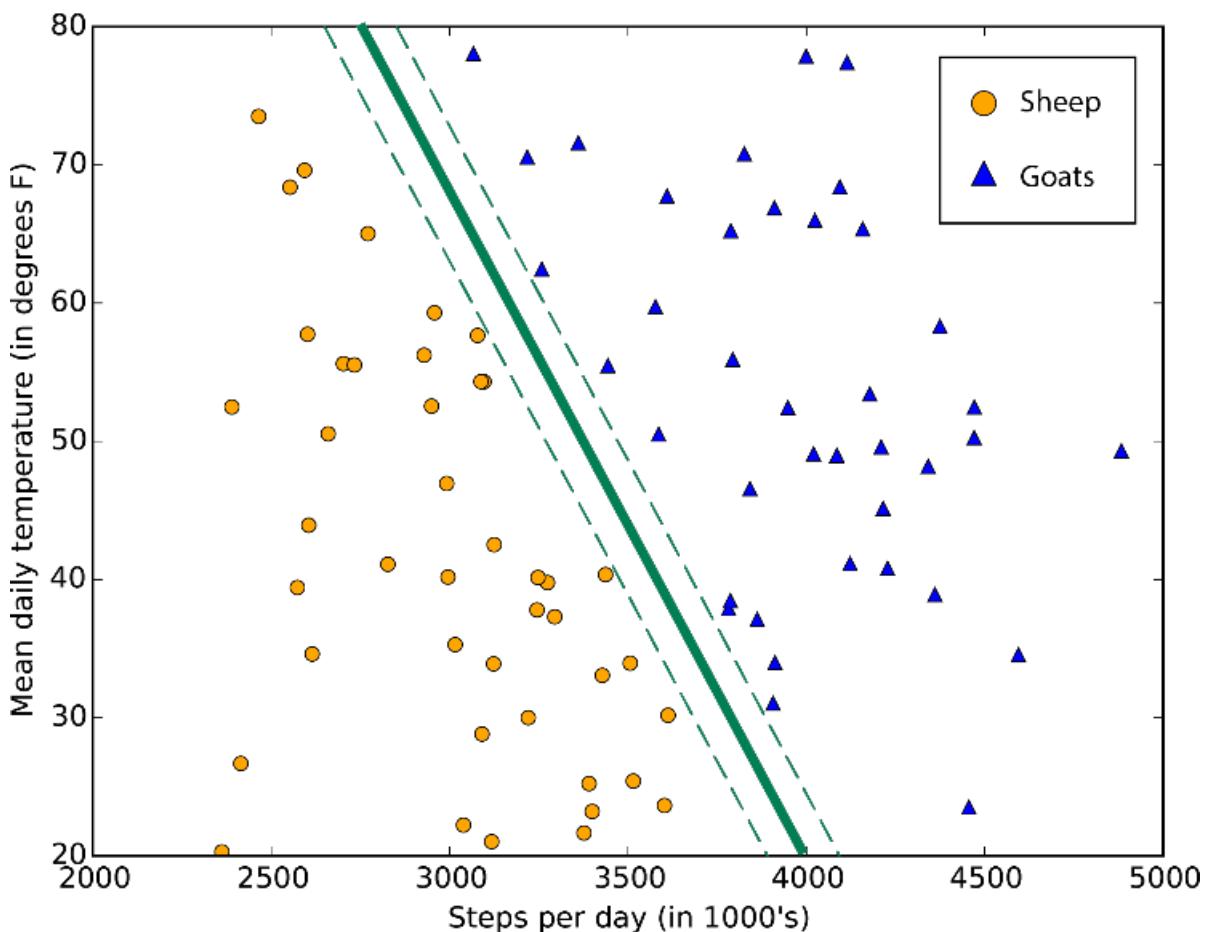


The boundaries learned by neural networks can be complex and irregular

The [two-class averaged perceptron](#) is neural networks' answer to skyrocketing training times. It uses a network structure that gives linear class boundaries. It is almost primitive by today's standards, but it has a long history of working robustly and is small enough to learn quickly.

SVMs

Support vector machines (SVMs) find the boundary that separates classes by as wide a margin as possible. When the two classes can't be clearly separated, the algorithms find the best boundary they can. As written in Azure Machine Learning Studio, the [two-class SVM](#) does this with a straight line only (in SVM-speak, it uses a linear kernel). Because it makes this linear approximation, it is able to run fairly quickly. Where it really shines is with feature-intense data, like text or genomic data. In these cases SVMs are able to separate classes more quickly and with less overfitting than most other algorithms, in addition to requiring only a modest amount of memory.



A typical support vector machine class boundary maximizes the margin separating two classes

Another product of Microsoft Research, the [two-class locally deep SVM](#) is a non-linear variant of SVM that retains most of the speed and memory efficiency of the linear version. It is ideal for cases where the linear approach doesn't give accurate enough answers. The developers kept it fast by breaking down the problem into a number of small linear SVM problems. Read the [full description](#) for the details on how they pulled off this trick.

Using a clever extension of nonlinear SVMs, the [one-class SVM](#) draws a boundary that tightly outlines the entire data set. It is useful for anomaly detection. Any new data points that fall far outside that boundary are unusual enough to be noteworthy.

Bayesian methods

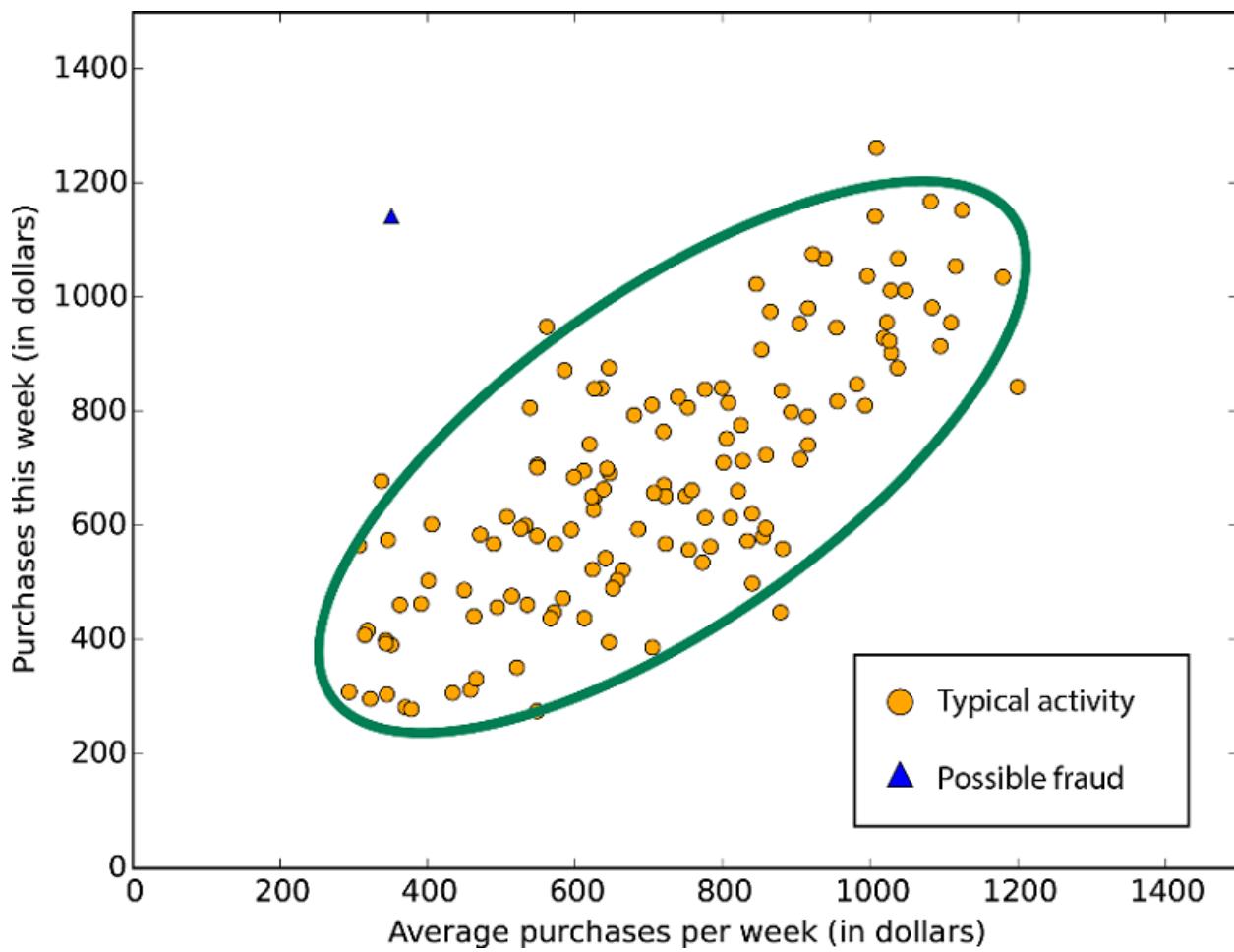
Bayesian methods have a highly desirable quality: they avoid overfitting. They do this by making some assumptions beforehand about the likely distribution of the answer. Another byproduct of this approach is that they have very few parameters. Azure Machine Learning Studio has Bayesian algorithms for both classification ([Two-class Bayes' point machine](#)) and regression ([Bayesian linear regression](#)). Note that these assume that the data can be split or fit with a straight line.

On a historical note, Bayes' point machines were developed at Microsoft Research. They have some exceptionally beautiful theoretical work behind them. The interested student is directed to the [original article in JMLR](#) and an [insightful blog by Chris Bishop](#).

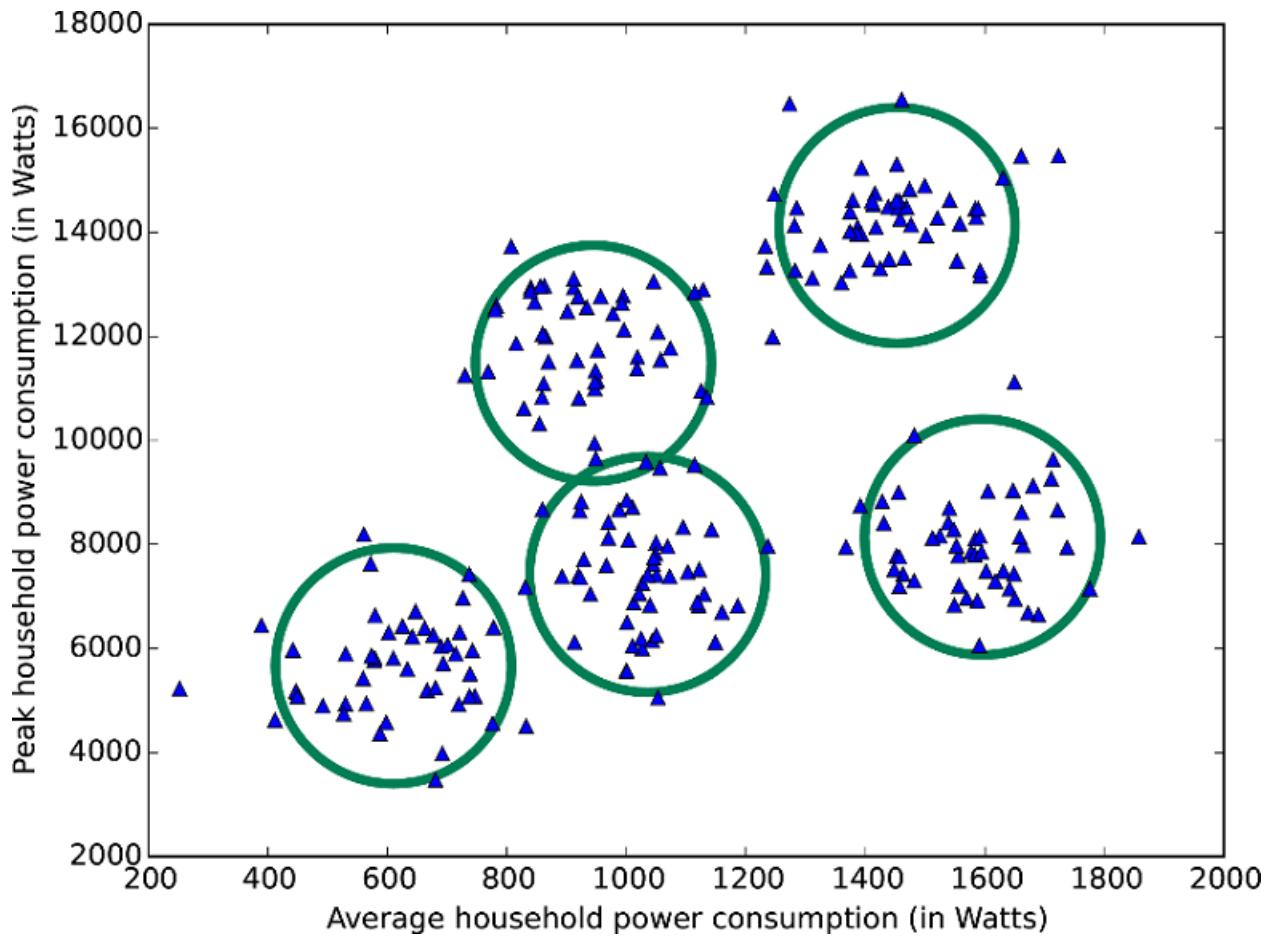
Specialized algorithms

If you have a very specific goal you may be in luck. Within the Azure Machine Learning Studio collection, there are algorithms that specialize in:

- rank prediction ([ordinal regression](#)),
- count prediction ([Poisson regression](#)),
- anomaly detection (one based on [principal components analysis](#) and one based on [support vector machines](#))
- clustering ([K-means](#))

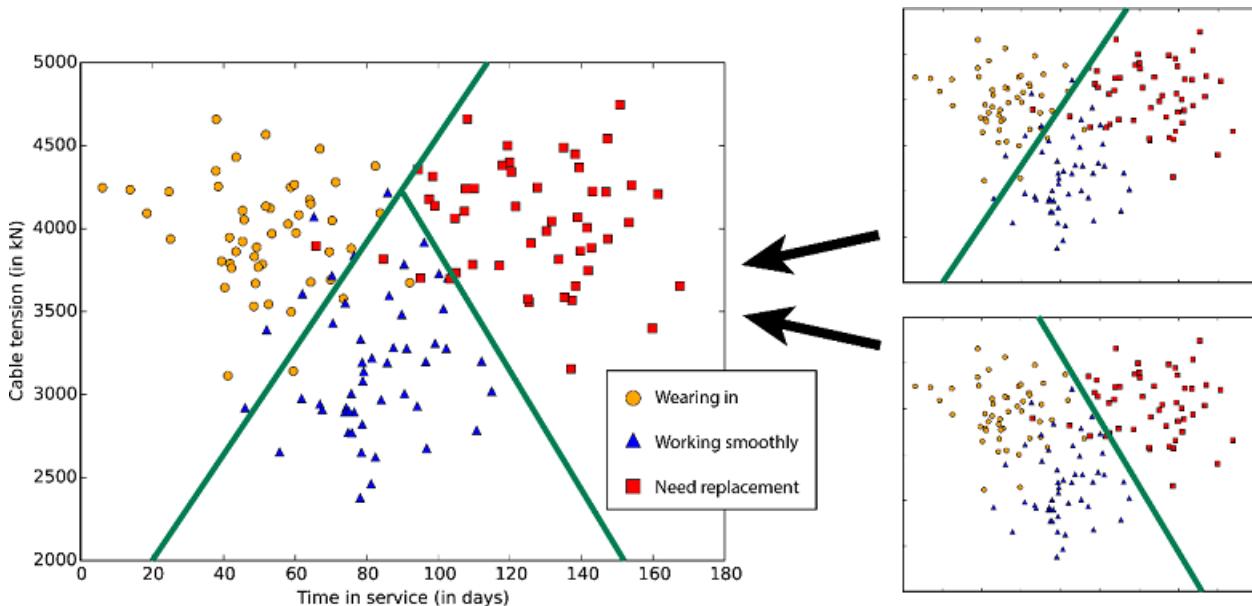


PCA-based anomaly detection - the vast majority of the data falls into a stereotypical distribution; points deviating dramatically from that distribution are suspect



A data set is grouped into five clusters using K-means

There is also an ensemble [one-v-all multiclass classifier](#), which breaks the N-class classification problem into N-1 two-class classification problems. The accuracy, training time, and linearity properties are determined by the two-class classifiers used.



