

Contents

Workshop Summary	3
Azure Services	3
Architecture	3
Data Sources	3
Activity 1 – Azure Data Lake (ADLS)	4
Create the Resource Group	4
Create the Storage Account	6
Configure the ADLS storage	10
Activity 2 – Azure Data Explorer (ADX)	11
Create ADX Cluster	11
Create ADX Database	13
Activity 3 – Data Factory (ADF)	16
Overview	16
Publishing in ADF	16
Create the Data Factory	17
Create ADF Pipeline	19
Create Pipeline Linked Services	22
ADF ‘Httplinkedservice’ linked service	22
ADF ‘adlslinkedservice’ linked service	23
ADF ‘demoadadxsink’ linked service	25
Create Pipeline Datasets	26
ADF ‘cvsink’ dataset	26
ADF ‘csvsource’ dataset	28
ADF ‘adxsink’ dataset	30
Create Pipeline Activities	31
ADF ‘Copy COVID data to ADLS’ activity	31
ADF ‘ADX – DDL drop extents’ activity	34
ADF ‘Import Daily File to ADX’ activity	36
Create Pipeline Trigger	38
Activity 4 – Security & Access	41
Terminology Explained	41
Application Registration	41
Managed Identities	41
Service Principal	41

COV_19 Analytics with Azure Data Explorer (ADX)

ADF – Security Configuration (Linked Service).....	41
Httplinkedservice	41
Adlslinkedservice	42
demoadxsink	42
ADLS – Security Configuration	43
ADX – Security Configuration (cluster)	44
ADX – Security Configuration (database).....	45
Further Reading	45
ADX Dashboard – Sharing	47
Activity 5 – ADF Pipeline	48
Start/activate the ADF trigger.....	48
Monitor ADF Pipeline execution.....	49
Activity 6 – Analytics via KQL	51
ADX KQL queries	51
Basic Analytics.....	52
Activity 7 – Analytics via ADX Dashboards.....	56
Create Dashboard	56
Data sources.....	57
Create Parameters	57
Date Parameter.....	57
Time Range Parameter	59
Country Parameter	61
Dashboard Visualisations.....	64
Remaining Dashboard Visualisations	66
Activity 8 – Wrap-up	69
Clean-up	69

Workshop Summary

The first workshop, "Workshop 1", is intended to produce a data pipeline that ingests a publicly available COVID-19 dataset, via ADF, to ADLS. This dataset is then loaded into ADX by fully replacing the existing data during each iteration.

Azure Services

The following Azure services are introduced during this workshop to the architecture:

Azure Data Factory (ADF) [[link](#)]

Azure Data Lake (ADLS) [[link](#)]

Azure Data Explorer (ADX) [[link](#)]

Architecture

The diagram “COVID-19 HLD Architecture.pdf” shows the architecture that will be built. A copy of this can be found in the /downloads f



Data Sources

The COVID-19 datasets used in this workshop are publicly available. They are located here:

<https://azure.microsoft.com/en-in/services/open-datasets/catalog/?q=covid>

For this workshop, the “COVID-19 Data Lake” datasets were chosen:

<https://azure.microsoft.com/en-in/services/open-datasets/catalog/covid-19-data-lake/>

and then the “Bing COVID-19 Data”:

<https://azure.microsoft.com/en-in/services/open-datasets/catalog/bing-covid-19-data/>

We will be using the daily JSON file, in CSV format:

https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.json

Activity 1 – Azure Data Lake (ADLS)

As per the diagram “Azure data Factory-Security.pdf” we are creating an Azure Data Lake (Gen-2) account.

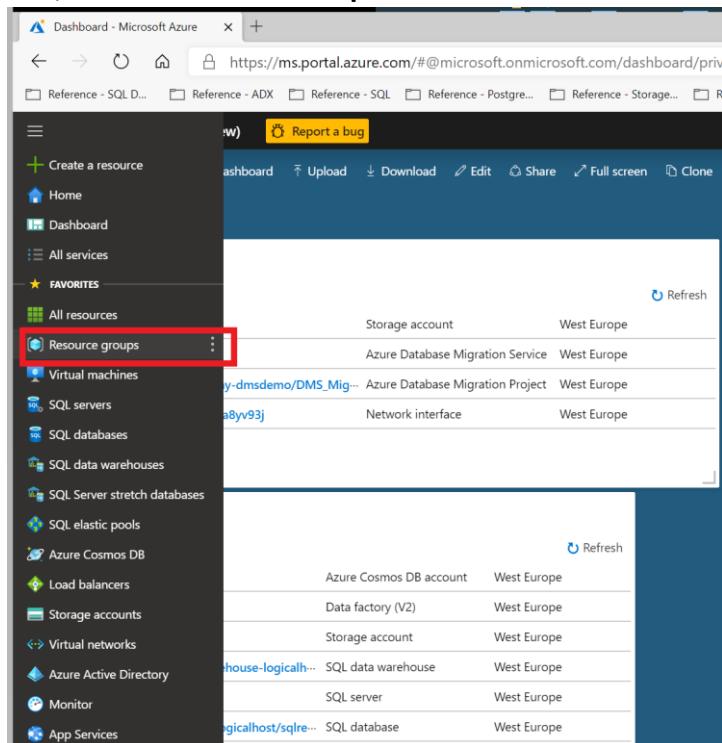
As part of this activity it maybe necessary to also refer to ‘[Activity 4 – Security & Access](#)’.

Create the Resource Group

Define the Azure Resource Group that the Azure services will be created in.

NOTE: *Creating the Resource Group can also be achieved via PowerShell [here](#).*

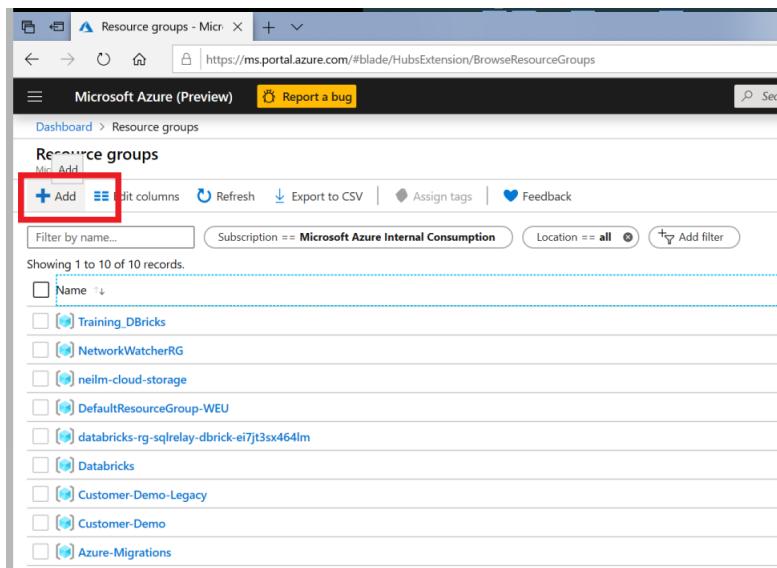
1. Sign in to the [Azure portal](#).
2. In the Azure portal, select **Resource Groups**:



The screenshot shows the Microsoft Azure portal dashboard. On the left, there is a sidebar with various service links: Create a resource, Home, Dashboard, All services, Favorites, All resources, and Resource groups. The 'Resource groups' link is highlighted with a red box. The main content area displays a table of existing resource groups, including Storage account, Azure Database Migration Service, Azure Cosmos DB account, Data factory (V2), and others, all located in West Europe. There are refresh buttons at the top and bottom of the table.

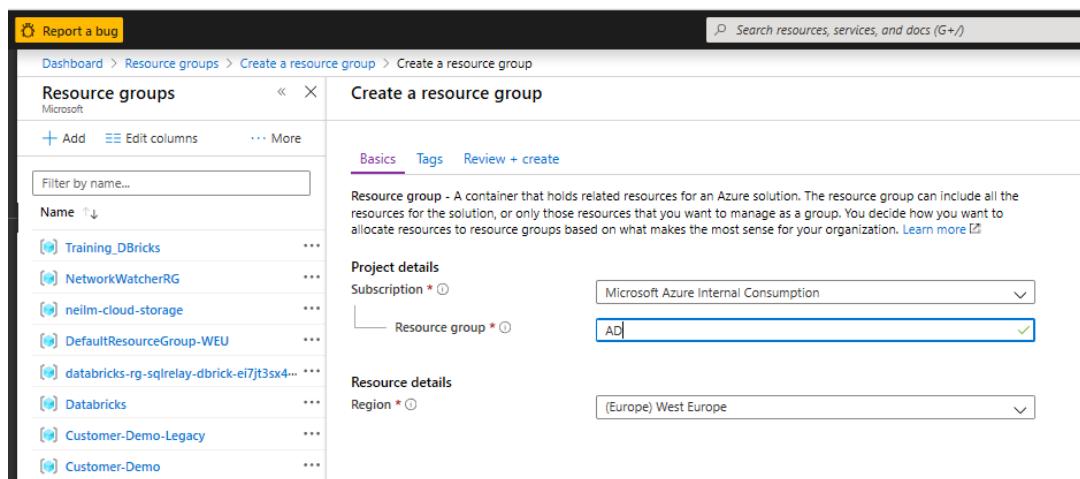
3. When the blade loads, select **Add**:

COV_19 Analytics with Azure Data Explorer (ADX)



The screenshot shows the 'Resource groups' blade in the Azure portal. At the top, there's a search bar and filter options for 'Subscription' (Microsoft Azure Internal Consumption) and 'Location' (all). Below that is a table header with columns for Name, Type, Status, and Last activity. A red box highlights the '+ Add' button in the top-left corner of the table header area.

4. Enter the Resource Group Details:



The screenshot shows the 'Create a resource group' blade. On the left is a sidebar with a list of existing resource groups. The main area has tabs for 'Basics', 'Tags', and 'Review + create'. Under 'Project details', the 'Subscription' is set to 'Microsoft Azure Internal Consumption' and the 'Resource group' is set to 'AD'. Under 'Resource details', the 'Region' is set to '(Europe) West Europe'.

Property	Description	Required
Subscription	The value must be set to the subscription the ADX resource group is to be created in.	Yes
Resource Group	The name of the workshop resource group. Enter a name of your choice.	Yes
Region	Enter the Azure region your resource group is to be located in. This should be a region that supports the Azure services required in this workshop and located as close as possible to your geographic location.	Yes

5. Select **Create**.

COV_19 Analytics with Azure Data Explorer (ADX)

After a few minutes you will have created the resource group into which the Azure services can be deployed to.

Create the Storage Account

The data lake storage account will serve as the *data lake* for both this and all the other workshops.

An explanation of Azure Data Lake (ADLS) resource is here [[link](#)].

