­

IoC Bootcamp

Data Science and Cloud Systems

Birmingham City University

**Lab 1.1:**Data Factory

# Introduction

Data Factory is a cloud data integration service used to compose data storage, movement, and processing services into automated data pipeline. Azure Data Factory is a cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation. Azure Data Factory does not store any data itself. It allows you to create data-driven workflows to orchestrate the movement of data between supported data stores and processing of data using compute services in other regions or in an on-premises environment. It also allows you to monitor and manage workflows using both programmatic and UI mechanisms

## 4 key components in Data Factory

Data Factory has four key components that work together to define input and output data, processing events, and the schedule and resources required to execute the desired data flow:

* Datasets represent data structures within the data stores. An input dataset represents the input for an activity in the pipeline. An output dataset represents the output for the activity. For example, an Azure Blob dataset specifies the blob container and folder in the Azure Blob Storage from which the pipeline should read the data. Or an Azure SQL Table dataset specifies the table to which the output data is written by the activity.
* Pipeline is a group of activities. They are used to group activities into a unit that together performs a task. A data factory may have one or more pipelines. For example, a pipeline could contain a group of activities that ingests data from an Azure blob and then runs a Hive query on an HDInsight cluster to partition the data.
* Activities define the actions to perform on your data. Currently, Data Factory supports two types of activities: data movement and data transformation.
* Linked services define the information needed for Data Factory to connect to external resources. For example, an Azure Storage linked service specifies a connection string to connect to the Azure Storage account.

The following schema shows us the relationships between the Dataset, Activity, Pipeline, and Linked Services components:

# Diagram Description automatically generated

## Lab Architectural:

Graphical user interface

Description automatically generated with low confidence

# Objectives

* To create a pipeline to enable the flow of data from an application to a data warehouse.
* To create a Blob storage as a service to store large amount of unstructured data.
* To create a linking service that contain connection strings that Data Factory uses at runtime to connect to your Azure Storage and Azure SQL Database, respectively.
* To create SQL database.

# Requirements

* Resource Group
* Data factory resource
* SQL database resource
* Storage account resource

Lab Instructions

After configuring and creating the required resources for this lab, now we will start configure and second part of the system by creating Pipeline, Blob storage and other services that are needed for this lab.

## Step 5:

The first step in this lab is to use a Blob storage. In order to create a blob storage, we need to go to the storage account, blob service and then click on Container **(Figure 1)** and then click on **+container (Figure 2)**

Graphical user interface, application

Description automatically generated

Figure 1 Storage account - Blob service

Graphical user interface, text, application, email

Description automatically generated

Figure 2 Create Blob service – Container

Once add the name of the container click create and go to the container, then click on upload to select the IOTtemp.csv (the IOTtemp.csv is available at the moodle) then click upload to upload the dataset (Figure 3).

If you click on the uploaded dataset, you can explore the data inside by clicking on **Edit** (Figure 4).

Graphical user interface, text, application, email

Description automatically generated

Figure 3 Upload the CSV file

Graphical user interface, text, application

Description automatically generated

Figure 4 Explore dataset rows

## Step 6:

Now, we need to start configuring the SQL database and create the tables where the data will be stored.

To create the SQL table, follow these steps:

1. Go to resource group and click on SQL database.

Graphical user interface, text, application, email

Description automatically generated

Figure 5 Resource group

1. Click on query editor (Figure 6) and enter the username and password (Figure 7) you created in Lab 10a step 3.

Graphical user interface, text, application, Word

Description automatically generated

Figure 6 SQL database

## Graphical user interface, Word Description automatically generated

Figure 7 Query editor

**Note: if you have an issue login to the database query editor, check the firewall on SQL database.**

When you login to the Query Editor, use the following SQL query to create the tables then click run to run the query.

CREATE TABLE smartsystems(

id varchar(100),

room\_id varchar(100),

noted\_date DATETIME,

temp decimal(3),

out\_in varchar(100)

);

Graphical user interface, text, application, email

Description automatically generated

Figure 8 create tables using a query

## Step 7:

At this step you need to go to the data factory to start creating pipeline and copying the data to the SQL table.