A pair of two-class classifiers combine to form a three-class classifier

Azure Machine Learning Studio also includes access to a powerful machine learning framework under the title of [Vowpal Wabbit](#). VW defies categorization here, since it can learn both classification and regression problems and can even learn from partially unlabeled data. You can configure it to use any one of a number of learning algorithms, loss functions, and optimization algorithms. It was designed from the ground up to be efficient, parallel, and extremely fast. It handles ridiculously large feature sets with little apparent effort. Started and led by Microsoft Research's own John Langford, VW is a Formula One entry in a field of stock car algorithms. Not every problem fits VW, but if yours does, it may be worth your while to climb the learning curve on its interface. It's also available as [stand-alone open source code](#) in several languages.

Next Steps

- To download an easy-to-understand infographic overview of machine learning basics to learn about popular algorithms used to answer common machine learning questions, see [Machine learning basics with algorithm examples](#).
- For a list by category of all the machine learning algorithms available in Machine Learning Studio, see [Initialize Model](#) in the Machine Learning Studio Algorithm and Module Help.
- For a complete alphabetical list of algorithms and modules in Machine Learning Studio, see [A-Z list of Machine Learning Studio modules](#) in Machine Learning Studio Algorithm and Module Help.

Machine learning algorithm cheat sheet for Azure Machine Learning Studio

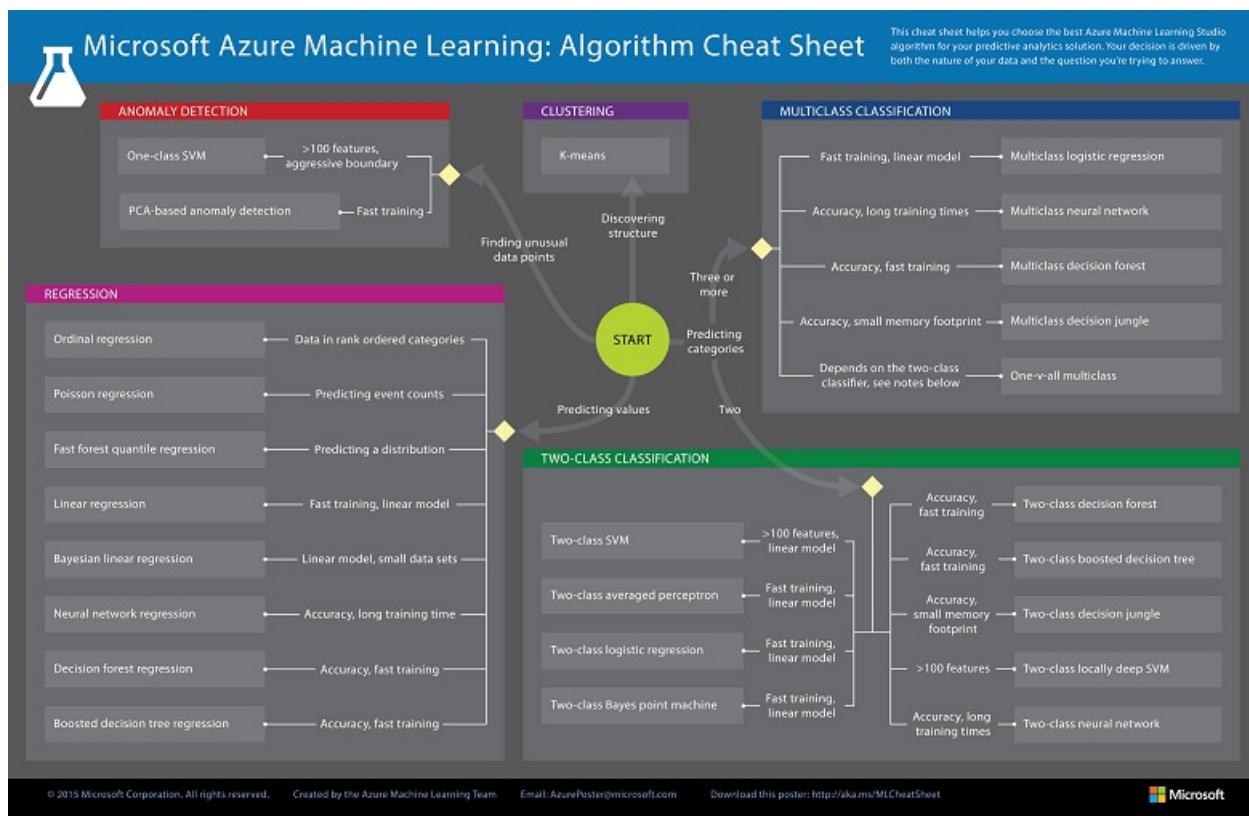
3/12/2019 • 5 minutes to read

The **Azure Machine Learning Studio Algorithm Cheat Sheet** helps you choose the right algorithm for a predictive analytics model.

Azure Machine Learning Studio has a large library of algorithms from the **regression**, **classification**, **clustering**, and **anomaly detection** families. Each is designed to address a different type of machine learning problem.

Download: Machine learning algorithm cheat sheet

Download the cheat sheet here: [Machine Learning Algorithm Cheat Sheet \(11x17 in.\)](#)



Download and print the Machine Learning Studio Algorithm Cheat Sheet in tabloid size to keep it handy and get help choosing an algorithm.

NOTE

For help in using this cheat sheet for choosing the right algorithm, plus a deeper discussion of the different types of machine learning algorithms and how they're used, see [How to choose algorithms for Microsoft Azure Machine Learning Studio](#).

Notes and terminology definitions for the Machine Learning Studio algorithm cheat sheet

- The suggestions offered in this algorithm cheat sheet are approximate rules-of-thumb. Some can be bent, and some can be flagrantly violated. This is intended to suggest a starting point. Don't be afraid to run a

head-to-head competition between several algorithms on your data. There is simply no substitute for understanding the principles of each algorithm and the system that generated your data.

- Every machine learning algorithm has its own style or *inductive bias*. For a specific problem, several algorithms may be appropriate and one algorithm may be a better fit than others. But it's not always possible to know beforehand which is the best fit. In cases like these, several algorithms are listed together in the cheat sheet. An appropriate strategy would be to try one algorithm, and if the results are not yet satisfactory, try the others. Here's an example from the [Azure AI Gallery](#) of an experiment that tries several algorithms against the same data and compares the results: [Compare Multi-class Classifiers: Letter recognition](#).
- There are three main categories of machine learning: **supervised learning**, **unsupervised learning**, and **reinforcement learning**.
 - In **supervised learning**, each data point is labeled or associated with a category or value of interest. An example of a categorical label is assigning an image as either a 'cat' or a 'dog'. An example of a value label is the sale price associated with a used car. The goal of supervised learning is to study many labeled examples like these, and then to be able to make predictions about future data points. For example, identifying new photos with the correct animal or assigning accurate sale prices to other used cars. This is a popular and useful type of machine learning. All of the modules in Azure Machine Learning Studio are supervised learning algorithms except for [K-Means Clustering](#).
 - In **unsupervised learning**, data points have no labels associated with them. Instead, the goal of an unsupervised learning algorithm is to organize the data in some way or to describe its structure. This can mean grouping it into clusters, as K-means does, or finding different ways of looking at complex data so that it appears simpler.
 - In **reinforcement learning**, the algorithm gets to choose an action in response to each data point. It is a common approach in robotics, where the set of sensor readings at one point in time is a data point, and the algorithm must choose the robot's next action. It's also a natural fit for Internet of Things applications. The learning algorithm also receives a reward signal a short time later, indicating how good the decision was. Based on this, the algorithm modifies its strategy in order to achieve the highest reward. Currently there are no reinforcement learning algorithm modules in Azure Machine Learning studio.
- **Bayesian methods** make the assumption of statistically independent data points. This means that the unmodeled variability in one data point is uncorrelated with others, that is, it can't be predicted. For example, if the data being recorded is the number of minutes until the next subway train arrives, two measurements taken a day apart are statistically independent. However, two measurements taken a minute apart are not statistically independent - the value of one is highly predictive of the value of the other.
- **Boosted decision tree regression** takes advantage of feature overlap or interaction among features. That means that, in any given data point, the value of one feature is somewhat predictive of the value of another. For example, in daily high/low temperature data, knowing the low temperature for the day allows you to make a reasonable guess at the high. The information contained in the two features is somewhat redundant.
- Classifying data into more than two categories can be done either by using an inherently multi-class classifier, or by combining a set of two-class classifiers into an **ensemble**. In the ensemble approach, there is a separate two-class classifier for each class - each one separates the data into two categories: "this class" and "not this class." Then these classifiers vote on the correct assignment of the data point. This is the operational principle behind [One-vs-All Multiclass](#).
- Several methods, including logistic regression and the Bayes point machine, assume **linear class boundaries**. That is, they assume that the boundaries between classes are approximately straight lines (or hyperplanes in the more general case). Often this is a characteristic of the data that you don't know until after you've tried to separate it, but it's something that typically can be learned by visualizing beforehand. If

the class boundaries look very irregular, stick with decision trees, decision jungles, support vector machines, or neural networks.

- Neural networks can be used with categorical variables by creating a **dummy variable** for each category, setting it to 1 in cases where the category applies, 0 where it doesn't.

Next steps

- For a downloadable infographic that describes algorithms and provides examples, see [Downloadable Infographic: Machine learning basics with algorithm examples](#).
- For a list by category of all the machine learning algorithms available in Machine Learning Studio, see [Initialize Model](#) in the Machine Learning Studio Algorithm and Module Help.
- For a complete alphabetical list of algorithms and modules in Machine Learning Studio, see [A-Z list of Machine Learning Studio modules](#) in Machine Learning Studio Algorithm and Module Help.

Deploy models to production to play an active role in making business decisions

1/30/2019 • 2 minutes to read

Production deployment enables a model to play an active role in a business. Predictions from a deployed model can be used for business decisions.

Production platforms

There are various approaches and platforms to put models into production. Here are a few options:

- [Where to deploy models with Azure Machine Learning service](#)
- [Deployment of a model in SQL-server](#)
- [Microsoft Machine Learning Server](#)

NOTE

Prior to deployment, one has to insure the latency of model scoring is low enough to use in production.

NOTE

For deployment using Azure Machine Learning Studio, see [Deploy an Azure Machine Learning web service](#).

A/B testing

When multiple models are in production, it can be useful to perform [A/B testing](#) to compare performance of the models.

Next steps

Walkthroughs that demonstrate all the steps in the process for **specific scenarios** are also provided. They are listed and linked with thumbnail descriptions in the [Example walkthroughs](#) article. They illustrate how to combine cloud, on-premises tools, and services into a workflow or pipeline to create an intelligent application.

Machine Learning Anomaly Detection API

3/12/2019 • 10 minutes to read

Overview

[Anomaly Detection API](#) is an example built with Azure Machine Learning that detects anomalies in time series data with numerical values that are uniformly spaced in time.

This API can detect the following types of anomalous patterns in time series data:

- **Positive and negative trends:** For example, when monitoring memory usage in computing an upward trend may be of interest as it may be indicative of a memory leak,
- **Changes in the dynamic range of values:** For example, when monitoring the exceptions thrown by a cloud service, any changes in the dynamic range of values could indicate instability in the health of the service, and
- **Spikes and Dips:** For example, when monitoring the number of login failures in a service or number of checkouts in an e-commerce site, spikes or dips could indicate abnormal behavior.

These machine learning detectors track such changes in values over time and report ongoing changes in their values as anomaly scores. They do not require adhoc threshold tuning and their scores can be used to control false positive rate. The anomaly detection API is useful in several scenarios like service monitoring by tracking KPIs over time, usage monitoring through metrics such as number of searches, numbers of clicks, performance monitoring through counters like memory, CPU, file reads, etc. over time.

The Anomaly Detection offering comes with useful tools to get you started.

- The [web application](#) helps you evaluate and visualize the results of anomaly detection APIs on your data.

NOTE

Try [IT Anomaly Insights solution](#) powered by [this API](#)

API Deployment

In order to use the API, you must deploy it to your Azure subscription where it will be hosted as an Azure Machine Learning web service. You can do this from the [Azure AI Gallery](#). This will deploy two Azure Machine Learning studio Web Services (and their related resources) to your Azure subscription - one for anomaly detection with seasonality detection, and one without seasonality detection. Once the deployment has completed, you will be able to manage your APIs from the [Azure Machine Learning studio web services](#) page. From this page, you will be able to find your endpoint locations, API keys, as well as sample code for calling the API. More detailed instructions are available [here](#).

Scaling the API

By default, your deployment will have a free Dev/Test billing plan which includes 1,000 transactions/month and 2 compute hours/month. You can upgrade to another plan as per your needs. Details on the pricing of different plans are available [here](#) under "Production Web API pricing".

Managing AML Plans

You can manage your billing plan [here](#). The plan name will be based on the resource group name you chose when

deploying the API, plus a string that is unique to your subscription. Instructions on how to upgrade your plan are available [here](#) under the "Managing billing plans" section.

API Definition

The web service provides a REST-based API over HTTPS that can be consumed in different ways including a web or mobile application, R, Python, Excel, etc. You send your time series data to this service via a REST API call, and it runs a combination of the three anomaly types described below.

Calling the API

In order to call the API, you will need to know the endpoint location and API key. Both of these, along with sample code for calling the API, are available from the [Azure Machine Learning studio web services](#) page. Navigate to the desired API, and then click the "Consume" tab to find them. Note that you can call the API as a Swagger API (i.e. with the URL parameter `format=swagger`) or as a non-Swagger API (i.e. without the `format` URL parameter). The sample code uses the Swagger format. Below is an example request and response in non-Swagger format. These examples are to the seasonality endpoint. The non-seasonality endpoint is similar.

