# Constructing On-Premise Data Pipeline into Azure for AML operationalization

Data driven decision making lies at the center of all fast growing companies. A vast majority of companies use SQL servers for data storage, aggregation and number crunching. This data can also be used to forecast machine breakdowns, detect anomaly, or to manage cost, etc. While data analytics can give very precise answers to static data questions but it lacks ability to predict trends or anomalies in various processes, with continuously improving high degree of accuracy. Machine leaning fills this gap.

## Use-case

This document helps you create an E2E deployment ready data pipeline for consuming an Azure Machine Learning (AML) solution in your on-premise SQL server. Here we take as an example, a major ATM machine manufacturer that wants to predict ATM machine failure and determine the root cause. This would help the company minimize ‘customer interrupts’ during a transaction, and reduce their cost of servicing (by sending right skilled technician and correct replacement parts).

This document will describe:

* Overall approach for operationalizing an on-premise to azure data pipeline
* An Ingress data pipeline that will bring in data for training and then subsequently for scoring (after a trained model is ready and published as a web service).
* An AML data pipeline to send incoming data (from on-premise data center in to Azure) for AML batch scoring and store the scores.
* An Egress data pipeline that will send the scores back to the on-premise data center (SQL server)
* Create sample data source in your SQL server wherever needed.
* A PowerBI dashboard to show results of scoring.

By the end of the document you should be able to create a fully operationalized data pipeline in Azure that will (1) connect to your on-premise SQL server data source, (2) call AML batch execution service endpoint for scoring, and (3) copy the scores back to your on-premise SQL server data source. This pipeline will be scheduled to run automatically on an hourly schedule for you (until you stop it).

## Big Picture

Building Azure Machine Learning model requires historical data, and perhaps lots of it based on the quality of the data. This is the reason why it is important to think about building data pipelines not just for building the model but also for getting scores from the model and integrating the results with your current business processes. For our current use case we have divided this pipeline into 3 discrete parts each serving a specific objective and can be tested separately if desired. We start by giving you an over view of the big picture below. And then we will briefly define various Azure platform services and tools we will use. Finally we will list step by step instructions you can follow to create an E2E data pipeline.

