

This approach provides an update on how you can use Azure Data Factory to integrate with Azure Machine Learning using the latest Azure Data Factory **AzureMLBatchExecution activity**. With Azure Data Factory, you can create big data pipelines (eg. Pig and Hive) to ingest and process data from various data sources and use an Azure Machine Learning web service to predict customer behaviors. Consider the following example:

Imagine you are a mobile service operator and you want to identify the subscribers who are likely to cancel their subscription within the next few months. This will help you transform the raw data into intelligent action you can take to figure out diverse ways of retaining the customers that might churn. The example below outlines an end-to-end solution to figure out the mobile customers that will likely churn within the month.

ADF enables you to easily orchestrate all the processes and resources needed for the example including ingesting data into Azure blob storage, managing the HDInsight cluster and job execution, executing the Machine Learning process and finally publishing the end results to Azure blob storage. The Hive queries that run on HDInsight aggregate the call details (monthly minutes spent on the call per customer), before merging it with the customer profile. This combined data can then be used as inputs to the Azure Machine Learning model for figuring out customer churn (shown in the diagram below).

In the diagram, you can see the AggregateMobileCustomerUsage pipeline which runs a Hive job to perform the aggregation on a HDInsight cluster. Once the data has been aggregated, it is used as inputs to the PredictCustomerChurnPipeline which invokes the Azure Machine Learning model.

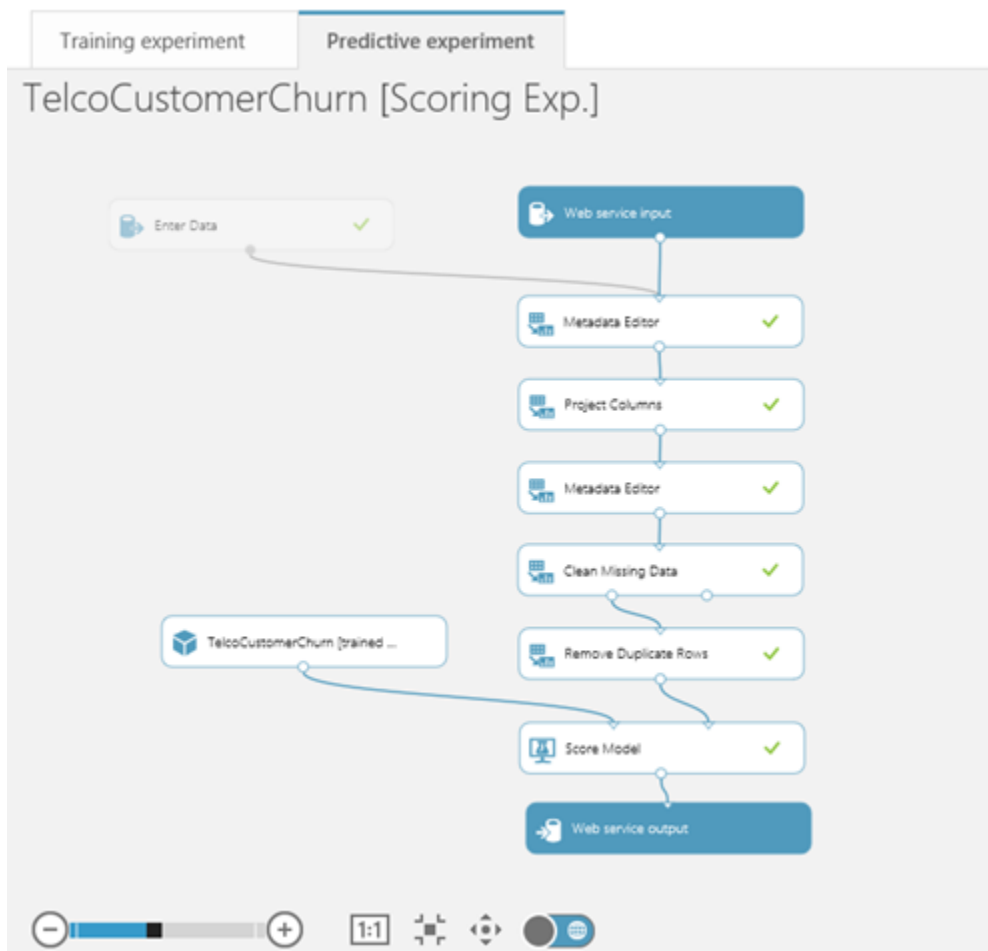


There are two key parts to this solution that we will look at in more detail:

1. **Azure Machine Learning:** Deploying the customer churn model
2. **Azure Data Factory:** Consuming the published Azure Machine Learning web service

Azure Machine Learning – Deploying the customer churn model

Let us explore the predictive experiment the data scientist published. Notice the Web Service Input and Web Service Output (in the picture below) shown in the predictive experiment. The Web Service Input enables you to specify the data will be used as inputs to the published web service and generate a predicted result of the likelihood of churn for each customer. The Web Service Output specifies the Azure Blob storage where predicted results will be written. When you create a predictive experiment using Azure Machine Learning, the Web Service Input and Output are automatically created for you. Refer to the article on [deploying an Azure Machine Learning web service](#) to learn more.