Sample Request Body

The request contains two objects: `Inputs` and `GlobalParameters`. In the example request below, some parameters are sent explicitly while others are not (scroll down for a full list of parameters for each endpoint). Parameters that are not sent explicitly in the request will use the default values given below.

```
{  
    "Inputs": {  
        "input1": {  
            "ColumnNames": ["Time", "Data"],  
            "Values": [  
                ["5/30/2010 18:07:00", "1"],  
                ["5/30/2010 18:08:00", "1.4"],  
                ["5/30/2010 18:09:00", "1.1"]  
            ]  
        }  
    },  
    "GlobalParameters": {  
        "tspikedetector.sensitivity": "3",  
        "zspikedetector.sensitivity": "3",  
        "bileveldetector.sensitivity": "3.25",  
        "detectors.spikesdips": "Both"  
    }  
}
```

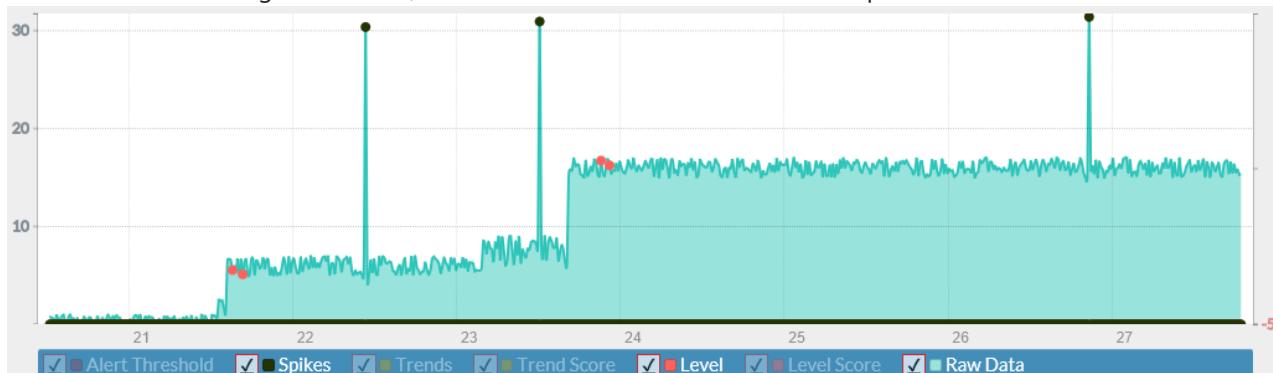
Sample Response

Note that, in order to see the `ColumnNames` field, you must include `details=true` as a URL parameter in your request. See the tables below for the meaning behind each of these fields.

```
{
  "Results": {
    "output1": {
      "type": "table",
      "value": {
        "Values": [
          ["5/30/2010 6:07:00 PM", "1", "1", "0", "0", "-0.687952590518378", "0", "-0.687952590518378", "0", "-0.687952590518378", "0"],
          ["5/30/2010 6:08:00 PM", "1.4", "1.4", "0", "0", "-1.07030497733224", "0", "-0.884548154298423", "0", "-1.07030497733224", "0"],
          ["5/30/2010 6:09:00 PM", "1.1", "1.1", "0", "0", "-1.30229513613974", "0", "-1.173800281031", "0", "-1.30229513613974", "0"]
        ],
        "ColumnNames": ["Time", "OriginalData", "ProcessedData", "TSpike", "ZSpike", "BiLevelChangeScore", "BiLevelChangeAlert", "PosTrendScore", "PosTrendAlert", "NegTrendScore", "NegTrendAlert"],
        "ColumnTypes": ["DateTime", "Double", "Double", "Double", "Double", "Double", "Int32", "Double", "Int32", "Double", "Int32"]
      }
    }
  }
}
```

Score API

The Score API is used for running anomaly detection on non-seasonal time series data. The API runs a number of anomaly detectors on the data and returns their anomaly scores. The figure below shows an example of anomalies that the Score API can detect. This time series has 2 distinct level changes, and 3 spikes. The red dots show the time at which the level change is detected, while the black dots show the detected spikes.



Detectors

The anomaly detection API supports detectors in 3 broad categories. Details on specific input parameters and outputs for each detector can be found in the following table.

Detector Category	Detector	Description	Input Parameters	Outputs
Spike Detectors	TSpike Detector	Detect spikes and dips based on far the values are from first and third quartiles	<i>tspikedetector:sensitivity</i> : takes integer value in the range 1-10, default: 3; Higher values will catch more extreme values thus making it less sensitive	TSpike: binary values – '1' if a spike/dip is detected, '0' otherwise

Detector Category	Detector	Description	Input Parameters	Outputs
Spike Detectors	ZSpike Detector	Detect spikes and dips based on how far the datapoints are from their mean	<code>zspikedetector.sensitivity</code> : take integer value in the range 1-10, default: 3; Higher values will catch more extreme values making it less sensitive	ZSpike: binary values – '1' if a spike/dip is detected, '0' otherwise
Slow Trend Detector	Slow Trend Detector	Detect slow positive trend as per the set sensitivity	<code>trenddetector.sensitivity</code> : threshold on detector score (default: 3.25, 3.25 – 5 is a reasonable range to select this from; The higher the less sensitive)	tscore: floating number representing anomaly score on trend
Level Change Detectors	Bidirectional Level Change Detector	Detect both upward and downward level change as per the set sensitivity	<code>bileveldetector.sensitivity</code> : threshold on detector score (default: 3.25, 3.25 – 5 is a reasonable range to select this from; The higher the less sensitive)	rpscore: floating number representing anomaly score on upward and downward level change

Parameters

More detailed information on these input parameters is listed in the table below:

INPUT PARAMETERS	DESCRIPTION	DEFAULT SETTING	TYPE	VALID RANGE	SUGGESTED RANGE
<code>detectors.history.window</code>	History (in # of data points) used for anomaly score computation	500	integer	10-2000	Time-series dependent
<code>detectors.spikesdips</code>	Whether to detect only spikes, only dips, or both	Both	enumerated	Both, Spikes, Dips	Both
<code>bileveldetector.sensitivity</code>	Sensitivity for bidirectional level change detector.	3.25	double	None	3.25-5 (Lesser values mean more sensitive)
<code>trenddetector.sensitivity</code>	Sensitivity for positive trend detector.	3.25	double	None	3.25-5 (Lesser values mean more sensitive)
<code>tspikedetector.sensitivity</code>	Sensitivity for TSpike Detector	3	integer	1-10	3-5 (Lesser values mean more sensitive)

INPUT PARAMETERS	DESCRIPTION	DEFAULT SETTING	TYPE	VALID RANGE	SUGGESTED RANGE
zspikedetector.sensitivity	Sensitivity for ZSpike Detector	3	integer	1-10	3-5 (Lesser values mean more sensitive)
postprocess.tailRows	Number of the latest data points to be kept in the output results	0	integer	0 (keep all data points), or specify number of points to keep in results	N/A

Output

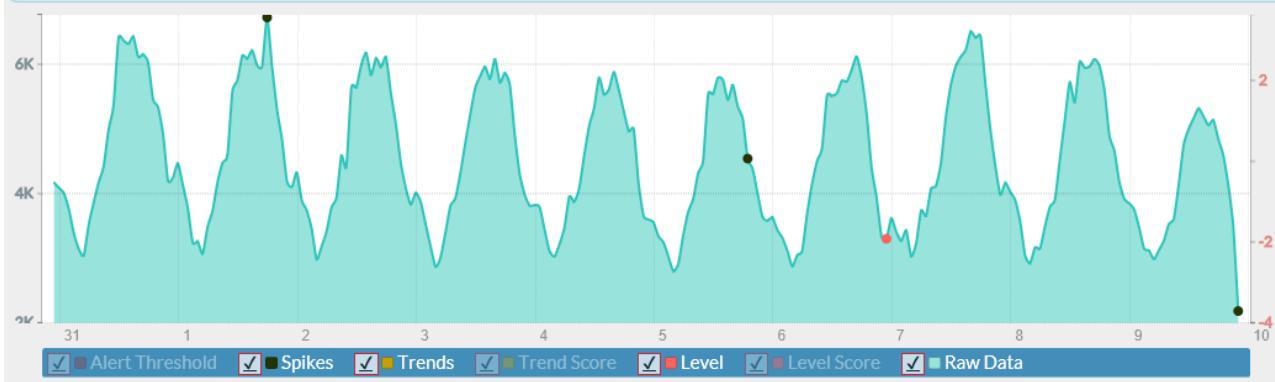
The API runs all detectors on your time series data and returns anomaly scores and binary spike indicators for each point in time. The table below lists outputs from the API.

OUTPUTS	DESCRIPTION
Time	Timestamps from raw data, or aggregated (and/or) imputed data if aggregation (and/or) missing data imputation is applied
Data	Values from raw data, or aggregated (and/or) imputed data if aggregation (and/or) missing data imputation is applied
TSpike	Binary indicator to indicate whether a spike is detected by TSpike Detector
ZSpike	Binary indicator to indicate whether a spike is detected by ZSpike Detector
rpscore	A floating number representing anomaly score on bidirectional level change
rpalert	1/0 value indicating there is a bidirectional level change anomaly based on the input sensitivity
tscore	A floating number representing anomaly score on positive trend
talert	1/0 value indicating there is a positive trend anomaly based on the input sensitivity

ScoreWithSeasonality API

The ScoreWithSeasonality API is used for running anomaly detection on time series that have seasonal patterns. This API is useful to detect deviations in seasonal patterns. The following figure shows an example of anomalies detected in a seasonal time series. The time series has one spike (the 1st black dot), two dips (the 2nd black dot and one at the end), and one level change (red dot). Note that both the dip in the middle of the time series and the level change are only discernable after seasonal components are removed from the series.

With seasonality detection



Detectors

The detectors in the seasonality endpoint are similar to the ones in the non-seasonality endpoint, but with slightly different parameter names (listed below).

Parameters

More detailed information on these input parameters is listed in the table below:

INPUT PARAMETERS	DESCRIPTION	DEFAULT SETTING	TYPE	VALID RANGE	SUGGESTED RANGE
preprocess.aggregationInterval	Aggregation interval in seconds for aggregating input time series	0 (no aggregation is performed)	integer	0: skip aggregation, > 0 otherwise	5 minutes to 1 day, time-series dependent
preprocess.aggregationFunc	Function used for aggregating data into the specified AggregationInterval	mean	enumerated	mean, sum, length	N/A
preprocess.replaceMissing	Values used to impute missing data	lkv (last known value)	enumerated	zero, lkv, mean	N/A
detectors.historyWindow	History (in # of data points) used for anomaly score computation	500	integer	10-2000	Time-series dependent
detectors.spikesdips	Whether to detect only spikes, only dips, or both	Both	enumerated	Both, Spikes, Dips	Both
bileveldetector.sensitivity	Sensitivity for bidirectional level change detector.	3.25	double	None	3.25-5 (Lesser values mean more sensitive)
postrenddetector.sensitivity	Sensitivity for positive trend detector.	3.25	double	None	3.25-5 (Lesser values mean more sensitive)

INPUT PARAMETERS	DESCRIPTION	DEFAULT SETTING	TYPE	VALID RANGE	SUGGESTED RANGE
negtrenddetector.sensitivity	Sensitivity for negative trend detector.	3.25	double	None	3.25-5 (Lesser values mean more sensitive)
tspikedetector.sensitivity	Sensitivity for TSpikes Detector	3	integer	1-10	3-5 (Lesser values mean more sensitive)
zspikedetector.sensitivity	Sensitivity for ZSpikes Detector	3	integer	1-10	3-5 (Lesser values mean more sensitive)
seasonality.enable	Whether seasonality analysis is to be performed	true	boolean	true, false	Time-series dependent
seasonality.numSeasonality	Maximum number of periodic cycles to be detected	1	integer	1, 2	1-2
seasonality.transform	Whether seasonal (and) trend components shall be removed before applying anomaly detection	deseason	enumerated	none, deseason, deseasontrend	N/A
postprocess.tailRows	Number of the latest data points to be kept in the output results	0	integer	0 (keep all data points), or specify number of points to keep in results	N/A

Output

The API runs all detectors on your time series data and returns anomaly scores and binary spike indicators for each point in time. The table below lists outputs from the API.

OUTPUTS	DESCRIPTION
Time	Timestamps from raw data, or aggregated (and/or) imputed data if aggregation (and/or) missing data imputation is applied
OriginalData	Values from raw data, or aggregated (and/or) imputed data if aggregation (and/or) missing data imputation is applied
ProcessedData	<p>Either of the following:</p> <ul style="list-style-type: none"> • Seasonally adjusted time series if significant seasonality has been detected and deseason option selected; • seasonally adjusted and detrended time series if significant seasonality has been detected and deseasontrend option selected • otherwise, this is the same as OriginalData

OUTPUTS	DESCRIPTION
TSpike	Binary indicator to indicate whether a spike is detected by TSpike Detector
ZSpike	Binary indicator to indicate whether a spike is detected by ZSpike Detector
BiLevelChangeScore	A floating number representing anomaly score on level change
BiLevelChangeAlert	1/0 value indicating there is a level change anomaly based on the input sensitivity
PosTrendScore	A floating number representing anomaly score on positive trend
PosTrendAlert	1/0 value indicating there is a positive trend anomaly based on the input sensitivity
NegTrendScore	A floating number representing anomaly score on negative trend
NegTrendAlert	1/0 value indicating there is a negative trend anomaly based on the input sensitivity

Azure AI guide for predictive maintenance solutions

3/12/2019 • 42 minutes to read

Summary

Predictive maintenance (**PdM**) is a popular application of predictive analytics that can help businesses in several industries achieve high asset utilization and savings in operational costs. This guide brings together the business and analytical guidelines and best practices to successfully develop and deploy PdM solutions using the [Microsoft Azure AI platform](#) technology.

For starters, this guide introduces industry-specific business scenarios and the process of qualifying these scenarios for PdM. The data requirements and modeling techniques to build PdM solutions are also provided. The main content of the guide is on the data science process - including the steps of data preparation, feature engineering, model creation, and model operationalization. To complement these key concepts, this guide lists a set of solution templates to help accelerate PdM application development. The guide also points to useful training resources for the practitioner to learn more about the AI behind the data science.

Data Science guide overview and target audience

The first half of this guide describes typical business problems, the benefits of implementing PdM to address these problems, and lists some common use cases. Business decision makers (BDMs) will benefit from this content. The second half explains the data science behind PdM, and provides a list of PdM solutions built using the principles outlined in this guide. It also provides learning paths and pointers to training material. Technical decision makers (TDMs) will find this content useful.

START WITH ...	IF YOU ARE ...
Business case for predictive maintenance	a business decision maker (BDM) looking to reduce downtime and operational costs, and improve utilization of equipment
Data Science for predictive maintenance	a technical decision maker (TDM) evaluating PdM technologies to understand the unique data processing and AI requirements for predictive maintenance
Solution templates for predictive maintenance	a software architect or AI Developer looking to quickly stand up a demo or a proof-of-concept
Training resources for predictive maintenance	any or all of the above, and want to learn the foundational concepts behind the data science, tools, and techniques.

Prerequisite knowledge

The BDM content does not expect the reader to have any prior data science knowledge. For the TDM content, basic knowledge of statistics and data science is helpful. Knowledge of Azure Data and AI services, Python, R, XML, and JSON is recommended. AI techniques are implemented in Python and R packages. Solution templates are implemented using Azure services, development tools, and SDKs.

Business case for predictive maintenance

Businesses require critical equipment to be running at peak efficiency and utilization to realize their return on capital investments. These assets could range from aircraft engines, turbines, elevators, or industrial chillers - that cost millions - down to everyday appliances like photocopiers, coffee machines, or water coolers.

- By default, most businesses rely on *corrective maintenance*, where parts are replaced as and when they fail. Corrective maintenance ensures parts are used completely (therefore not wasting component life), but costs the business in downtime, labor, and unscheduled maintenance requirements (off hours, or inconvenient locations).
- At the next level, businesses practice *preventive maintenance*, where they determine the useful lifespan for a part, and maintain or replace it before a failure. Preventive maintenance avoids unscheduled and catastrophic failures. But the high costs of scheduled downtime, under-utilization of the component before its full lifetime of use, and labor still remain.
- The goal of *predictive maintenance* is to optimize the balance between corrective and preventative maintenance, by enabling *just in time* replacement of components. This approach only replaces those components when they are close to a failure. By extending component lifespans (compared to preventive maintenance) and reducing unscheduled maintenance and labor costs (over corrective maintenance), businesses can gain cost savings and competitive advantages.