To start these configurations, follow these steps:

1. Go to the resource group and click on data factory.
2. Click on Author & Monitor

A picture containing text, screenshot, computer, computer

Description automatically generated

Figure 9 Data factory Resource

A picture containing text, screenshot, parking

Description automatically generated

Figure 10 Author & Monitor

When you enter Author and Monitor, go to Author and then click on the Plus sign then add a pipeline.

* Change the default name to **Blob to SQL.**

## Graphical user interface, application, Word Description automatically generated

Figure 11 Add a pipeline

## A picture containing text, screenshot, computer, computer Description automatically generated

Figure 12 Configure the pipeline

The next step is to connect to Blob storage. To do that, follow these steps:

1. Go to Manage , Linked Services and click on **New**.
2. Add Azure Blob Storage.

## A screenshot of a computer Description automatically generated with medium confidence

Figure 13 add a blob storage through linked services

1. Change the name of Blob Storage to **InputBlob**
2. Select the Azure subscription and storage account.
3. Test the connection before creating the blob storage.

Graphical user interface, text, application, email

Description automatically generated

Figure 14 configure Azure linked service - Blob Storage

## Step 8:

The next step is to define the data. When we pull the data from the data storage we need to know how does this data looks like. To do that, Follow these steps:

1. Go to Author and click on + then add a dataset (Figure 15)
2. Select **Azure Blob Storage** and then choose **CSV** as a format (Figure 16).
3. Choose a name for the properties and select the **InputBlob** that you created.
4. Select the path of your dataset.
5. If the first row of your dataset contains headers, then select First row as a header
6. Once you create the dataset you can test to connection to ensure it is connected (Figure 18).

Graphical user interface, application

Description automatically generated

Figure 15 Create a dataset

Graphical user interface, application

Description automatically generated

Figure 16 Azure blob storage

Graphical user interface, application, Word

Description automatically generated

Figure 17 Selecting CSV as a format and adding the path of the dataset

Graphical user interface, text, application

Description automatically generated

Figure 18 Dataset testing the connection

## Step 9:

The next step is to setup the connection between the data and SQL. To create the connection between them follow these steps:

1. Add a new linked service and select Azure SQL Database (Figure 19).

Graphical user interface, application, Teams

Description automatically generated

Figure 19 Create output for the Data to SQL - Azure SQL database

1. Change the name to OutputSQL.
2. Select the server’s name and database name.
3. provide your username and password that you created when you created the SQL database in step 3.
4. Test the connection and then create.

Graphical user interface, text, application

Description automatically generated

Based on our diagram we need another dataset as an output to the SQL database, since we pull the data and have it in the memory, then we need to copy the data to SQL dataset.

To create a new dataset, follow these steps:

1. Go to Author and click + then select Dataset.
2. Select Azure SQL database
3. Select the tables name that you created through the query editor.

Graphical user interface, application

Description automatically generated

Figure 20 create new database and set its properties

## Step 10:

Final step in this lab is to copy the data from the memory to the SQL database. In order to copy the data to the SQL database follow these steps:

1. Click on the pipeline and drag and drop Copy Data from the activates (Figure 21).
2. Change the name and click on the Source to select the source of the data.
3. Click on Sink to select the destination of the data (Figure 21).

Graphical user interface, application

Description automatically generated

Figure 21 Pipeline - copy data to SQL database

Graphical user interface

Description automatically generated

Figure 22 Copy Data - select Sink

The last step is to debug to ensure that everything is working correctly.

Graphical user interface, text

Description automatically generated

Figure 23 Copy data - debug

**Once the debug shows it has succeeded, click on publish all.**

Graphical user interface, text, application, email

Description automatically generated

Figure 24 Publish all

Now the data has been pushed to the SQL database. Finally, we need to check if the data was pushed correctly or not by running a query on the Query Editor.

1. Go to resource group and select SQL database
2. Go to Query editor and run the following query:

SELECT \* FROM smartsystems

That query should return the values inside the SQL database table(Figure 25).

Graphical user interface, text, application, email

Description automatically generated

Figure 25 Final query