­

IoC Bootcamp

Data Science and Cloud Systems

**Lab:**Data Factory

Birmingham City University

# 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

## Step 1:

First step is to create the resource group, if you already have one you may skip this step.

In order to create a resource group, follow these steps:

1. Click on resource group in the home page and then click create.

A computer screen capture

Description automatically generated with medium confidence

Figure Create resource Group

1. Add a name to the resource group and then click Review + Create.

Graphical user interface, text, application

Description automatically generated

Figure Adding a name to your resource group

Graphical user interface, text, application, email

Description automatically generated

Figure Resource Group

## STEP 2:

Once the resource group is ready the next step is to create a data factory resource.

* Click on create a resource, in the search bar write data factory and click on data factory.

A screenshot of a computer

Description automatically generated with medium confidence

Figure Add a resource

A screenshot of a computer

Description automatically generated

Figure Adding a data factory

Graphical user interface, text, application

Description automatically generated

Figure Create a data factory

Graphical user interface, text, application, Word

Description automatically generated

Figure Fix git issue

Graphical user interface, text, application, Word

Description automatically generated

Figure creating data factory - Validation Passed

## STEP 3:

In this step we will create the SQL database resource, to do that follow the following steps:

1. Add SQL database from resources (Figure 9).

A screenshot of a computer

Description automatically generated with medium confidence

Figure 9 Create SQL database resource

1. Select the resource group.
2. 3- Click on create server and insert the following details:
   * Server name
   * Server admin login
   * Password

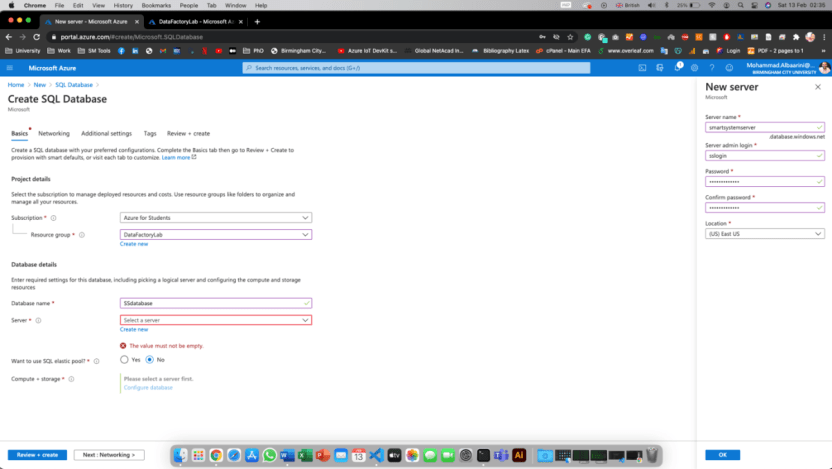


Figure 10 create new SQL database and server

* 1. 4- Click on configure database (Figure 11) and follow these steps:
* Select standard option
* Set Database Transaction Unit (DTU) to 20
* Data Max size to 250GB
* Click Apply
  1. 5- Click Review + Create (Figure 12)

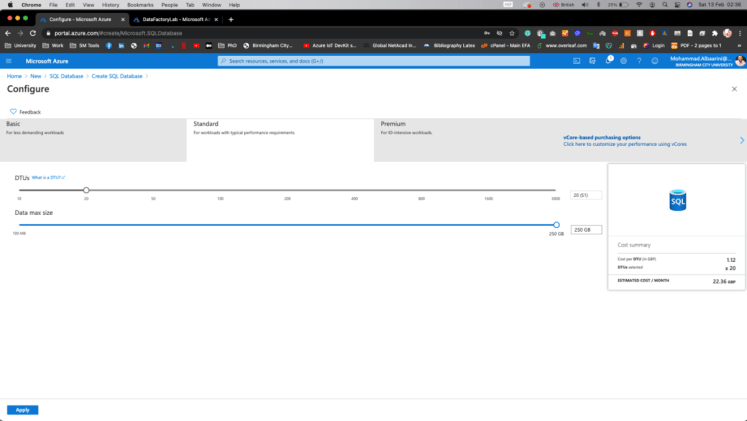


Figure 11 Configure database Server

A picture containing text, screenshot, computer, computer

Description automatically generated

Figure 12 Create and Review

## STEP 4:

After creating and configuring the SQL database, the last resource is important to create is a storage account. To create a storage account, follow these steps:

1. Add a new resource and select storage account and click create.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 13 create storage account

1. Select the resource group and add a name to the storage account then click review + create.

A picture containing text, screenshot, computer, computer

Description automatically generated

Figure 14 select the resource group and create the storage account.