Business problems in PdM

Businesses face high operational risk due to unexpected failures and have limited insight into the root cause of problems in complex systems. Some of the key business questions are:

- Detect anomalies in equipment or system performance or functionality.
- Predict whether an asset may fail in the near future.
- Estimate the remaining useful life of an asset.
- Identify the main causes of failure of an asset.
- Identify what maintenance actions need to be done, by when, on an asset.

Typical goal statements from PdM are:

- Reduce operational risk of mission critical equipment.
- Increase rate of return on assets by predicting failures before they occur.
- Control cost of maintenance by enabling just-in-time maintenance operations.
- Lower customer attrition, improve brand image, and lost sales.
- Lower inventory costs by reducing inventory levels by predicting the reorder point.
- Discover patterns connected to various maintenance problems.
- Provide KPIs (key performance indicators) such as health scores for asset conditions.
- Estimate remaining lifespan of assets.
- Recommend timely maintenance activities.
- Enable just in time inventory by estimating order dates for replacement of parts.

These goal statements are the starting points for:

- *data scientists* to analyze and solve specific predictive problems.
- *cloud architects and developers* to put together an end to end solution.

Qualifying problems for predictive maintenance

It is important to emphasize that not all use cases or business problems can be effectively solved by PdM. There are three important qualifying criteria that need to be considered during problem selection:

- The problem has to be predictive in nature; that is, there should be a target or an outcome to predict. The problem should also have a clear path of action to prevent failures when they are detected.
- The problem should have a record of the operational history of the equipment that contains *both good and bad outcomes*. The set of actions taken to mitigate bad outcomes should also be available as part of these records. Error reports, maintenance logs of performance degradation, repair, and replace logs are also important. In addition, repairs undertaken to improve them, and replacement records are also useful.

- The recorded history should be reflected in *relevant* data that is of *sufficient* enough quality to support the use case. For more information about data relevance and sufficiency, see [Data requirements for predictive maintenance](#).
- Finally, the business should have domain experts who have a clear understanding of the problem. They should be aware of the internal processes and practices to be able to help the analyst understand and interpret the data. They should also be able to make the necessary changes to existing business processes to help collect the right data for the problems, if needed.

Sample PdM use cases

This section focuses on a collection of PdM use cases from several industries such as Aerospace, Utilities, and Transportation. Each section starts with a business problem, and discusses the benefits of PdM, the relevant data surrounding the business problem, and finally the benefits of a PdM solution.

BUSINESS PROBLEM	BENEFITS FROM PDM
Aviation	
<i>Flight delay and cancellations</i> due to mechanical problems. Failures that cannot be repaired in time may cause flights to be canceled, and disrupt scheduling and operations.	PdM solutions can predict the probability of an aircraft being delayed or canceled due to mechanical failures.
<i>Aircraft engine parts failure:</i> Aircraft engine part replacements are among the most common maintenance tasks within the airline industry. Maintenance solutions require careful management of component stock availability, delivery, and planning	Being able to gather intelligence on component reliability leads to substantial reduction on investment costs.
Finance	
<i>ATM failure</i> is a common problem within the banking industry. The problem here is to report the probability that an ATM cash withdrawal transaction gets interrupted due to a paper jam or part failure in the cash dispenser. Based on predictions of transaction failures, ATMs can be serviced proactively to prevent failures from occurring.	Rather than allow the machine to fail midway through a transaction, the desirable alternative is to program the machine to deny service based on the prediction.
Energy	
<i>Wind turbine failures:</i> Wind turbines are the main energy source in environmentally responsible countries, and involve high capital costs. A key component in wind turbines is the generator motor. Its failure renders the turbine ineffective. It is also highly expensive to fix.	Predicting KPIs such as MTTF (mean time to failure) can help the energy companies prevent turbine failures, and ensure minimal downtime. Failure probabilities will inform technicians to monitor turbines that are likely to fail soon, and schedule time-based maintenance regimes. Predictive models provide insights into different factors that contribute to the failure, which helps technicians better understand the root causes of problems.
<i>Circuit breaker failures:</i> Distribution of electricity to homes and businesses requires power lines to be operational at all times to guarantee energy delivery. Circuit breakers help limit or avoid damage to power lines during overloading or adverse weather conditions. The business problem here is to predict circuit breaker failures.	PdM solutions help reduce repair costs and increase the lifespan of equipment such as circuit breakers. They help improve the quality of the power network by reducing unexpected failures and service interruptions.
Transportation and logistics	

BUSINESS PROBLEM	BENEFITS FROM PDM
<i>Elevator door failures:</i> Large elevator companies provide a full stack service for millions of functional elevators around the world. Elevator safety, reliability, and uptime are the main concerns for their customers. These companies track these and various other attributes via sensors, to help them with corrective and preventive maintenance. In an elevator, the most prominent customer problem is malfunctioning elevator doors. The business problem in this case is to provide a knowledge base predictive application that predicts the potential causes of door failures.	Elevators are capital investments for potentially a 20-30 year lifespan. So each potential sale can be highly competitive; hence expectations for service and support are high. Predictive maintenance can provide these companies with an advantage over their competitors in their product and service offerings.
<i>Wheel failures:</i> Wheel failures account for half of all train derailments and cost billions to the global rail industry. Wheel failures also cause rails to deteriorate, sometimes causing the rail to break prematurely. Rail breaks lead to catastrophic events such as derailments. To avoid such instances, railways monitor the performance of wheels and replace them in a preventive manner. The business problem here is the prediction of wheel failures.	Predictive maintenance of wheels will help with just-in-time replacement of wheels
<i>Subway train door failures:</i> A major reason for delays in subway operations is door failures of train cars. The business problem here is to predict train door failures.	Early awareness of a door failure, or the number of days until a door failure, will help the business optimize train door servicing schedules.

The next section gets into the details of how to realize the PdM benefits discussed above.

Data Science for predictive maintenance

This section provides general guidelines of data science principles and practice for PdM. It is intended to help a TDM, solution architect, or a developer understand the prerequisites and process for building end-to-end AI applications for PdM. You can read this section along with a review of the demos and proof-of-concept templates listed in [Solution Templates for predictive maintenance](#). You can then use these principles and best practices to implement your PdM solution in Azure.

NOTE

This guide is NOT intended to teach the reader Data Science. Several helpful sources are provided for further reading in the section for [training resources for predictive maintenance](#). The [solution templates](#) listed in the guide demonstrate some of these AI techniques for specific PdM problems.

Data requirements for predictive maintenance

The success of any learning depends on (a) the quality of what is being taught, and (b) the ability of the learner. Predictive models learn patterns from historical data, and predict future outcomes with certain probability based on these observed patterns. A model's predictive accuracy depends on the relevancy, sufficiency, and quality of the training and test data. The new data that is 'scored' using this model should have the same features and schema as the training/test data. The feature characteristics (type, density, distribution, and so on) of new data should match that of the training and test data sets. The focus of this section is on such data requirements.

Relevant data

First, the data has to be *relevant to the problem*. Consider the *wheel failure* use case discussed above - the training data should contain features related to the wheel operations. If the problem was to predict the failure of the *traction system*, the training data has to encompass all the different components for the traction system. The first case targets a specific component whereas the second case targets the failure of a larger subsystem. The general

recommendation is to design prediction systems about specific components rather than larger subsystems, since the latter will have more dispersed data. The domain expert (see [Qualifying problems for predictive maintenance](#)) should help in selecting the most relevant subsets of data for the analysis. The relevant data sources are discussed in greater detail in [Data preparation for predictive maintenance](#).

Sufficient data

Two questions are commonly asked with regard to failure history data: (1) "How many failure events are required to train a model?" (2) "How many records is considered as "enough"?" There are no definitive answers, but only rules of thumb. For (1), more the number of failure events, better the model. For (2), and the exact number of failure events depends on the data and the context of the problem being solved. But on the flip side, if a machine fails too often then the business will replace it, which will reduce failure instances. Here again, the guidance from the domain expert is important. However, there are methods to cope with the issue of *rare events*. They are discussed in the section [Handling imbalanced data](#).

Quality data

The quality of the data is critical - each predictor attribute value must be *accurate* in conjunction with the value of the target variable. Data quality is a well-studied area in statistics and data management, and hence out of scope for this guide.

NOTE

There are several resources and enterprise products to deliver quality data. A sample of references is provided below:

- Dasu, T, Johnson, T, *Exploratory Data Mining and Data Cleaning*, Wiley, 2003.
- [Exploratory Data Analysis, Wikipedia](#)
- [Hellerstein, J, Quantitative Data Cleaning for Large Databases](#)
- [de Jonge, E, van der loo, M, Introduction to Data Cleaning with R](#)

Data preparation for predictive maintenance

Data sources

The relevant data sources for predictive maintenance include, but are not limited to:

- Failure history
- Maintenance/repair history
- Machine operating conditions
- Equipment metadata

Failure history

Failure events are rare in PdM applications. However, when building prediction models, the algorithm needs to learn about a component's normal operational pattern, as well as its failure patterns. So the training data should contain sufficient number of examples from both categories. Maintenance records and parts replacement history are good sources to find failure events. With the help of some domain knowledge, anomalies in the training data can also be defined as failures.

Maintenance/repair history

Maintenance history of an asset contains details about components replaced, repair activities performed etc. These events record degradation patterns. Absence of this crucial information in the training data can lead to misleading model results. Failure history can also be found within maintenance history as special error codes, or order dates for parts. Additional data sources that influence failure patterns should be investigated and provided by domain experts.

Machine operating conditions

Sensor based (or other) streaming data of the equipment in operation is an important data source. A key

assumption in PdM is that a machine's health status degrades over time during its routine operation. The data is expected to contain time-varying features that capture this aging pattern, and any anomalies that lead to degradation. The temporal aspect of the data is required for the algorithm to learn the failure and non-failure patterns over time. Based on these data points, the algorithm learns to predict how many more units of time a machine can continue to work before it fails.

Static feature data

Static features are metadata about the equipment. Examples are the equipment make, model, manufactured date, start date of service, location of the system, and other technical specifications.

Examples of relevant data for the [sample PdM use cases](#) are tabulated below:

USE CASE	EXAMPLES OF RELEVANT DATA
<i>Flight delay and cancellations</i>	Flight route information in the form of flight legs and page logs. Flight leg data includes routing details such as departure/arrival date, time, airport, layovers etc. Page log includes a series of error and maintenance codes recorded by the ground maintenance personnel.
<i>Aircraft engine parts failure</i>	Data collected from sensors in the aircraft that provide information on the condition of the various parts. Maintenance records help identify when component failures occurred and when they were replaced.
<i>ATM Failure</i>	Sensor readings for each transaction (depositing cash/check) and dispensing of cash. Information on gap measurement between notes, note thickness, note arrival distance, check attributes etc. Maintenance records that provide error codes, repair information, last time the cash dispenser was refilled.
<i>Wind turbine failure</i>	Sensors monitor turbine conditions such as temperature, wind direction, power generated, generator speed etc. Data is gathered from multiple wind turbines from wind farms located in various regions. Typically, each turbine will have multiple sensor readings relaying measurements at a fixed time interval.
<i>Circuit breaker failures</i>	Maintenance logs that include corrective, preventive, and systematic actions. Operational data that includes automatic and manual commands sent to circuit breakers such as for open and close actions. Device metadata such as date of manufacture, location, model, etc. Circuit breaker specifications such as voltage levels, geolocation, ambient conditions.
<i>Elevator door failures</i>	Elevator metadata such as type of elevator, manufactured date, maintenance frequency, building type, and so on. Operational information such as number of door cycles, average door close time. Failure history with causes.
<i>Wheel failures</i>	Sensor data that measures wheel acceleration, braking instances, driving distance, velocity etc. Static information on wheels like manufacturer, manufactured date. Failure data inferred from part order database that track order dates and quantities.
<i>Subway train door failures</i>	Door opening and closing times, other operational data such as current condition of train doors. Static data would include asset identifier, time, and condition value columns.

Data types

Given the above data sources, the two main data types observed in PdM domain are:

- *Temporal data*: Operational telemetry, machine conditions, work order types, priority codes that will have timestamps at the time of recording. Failure, maintenance/repair, and usage history will also have timestamps associated with each event.
- *Static data*: Machine features and operator features in general are static since they describe the technical specifications of machines or operator attributes. If these features could change over time, they should also have timestamps associated with them.

Predictor and target variables should be preprocessed/transformed into [numerical, categorical, and other data types](#) depending on the algorithm being used.

Data preprocessing

As a prerequisite to *feature engineering*, prepare the data from various streams to compose a schema from which it is easy to build features. Visualize the data first as a table of records. Each row in the table represents a training instance, and the columns represent *predictor* features (also called independent attributes or variables). Organize the data such that the last column(s) is the *target* (dependent variable). For each training instance, assign a *label* as the value of this column.

For temporal data, divide the duration of sensor data into time units. Each record should belong to a time unit for an asset, *and should offer distinct information*. Time units are defined based on business needs in multiples of seconds, minutes, hours, days, months, and so on. The time unit *does not have to be the same as the frequency of data collection*. If the frequency is high, the data may not show any significant difference from one unit to the other. For example, assume that ambient temperature was collected every 10 seconds. Using that same interval for training data only inflates the number of examples without providing any additional information. For this case, a better strategy would be to use average the data over 10 minutes, or an hour based on the business justification.

For static data,

- *Maintenance records*: Raw maintenance data has an asset identifier and timestamp with information on maintenance activities that have been performed at a given point in time. Transform maintenance activities into *categorical* columns, where each category descriptor uniquely maps to a specific maintenance action. The schema for maintenance records would include asset identifier, time, and maintenance action.
- *Failure records*: Failures or failure reasons can be recorded as specific error codes or failure events defined by specific business conditions. In cases where the equipment has multiple error codes, the domain expert should help identify the ones that are pertinent to the target variable. Use the remaining error codes or conditions to construct *predictor* features that correlate with these failures. The schema for failure records would include asset identifier, time, failure, or failure reason - if available.
- *Machine and operator metadata*: Merge the machine and operator data into one schema to associate an asset with its operator, along with their respective attributes. The schema for machine conditions would include asset identifier, asset features, operator identifier, and operator features.

Other data preprocessing steps include *handling missing values* and *normalization* of attribute values. A detailed discussion is beyond the scope of this guide - see the next section for some useful references.

With the above preprocessed data sources in place, the final transformation before feature engineering is to join the above tables based on the asset identifier and timestamp. The resulting table would have null values for the failure column when machine is in normal operation. These null values can be imputed by an indicator for normal operation. Use this failure column to create *labels for the predictive model*. For more information, see the section on [modeling techniques for predictive maintenance](#).

Feature engineering

Feature engineering is the first step prior to modeling the data. Its role in the data science process [is described here](#). A *feature* is a predictive attribute for the model - such as temperature, pressure, vibration, and so on. For PdM, feature engineering involves abstracting a machine's health over historical data collected over a sizable duration. In that sense, it is different from its peers such as remote monitoring, anomaly detection, and failure detection.

Time windows

Remote monitoring entails reporting the events that happen as of *points in time*. Anomaly detection models evaluate (score) incoming streams of data to flag anomalies as of points in time. Failure detection classifies failures to be of specific types as they occur points in time. In contrast, PdM involves predicting failures over a *future time period*, based on features that represent machine behavior over *historical time period*. For PdM, feature data from individual points of time are too noisy to be predictive. So the data for each feature needs to be *smoothed* by aggregating data points over time windows.