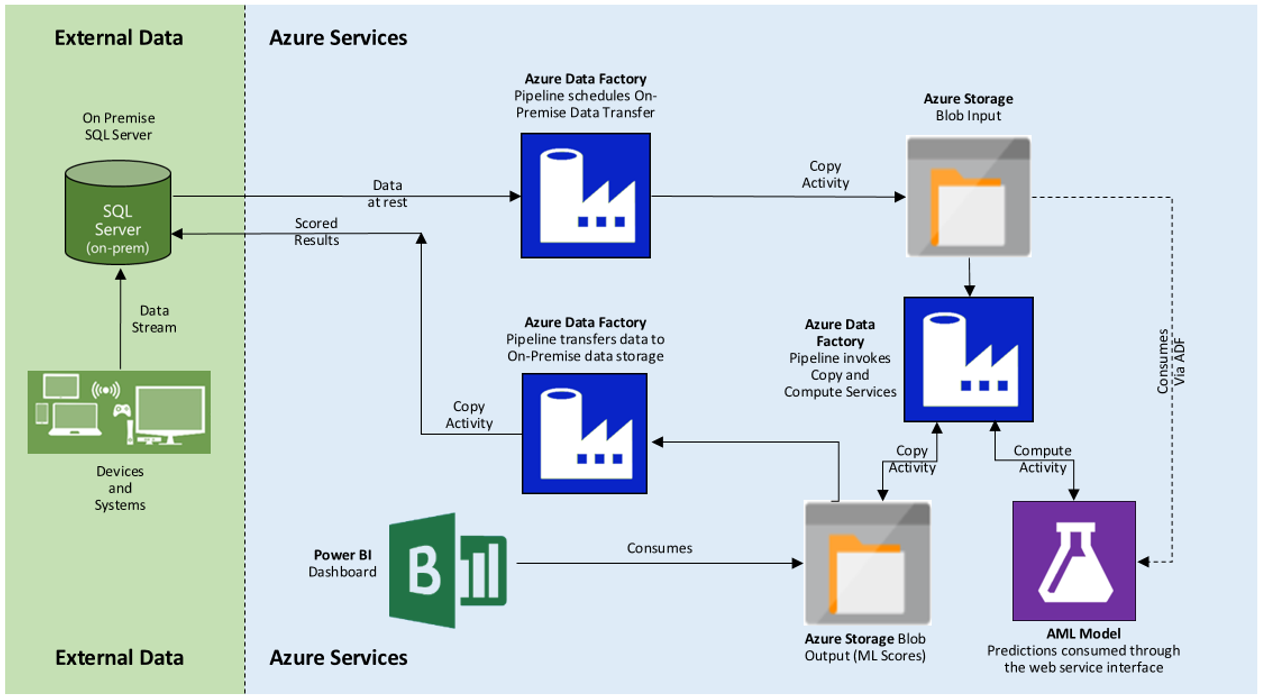


Figure 1: Overall Architecture of E2E data pipeline for On-Premise data source

In the architecture shown above, we use Azure Data Factory (ADF) to move the data between on premise database and Azure as well as for calling AML web service endpoint for scoring batches of input data. An ADF can schedule execution of certain predefined activities (such as, copy or compute activity) and also allow for adding custom activities.

The data pipeline is divided in to three section as follows:

*Azure Data Factory is a service for orchestrating data storage, processing, and movement.*

1. We use an ADF to first used to copy data from an on-premise SQL Server to an Azure Blob Storage (we will refer this as *Ingress OnPrem Pipeline*).
2. The ADF then invokes ‘compute’ activity using AML model (using AML Batch Execution Service endpoint) for prediction and copies the predicted results to an output blob location (we will refer this as *AML Scoring Pipeline*).
3. The scored data from the output blob location is then copied to the on-premise SQL server for integration with the rest of the business process (we will refer this as *Egress OnPrem Pipeline*).

The scored data is also consumed by a dashboard that is published in Azure cloud using PowerBI. This dashboard shows hourly and daily ATM machine health scores. The dashboard is available anywhere from over the internet.

## Walkthrough

### Create ADF and Setup Data Management Gateway

First few steps in creating an E2E data pipeline to connect to an on-premise SQL server is to create an *Azure Data Factory* and install *Data Management Gateway.* If you are not familiar with the process you can find more detailed instructions [here](http://azure.microsoft.com/en-us/documentation/articles/data-factory-use-onpremises-datasources/). Follow steps to [create an ADF](http://azure.microsoft.com/en-us/documentation/articles/data-factory-use-onpremises-datasources/#step-1-create-an-azure-data-factory) and [install the data management gateway](http://azure.microsoft.com/en-us/documentation/articles/data-factory-use-onpremises-datasources/#step-2-create-a-data-management-gateway) before returning to this document. Use the following values for ‘ResourceGruopName’ and ‘DataFactoryName’ as you follow instructions for creating ADF.

-ResourceGroupName OnPremPipelineRG

-DataFactoryName OnPremPipeline

### Create Sample Input Data

Next, we create sample tables and insert some data to those to provide as input to the data pipelines in the next section. Use the [SQL script](SQLScripts/CreateTableAndInsertData.sql) provided with this package to create required input and output tables and sample insert data.

### Create Data Pipelines

Finally, we will create 3 data pipelines (described earlier in this document) all in within the contest of the ADF we just created.

### The Ingress On-Premise Pipeline

The ingress pipeline is the first section of the overall data pipeline we will create in this document. Figure below highlights the ingress portion from the overall architecture.