1. Within the resource Group you just created, select '+ Add' or 'Create Resource'

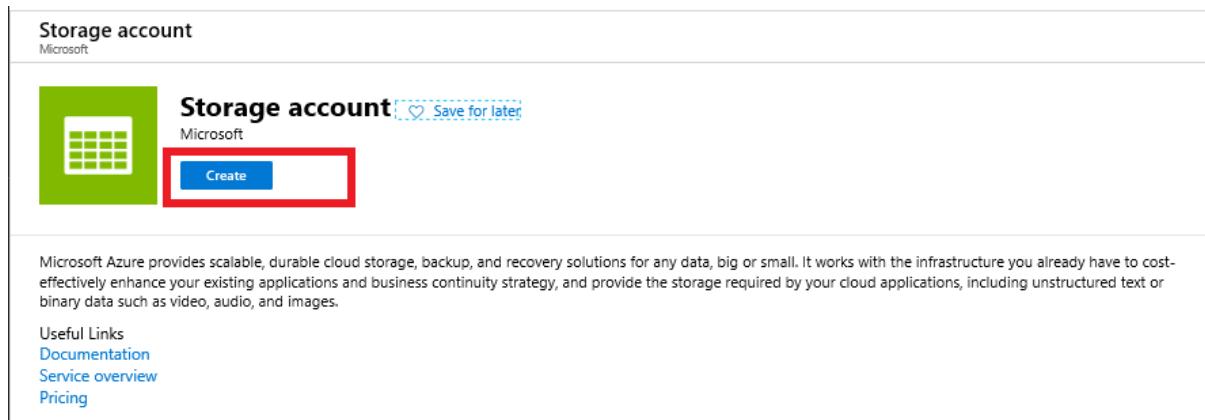
The screenshot shows the Azure Resource Groups blade. On the left, there's a list of resource groups: ADX (selected and highlighted with a red box), Azure-Migrations, Customer-Demo, Customer-Demo-Legacy, and Databricks. On the right, the details for the ADX resource group are shown, including its subscription information (Microsoft Azure Internal Consumption, Subscription ID: 7795b573-5882-4d6c-9b18-6b58d2ab218b), tags (none), and settings. At the top right, there's a '+ Add' button, which is also highlighted with a red box.

2. Type **blob** or **storage** in the search window (see below) and then select **Storage Account** from the results displayed:

The screenshot shows the Azure Marketplace search results for 'blob'. A search bar at the top contains the text 'blob' and is also highlighted with a red box. Below the search bar, there are two tabs: 'Azure Marketplace' (selected) and 'See all'. To the right, under the heading 'Popular', there is a grid of tiles representing various Azure services, each with an icon and a link to a 'Quickstart tutorial'. The services listed are: Windows Server 2016 Datacenter, Ubuntu Server 18.04 LTS, Web App, SQL Database, Function App, Azure Cosmos DB, Kubernetes Service, and DevOps Project.

3. On the Storage Account blade, select **Create**:

COV_19 Analytics with Azure Data Explorer (ADX)



Storage account

Microsoft

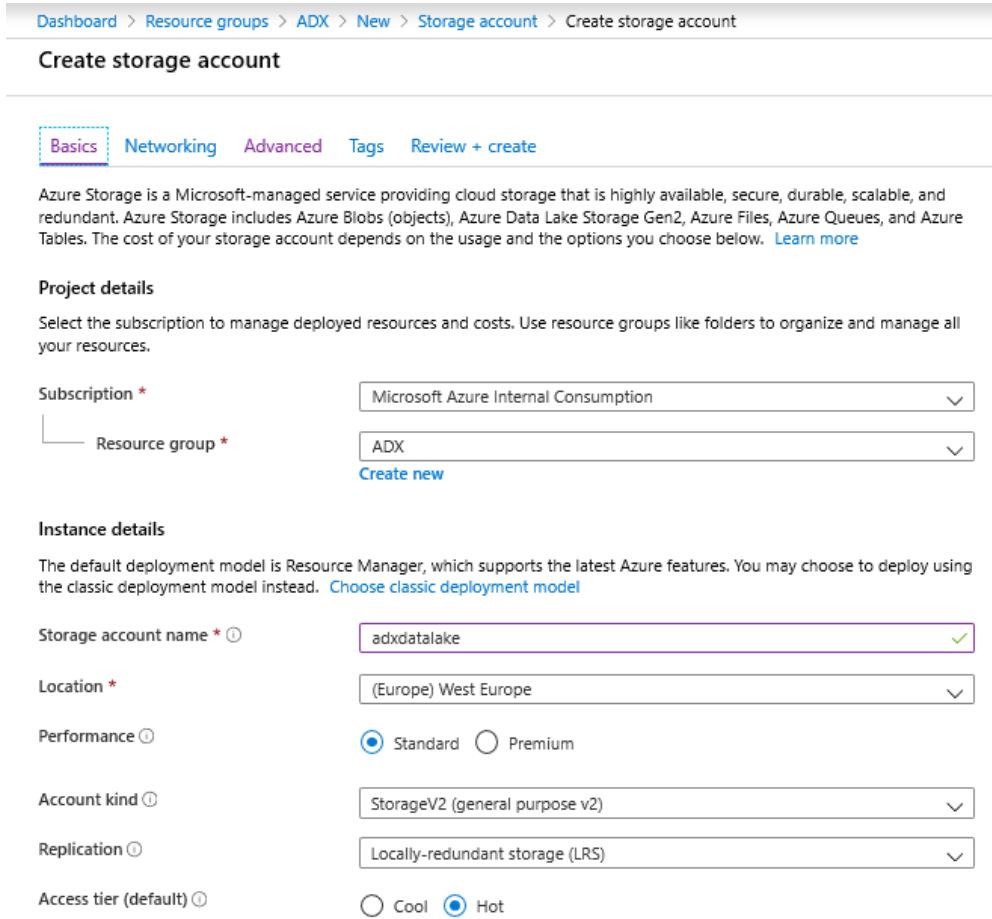
Storage account Save for later

Create

Microsoft Azure provides scalable, durable cloud storage, backup, and recovery solutions for any data, big or small. It works with the infrastructure you already have to cost-effectively enhance your existing applications and business continuity strategy, and provide the storage required by your cloud applications, including unstructured text or binary data such as video, audio, and images.

Useful Links
Documentation
Service overview
Pricing

4. On the **Create storage account** blade, complete the Basics blade (see below)



Dashboard > Resource groups > ADX > New > Storage account > Create storage account

Create storage account

Basics Networking Advanced Tags Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * Microsoft Azure Internal Consumption

Resource group * ADX [Create new](#)

Instance details

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. [Choose classic deployment model](#)

Storage account name * adxdatalake

Location * (Europe) West Europe

Performance Standard Premium

Account kind StorageV2 (general purpose v2)

Replication Locally-redundant storage (LRS)

Access tier (default) Cool Hot

Complete the blade by entering the details as below:

Property	Description	Required
Subscription	The value must be set to the subscription the ADX resource group is created in.	Yes
Resource Group	The name of the workshop resource group you created in 'Activity 1.1'.	Yes

COV_19 Analytics with Azure Data Explorer (ADX)

Property	Description	Required
Storage account name	Enter <for example> demodataalakeadls	Yes
Location	Select your chosen region to match the resource group <i>ADX</i> you have created.	Yes
Performance	Enter Standard	Yes
Account Kind	Select Storage v2 (general purpose v2)	Yes
Replication	Select Locally-redundant storage (LRS)	Yes

5. Select **Advanced**, this is where the blob storage account had the ADLS data lake attributes enabled:

Dashboard > Resource groups > ADX > New > Storage account > Create storage account

Create storage account

Basics Networking Advanced Tags Review + create

Security

Secure transfer required Enabled Disabled

Azure Files

Large file shares Enabled Disabled

Data protection

Blob soft delete Enabled Disabled

⚠ Blob soft delete and hierarchical namespace cannot be enabled simultaneously.

Data Lake Storage Gen2

Hierarchical namespace Enabled Disabled

Complete the blade by entering the details as below:

Property	Description	Required
Security	Select Enabled	Yes

Property	Description	Required
Azure Files	Select Disabled .	Yes
Hierarchical Namespace	Select Enabled .	Yes

6. Select **Review + Create**. The blade below will appear, select **Create**:

The screenshot shows the 'Create storage account' blade in the Azure portal. At the top, a green bar indicates 'Validation passed'. Below it, the 'Review + create' tab is selected. The blade is divided into sections: Basics, Networking, Advanced, and Tags. Under Basics, settings include Subscription (Microsoft Azure Internal Consumption), Resource group (ADX), Location ((Europe) West Europe), Storage account name (adxdatalake), Deployment model (Resource manager), Account kind (StorageV2 (general purpose v2)), Replication (Locally-redundant storage (LRS)), Performance (Standard), and Access tier (default) (Hot). Under Networking, Connectivity method is set to Public endpoint (all networks). Under Advanced, Secure transfer required, Hierarchical namespace, Blob soft delete, and Large file shares are all set to Enabled. At the bottom, there are buttons for < Previous, Next >, and Download a template for automation, with the 'Create' button highlighted by a red box.

After a short wait you will be notified that the ADLS storage has been created. You can go to the resource to review its setup or select **Resource Groups**, then select your resource group, to see the **demodatalakeadls** resource you have just created:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the Azure portal interface for the 'DataDemo' resource group. On the left, there's a navigation sidebar with options like Overview, Activity log, Access control (IAM), Tags, Events, Settings, Quickstart, Deployments, Policies, Properties, and Law. The main area displays resource details: Subscription (Microsoft Azure Internal Consumption), Subscription ID (7795b573-5882-4d6c-9b18-6b58d2ab218b), and Tags (Click here to add tags). Below this is a search bar and a filter section. The main list shows three resources: 'demodatafactory' (Data factory (V2)), 'demodataclusteradr' (Azure Data Explorer Cluster), and 'demodatalakeadls' (Storage account). The 'demodatalakeadls' row is highlighted with a red box.

Notice that **demodatalakeadls** has the type: Storage Account.

Configure the ADLS storage

A root container is required for files to be landed to in a hierarchical structure. Files will be landed to ADLS via an Azure Data Factory Pipeline activity.

1. Select the resource group in the Azure portal
2. Select **demodatalakeadls** storage account
3. Select **Containers** under **Data Lake Storage**
4. Select **+ Container** to add a new container, as below:
5. Enter **landingzone** and select **OK**;
6. Repeat and create folders **stagingzone** and **loggingzone**

The container structure should match the image below once completed. Also note the **Public Access Level** and **Lease State** settings:

The screenshot shows the 'Containers' page for the 'demodatalakeadls' storage account. On the left, there's a navigation sidebar with Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, and Data transfer. The main area shows a table of containers. The table has columns for Name, Last modified, Public access level, and Lease state. Three containers are listed: 'landingzone' (Last modified 6/23/2020, 3:39:10 PM, Private, Available), 'loggingzone' (Last modified 6/25/2020, 10:30:02 AM, Private, Available), and 'stagingzone' (Last modified 6/25/2020, 10:28:32 AM, Private, Available).

Name	Last modified	Public access level	Lease state
landingzone	6/23/2020, 3:39:10 PM	Private	Available
loggingzone	6/25/2020, 10:30:02 AM	Private	Available
stagingzone	6/25/2020, 10:28:32 AM	Private	Available

The JSON document 'demodatalakeadls.json' has the configuration settings for **demodatalakeadls** is located in the /downloads folder.