After the experiment has been deployed as a web service, you can view the dashboard of the web service to get the API key and the Batch URI as shown below. The API key and Batch URI are then used in the next step where you create an Azure Data Factory linked service.

telcocustomerchurn [scoring exp.]

DASHBOARD CONFIGURATION

General

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

Default Endpoint

API HELP PAGE	TEST	APPS
REQUEST/RESPONSE	Test	Download Excel Workbook
BATCH EXECUTION		

Additional endpoints

Number of additional endpoints created for this web service: 0

[Manage endpoints in Azure management portal](#)

Batch Execution API Documentation for TelcoCustomerChurn [Scoring Exp.]

Updated: 08/16/2015 22:05

No description provided for this web service.

- [Previous version of this API](#)
- [Submit a Job](#)
- [Start a Job](#)
- [Get Job Status](#)
- [Delete a Job](#)
- [Sample Code](#)

Submit (but not start) a Batch Execution job

Request

Method	Request URI	HTTP Version
POST	https://usouthcentral.services.azureml.net/workspaces/cob85c748974468a8550a20fcfc8569e/services/146b153f933740d3b364795070084963/jobs?api-version=2.0	HTTP/1.1

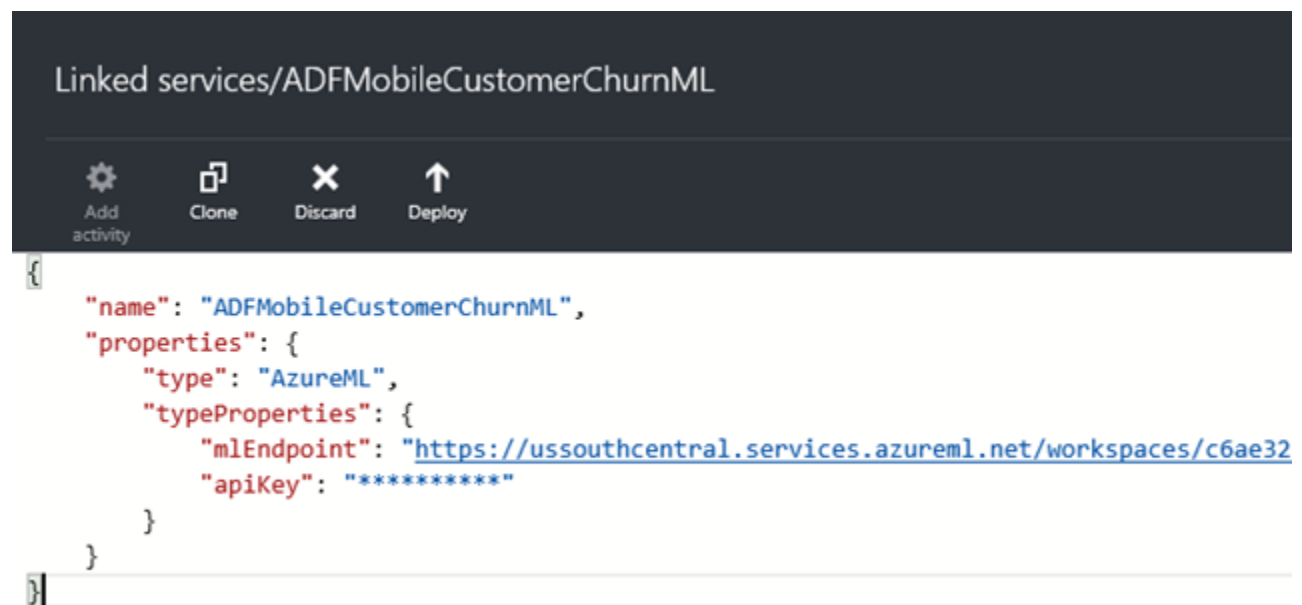
Request Headers

Request Header	Description
----------------	-------------

Azure Data Factory - Consuming the published Azure Machine Learning web service

With the Batch URI and API Key, you can use the **AzureMLBatchExecution activity** in an Azure Data Factory pipeline to score your input data and schedule it to run on a regular basis.

To specify the information needed to connect to Azure Machine Learning, you will need to define a [linked service](#). A linked service, called **ADFMobileCustomerChurnML**, is shown below. In the linked service, you will use the **mlEndpoint property** to point to the Batch URI obtained in the prior step. You will also specify the API key needed to access the deployed Azure Machine Learning web service.