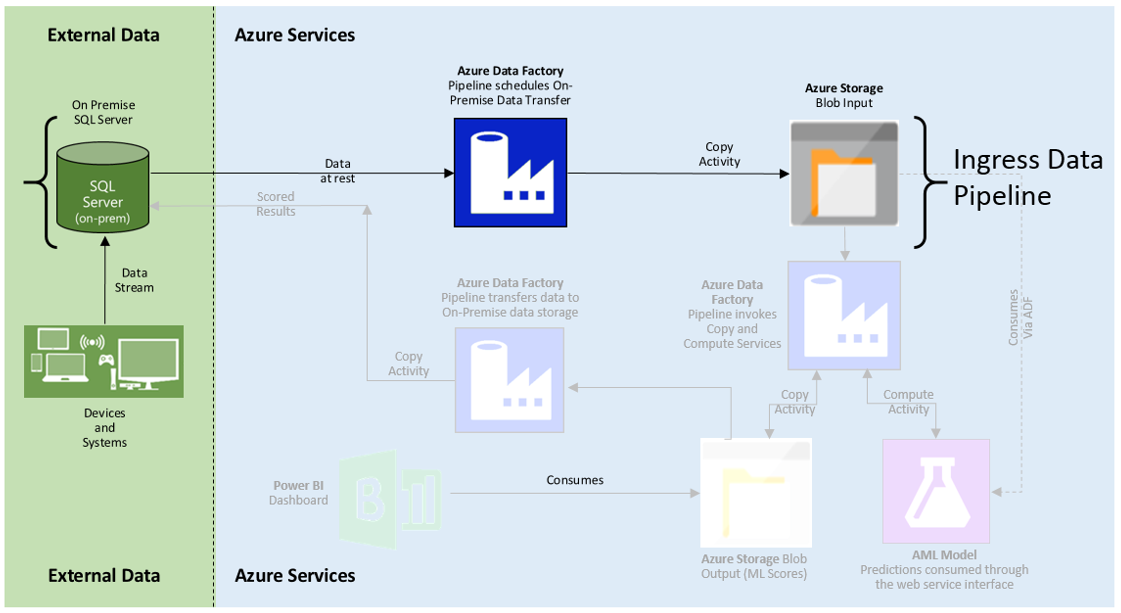


Figure 2: Ingress On-Premise Pipeline

This pipeline connects your company’s on-premise SQL Server (in your data center or lab) to Azure Cloud via the data management gateway installed in the one of the previous steps. The *ADF* [*pipeline*](http://azure.microsoft.com/en-us/documentation/articles/data-factory-introduction/#terminology) (shown in blue in the picture above) itself will run a *copy* activity to copy the data between SQL server database(s) and Azure Blob Storage. This is a good time to quickly take a break and review some of the ADF terminology [here](http://azure.microsoft.com/en-us/documentation/articles/data-factory-introduction/#terminology) and then install [Azure PowerShell](http://azure.microsoft.com/en-us/documentation/articles/powershell-install-configure/#Install) before continuing further.

Next we will use ADF [pipeline configuration scripts](1_IngressOnPremPipeline/IngressPipeline.json) that came with this document package to create the Ingress Data Pipeline.

Creating Input and Output tables and the pipeline:

1. On-premise SQL Table definition: Open PowerShell and run the following command and provide it the location of [InputOnPremSQLTable](1_IngressOnPremPipeline/InputOnPremSQLTable.json) definition on your computer to create an ADF schema (table) mapping for your on-premise SQL table containing the input data for the machine learning model.

a) Switch to correct PowerShell mode:

Switch-AzureMode AzureResourceManager

b) Add your Azure Subscription Account:

Add-AzureAccount

c) Create SQL Table Schema:

New-AzureDataFactoryTable -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [InputOnPremSQLTable](1_IngressOnPremPipeline/InputOnPremSQLTable.json) definition >

1. Next, in the PowerShell run the following command and provide it the location of [RawOutputBlobTable](1_IngressOnPremPipeline/RawOutputBlobTable.json) definition on your computer to create an ADF schema (table) mapping for Azure Blob Store that stores the data copied from your on-premise SQL table.
2. Create Blob Storage Schema:

New-AzureDataFactoryTable -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [RawOutputBlobTable](1_IngressOnPremPipeline/RawOutputBlobTable.json) definition >

1. Finally, in the PowerShell run the following command and provide it the location of [IngressPipeline](1_IngressOnPremPipeline/IngressPipeline.json) definition on your computer to create an ADF pipeline that connects your on-premise SQL server table to the Azure blob storage defined above.
2. Create the Ingress data pipeline:

New-AzureDataFactoryPipeline -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [IngressPipeline](1_IngressOnPremPipeline/IngressPipeline.json) definition >