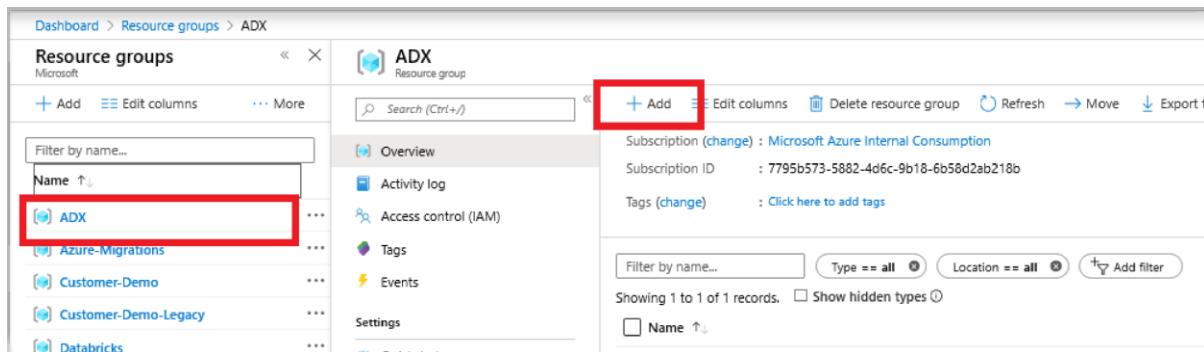
Activity 2 – Azure Data Explorer (ADX)

Create ADX Cluster

As per the diagram “Lab1 - Activity 4 – Azure Data Explorer.pdf” we are creating the ADX cluster.

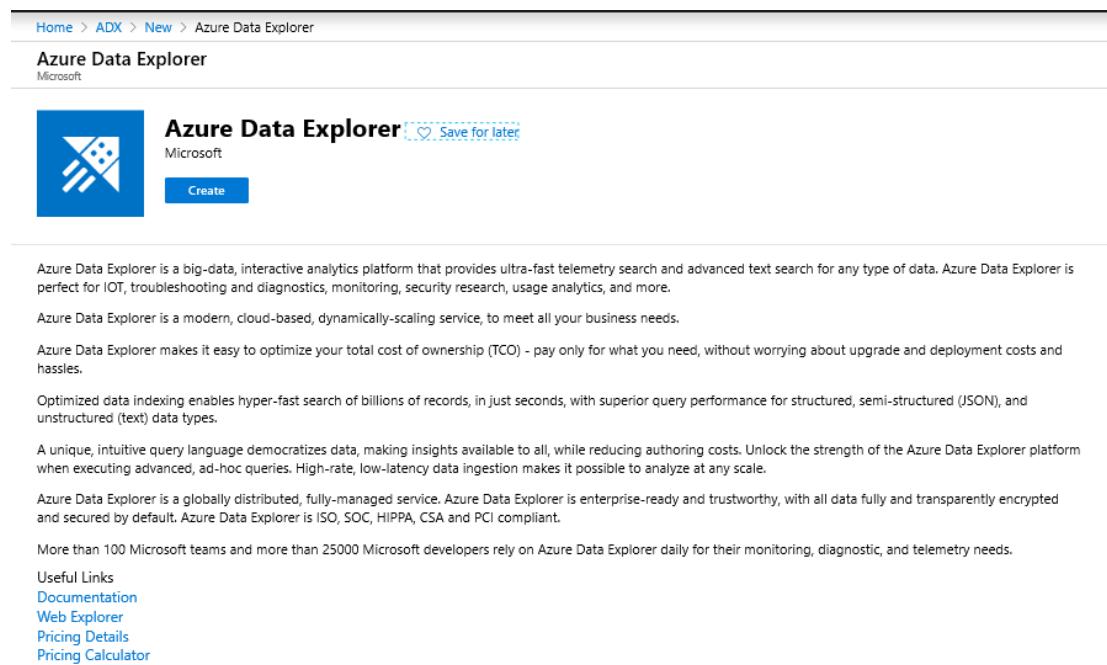
As part of this activity it may be necessary to also refer to ‘[Activity 4 – Security & Access](#)’.

1. Within the ADX resource Group, select **+ Add or Create Resource**



The screenshot shows the Azure Resource Groups blade for the 'ADX' resource group. On the left, there's a list of other resource groups: 'Azure-Migrations', 'Customer-Demo', 'Customer-Demo-Legacy', and 'Databricks'. A red box highlights the 'Add' button at the top right of the blade. The main area displays the 'Overview' section with details like Subscription ID and Tags.

2. Type **data explorer** in the search window (see below) and then select **Azure Data Explorer** from the results displayed. The following blade will appear:



The screenshot shows the 'Azure Data Explorer' creation blade. It features a summary section with a blue icon, the service name 'Azure Data Explorer', a 'Save for later' link, and a prominent 'Create' button. Below this, there's a detailed description of the service, followed by sections for TCO optimization, indexing performance, query language, security, and developer adoption.

3. Select **Create** and the following blade will appear:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the 'Create an Azure Data Explorer Cluster' page. At the top, there's a breadcrumb navigation: Home > ADX > New > Azure Data Explorer > Create an Azure Data Explorer Cluster. Below the breadcrumb, the title 'Create an Azure Data Explorer Cluster' is displayed. A horizontal navigation bar at the top of the form includes tabs: Basics * (highlighted with a blue border), Configurations, Tags, and Review + create.

PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *: Microsoft Azure Internal Consumption

Resource group * ⓘ: ADX (with a 'Create new' link)

CLUSTER DETAILS

Cluster name * ⓘ: adxlab1

Region * ⓘ: West Europe

Availability zones ⓘ: None

Compute specifications (View full pricing details) *: Dev(No SLA)_Standard_D11_v2 (2 vCPUs, 75 GB Cache, 14 GB RAM, 1 Instance)

Complete the **Basics** blade by entering the details as below:

Property	Description	Required
Subscription	The value must be set to the subscription the ADX resource group is created in.	Yes
Resource Group	The name of the workshop resource group.	Yes
Name	Enter demoadxclusteruk	Yes
Region	Select your chosen region to match the resource group <i>ADX</i> you have created.	Yes
Availability zones	Select None	Yes
Compute specifications	<p>The options available in this listbox will depend upon your Region selection. As this is a workshop, select a low spec configuration to work with. For a PoC, MVP or production installations you will need a higher specification.</p> <p>Select 'Dev(No SLA)' or the smallest specification in your listbox.</p>	Yes

COV_19 Analytics with Azure Data Explorer (ADX)

4. Select **Review + Create**, then select **Create** to build the ADX cluster. After a short wait you will be notified that the ADX cluster has been created. You can go to the resource to review its setup or select **Resource Groups**, then select **your resource group**, to see the **demoadxclusteruk** resource you have just created:

The screenshot shows the Azure portal's 'Resource groups' blade for the 'DataDemo' group. It lists three resources: 'demoadffdatafactory' (Data factory V2), 'demoadxclusteruk' (Azure Data Explorer Cluster), and 'demodatahubeds'. The 'demoadxclusteruk' row is highlighted with a red box.

5. Select **Properties** to see the ADX cluster URI. This will be needed later in this workshop.

Create ADX Database

We will now create the ADX database for the COVID-19 data to be ingested and loaded into.

The KQL script to for the DDL [KQL_covidstaging.kql] is located in the /downloads folder.

1. In the workshop resource group, select the ADX cluster resource you created [**demoadxclusteruk**]
2. When the ADX resource blade appears, select '**+ Add database**':

The screenshot shows the 'demoadxclusteruk' ADX cluster blade. On the left, the 'Overview' tab is selected. On the right, there is a table titled 'Databases from A-Z' with columns 'DATABASE' and 'SIZE'. Below the table, there is a large blue button labeled '+ Add database'.

3. Enter '**covid**' for the database name and complete the tab as below:

COV_19 Analytics with Azure Data Explorer (ADX)

Azure Data Explorer Database

Create new database

Admin Microsoft

Database name * covid ✓

Retention period (in days) 365 ✓
Unlimited

Cache period (in days) 31 ✓
Unlimited



4. Select **Create** to complete the database creation.
5. Once the database has been created, select **covid** under **Databases from A-Z** or **Data:**

Dashboard > DataDemo >

demoadxclusteruk Azure Data Explorer Cluster

Search (Ctrl+ /) Add database Stop Refresh Move Delete Feedback

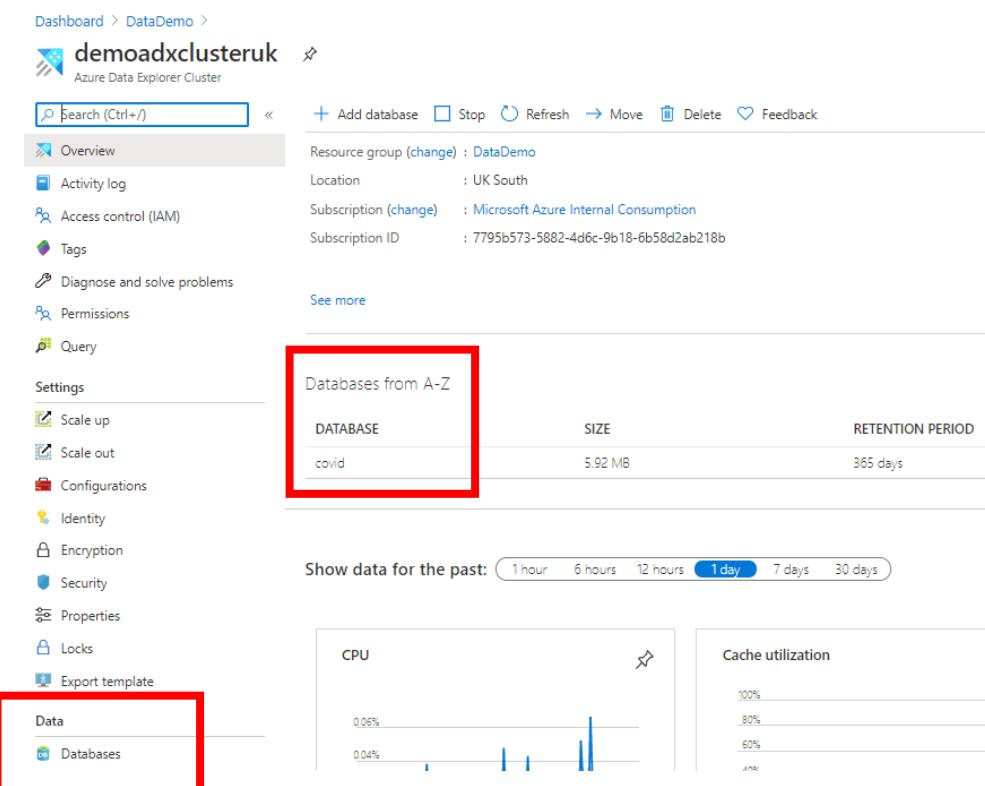
Overview Activity log Access control (IAM) Tags Diagnose and solve problems Permissions Query Settings Scale up Scale out Configurations Identity Encryption Security Properties Locks Export template Data Databases