Lag features

The business requirements define how far the model has to predict into the future. In turn, this duration helps define 'how far back the model has to look' to make these predictions. This 'looking back' period is called the *lag*, and features engineered over this lag period are called *lag features*. This section discusses lag features that can be constructed from data sources with timestamps, and feature creation from static data sources. Lag features are typically *numerical* in nature.

IMPORTANT

The window size is determined via experimentation, and should be finalized with the help of a domain expert. The same caveat holds for the selection and definition of lag features, their aggregations, and the type of windows.

Rolling aggregates

For each record of an asset, a rolling window of size "W" is chosen as the number of units of time to compute the aggregates. Lag features are then computed using the W periods *before the date* of that record. In Figure 1, the blue lines show sensor values recorded for an asset for each unit of time. They denote a rolling average of feature values over a window of size $W=3$. The rolling average is computed over all records with timestamps in the range t_1 (in orange) to t_2 (in green). The value for W is typically in minutes or hours depending on the nature of the data. But for certain problems, picking a large W (say 12 months) can provide the whole history of an asset until the time of the record.

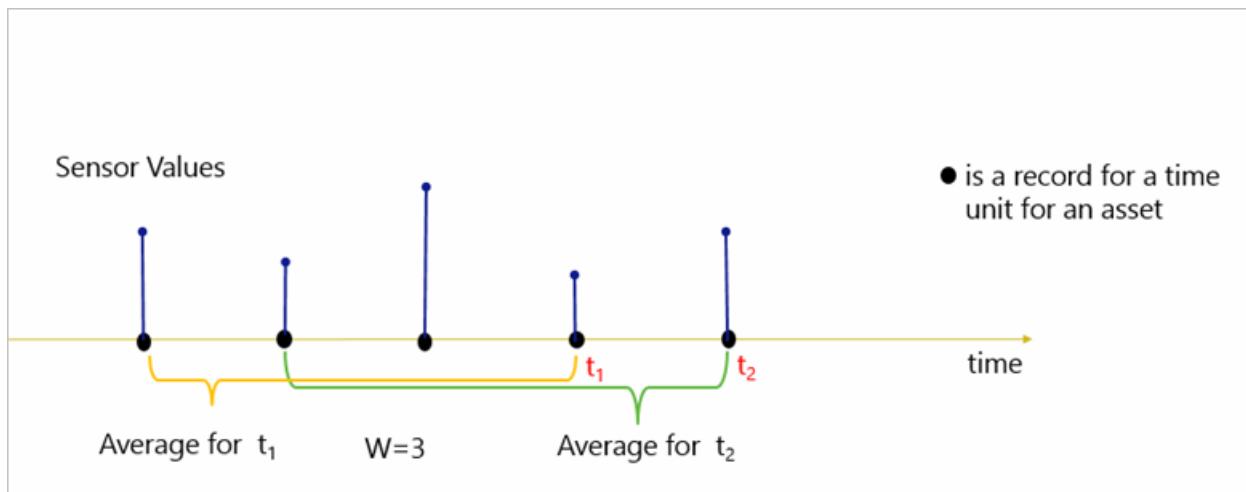


Figure 1. Rolling aggregate features

Examples of rolling aggregates over a time window are count, average, CUMESUM (cumulative sum) measures, min/max values. In addition, variance, standard deviation, and count of outliers beyond N standard deviations are often used. Examples of aggregates that may be applied for the [use cases](#) in this guide are listed below.

- *Flight delay*: count of error codes over the last day/week.
- *Aircraft engine part failure*: rolling means, standard deviation, and sum over the past day, week etc. This metric should be determined along with the business domain expert.
- *ATM failures*: rolling means, median, range, standard deviations, count of outliers beyond three standard deviations, upper and lower CUMESUM.
- *Subway train door failures*: Count of events over past day, week, two weeks etc.
- *Circuit breaker failures*: Failure counts over past week, year, three years etc.

Another useful technique in PdM is to capture trend changes, spikes, and level changes using algorithms that detect anomalies in data.

Tumbling aggregates

For each labeled record of an asset, a window of size W_k is defined, where k is the number of windows of size W . Aggregates are then created over k *tumbling windows* $W_{-k}, W_{-(k-1)}, \dots, W_{-2}, W_{-1}$ for the periods before a record's timestamp. k can be a small number to capture short-term effects, or a large number to capture long-term degradation patterns. (see Figure 2).

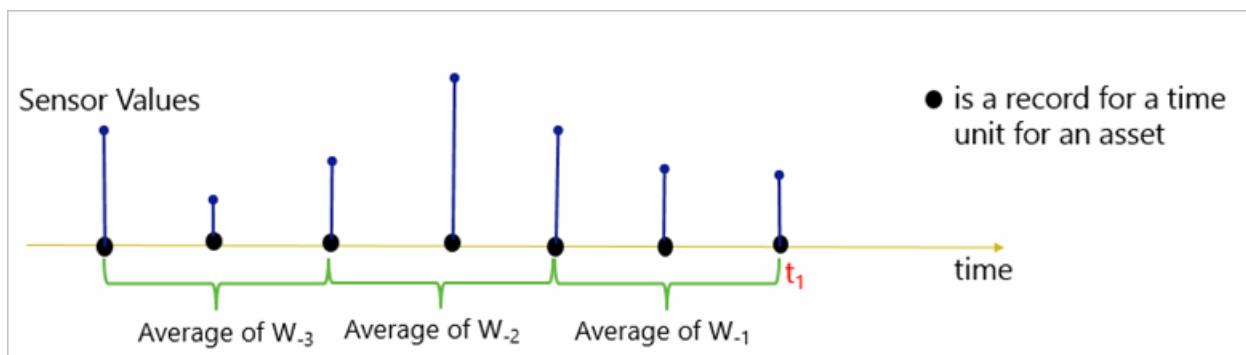


Figure 2. Tumbling aggregate features

For example, lag features for the wind turbines use case may be created with $W=1$ and $k=3$. They imply the lag for each of the past three months using top and bottom outliers.

Static features

Technical specifications of the equipment such as date of manufacture, model number, location, are some examples of static features. They are treated as *categorical* variables for modeling. Some examples for the circuit breaker use case are voltage, current, power capacity, transformer type, and power source. For wheel failures, the type of tire wheels (alloy vs steel) is an example.

The data preparation efforts discussed so far should lead to the data being organized as shown below. Training, test, and validation data should have this logical schema (this example shows time in units of days).

ASSET ID	TIME	LABEL
A123	Day 1	...
A123	Day 2	...
...
B234	Day 1	...
B234	Day 2	...
...

The last step in feature engineering is the **labeling** of the target variable. This process is dependent on the modeling technique. In turn, the modeling technique depends on the business problem and nature of the available data. Labeling is discussed in the next section.

IMPORTANT

Data preparation and feature engineering are as important as modeling techniques to arrive at successful PdM solutions. The domain expert and the practitioner should invest significant time in arriving at the right features and data for the model. A small sample from many books on feature engineering are listed below:

- Pyle, D. Data Preparation for Data Mining (The Morgan Kaufmann Series in Data Management Systems), 1999
- Zheng, A., Casari, A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, O'Reilly, 2018.
- Dong, G. Liu, H. (Editors), Feature Engineering for Machine Learning and Data Analytics (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series), CRC Press, 2018.

Modeling techniques for predictive maintenance

This section discusses the main modeling techniques for PdM problems, along with their specific label construction methods. Notice that a single modeling technique can be used across different industries. The modeling technique is paired to the data science problem, rather than the context of the data at hand.

IMPORTANT

The choice of labels for the failure cases and the labeling strategy should be determined in consultation with the domain expert.

Binary classification

Binary classification is used to *predict the probability that a piece of equipment fails within a future time period* - called the *future horizon period X*. X is determined by the business problem and the data at hand, in consultation with the domain expert. Examples are:

- *minimum lead time* required to replace components, deploy maintenance resources, perform maintenance to avoid a problem that is likely to occur in that period.
- *minimum count of events* that can happen before a problem occurs.

In this technique, two types of training examples are identified. A positive example, *which indicates a failure*, with label = 1. A negative example, which indicates normal operations, with label = 0. The target variable, and hence the label values, are *categorical*. The model should identify each new example as likely to fail or work normally over the next X time units.

Label construction for binary classification

The question here is: "What is the probability that the asset will fail in the next X units of time?" To answer this question, label X records prior to the failure of an asset as "about to fail" (label = 1), and label all other records as being "normal" (label = 0). (see Figure 3).

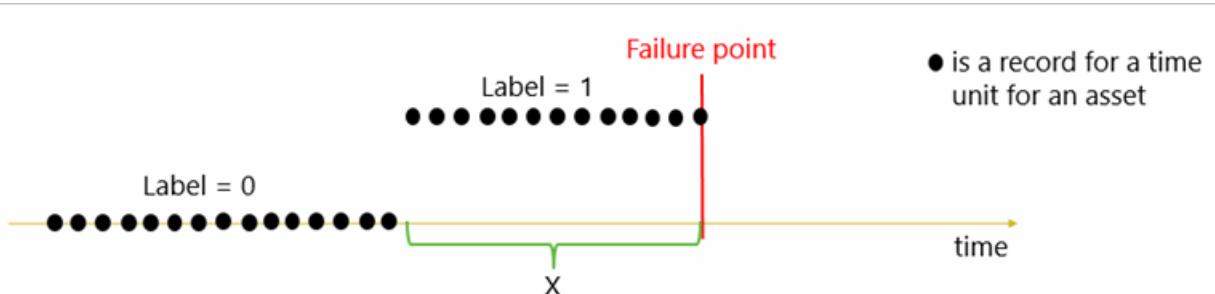


Figure 3. Labeling for binary classification

Examples of labeling strategy for some of the use cases are listed below.

- *Flight delays:* X may be chosen as 1 day, to predict delays in the next 24 hours. Then all flights that are within 24 hours before failures are labeled as 1.
- *ATM cash dispense failures:* A goal may be to determine failure probability of a transaction in the next one hour. In that case, all transactions that happened within the past hour of the failure are labeled as 1. To predict failure probability over the next N currency notes dispensed, all notes dispensed within the last N notes of a failure are labeled as 1.
- *Circuit breaker failures:* The goal may be to predict the next circuit breaker command failure. In that case, X is chosen to be one future command.
- *Train door failures:* X may be chosen as two days.
- *Wind turbine failures:* X may be chosen as two months.

Regression for predictive maintenance

Regression models are used to *compute the remaining useful life (RUL) of an asset*. RUL is defined as the amount of time that an asset is operational before the next failure occurs. Each training example is a record that belongs to a time unit nY for an asset, where n is the multiple. The model should calculate the RUL of each new example as a *continuous number*. This number denotes the period of time remaining before the failure.

Label construction for regression

The question here is: "What is the remaining useful life (RUL) of the equipment?" For each record prior to the failure, calculate the label to be the number of units of time remaining before the next failure. In this method, labels are continuous variables. (See Figure 4)

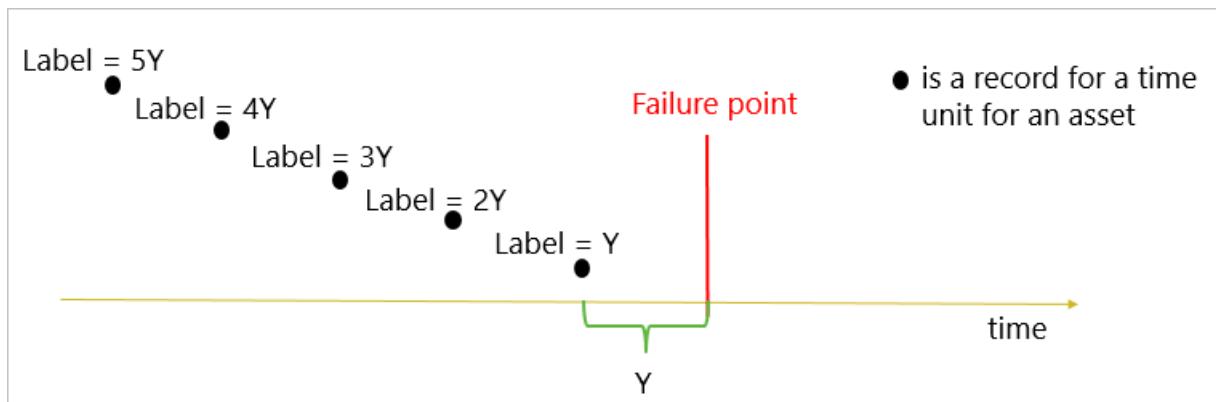


Figure 4. Labeling for regression

For regression, labeling is done with reference to a failure point. Its calculation is not possible without knowing how long the asset has survived before a failure. So in contrast to binary classification, assets without any failures in the data cannot be used for modeling. This issue is best addressed by another statistical technique called [Survival Analysis](#). But potential complications may arise when applying this technique to PdM use cases that involve time-varying data with frequent intervals. For more information on Survival Analysis, see [this one-pager](#).

Multi-class classification for predictive maintenance

Multi-class classification techniques can be used in PdM solutions for two scenarios:

- *Predict two future outcomes:* The first outcome is *a range of time to failure* for an asset. The asset is assigned to one of multiple possible periods of time. The second outcome is the likelihood of failure in a future period due to *one of the multiple root causes*. This prediction enables the maintenance crew to watch for symptoms and plan maintenance schedules.
- *Predict the most likely root cause* of a given failure. This outcome recommends the right set of maintenance actions to fix a failure. A ranked list of root causes and recommended repairs can help technicians prioritize their repair actions after a failure.

Label construction for multi-class classification

The question here is: "What is the probability that an asset will fail in the next nZ units of time where n is the number of periods?" To answer this question, label nZ records prior to the failure of an asset using buckets of time ($3Z, 2Z, Z$). Label all other records as "normal" (label = 0). In this method, the target variable holds *categorical* values. (See Figure 5).

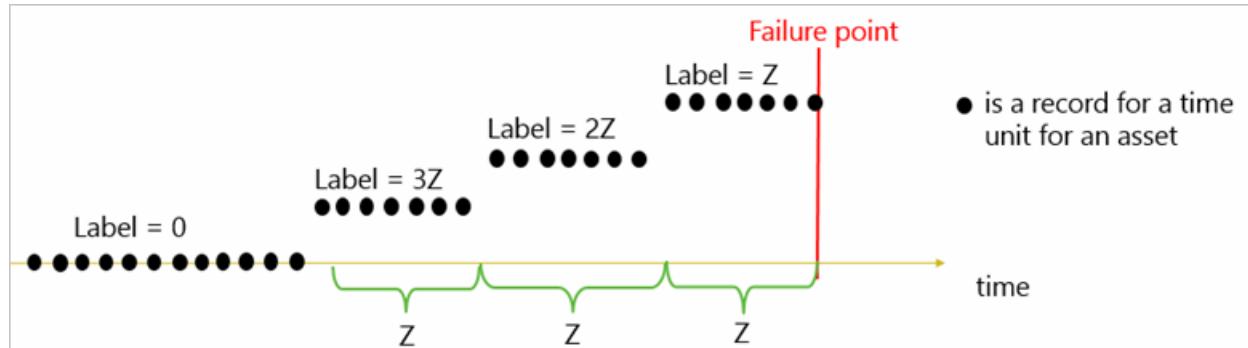


Figure 5. Labeling for multi-class classification for failure time prediction

The question here is: "What is the probability that the asset will fail in the next X units of time due to root cause/problem P_i ?" where i is the number of possible root causes. To answer this question, label X records prior to the failure of an asset as "about to fail due to root cause P_i " (label = P_i). Label all other records as being "normal" (label = 0). In this method also, labels are categorical (See Figure 6).