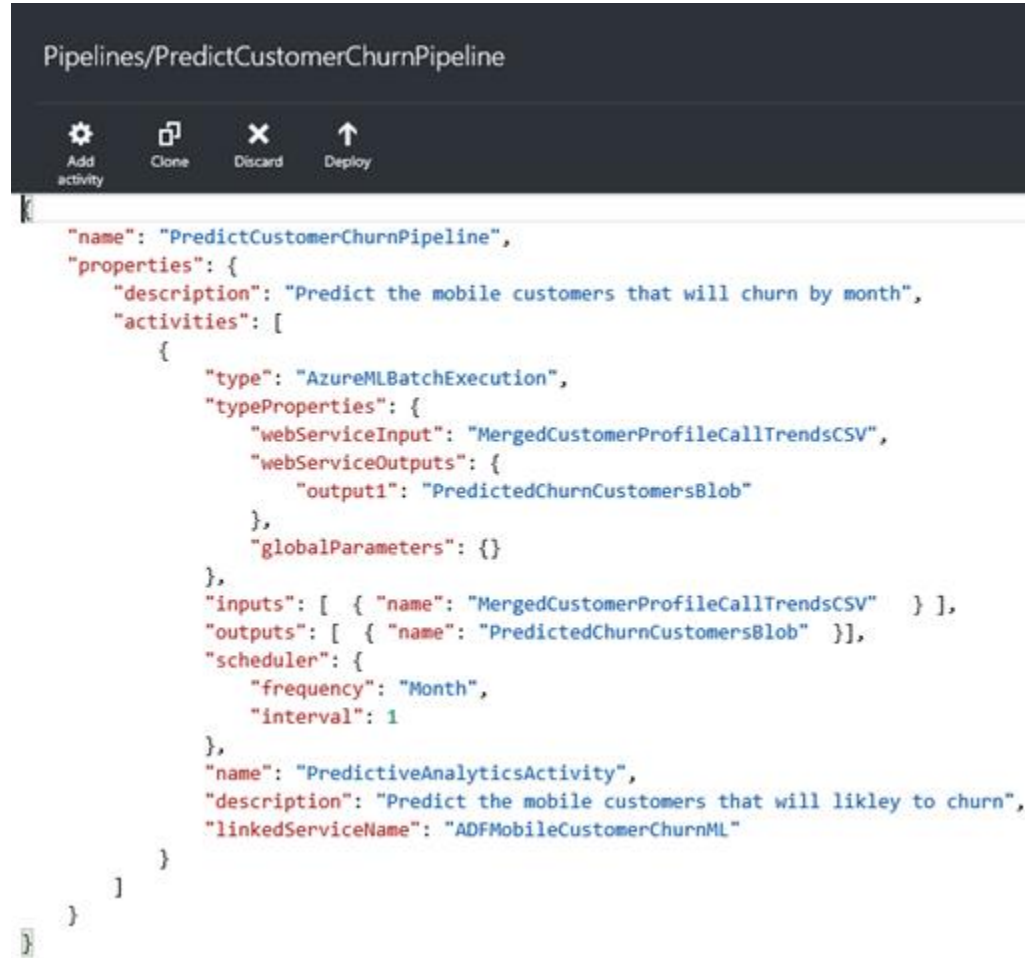


Next, you will use the linked service in an Azure Data Factory pipeline. In the pipeline (shown below), you will see the **AzureMLBatchExecution** activity is referring to the linked service, **ADFMobileCustomerChurnML**. The activity is scheduled to run on a monthly basis and is started only when the input dependencies ("inputs": [{ "name" : MergedCustomerProfileCallTrendsCSV"}]) are ready. This enables you to invoke the web service only when upstream data processing by the big data pipelines have been completed.

You will also notice the activity definition specifies data required by the web service input of the deployed web service is provided by the **MergedCustomerProfileCallTrendsCSV dataset**. The **MergedCustomerProfileCallTrendsCSV dataset** refers to processed stored in

Azure blob storage. After batch scoring has been performed by the published Azure Machine Learning web service, the data (with predicted results on whether the customer will likely churn) is written to the storage location specified by the **PredictedChurnCustomersBlob** dataset.

Find out more on how you can create [Azure Data Factory pipelines that integrate with Azure Machine Learning](#).



The screenshot shows the Azure Data Factory Pipelines editor interface. At the top, the title bar reads "Pipelines/PredictCustomerChurnPipeline". Below the title bar is a toolbar with four icons: a gear for "Add activity", a square with a plus sign for "Clone", an 'X' for "Discard", and an upward arrow for "Deploy". The main area of the editor displays the JSON definition for the pipeline, which is a Predictive Analytics Activity. The JSON is as follows:

```
{
  "name": "PredictCustomerChurnPipeline",
  "properties": {
    "description": "Predict the mobile customers that will churn by month",
    "activities": [
      {
        "type": "AzureMLBatchExecution",
        "typeProperties": {
          "webServiceInput": "MergedCustomerProfileCallTrendsCSV",
          "webServiceOutputs": {
            "output1": "PredictedChurnCustomersBlob"
          },
          "globalParameters": {}
        },
        "inputs": [ { "name": "MergedCustomerProfileCallTrendsCSV" } ],
        "outputs": [ { "name": "PredictedChurnCustomersBlob" } ],
        "scheduler": {
          "frequency": "Month",
          "interval": 1
        },
        "name": "PredictiveAnalyticsActivity",
        "description": "Predict the mobile customers that will likely to churn",
        "linkedServiceName": "ADFMobileCustomerChurnML"
      }
    ]
  }
}
```