Databases from A-Z

DATABASE	SIZE	RETENTION PERIOD
covid	5.92 MB	365 days

Show data for the past: 1 hour 6 hours 12 hours 1 day 7 days 30 days

CPU Cache utilization



6. When the **covid** database blade appears, select **Query**:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the Azure Data Explorer Database settings page for the 'covid' database in the 'demoadxclusteruk' resource group. The 'Query' tab is highlighted with a red box. The 'Overview' tab is also visible. The database details include:

- Resource group (change) : DataDemo
- Location : UK South
- Subscription (change) : Microsoft Azure Internal Consumption
- Subscription ID : 7795b573-5882-4d6c-9b18-6b58d2ab218b

- When the query window appears, open/paste the [KQL_covidstaging.kql] KQL into the query window and then select **Run** to create the table. The newly created table will appear under the **covid** database outline:

The screenshot shows the Azure Data Explorer Query editor with the following KQL code:

```
1 .drop table covidstaging;
2
3 .create table covidstaging
4 (
5     S_id: string,
6     S_updated: datetime ,
7     S_confirmed: int ,
8     S_confirmed_change: int ,
9     S_deaths: int ,
10    S_deaths_change: int,
11    S_recovered: int,
12    S_recovered_change: int,
13    latitude: string,
14    longitude: string,
15    iso2: string,
16    iso3: string,
17    country_region: string,
18    admin_region_1: string,
19    iso_subdivision: string,
20    admin_region_2: string,
21    load_time: string
22 );
23
24 .create-or-alter table covidstaging ingestion csv mapping "covidstagingdatamapping"
25 '['
26     {"column": "S_id", "DataType": "string", "Properties": {"Ordinal": "0"}},'
27     {"column": "S_updated", "DataType": "datetime", "Properties": {"Ordinal": "1"}},'
28     {"column": "S_confirmed", "DataType": "int", "Properties": {"Ordinal": "2"}},'
29     {"column": "S_confirmed_change", "DataType": "int", "Properties": {"Ordinal": "3"}},'
30     {"column": "S_deaths", "DataType": "int", "Properties": {"Ordinal": "4"}},'
31     {"column": "S_deaths_change", "DataType": "int", "Properties": {"Ordinal": "5"}},'
32     {"column": "S_recovered", "DataType": "int", "Properties": {"Ordinal": "6"}},'
33     {"column": "S_recovered_change", "DataType": "int", "Properties": {"Ordinal": "7"}},'
34     {"column": "latitude", "DataType": "string", "Properties": {"Ordinal": "8"}},'
35     {"column": "longitude", "DataType": "string", "Properties": {"Ordinal": "9"}},'
```

Activity 3 – Data Factory (ADF)

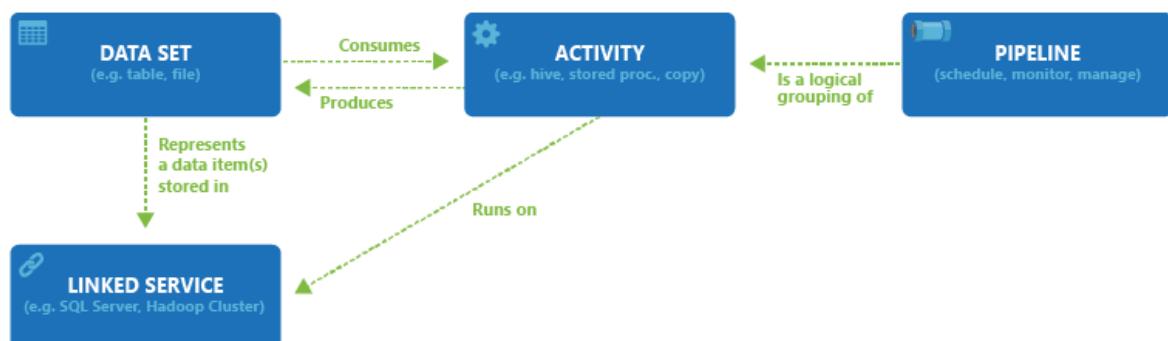
Overview

Data Factory provides a single hybrid data integration service for all skill levels. It is a cloud-based data integration service that allows you to create data-driven workflows in the cloud that orchestrate and automate data movement and data transformation.

An explanation of Azure Data Factory (ADF) is here [[link](#)].

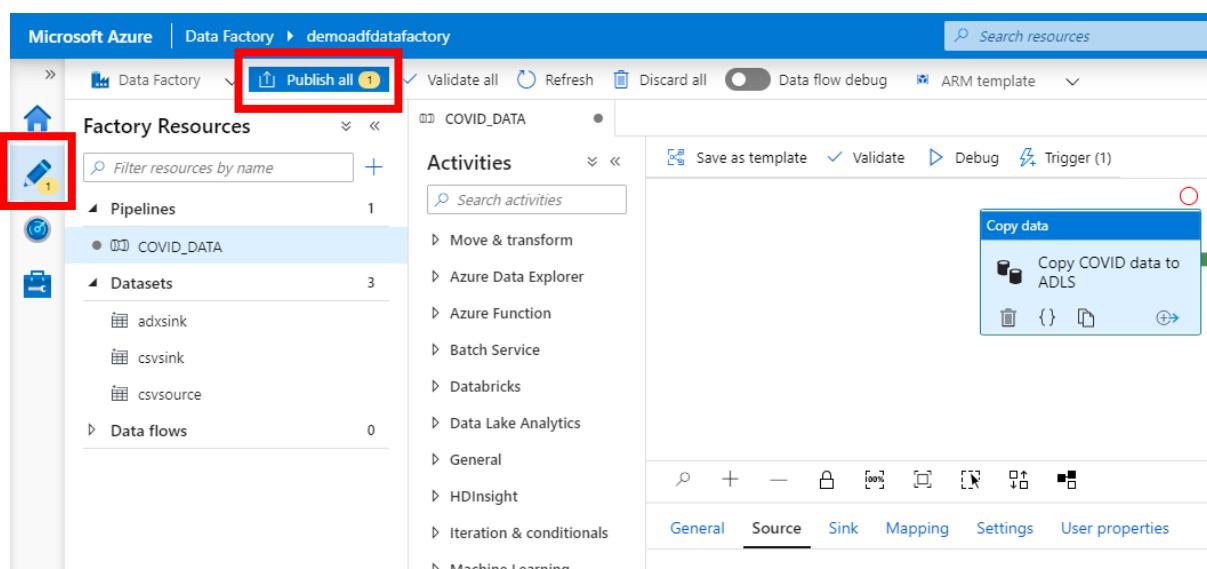
For an explanation of Azure Data Factory (ADF) terminology used within the next sections [Activities, Datasets, Pipeline, Linked Service), please see this link:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services>



Publishing in ADF

ADF runs as a service. When working with Pipelines, Datasets, Activities and all objects within ADF, you need to ‘Publish’ (or save) your changes. You can see when ADF requires that changes need to be published(saved) as a yellow count of unsaved changes is displayed on the ADF resources blade, as shown in the image below:



Through the course of the creating the ADF pipeline, remember to ‘Publish’ all changes.

As part of this activity it may be necessary to also refer to ‘[Activity 4 – Security & Access](#)’.

COV_19 Analytics with Azure Data Explorer (ADX)

Create the Data Factory

As per the diagram “Azure data Factory-Security.pdf” we are creating an Azure Data Factory (ADF). The ADF activity will be triggered by a schedule. The ADF pipeline will move the COVID-19 data file to the **demodatalakeadls** and deposit it in the **/landingzone/..** hierarchical structure based on the Year/Month/Day of the file pull.

1. Within the ADX resource Group, select **+ Add or Create Resource**

The screenshot shows the Azure Resource Groups blade. On the left, there is a list of resource groups: Microsoft, ADX (selected and highlighted with a red box), Azure-Migrations, Customer-Demo, Customer-Demo-Legacy, and Databricks. On the right, the details for the ADX resource group are displayed, including its subscription information (Microsoft Azure Internal Consumption, Subscription ID: 7795b573-5882-4d6c-9b18-6b58d2ab218b), tags (Click here to add tags), and various management options like Overview, Activity log, Access control (IAM), Tags, Events, and Settings. A search bar at the top is also visible.

2. Type **data factory** in the search window (see below) and then select **Data Factory** from the results displayed. The following blade should be displayed:

The screenshot shows the Data Factory blade. At the top, the navigation path is Home > ADX > New > Data Factory. Below this, there is a section titled "Data Factory" with a Microsoft logo and a "Create" button. A description of Azure Data Factory follows, highlighting its features such as composing data storage, movement, and processing services into data flow pipelines, enhanced HDInsight integration, scheduling data pipelines, new data connectors, integration with Azure Machine Learning and Azure Batch, globally deployed data movement, and Visual Studio plug-in support. At the bottom, there is a "Useful Links" section with links to Documentation, Service overview, and Pricing details.

3. Select **Create**, the **New data factory** blade will appear:

COV_19 Analytics with Azure Data Explorer (ADX)

Home > ADX > New > Data Factory > New data factory

New data factory

Name *
adxdatafactoryservice

Version ⓘ
V2

Subscription *
Microsoft Azure Internal Consumption

Resource Group *
ADX
[Create new](#)

Location * ⓘ
West Europe

Enable GIT ⓘ

[Create](#)

Complete the blade by entering the details as below:

Property	Description	Required
Name	Enter demoadfdatafactory	Yes
Version	Enter V2	Yes
Subscription	The value must be set to the subscription the ADX resource group is created in.	Yes
Resource Group	The name of the workshop resource group you have chosen.	Yes
Region	Select your chosen region to match the resource group <i>ADX</i> you have created.	Yes
Enable GIT	Uncheck this box. GIT integration allows you to utilise in a DevOps CI/CD environment. This integration is beyond the scope of this workshop, so will not be configured. Further details can be found here if this is of interest. https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment	Yes

NOTE: *GIT integration allows you to utilise in a DevOps CI/CD environment. This integration is beyond the scope of this workshop, so will not be configured. Further details can be found here - <https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment>.*

COV_19 Analytics with Azure Data Explorer (ADX)

4. Select **Create** to build the ADF. After a short wait you will be notified that the Azure Data Factory has been created. You can go to the resource to review its setup or select **Resource Groups**, then select **your resource group**, to see the **demoadfdatafactory** resource you have just created:

The screenshot shows the Azure Resource Groups blade. The left sidebar shows 'DataDemo' as the selected resource group. The main area displays a table of resources. One row for 'demoadfdatafactory' is highlighted with a red box. The table columns include Name, Type, Location, and Storage account.