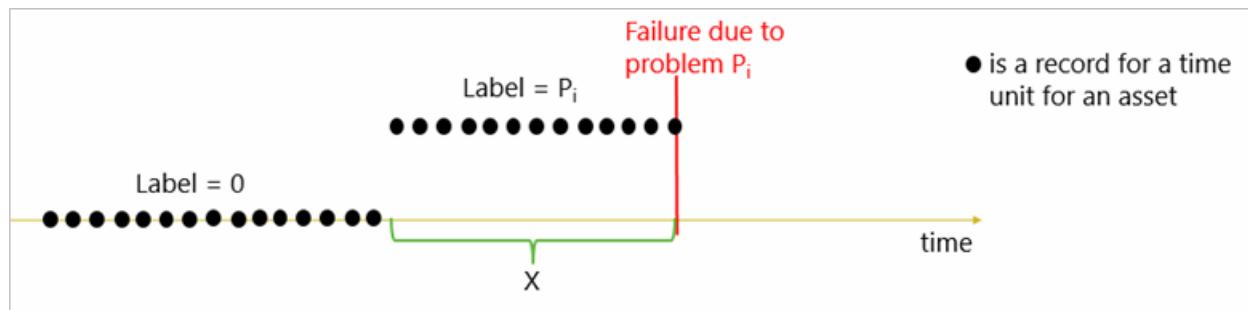


Figure 6. Labeling for multi-class classification for root cause prediction

The model assigns a failure probability due to each P_i as well as the probability of no failure. These probabilities can be ordered by magnitude to allow prediction of the problems that are most likely to occur in the future.

The question here is: "What maintenance actions do you recommend after a failure?" To answer this question, labeling *does not require a future horizon to be picked*, because the model is not predicting failure in the future. It is just predicting the most likely root cause *once the failure has already happened*.

Training, validation, and testing methods for predictive maintenance

The [Team Data Science Process](#) provides a full coverage of the model train-test-validate cycle. This section discusses aspects unique to PdM.

Cross validation

The goal of [cross validation](#) is to define a data set to "test" the model in the training phase. This data set is called the *validation set*. This technique helps limit problems like *overfitting* and gives an insight on how the model will generalize to an independent data set. That is, an unknown data set, which could be from a real problem. The training and testing routine for PdM needs to take into account the time varying aspects to better generalize on unseen future data.

Many machine learning algorithms depend on a number of hyperparameters that can change the model performance significantly. The optimal values of these hyperparameters are not computed automatically when training the model. They should be specified by the data scientist. There are several ways of finding good values of hyperparameters.

The most common one is *k-fold cross-validation* that splits the examples randomly into k folds. For each set of hyperparameters values, run the learning algorithm k times. At each iteration, use the examples in the current fold as a validation set, and the rest of the examples as a training set. Train the algorithm over training examples and compute the performance metrics over validation examples. At the end of this loop, compute the average of k performance metrics. For each set of hyperparameter values, choose the ones that have the best average performance. The task of choosing hyperparameters is often experimental in nature.

In PdM problems, data is recorded as a time series of events that come from several data sources. These records may be ordered according to the time of labeling. Hence, if the dataset is split *randomly* into training and validation set, *some of the training examples may be later in time than some of validation examples*. Future performance of hyperparameter values will be estimated based on some data that arrived *before* model was trained. These estimations might be overly optimistic, especially if the time-series is not stationary and evolves over time. As a result, the chosen hyperparameter values might be suboptimal.

The recommended way is to split the examples into training and validation set in a *time-dependent* manner, where all validation examples are later in time than all training examples. For each set of hyperparameter values, train the algorithm over the training data set. Measure the model's performance over the same validation set. Choose hyperparameter values that show the best performance. Hyperparameter values chosen by train/validation split result in better future model performance than with the values chosen randomly by cross-validation.

The final model can be generated by training a learning algorithm over entire training data using the best hyperparameter values.

Testing for model performance

Once a model is built, an estimate of its future performance on new data is required. A good estimate is the performance metric of hyperparameter values computed over the validation set, or an average performance metric computed from cross-validation. These estimations are often overly optimistic. The business might often have some additional guidelines on how they would like to test the model.

The recommended way for PdM is to split the examples into training, validation, and test data sets in a *time-dependent* manner. All test examples should be later in time than all the training and validation examples. After the split, generate the model and measure its performance as described earlier.

When time-series are stationary and easy to predict, both random and time-dependent approaches generate similar estimations of future performance. But when time-series are non-stationary, and/or hard to predict, the time-dependent approach will generate more realistic estimates of future performance.

Time-dependent split

This section describes best practices to implement time-dependent split. A time-dependent two-way split between training and test sets is described below.

Assume a stream of timestamped events such as measurements from various sensors. Define features and labels of training and test examples over time frames that contain multiple events. For example, for binary classification, create features based on past events, and create labels based on future events within "X" units of time in the future (see the sections on [feature engineering](#) and modeling techniques). Thus, the labeling time frame of an example comes later than the time frame of its features.

For time-dependent split, pick a *training cutoff time* T_c at which to train a model, with hyperparameters tuned using historical data up to T_c . To prevent leakage of future labels that are beyond T_c into the training data, choose the latest time to label training examples to be X units before T_c . In the example shown in Figure 7, each square represents a record in the data set where features and labels are computed as described above. The figure shows the records that should go into training and testing sets for $X=2$ and $W=3$:



Figure 7. Time-dependent split for binary classification

The green squares represent records belonging to the time units that can be used for training. Each training example is generated by considering the past three periods for feature generation, and two future periods for labeling before T_c . When any part of the two future periods is beyond T_c , exclude that example from the training data set because no visibility is assumed beyond T_c .

The black squares represent the records of the final labeled data set that should not be used in the training data set, given the above constraint. These records will also not be used in testing data, since they are before T_c . In addition, their labeling time frames partially depend on the training time frame, which is not ideal. Training and test data should have separate labeling time frames to prevent label information leakage.

The technique discussed so far allows for overlap between training and testing examples that have timestamps near T_c . A solution to achieve greater separation is to exclude examples that are within W time units of T_c from the test set. But such an aggressive split depends on ample data availability.

Regression models used for predicting RUL are more severely affected by the leakage problem. Using the random split method leads to extreme over-fitting. For regression problems, the split should be such that the records belonging to assets with failures before T_c go into the training set. Records of assets that have failures after the cutoff go into the test set.

Another best practice for splitting data for training and testing is to use a split by asset ID. The split should be such that none of the assets used in the training set are used in testing the model performance. Using this approach, a model has a better chance of providing more realistic results with new assets.

Handling imbalanced data

In classification problems, if there are more examples of one class than of the others, the data set is said to be *imbalanced*. Ideally, enough representatives of each class in the training data are preferred to enable differentiation between different classes. If one class is less than 10% of the data, the data is deemed to be imbalanced. The underrepresented class is called a *minority class*.

Many PdM problems face such imbalanced datasets, where one class is severely underrepresented compared to the other class, or classes. In some situations, the minority class may constitute only 0.001% of the total data points. Class imbalance is not unique to PdM. Other domains where failures and anomalies are rare occurrences face a similar problem, for example, fraud detection and network intrusion. These failures make up the minority class examples.

With class imbalance in data, performance of most standard learning algorithms is compromised, since they aim to minimize the overall error rate. For a data set with 99% negative and 1% positive examples, a model can be shown to have 99% accuracy by labeling all instances as negative. But the model will mis-classify all positive examples; so even if its accuracy is high, the algorithm is not a useful one. Consequently, conventional evaluation metrics such as *overall accuracy on error rate* are insufficient for imbalanced learning. When faced with imbalanced datasets, other metrics are used for model evaluation:

- Precision
- Recall

- F1 scores
- Cost adjusted ROC (receiver operating characteristics)

For more information about these metrics, see [model evaluation](#).

However, there are some methods that help remedy class imbalance problem. The two major ones are *sampling techniques* and *cost sensitive learning*.

Sampling methods

Imbalanced learning involves the use of sampling methods to modify the training data set to a balanced data set. Sampling methods are not to be applied to the test set. Although there are several sampling techniques, most straight forward ones are *random oversampling* and *under sampling*.

Random oversampling involves selecting a random sample from minority class, replicating these examples, and adding them to training data set. Consequently, the number of examples in minority class is increased, and eventually balance the number of examples of different classes. A drawback of oversampling is that multiple instances of certain examples can cause the classifier to become too specific, leading to over-fitting. The model may show high training accuracy, but its performance on unseen test data may be suboptimal.

Conversely, *random under sampling* is selecting a random sample from a majority class and removing those examples from training data set. However, removing examples from majority class may cause the classifier to miss important concepts pertaining to the majority class. *Hybrid sampling* where minority class is over-sampled and majority class is under-sampled at the same time is another viable approach.

There are many sophisticated sampling techniques. The technique chosen depends on the data properties and results of iterative experiments by the data scientist.

Cost sensitive learning

In PdM, failures that constitute the minority class are of more interest than normal examples. So the focus is mainly on the algorithm's performance on failures. Incorrectly predicting a positive class as a negative class can cost more than vice-versa. This situation is commonly referred as unequal loss or asymmetric cost of mis-classifying elements to different classes. The ideal classifier should deliver high prediction accuracy over the minority class, without compromising on the accuracy for the majority class.

There are multiple ways to achieve this balance. To mitigate the problem of unequal loss, assign a high cost to misclassification of the minority class, and try to minimize the overall cost. Algorithms like *SVMs (Support Vector Machines)* adopt this method inherently, by allowing cost of positive and negative examples to be specified during training. Similarly, boosting methods such as *boosted decision trees* usually show good performance with imbalanced data.

Model evaluation

Mis-classification is a significant problem for PdM scenarios where the cost of false alarms to the business is high. For instance, a decision to ground an aircraft based on an incorrect prediction of engine failure can disrupt schedules and travel plans. Taking a machine offline from an assembly line can lead to loss of revenue. So model evaluation with the right performance metrics against new test data is critical.

Typical performance metrics used to evaluate PdM models are discussed below:

- [Accuracy](#) is the most popular metric used for describing a classifier's performance. But accuracy is sensitive to data distributions, and is an ineffective measure for scenarios with imbalanced data sets. Other metrics are used instead. Tools like [confusion matrix](#) are used to compute and reason about accuracy of the model.
- [Precision](#) of PdM models relate to the rate of false alarms. Lower precision of the model generally corresponds to a higher rate of false alarms.
- [Recall](#) rate denotes how many of the failures in the test set were correctly identified by the model. Higher recall rates mean the model is successful in identifying the true failures.

- [F1 score](#) is the harmonic average of precision and recall, with its value ranging between 0 (worst) to 1 (best).

For binary classification,

- [Receiver operating curves \(ROC\)](#) is also a popular metric. In ROC curves, model performance is interpreted based on one fixed operating point on the ROC.
- But for PdM problems, *decile tables* and *lift charts* are more informative. They focus only on the positive class (failures), and provide a more complex picture of the algorithm performance than ROC curves.
 - *Decile tables* are created using test examples in a descending order of failure probabilities. The ordered samples are then grouped into deciles (10% of the samples with highest probability, then 20%, 30%, and so on). The ratio (true positive rate)/(random baseline) for each decile helps estimate the algorithm performance at each decile. The random baseline takes on values 0.1, 0.2, and so on.
 - *Lift charts* plot the decile true positive rate versus random true positive rate for all deciles. The first deciles are usually the focus of results, since they show the largest gains. First deciles can also be seen as representative for "at risk", when used for PdM.

Model operationalization for predictive maintenance

The benefit the data science exercise is realized only when the trained model is made operational. That is, the model must be deployed into the business systems to make predictions based on new, previously unseen, data. The new data must exactly conform to the *model signature* of the trained model in two ways:

- all the features must be present in every logical instance (say a row in a table) of the new data.
- the new data must be pre-processed, and each of the features engineered, in exactly the same way as the training data.

The above process is stated in many ways in academic and industry literature. But all the following statements mean the same thing:

- *Score new data* using the model
- *Apply the model* to new data
- *Operationalize* the model
- *Deploy* the model
- *Run the model* against new data

As stated earlier, model operationalization for PdM is different from its peers. Scenarios involving anomaly detection and failure detection typically implement *online scoring* (also called *real time scoring*). Here, the model *scores* each incoming record, and returns a prediction. For anomaly detection, the prediction is an indication that an anomaly occurred (Example: One-class SVM). For failure detection, it would be the type or class of failure.

In contrast, PdM involves *batch scoring*. To conform to the model signature, the features in the new data must be engineered in the same manner as the training data. For the large datasets that is typical for new data, features are aggregated over time windows and scored in batch. Batch scoring is typically done in distributed systems like [Spark](#) or [Azure Batch](#). There are a couple of alternatives - both suboptimal:

- Streaming data engines support aggregation over windows in memory. So it could be argued that they support online scoring. But these systems are suitable for dense data in narrow windows of time, or sparse elements over wider windows. They may not scale well for the dense data over wider time windows, as seen in PdM scenarios.
- If batch scoring is not available, the solution is to adapt online scoring to handle new data in small batches at a time.

Solution templates for predictive maintenance

The final section of this guide provides a list of PdM solution templates, tutorials, and experiments implemented in

Azure. These PdM applications can be deployed into an Azure subscription within minutes in some cases. They can be used as proof-of-concept demos, sandboxes to experiment with alternatives, or accelerators for actual production implementations. These templates are located in the [Azure AI Gallery](#) or [Azure GitHub](#). These different samples will be rolled into this solution template over time.

#	TITLE	DESCRIPTION
2	Azure Predictive Maintenance Solution Template	An open-source solution template which demonstrates ML modeling and a complete Azure infrastructure capable of supporting Predictive Maintenance scenarios in the context of IoT remote monitoring.
3	Deep Learning for Predictive Maintenance	Azure Notebook with a demo solution of using LSTM (Long Short-Term Memory) networks (a class of Recurrent Neural Networks) for Predictive Maintenance, with a blog post on this sample .
4	Predictive Maintenance Modeling Guide in R	PdM modeling guide with scripts in R.
5	Azure Predictive Maintenance for Aerospace	One of the first PdM solution templates based on Azure ML v1.0 for aircraft maintenance. This guide originated from this project.
6	Azure AI Toolkit for IoT Edge	AI in the IoT edge using TensorFlow; toolkit packages deep learning models in Azure IoT Edge-compatible Docker containers and expose those models as REST APIs.
7	Azure IoT Predictive Maintenance	Azure IoT Suite PCS - Preconfigured Solution. Aircraft maintenance PdM template with IoT Suite. Another document and walkthrough related to the same project.
8	Predictive Maintenance template using SQL Server R Services	Demo of remaining useful life scenario based on R services.
9	Predictive Maintenance Modeling Guide	Aircraft maintenance dataset feature engineered using R with experiments and datasets and Azure notebook and experiments in AzureML v1.0

Training resources for predictive maintenance

Microsoft Azure offers learning paths for the foundational concepts behind PdM techniques, besides content and training on general AI concepts and practice.