### The AML Scoring Pipeline

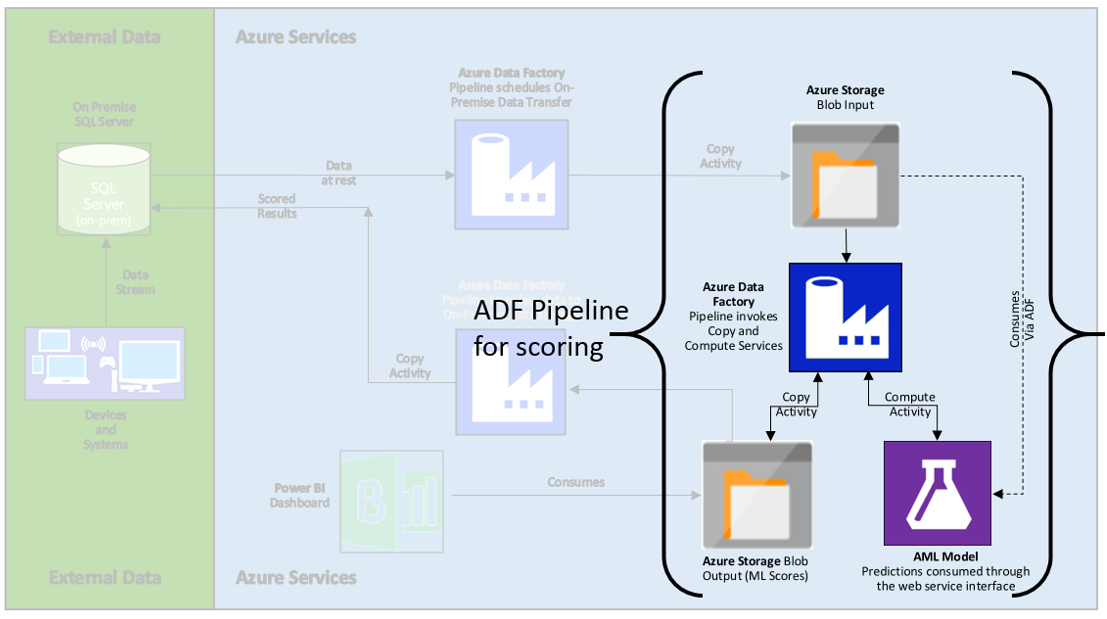


Figure 3: AML Scoring Pipeline

The AML scoring pipeline is show in Figure 3. This pipeline connects to the AML input blob location (defined as RawOutputBlobTable) and to the AML output blob location (defined as ScoredOutputBlobTable). Unlike the ingress pipeline, the AML scoring pipeline run a *compute* activity that calls in the AML web service batch execution endpoint, pass in the data from AML input blob location as a batch to the web service and writes the results back to the AML output blob location.

Next we will use ADF [pipeline configuration scripts](2_AMLScoringPipeline/AMLPipeline.json) that came with this document package to create the AML Scoring Pipeline.

Creating AML Linked Service and the pipeline:

In order to invoke AML as an Azure service we need to link the service to our AML Scoring Pipeline.

1. In the PowerShell run the following command and provide it the location of [AML Linked Service](2_AMLScoringPipeline/AMLLinkedService.json) definition on your computer to link your AML batch execution endpoint to this ADF pipeline.
2. Create Linked Service:

New-AzureDataFactoryPipeline -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [AML Linked Service](file:///C:\Users\mithal\Documents\StarterKit\Source\Azure-MachineLearning-DataScience\Misc\Advanced%20Analytics\OnPremDataSource\2_AMLScoringPipeline\AMLLinkedService.json) definition >

1. Finally, in the PowerShell run the following command and provide it the location of [AMLScoringPipeline](2_AMLScoringPipeline/AMLPipeline.json) definition on your computer to create an ADF pipeline that connects your on-premise SQL server table to the Azure blob storage defined above.
2. Create the AML scoring pipeline:

New-AzureDataFactoryPipeline -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [AMLScoringPipeline](file:///C:\Users\mithal\Documents\StarterKit\Source\Azure-MachineLearning-DataScience\Misc\Advanced%20Analytics\OnPremDataSource\2_AMLScoringPipeline\AMLPipeline.json) definition >

### The Egress Pipeline

The two pipelines created so far together bring the data into Azure for AML model and produce AML scores on an ongoing bases (stored in a temporally partitioned azure blob storage). Next pipeline in this process will allow us to copy the scored results back to your on-premise SQL Server. It is worth nothing that you can a chose different SQL database to write the score results back. Figure 4 shows the Egress Data Pipeline

Next we will use ADF [pipeline configuration scripts](3_EgressOnPremPipeline/EgressPipeline.json) that came with this document package to create the Egress Data Pipeline.

<Add Linked Service Definitions>

Creating Input and Output tables and the pipeline:

1. Create Blob Storage Schema: In the PowerShell run the following command and provide it the location of [ScoredOutputBlobTable](3_EgressOnPremPipeline/ScoredOutputBlobTable.json) definition on your computer to create an ADF schema (table) mapping for Azure Blob Store that stores the AML scores.
2. New-AzureDataFactoryTable -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [ScoredOutputBlobTable](3_EgressOnPremPipeline/ScoredOutputBlobTable.json) definition >
3. On-premise SQL Table definition: In PowerShell and run the following command and provide it the location of [OutputOnPremSQLTable](3_EgressOnPremPipeline/OutputOnPremSQLTable.json) definition on your computer to create an ADF schema (table) mapping for your on-premise SQL table where the scores from the machine learning model will be stored.

a) Create SQL Table Schema:

New-AzureDataFactoryTable -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [OutputOnPremSQLTable](3_EgressOnPremPipeline/OutputOnPremSQLTable.json) definition >

1. Finally, in the PowerShell run the following command and provide it the location of [EgressPipeline](3_EgressOnPremPipeline/EgressPipeline.json) definition on your computer to create an ADF pipeline that connects your on-premise SQL server table to the Azure blob storage to copy the scores back to your on-premise server.
2. Create the Egress data pipeline:

New-AzureDataFactoryPipeline -ResourceGroupName OnPremPipelineRG -DataFactoryName OnPremPipeline –File < location of [EgressPipeline](file:///C:\Users\mithal\Documents\StarterKit\Source\Azure-MachineLearning-DataScience\Misc\Advanced%20Analytics\OnPremDataSource\3_EgressOnPremPipeline\EgressPipeline.json) definition >

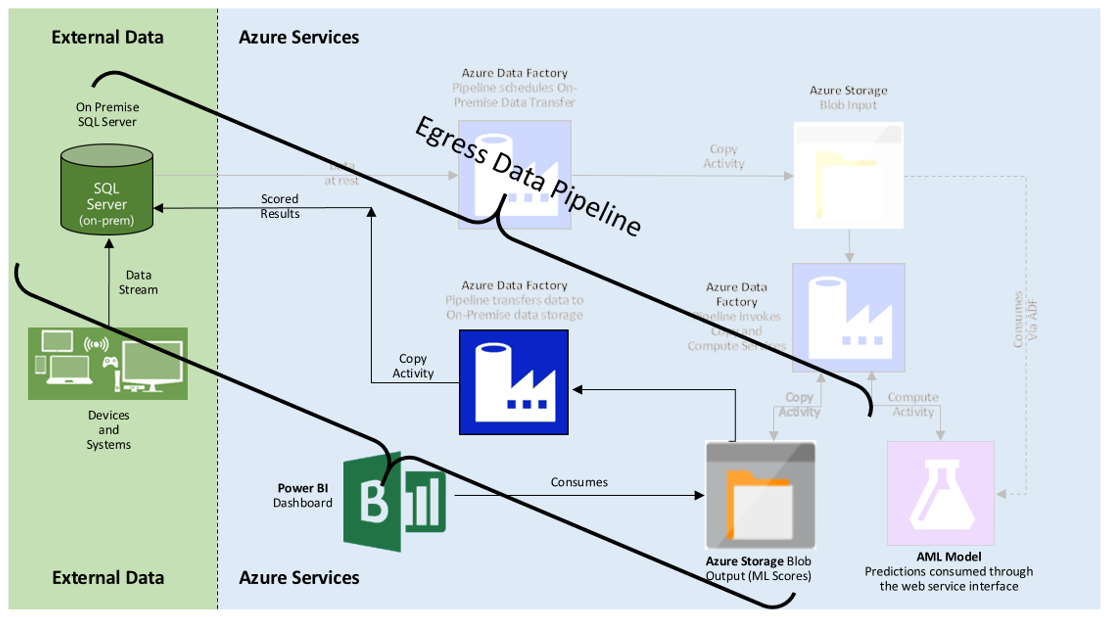


Figure 4: Egress On-Premise Pipeline

### Putting it all together - *Activating* the Pipeline

Will all the ADF pipelines in place we are ready to activate the ADF end to end flow. In the power shell run the following commands to activate the E2E *OnPremPipeline* ADF pipeline.

1. Set-AzureDataFactoryPipelineActivePeriod -ResourceGroupName OnPremPipelineRG-DataFactoryName OnPremPipeline-StartDateTime 2015-03-24T21:30:00 –EndDateTime 2015-03-24T21:45:00 –Name IngressPipeline
2. Set-AzureDataFactoryPipelineActivePeriod -ResourceGroupName OnPremPipelineRG-DataFactoryName OnPremPipelineRG-StartDateTime 2015-03-24T21:30:00 –EndDateTime 2015-03-24T21:45:00 –Name AMLPipeline
3. Set-AzureDataFactoryPipelineActivePeriod -ResourceGroupName OnPremPipelineRG-DataFactoryName OnPremPipelineRG-StartDateTime 2015-03-24T21:30:00 –EndDateTime 2015-03-24T21:45:00 –Name EgressPipeline

After you activate the pipelines you will see input data copied to Azure blob, scored by AML and finally AML scores available in your on-premise SQL server every hour. The pipelines are activated sequentially so any failure in one if the pipelines will stop the pipelines down the stream running and producing garbled results.