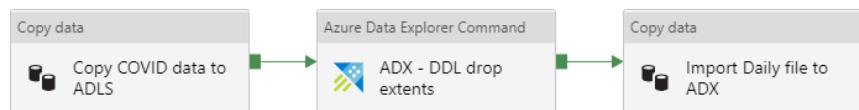
Name	Type	Location	Storage account
demoadfdatafactory	Data factory (V2)	UK South	
demoadiclusteruk	Azure Data Explorer Cluster	UK South	
demodatalakeads	Storage account	UK South	

Create ADF Pipeline

We will now create the ADF pipeline. The pipeline performs the following activities:

- Connects to the public COVID-19 server and ingests(pulls) the daily (csv) dataset to Azure
- Saves the COVID-19 data to a parameterised data (yyyy/MM/DD) driven hierarchy
- Loads the daily file into ADX

The steps above correspond to the image below:



1. Select the ADF created earlier:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the Azure Resource Group 'DataDemo'. On the left, there's a navigation menu with options like Overview, Activity log, Access control (IAM), Tags, Events, Quickstart, Deployments, Policies, Properties, and a '...' button. The main content area displays details for a 'Subscription (change)' resource. It includes fields for Subscription ID (7795b573-5882-4d6c-9b18-6b58d2ab218b), Tags (change), and a 'Click here to add tags'. Below this is a table listing resources: 'demoadfdatafactory' (Data factory (V2)), 'demoadxclusteruk' (Azure Data Explorer Cluster), and 'demodatalakeadls' (Storage account). A search bar at the top allows filtering by name, type, and location.

2. Select Author & Monitor:

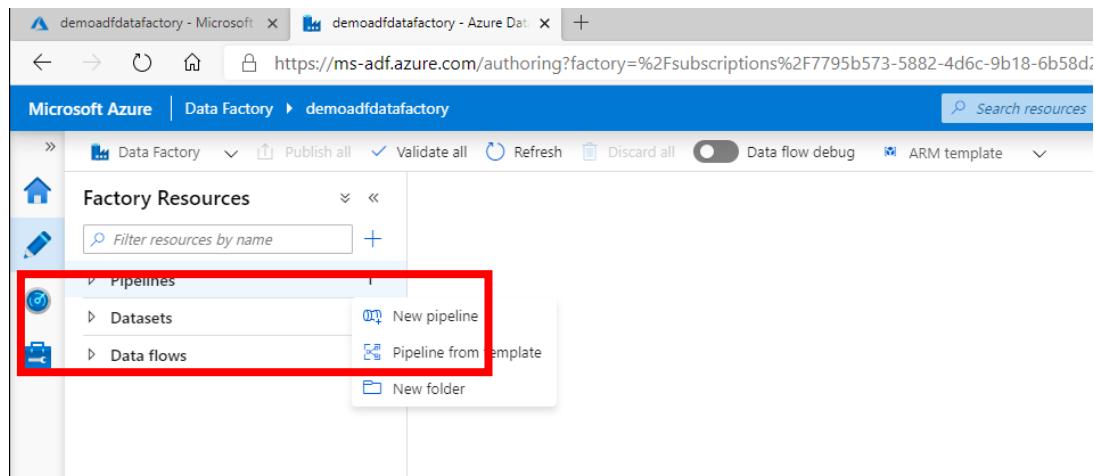
The screenshot shows the 'demoadfdatafactory' Data Factory (V2) overview page. The left sidebar includes links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings (Locks), General, and Properties. The main content area shows basic information: Status (Succeeded), Location (UK South), Subscription (change) (Microsoft Azure Internal Consumption), and Subscription ID (7795b573-5882-4d6c-9b18-6b58d2ab218b). At the bottom right, there are two buttons: 'Documentation' and 'Author & Monitor'. The 'Author & Monitor' button is highlighted with a red box.

3. Select Author icon:

The screenshot shows the Microsoft Azure Data Factory 'demoadfdatafactory' interface. On the left, there's a vertical toolbar with icons for Home, Pipelines, Data Flows, and Temp Tables. The main area has a blue background with the text 'Let's get started' and three circular icons: 'Create pipeline' (a blue cylinder with a plus sign), 'Create data flow' (two green cylinders with an arrow), and 'Create pipe from temp' (a network of nodes). The 'Create pipeline' button is highlighted with a red box.

4. Select Pipelines, then '...', then 'New Pipeline':

COV_19 Analytics with Azure Data Explorer (ADX)



COV_19 Analytics with Azure Data Explorer (ADX)

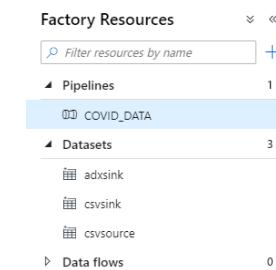
Create Pipeline Linked Services

ADF ‘Httplinkedservice’ linked service

This is the ADF connection to the COVID-19 public [http hosted] datasets.

The JSON file for this ADF activity is in the /download folder “Httplinkedservice.json”.

1. Select **Connections**:



2. Select '**+ New**

Complete the blade as per the screenshot below. Select ‘**Test Connection**’ to confirm the Linked Service is working correctly. Select ‘**Apply**’.

Edit linked service (HTTP)

Name *
Httplinkedservice

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Base URL *
https://pandemicdatalake.blob.core.windows.net

Server Certificate Validation
 Enable Disable

Authentication type *
Anonymous

Annotations
+ New

ADF 'adlslinkedservice' linked service

This is the ADF connection to ADLS. Note, on this blade, the 'Managed Identity details (name & object ID) are shown. These should be noted, as they will be required in the section '**'Security & Access'**'.

The JSON file for this is in the /download folder "adlslinkedservice.json".

1. Select '+ New'

Complete the blade as per the screenshot below. Select '**Test Connection**' to confirm the Linked Service is working correctly. Select '**Apply**':

COV_19 Analytics with Azure Data Explorer (ADX)

Edit linked service (Azure Data Lake Storage Gen2)

Info If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to test connection. Please make sure your self-hosted integration runtime is higher than version 4.0 if connecting via self-hosted integration runtime.

Name *
adlslinkedservice

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Authentication method
Managed Identity

Account selection method
 From Azure subscription Enter manually

URL *
https://demodatalakeadls.dfs.core.windows.net

Managed identity name: **demoadfdataproxy**
Managed identity object ID: **2ca7705e-9e33-4735-9f19-0ef48b5c7978**
Grant Data Factory service managed identity access to your Azure Data Lake Storage Gen2.
[Learn more](#)

Test connection
 To linked service To file path

Annotations
[New](#)

[Advanced](#)

Connection successful

Test connection

COV_19 Analytics with Azure Data Explorer (ADX)

ADF 'demoadxsink' linked service

This is the ADF connection to the ADX cluster. Note, the Service Principal details are required from the 'Security & Access' section.

The JSON file for this is in the /download folder "demoadxsink.json".

1. Select '+ New'

Complete the blade as per the screenshot below. Select 'Test Connection' to confirm the Linked Service is working correctly. Select 'Apply':

Edit linked service (Azure Data Explorer (Kusto))

Name *
demoadxsink

Description
|

Connect via integration runtime *
AutoResolveIntegrationRuntime

Account selection method
 From Azure subscription Enter manually

Endpoint *
https://demoadxclusteruk.eksouth.kusto.windows.net

Tenant *

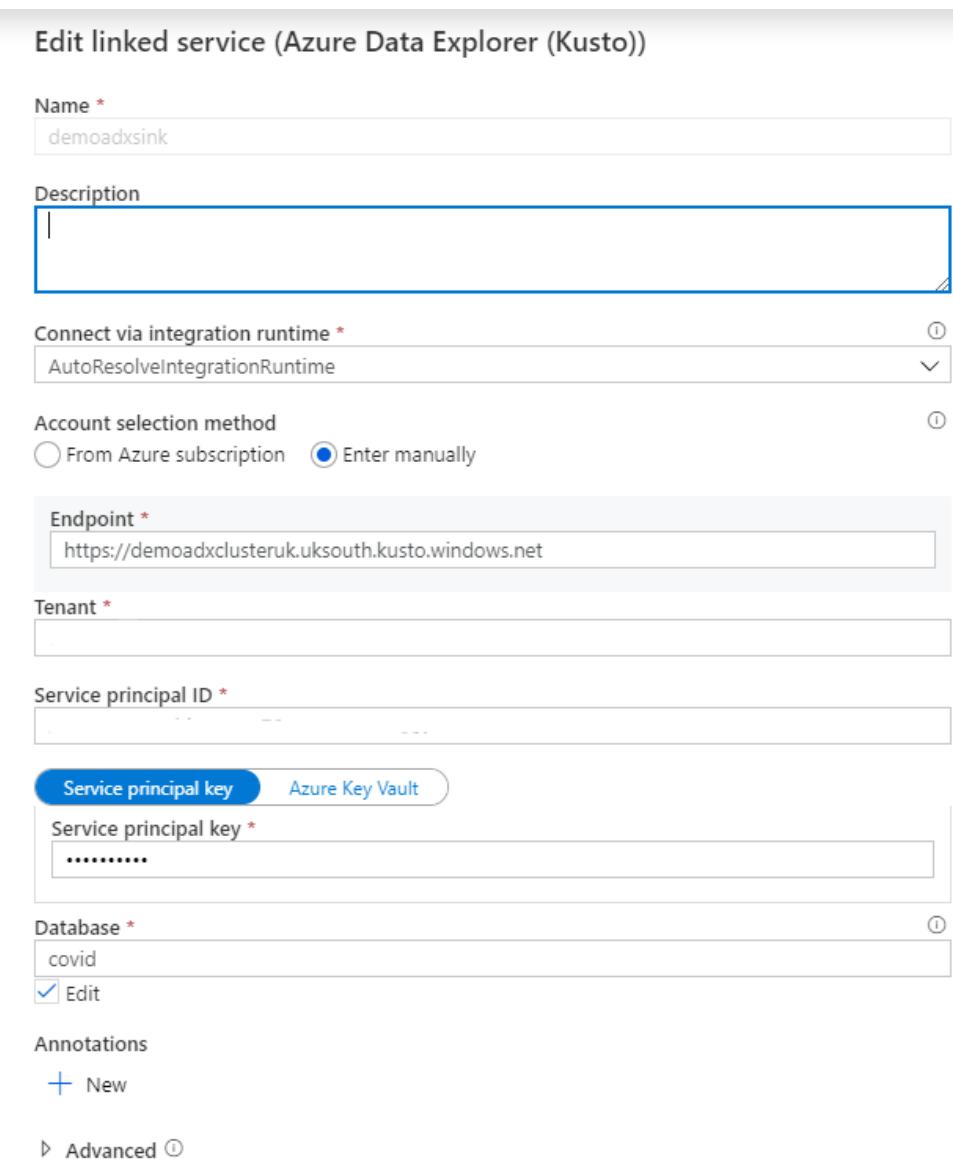
Service principal ID *

Service principal key Azure Key Vault
Service principal key *
.....

Database *
covid
 Edit

Annotations
+ New

Advanced (1)



Note: the 'Tenant' and 'Service Principal ID' fields have been blanked out. You need to substitute your Service Principal values here.