TRAINING RESOURCE	AVAILABILITY
Learning Path for PdM using Trees and Random Forest	Public

TRAINING RESOURCE	AVAILABILITY
Learning Path for PdM using Deep Learning	Public
AI Developer on Azure	Public
Microsoft AI School	Public
Azure AI Learning from GitHub	Public
LinkedIn Learning	Public
Microsoft AI YouTube Webinars	Public
Microsoft AI Show	Public
LearnAI@MS	Partners
Microsoft Partner Network	Partners

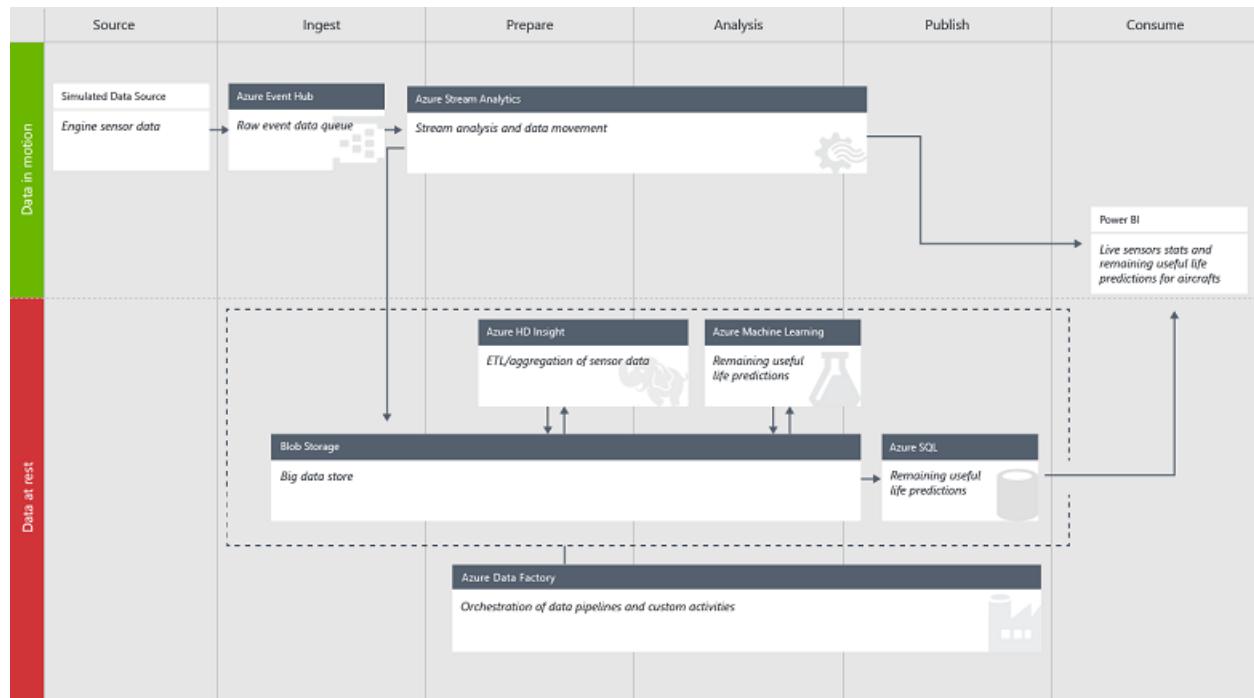
In addition, free MOOCS (massive open online courses) on AI are offered online by academic institutions like Stanford and MIT, and other educational companies.

Architecture of the Cortana Intelligence Solution Template for predictive maintenance in aerospace

1/30/2019 • 2 minutes to read

The diagram below provides an architectural overview of the [Cortana Intelligence Solution Template for predictive maintenance](#).

You can download a full-size version of the diagram here: [Architecture diagram: Solution Template for predictive maintenance](#).



Technical guide to the Cortana Intelligence Solution Template for predictive maintenance in aerospace

3/12/2019 • 16 minutes to read

IMPORTANT

This article has been deprecated. The discussion about Predictive Maintenance in Aerospace is still relevant, but for current information, refer to [Solution Overview for Business Audiences](#).

Solution templates are designed to accelerate the process of building an E2E demo on top of Cortana Intelligence Suite. A deployed template provisions your subscription with necessary Cortana Intelligence components and then builds the relationships between them. It also seeds the data pipeline with sample data from a data generator application, which you download and install on your local machine after you deploy the solution template. The data from the generator hydrates the data pipeline and start generating machine learning predictions, which can then be visualized on the Power BI dashboard.

The deployment process guides you through several steps to set up your solution credentials. Make sure you record the credentials such as solution name, username, and password that you provide during the deployment.

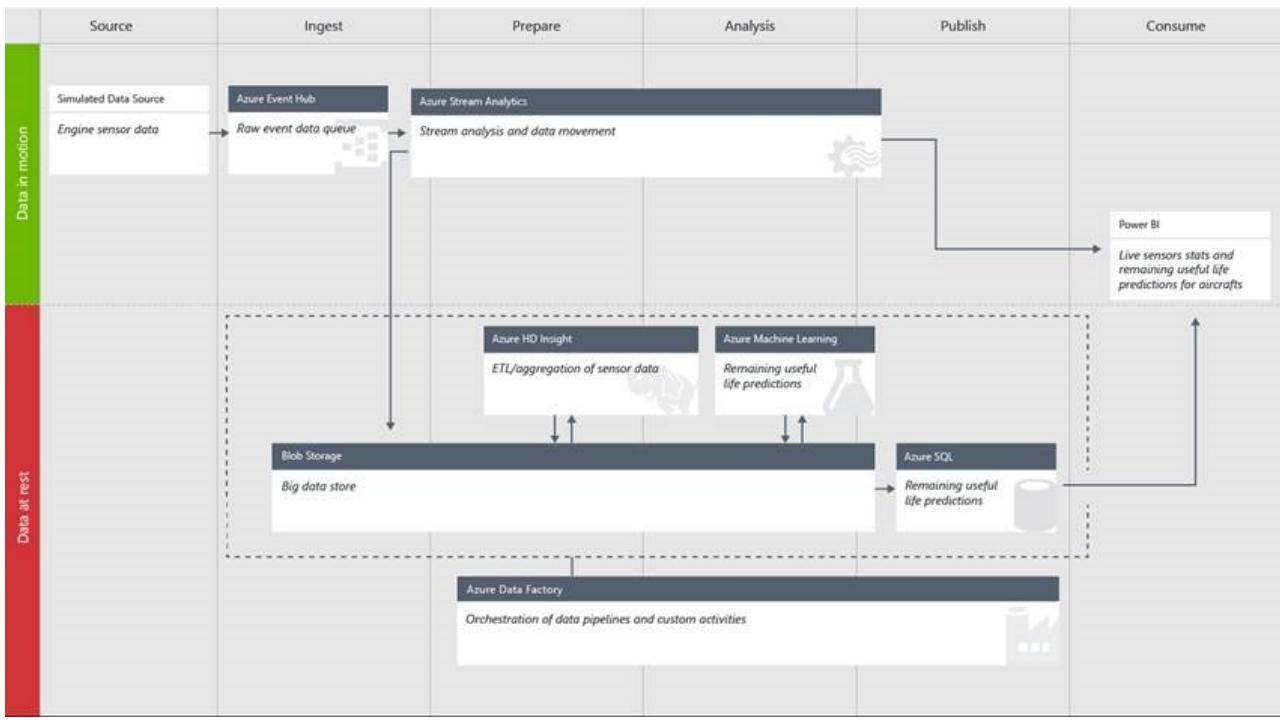
The goals of this article are to:

- Describe the reference architecture and components provisioned in your subscription.
- Demonstrate how to replace the sample data with your own data.
- Show how to modify the solution template.

TIP

You can download and print a [PDF version of this article](#).

Overview



When you deploy the solution, it activates Azure services within the Cortana Analytics Suite (including Event Hub, Stream Analytics, HDInsight, Data Factory, and Machine Learning). The architecture diagram shows how the Predictive Maintenance for Aerospace Solution Template is constructed. You can investigate these services in the Azure portal by clicking them in the solution template diagram created with the solution deployment (except for HDInsight, which is provisioned on demand when the related pipeline activities are required to run and are deleted afterwards). Download a [full-size version of the diagram](#).

The following sections describe the solution parts.

Data source and ingestion

Synthetic data source

For this template, the data source used is generated from a desktop application that you download and run locally after successful deployment.

To find the instructions to download and install this application, select the first node, Predictive Maintenance Data Generator, on the solution template diagram. The instructions are found in the Properties bar. This application feeds the [Azure Event Hub](#) service with data points, or events, used in the rest of the solution flow. This data source is derived from publicly available data from the [NASA data repository](#) using the [Turbofan Engine Degradation Simulation Data Set](#).

The event generation application populates the Azure Event Hub only while it's executing on your computer.

Azure Event Hub

The [Azure Event Hub](#) service is the recipient of the input provided by the Synthetic Data Source.

Data preparation and analysis

Azure Stream Analytics

Use [Azure Stream Analytics](#) to provide near real-time analytics on the input stream from the [Azure Event Hub](#) service. You then publish results onto a [Power BI](#) dashboard as well as archive all raw incoming events to the [Azure Storage](#) service for later processing by the [Azure Data Factory](#) service.

HDInsight custom aggregation

Run [Hive](#) scripts (orchestrated by Azure Data Factory) using HDInsight to provide aggregations on the raw events

archived using the Azure Stream Analytics service.

Azure Machine Learning

Make predictions on the remaining useful life (RUL) of a particular aircraft engine using the inputs received with [Azure Machine Learning Service](#) (orchestrated by Azure Data Factory).

Data publishing

Azure SQL Database

Use [Azure SQL Database](#) to store the predictions received by the Azure Machine Learning service, which are then consumed in the [Power BI](#) dashboard.

Data consumption

Power BI

Use [Power BI](#) to show a dashboard that contains aggregations and alerts provided by [Azure Stream Analytics](#), as well as RUL predictions stored in [Azure SQL Database](#) that were produced using [Azure Machine Learning](#).

How to bring in your own data

This section describes how to bring your own data to Azure, and what areas require changes for the data you bring into this architecture.

It's unlikely that your dataset matches the dataset used by the [Turbofan Engine Degradation Simulation Data Set](#) used for this solution template. Understanding your data and the requirements are crucial in how you modify this template to work with your own data.

The following sections discuss the parts of the template that require modifications when a new dataset is introduced.

Azure Event Hub

Azure Event Hub is generic; data can be posted to the hub in either CSV or JSON format. No special processing occurs in the Azure Event Hub, but it's important that you understand the data that's fed into it.

This document does not describe how to ingest your data, but you can easily send events or data to an Azure Event Hub using the Event Hub APIs.

Azure Stream Analytics

Use the Azure Stream Analytics service to provide near real-time analytics by reading from data streams and outputting data to any number of sources.

For the Predictive Maintenance for Aerospace Solution Template, the Azure Stream Analytics query consists of four sub queries, each query consuming events from the Azure Event Hub service, with outputs to four distinct locations. These outputs consist of three Power BI datasets and one Azure Storage location.

The Azure Stream Analytics query can be found by:

- Connect to the Azure portal
- Locating the Stream Analytics jobs  that were generated when the solution was deployed (*for example, **maintenancesa02asapbi** and **maintenancesa02asablob** for the predictive maintenance solution*)
- Selecting
 - **INPUTS** to view the query input
 - **QUERY** to view the query itself
 - **OUTPUTS** to view the different outputs

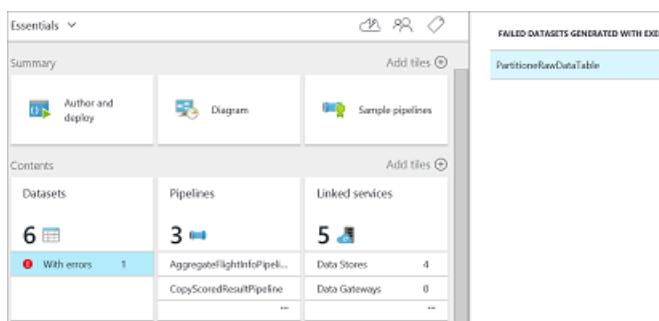
Information about Azure Stream Analytics query construction can be found in the [Stream Analytics Query Reference](#) on MSDN.

In this solution, the queries output three datasets with near real-time analytics information about the incoming data stream to a Power BI dashboard provided as part of this solution template. Because there's implicit knowledge about the incoming data format, these queries must be altered based on your data format.

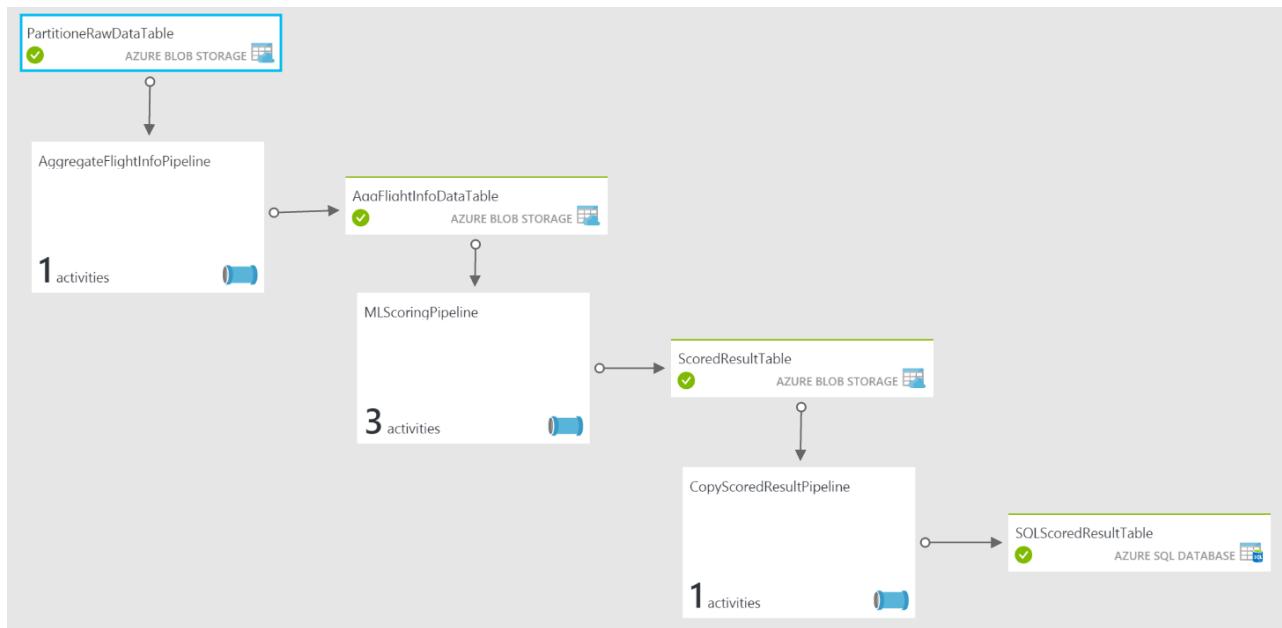
The query in the second Stream Analytics job **maintenancesa02asblob** simply outputs all [Event Hub](#) events to [Azure Storage](#) and hence requires no alteration regardless of your data format as the full event information is streamed to storage.

Azure Data Factory

The [Azure Data Factory](#) service orchestrates the movement and processing of data. In the Predictive Maintenance for Aerospace Solution Template, the data factory is made up of three [pipelines](#) that move and process the data using various technologies. Access your data factory by opening the Data Factory node at the bottom of the solution template diagram created with the deployment of the solution. Errors under your datasets are due to data factory being deployed before the data generator was started. Those errors can be ignored and do not prevent your data factory from functioning



This section discusses the necessary [pipelines](#) and [activities](#) contained in the [Azure Data Factory](#). Here is a diagram view of the solution.