COV_19 Analytics with Azure Data Explorer (ADX)

Create Pipeline Datasets

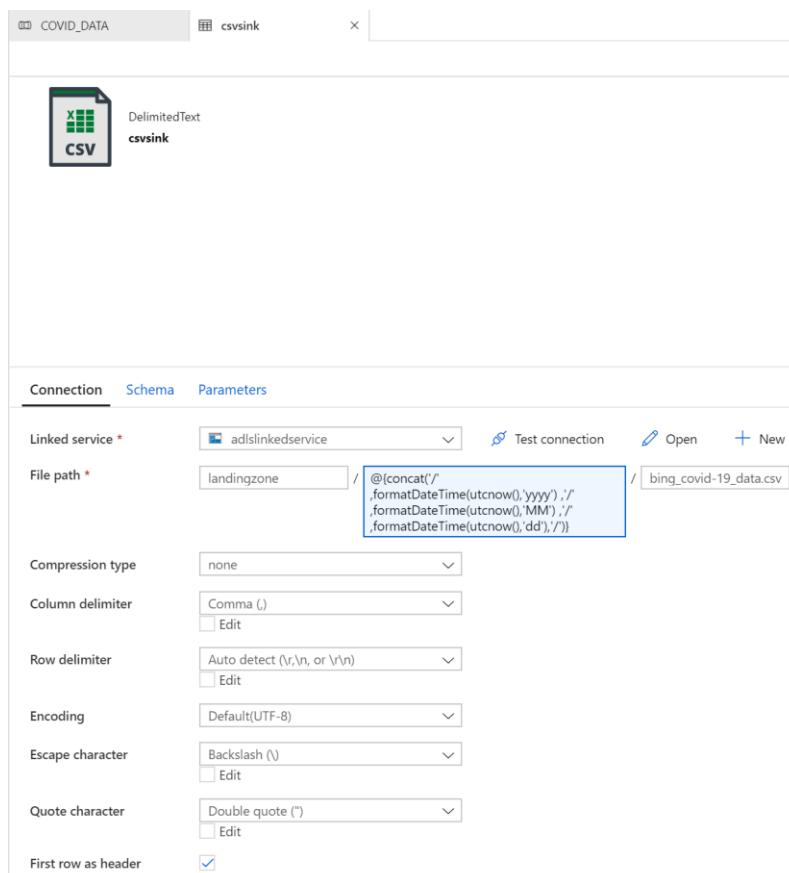
ADF 'csvsink' dataset

This dataset uses the '**adlslinkedservice**' to write the COVID-19 dataset into a date parameterised folder structure: landingzone/YYYY/MM/DD/.

The JSON file for this is in the /download folder "csvsink.json".

1. Complete the **Connection / Schema** tabs as per the following images.
2. **Connection tab:**

Select the **adlslinkedservice** Linked Service and then '**Test Connection**' to verify.



Note: for clarity, field values are:

- **File Path** is:

```
@{concat('/', formatDateTime(utcnow(), 'yyyy'), '/',
formatDateTime(utcnow(), 'MM'), '/', formatDateTime(utcnow(), 'dd'), '/')}
```

3. Select '**Preview Data**' to check the dataset is working correctly.
4. **Schema tab:**

Select '**Import Schema**' to verify the datafile/dataset (+linked service)

COV_19 Analytics with Azure Data Explorer (ADX)

Column name	Type
id	String
updated	String
confirmed	String
confirmed_change	String
deaths	String
deaths_change	String
recovered	String
recovered_change	String
latitude	String
longitude	String
iso2	String
iso3	String
country_region	String
admin_region_1	String
iso_subdivision	String
admin_region_2	String
load_time	String

COV_19 Analytics with Azure Data Explorer (ADX)

ADF 'csvsource' dataset

This dataset uses the '**Httplinkedservice**' to read the COVID-19 dataset from the public [<https://pandemicdatalake.blob.core.windows.net>] host server.

The JSON file for this is in the /download folder "csvsource.json".

1. Complete the **Connection / Schema** tabs as per the following images.
2. **Connection tab:**

Select the **Httplinkedservice** Linked Service and then '**Test Connection**' to verify.

The screenshot shows the 'Connection' tab of the 'csvsource' dataset in the Azure Data Explorer. The top navigation bar shows 'COVID_DATA' and 'csvsource'. The main area displays a CSV icon and the name 'csvsource'. Below this is a 'DelimitedText' section. The 'Connection' tab is selected, showing the following configuration:

Linked service *	Httplinkedservice	Test connection
Base URL	https://pandemicdatalake.blob.core.window	
Relative URL	/public/curated/covid-19/bing_covid-19_d	
Compression type	none	
Column delimiter	Comma (,)	<input type="checkbox"/> Edit
Row delimiter	Auto detect (\r,\n, or \r\n)	<input type="checkbox"/> Edit
Encoding	Default(UTF-8)	
Escape character	Backslash (\)	<input type="checkbox"/> Edit
Quote character	Double quote (")	<input type="checkbox"/> Edit
First row as header	<input checked="" type="checkbox"/>	
Null value		

Note: for clarity, field values are:

- **Base URL:** <https://pandemicdatalake.blob.core.windows.net>
- **Relative URL:** /public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.csv

3. **Schema tab:**

Select '**Import Schema**' to verify the datafile/dataset (+linked service)

COV_19 Analytics with Azure Data Explorer (ADX)

Column name	Type
id	String
updated	String
confirmed	String
confirmed_change	String
deaths	String
deaths_change	String
recovered	String
recovered_change	String
latitude	String
longitude	String
iso2	String
iso3	String
country_region	String
admin_region_1	String
iso_subdivision	String
admin_region_2	String
load_time	String

COV_19 Analytics with Azure Data Explorer (ADX)

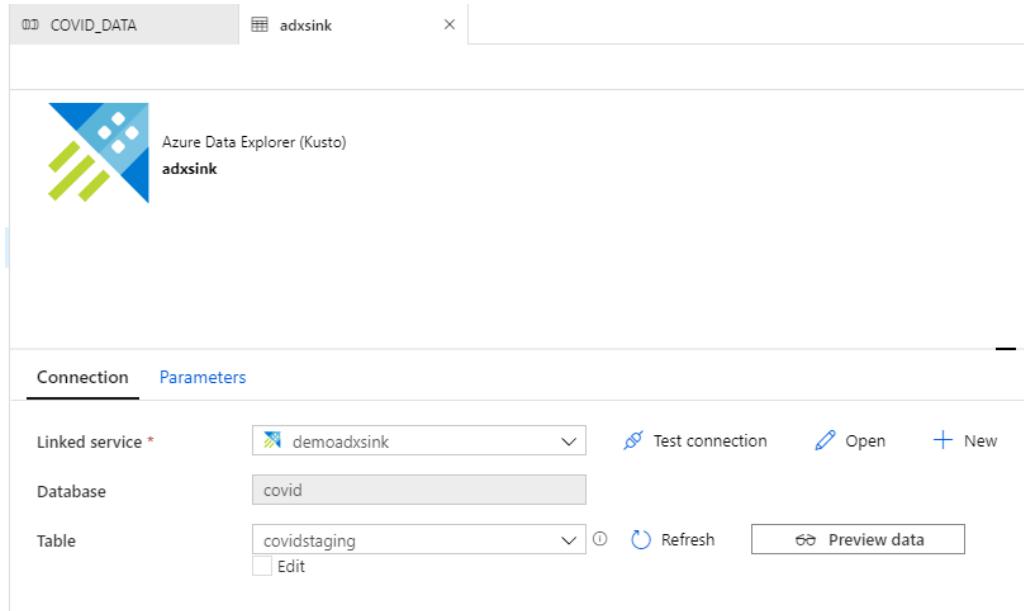
ADF 'adxsink' dataset

This dataset uses the '**demoadxsink**' to connect to the ADX cluster you have created.

The JSON file for this is in the /download folder “demoadxsink.json”.

1. Complete the tab as per the following images.
2. **Connection** tab:

Select the **demoadxsink** Linked Service and then '**Test Connection**' to verify.



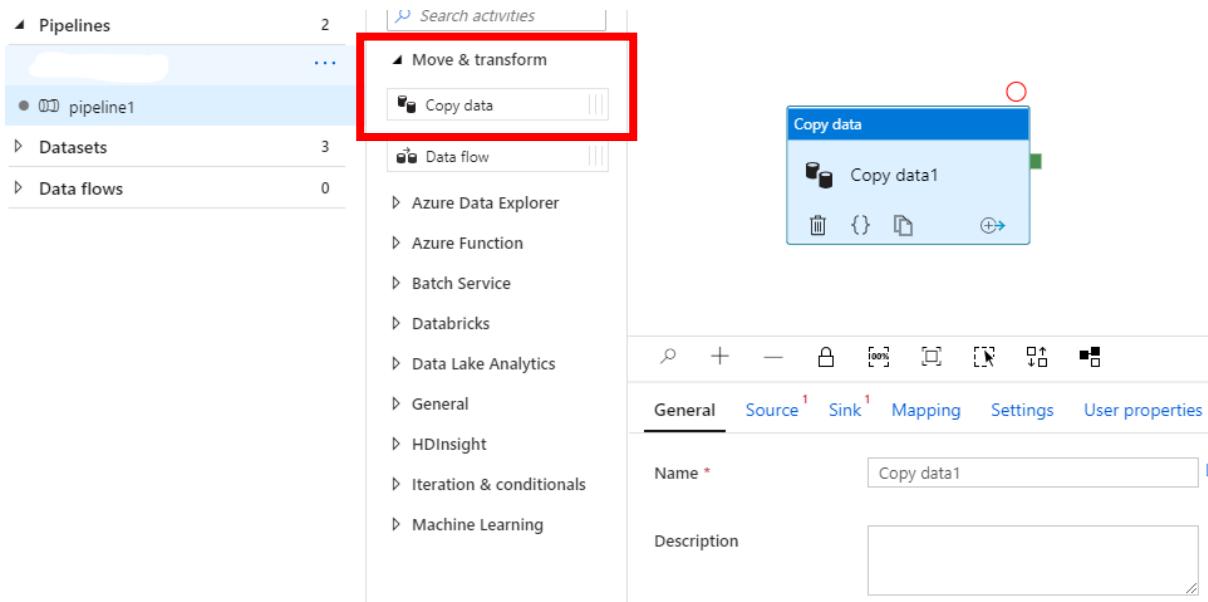
Create Pipeline Activities

ADF ‘Copy COVID data to ADLS’ activity

This activity connects to the COVID-19 public server and ingests the daily dataset to Azure ADLS.

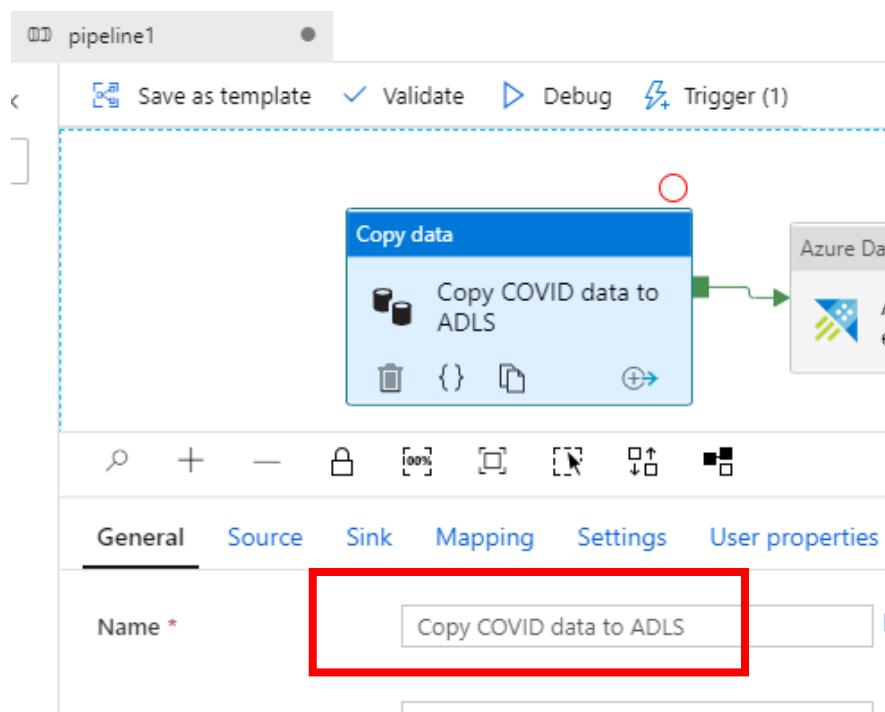
The JSON file for this ADF activity is in the /download folder “ADF - Copy COVID data to ADLS.json”.

1. Drag **Copy data** activity, from the **Move & transform** collection, and drag to the pipeline canvas:



2. Complete the **General / Source / Sink / Mapping** tabs as per the following images.

3. **General tab:**



COV_19 Analytics with Azure Data Explorer (ADX)

4. Source tab:

The screenshot shows the Azure Data Factory pipeline editor with a pipeline named "pipeline1". A "Copy data" activity is selected, which has a green arrow pointing to an "Azure Data" sink icon. The "Source" tab is active. Several fields are highlighted with red boxes: "Source dataset" (set to "csvsource"), "Request method" (set to "GET"), and "Max concurrent connections" (set to "8").

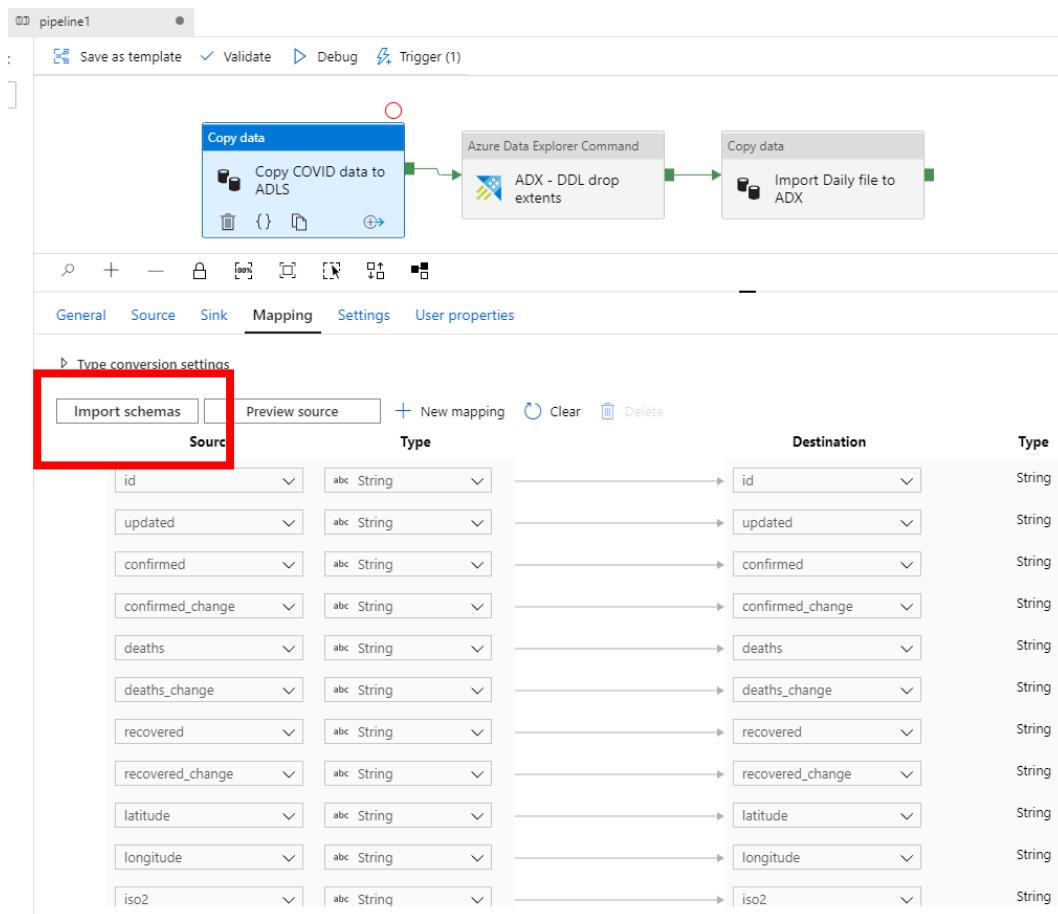
5. Sink tab:

The screenshot shows the Azure Data Factory pipeline editor with the same pipeline and activity setup as the previous screenshot. The "Sink" tab is active. Fields highlighted with red boxes include "Sink dataset" (set to "csvsink"), "Copy behavior" (set to "None"), and "Max concurrent connections" (set to "4").

6. Mapping tab

COV_19 Analytics with Azure Data Explorer (ADX)

Select 'Import Schemas' and then ensure all of the column definitions match:



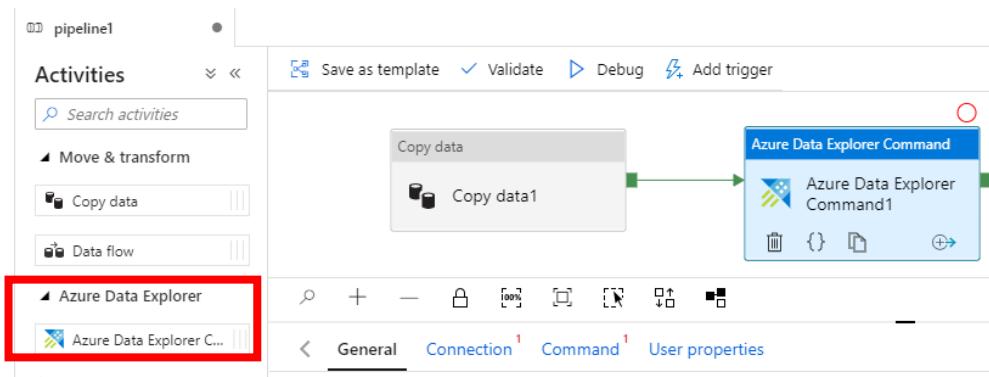
COV_19 Analytics with Azure Data Explorer (ADX)

ADF ‘ADX – DDL drop extents’ activity

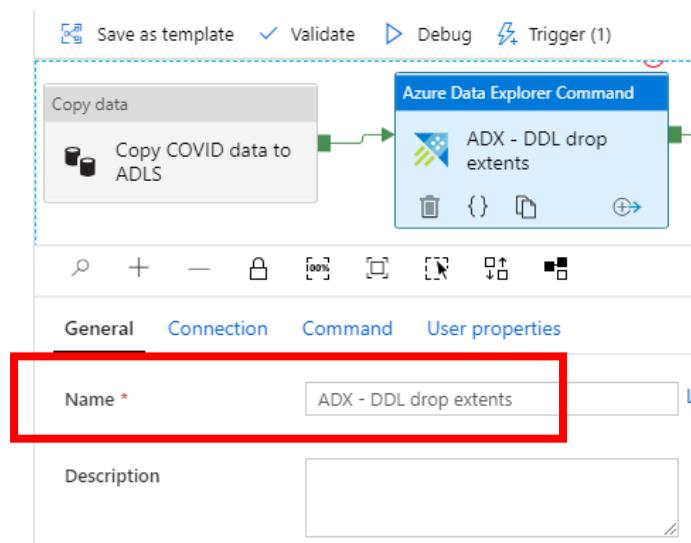
This activity runs native commands on the ADX cluster. For Workshop 1 it simply clears down the **covidstaging** table ready for the next data load using a KQL command.

The JSON file for this ADF activity is in the /download folder “**ADF - ADX - DDL drop extents.json**”.

1. Drag **Azure Data Explorer Command**, from the **Azure Data Explorer** collection, to the pipeline canvas. Connect the two activities as shown:

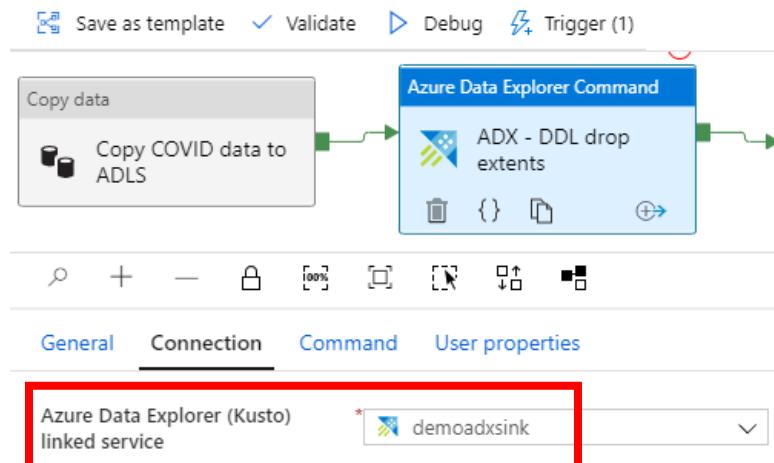


2. Complete the **General / Connection / Command** tabs as per the following images.
3. **General tab:**

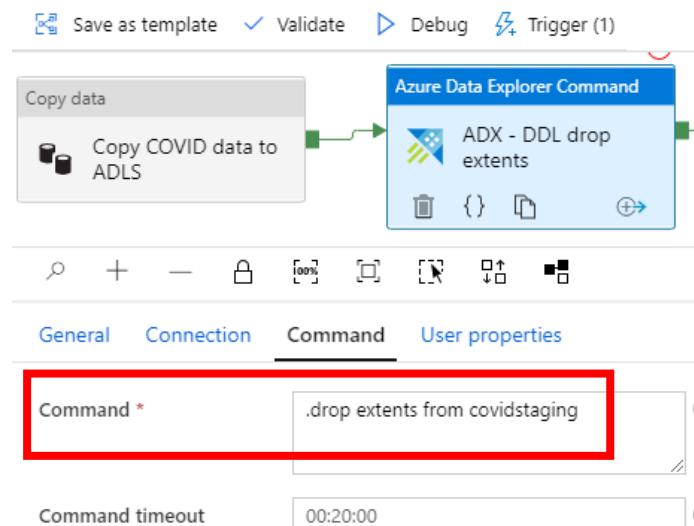


4. **Connection tab:**

COV_19 Analytics with Azure Data Explorer (ADX)



5. Command tab:



The KQL for the Command is:

.drop extents from covidstaging

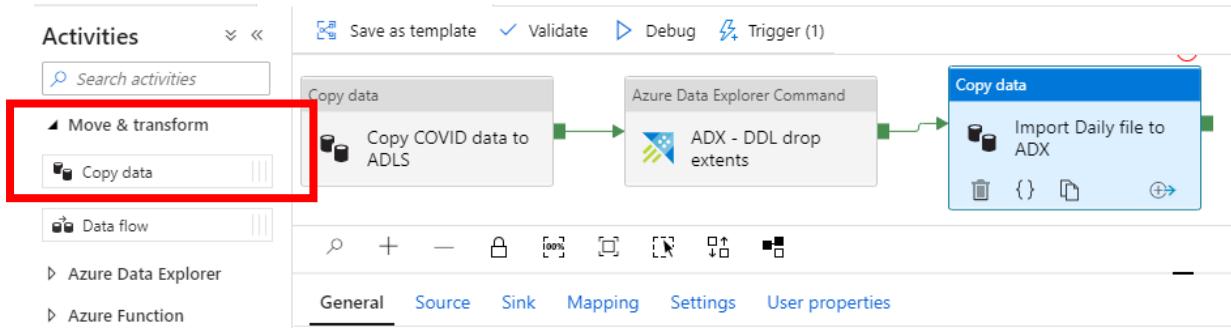
COV_19 Analytics with Azure Data Explorer (ADX)

ADF 'Import Daily File to ADX' activity

This activity loads the daily COVID-19 datafile into ADX.