Two of the pipelines of this factory contain [Hive](#) scripts used to partition and aggregate the data. When noted, the scripts are located in the [Azure Storage](#) account created during setup. Their location is: `maintenancesascript\script\hive\` (or [https://\[Your solution name\].blob.core.windows.net/maintenancesascript](https://[Your solution name].blob.core.windows.net/maintenancesascript)).

Similar to [Azure Stream Analytics](#) queries, the [Hive](#) scripts have implicit knowledge about the incoming data format and must be altered based on your data format.

AggregateFlightInfoPipeline

This [pipeline](#) contains a single activity - an [HDInsightHive](#) activity using a [HDInsightLinkedService](#) that runs a [Hive](#) script to partition the data put in [Azure Storage](#) during the [Azure Stream Analytics](#) job.

The [Hive](#) script for this partitioning task is [**AggregateFlightInfo.hql**](#)

MLScoringPipeline

This [pipeline](#) contains several activities whose end result is the scored predictions from the [Azure Machine Learning](#) experiment associated with this solution template.

Activities included are:

- [HDInsightHive](#) activity using an [HDInsightLinkedService](#) that runs a [Hive](#) script to perform aggregations and feature engineering necessary for the [Azure Machine Learning](#) experiment. The [Hive](#) script for this partitioning task is [**PrepareMLInput.hql**](#).
- [Copy](#) activity that moves the results from the [HDInsightHive](#) activity to a single [Azure Storage](#) blob accessed by the [AzureMLBatchScoring](#) activity.
- [AzureMLBatchScoring](#) activity calls the [Azure Machine Learning](#) experiment, with results put in a single [Azure Storage](#) blob.

CopyScoredResultPipeline

This [pipeline](#) contains a single activity - a [Copy](#) activity that moves the results of the [Azure Machine Learning](#) experiment from the [**MLScoringPipeline**](#) to the [Azure SQL Database](#) provisioned as part of the solution template installation.

Azure Machine Learning

The [Azure Machine Learning](#) experiment used for this solution template provides the Remaining Useful Life (RUL) of an aircraft engine. The experiment is specific to the data set consumed and requires modification or replacement specific to the data brought in.

For information about how the Azure Machine Learning experiment was created, see [Predictive Maintenance: Step 1 of 3, data preparation and feature engineering](#).

Monitor Progress

Once the Data Generator is launched, the pipeline begins to dehydrate, and the different components of your solution start kicking into action following the commands issued by the data factory. There are two ways to monitor the pipeline.

1. One of the Stream Analytics jobs writes the raw incoming data to blob storage. If you click on Blob Storage component of your solution from the screen you successfully deployed the solution and then click Open in the right panel, it takes you to the [Azure portal](#). Once there, click on Blobs. In the next panel, you see a list of Containers. Click on **maintenancesadata**. In the next panel is the **rawdata** folder. Inside the rawdata folder are folders with names such as hour=17, and hour=18. The presence of these folders indicates raw data is being generated on your computer and stored in blob storage. You should see csv files with finite sizes in MB in those folders.
2. The last step of the pipeline is to write data (for example predictions from machine learning) into SQL Database. You might have to wait a maximum of three hours for the data to appear in SQL Database. One way to monitor how much data is available in your SQL Database is through the [Azure portal](#). On the left panel locate **SQL DATABASES**  and click it. Then locate your database **pmaintainedb** and click on it. On the next page at the bottom, click on **MANAGE**



Here, you can click on New Query and query for the number of rows (for example select count(*) from

PMResult). As your database grows, the number of rows in the table should increase.

Power BI Dashboard

Set up a Power BI dashboard to visualize your Azure Stream Analytics data (hot path) and batch prediction results from Azure machine learning (cold path).

Set up the cold path dashboard

In the cold path data pipeline, the goal is to get the predictive RUL (remaining useful life) of each aircraft engine once it finishes a flight (cycle). The prediction result is updated every 3 hours for predicting the aircraft engines that have finished a flight during the past 3 hours.

Power BI connects to an Azure SQL database as its data source, where the prediction results are stored. Note: 1) On deploying your solution, a prediction will appear in the database within 3 hours. The pbix file that came with the Generator download contains some seed data so that you may create the Power BI dashboard right away. 2) In this step, the prerequisite is to download and install the free software [Power BI desktop](#).

The following steps guide you on how to connect the pbix file to the SQL Database that was spun up at the time of solution deployment containing data (for example, prediction results) for visualization.

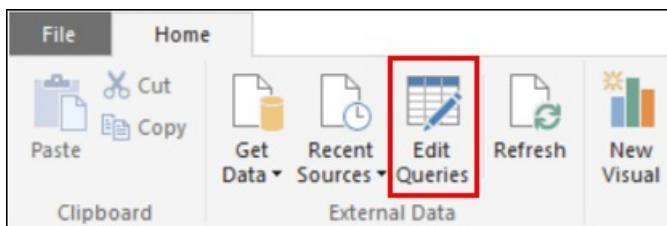
1. Get the database credentials.

You'll need **database server name, database name, user name and password** before moving to next steps. Here are the steps to guide you how to find them.

- Once '**Azure SQL Database**' on your solution template diagram turns green, click it and then click '**Open**'.
- You'll see a new browser tab/window which displays the Azure portal page. Click '**Resource groups**' on the left panel.
- Select the subscription you're using for deploying the solution, and then select '**YourSolutionName_ResourceGroup**'.
- In the new pop out panel, click the  icon to access your database. Your database name is next to this icon (for example, '**pmaintainededb**'), and the **database server name** is listed under the Server name property and should look similar to **YourSolutionName.database.windows.net**.
- Your database **username** and **password** are the same as the username and password previously recorded during deployment of the solution.

2. Update the data source of the cold path report file with Power BI Desktop.

- In the folder where you downloaded and unzipped the Generator file, double-click the **PowerBI\PredictiveMaintenanceAerospace.pbix** file. If you see any warning messages when you open the file, ignore them. On the top of the file, click '**Edit Queries**'.



- You'll see two tables, **RemainingUsefulLife** and **PMResult**. Select the first table and click  next to '**Source**' under '**APPLIED STEPS**' on the right '**Query Settings**' panel. Ignore any warning messages that appear.
- In the pop out window, replace '**Server**' and '**Database**' with your own server and database names, and then click '**OK**'. For server name, make sure you specify the port 1433 (**YourSolutionName.database.windows.net, 1433**). Leave the Database field as **pmaintainededb**.

Ignore the warning messages that appear on the screen.

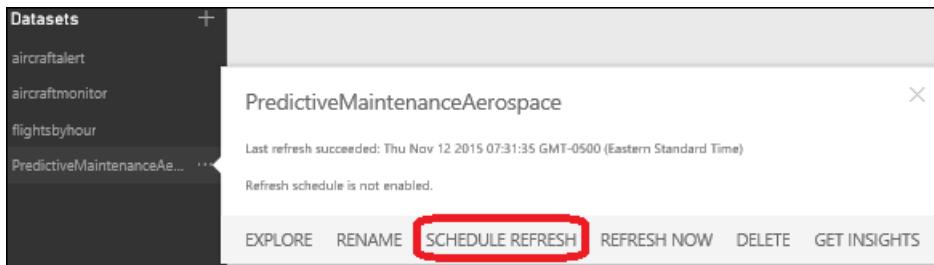
- In the next pop out window, you'll see two options on the left pane (**Windows** and **Database**). Click '**Database**', fill in your '**Username**' and '**Password**' (this is the username and password you entered when you first deployed the solution and created an Azure SQL database). In **Select which level to apply these settings to**, check database level option. Then click '**Connect**'.
- Click on the second table **PMResult** then click next to '**Source**' under '**APPLIED STEPS**' on the right '**Query Settings**' panel, and update the server and database names as in the above steps and click OK.
- Once you're guided back to the previous page, close the window. A message displays - click **Apply**. Lastly, click the **Save** button to save the changes. Your Power BI file has now established connection to the server. If your visualizations are empty, make sure you clear the selections on the visualizations to visualize all the data by clicking the eraser icon on the upper right corner of the legends. Use the refresh button to reflect new data on the visualizations. Initially, you only see the seed data on your visualizations as the data factory is scheduled to refresh every 3 hours. After 3 hours, you will see new predictions reflected in your visualizations when you refresh the data.

- (Optional) Publish the cold path dashboard to [Power BI online](#). Note that this step needs a Power BI account (or Office 365 account).

- Click '**Publish**' and few seconds later a window appears displaying "Publishing to Power BI Success!" with a green check mark. Click the link below "Open PredictiveMaintenanceAerospace.pbix in Power BI". To find detailed instructions, see [Publish from Power BI Desktop](#).
- To create a new dashboard: click the '+' sign next to the **Dashboards** section on the left pane. Enter the name "Predictive Maintenance Demo" for this new dashboard.
- Once you open the report, click to pin all the visualizations to your dashboard. To find detailed instructions, see [Pin a tile to a Power BI dashboard from a report](#). Go to the dashboard page and adjust the size and location of your visualizations and edit their titles. To find detailed instructions on how to edit your tiles, see [Edit a tile -- resize, move, rename, pin, delete, add hyperlink](#). Here is an example dashboard with some cold path visualizations pinned to it. Depending on how long you run your data generator, your numbers on the visualizations may be different.



- To schedule refresh of the data, hover your mouse over the **PredictiveMaintenanceAerospace** dataset, click and then choose **Schedule Refresh**.
- Note:** If you see a warning message, click **Edit Credentials** and make sure your database credentials are the same as those described in step 1.



- Expand the **Schedule Refresh** section. Turn on "keep your data up-to-date".
- Schedule the refresh based on your needs. To find more information, see [Data refresh in Power BI](#).

Setup hot path dashboard

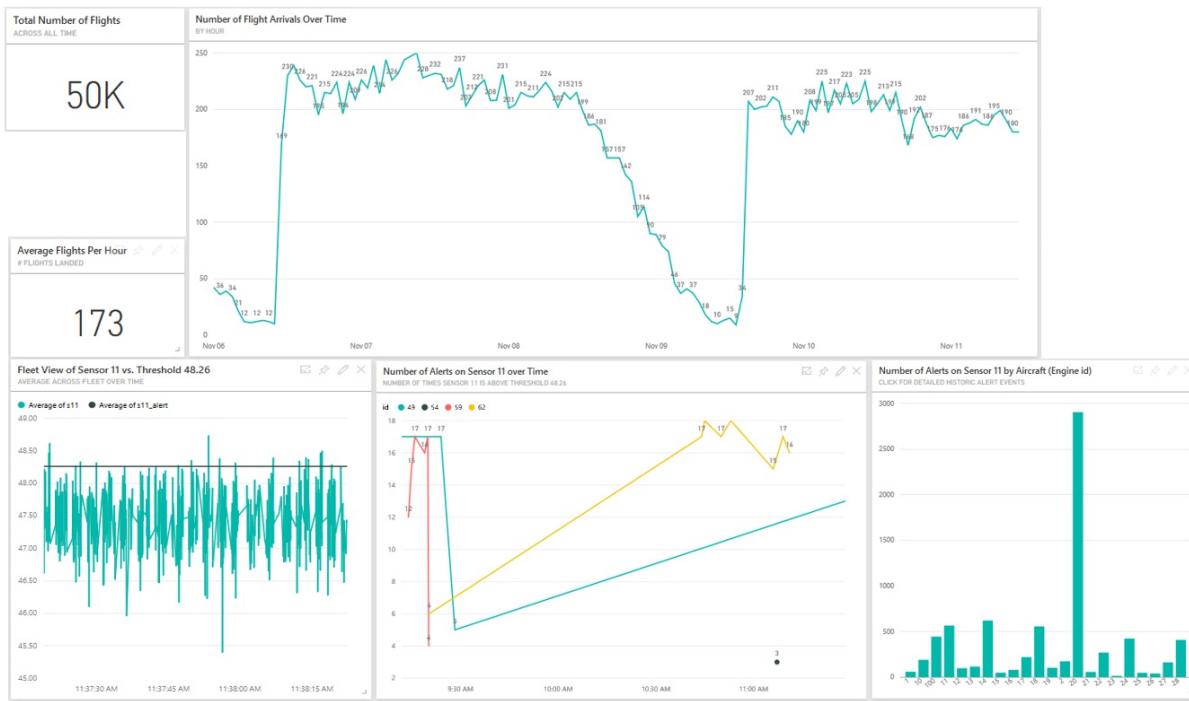
The following steps guide you how to visualize data output from Stream Analytics jobs that were generated at the time of solution deployment. A [Power BI online](#) account is required to perform the following steps. If you don't have an account, you can [create one](#).

1. Add Power BI output in Azure Stream Analytics (ASA).

- You must follow the instructions in [Azure Stream Analytics & Power BI: An analytics dashboard for real-time visibility of streaming data](#) to set up the output of your Azure Stream Analytics job as your Power BI dashboard.
- The ASA query has three outputs which are **aircraftmonitor**, **aircraftalert**, and **flightsbyhour**. You can view the query by clicking on query tab. Corresponding to each of these tables, you need to add an output to ASA. When you add the first output (**aircraftmonitor**) make sure the **Output Alias, Dataset Name** and **Table Name** are the same (**aircraftmonitor**). Repeat the steps to add outputs for **aircraftalert**, and **flightsbyhour**. Once you have added all three output tables and started the ASA job, you should get a confirmation message ("Starting Stream Analytics job maintenancesa02asapbi succeeded").

2. Log in to [Power BI online](#)

- On the left panel Datasets section in My Workspace, the **DATASET** names **aircraftmonitor**, **aircraftalert**, and **flightsbyhour** should appear. This is the streaming data you pushed from Azure Stream Analytics in the previous step. The dataset **flightsbyhour** may not show up at the same time as the other two datasets due to the nature of the SQL query behind it. However, it should show up after an hour.
 - Make sure the **Visualizations** pane is open and is shown on the right side of the screen.
3. Once you have the data flowing into Power BI, you can start visualizing the streaming data. Below is an example dashboard with some hot path visualizations pinned to it. You can create other dashboard tiles based on appropriate datasets. Depending on how long you run your data generator, your numbers on the visualizations may be different.



4. Here are some steps to create one of the tiles above – the "Fleet View of Sensor 11 vs. Threshold 48.26" tile:

- Click dataset **aircraftmonitor** on the left panel Datasets section.
- Click the **Line Chart** icon.
- Click **Processed** in the **Fields** pane so that it shows under "Axis" in the **Visualizations** pane.
- Click "s11" and "s11_alert" so that they both appear under "Values". Click the small arrow next to **s11** and **s11_alert**, change "Sum" to "Average".
- Click **SAVE** on the top and name the report "aircraftmonitor." The report named "aircraftmonitor" is shown in the **Reports** section in the **Navigator** pane on the left.
- Click the **Pin Visual** icon on the top right corner of this line chart. A "Pin to Dashboard" window may show up for you to choose a dashboard. Select "Predictive Maintenance Demo," then click "Pin."
- Hover the mouse over this tile on the dashboard, click the "edit" icon on the top right corner to change its title to "Fleet View of Sensor 11 vs. Threshold 48.26" and subtitle to "Average across fleet over time."

Delete your solution

Ensure that you stop the data generator when not actively using the solution as running the data generator will incur higher costs. Delete the solution if you are not using it. Deleting your solution deletes all the components provisioned in your subscription when you deployed the solution. To delete the solution, click your solution name in the left panel of the solution template, and then click **Delete**.

Cost estimation tools

The following two tools are available to help you better understand the total costs involved in running the Predictive Maintenance for Aerospace Solution Template in your subscription:

- [Microsoft Azure Cost Estimator Tool \(online\)](#)
- [Microsoft Azure Cost Estimator Tool \(desktop\)](#)