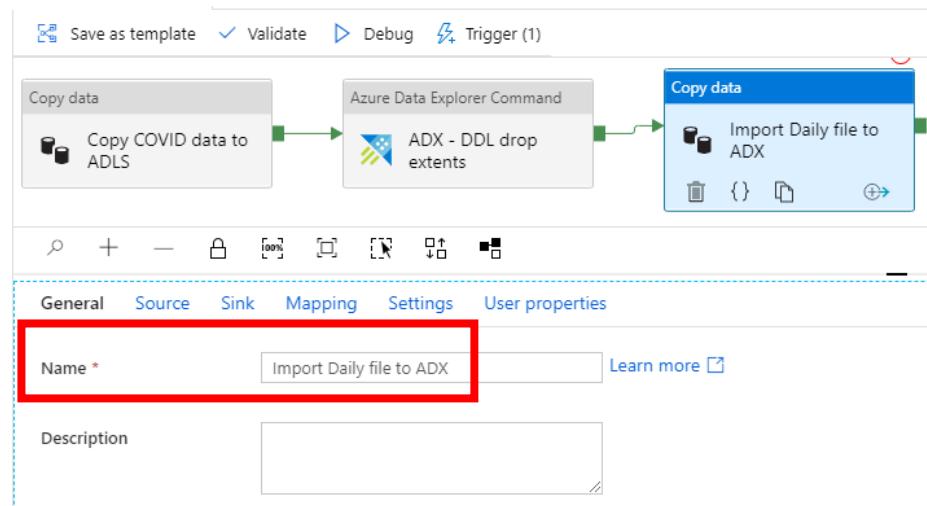
The JSON file for this ADF activity is in the /download folder “ADF - Import Daily file to ADX.json”.

1. Drag **Copy data** activity, from the **Move & transform** collection, and drag to the pipeline canvas. Connect the activities as shown:



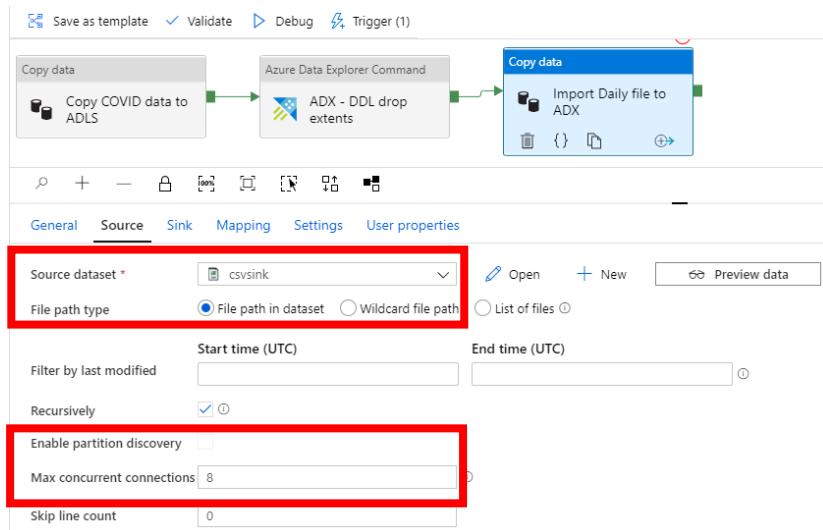
2. Complete the **General / Source / Sink / Mapping** tabs as per the following images.

3. **General tab:**

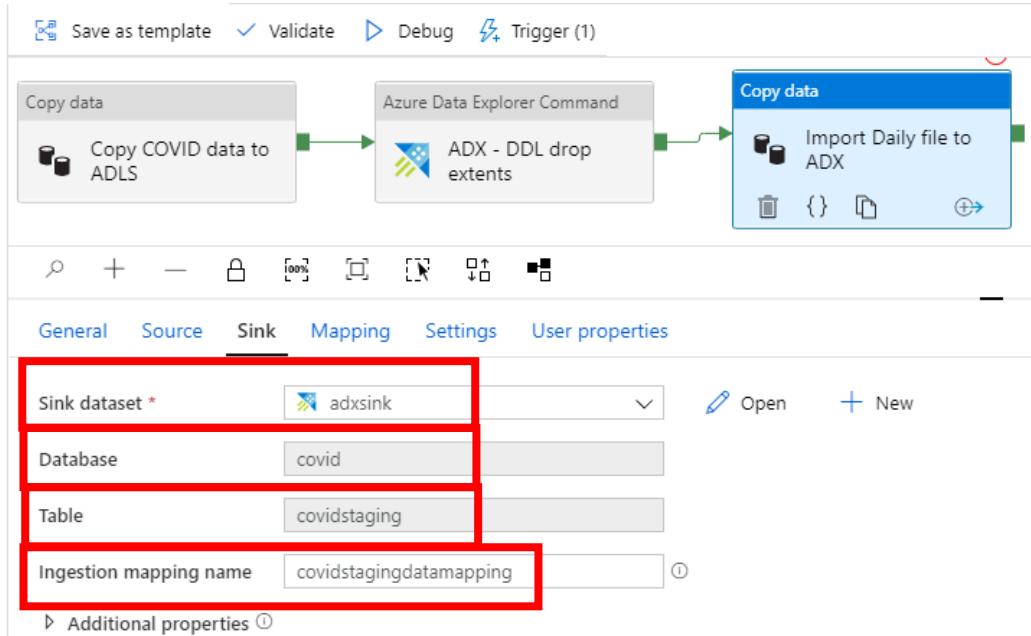


4. **Source tab:**

COV_19 Analytics with Azure Data Explorer (ADX)



5. Sink tab:



COV_19 Analytics with Azure Data Explorer (ADX)

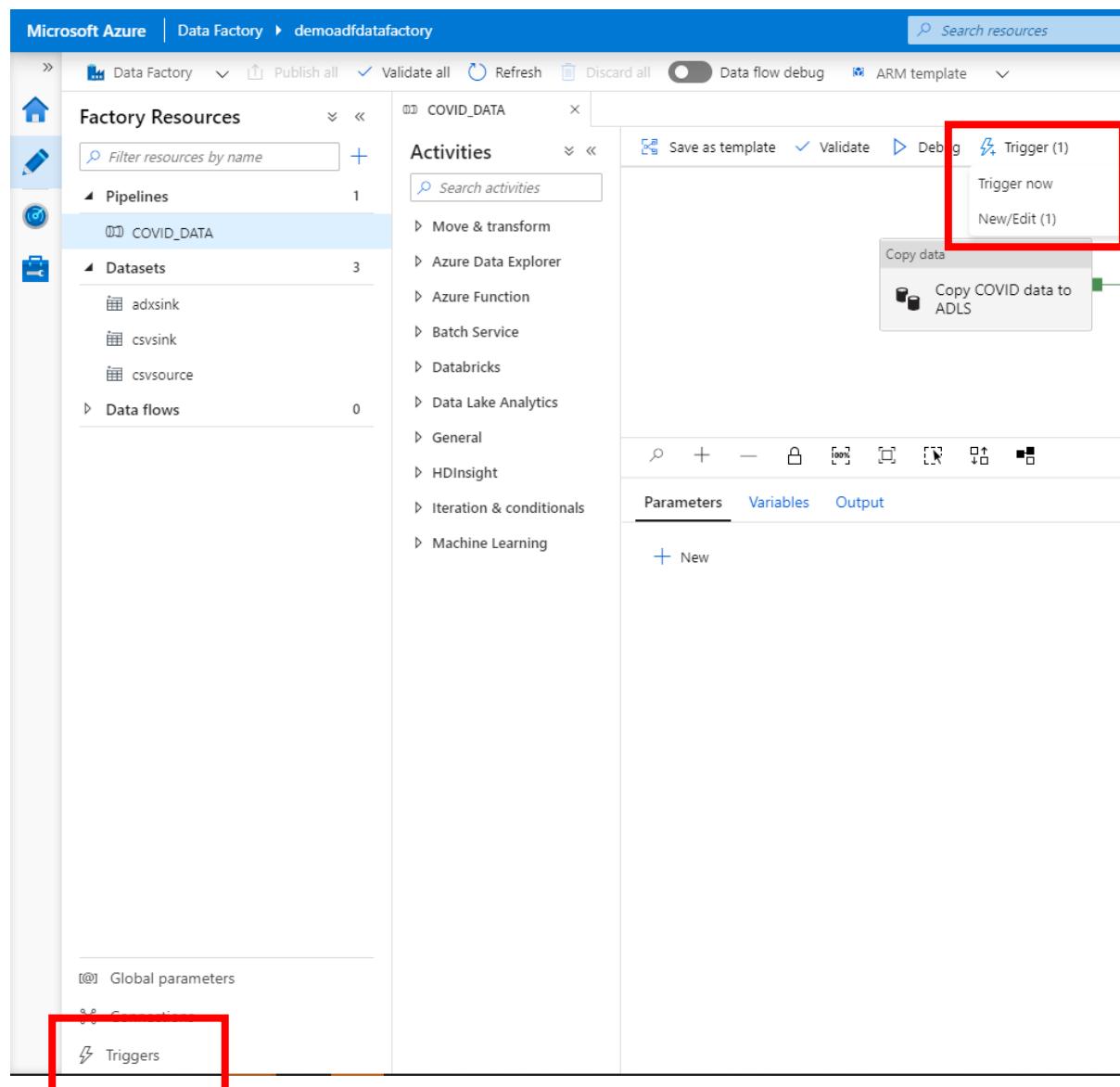
Create Pipeline Trigger

Triggers are how ADF Pipelines are executed. Triggers come in three forms:

- Event-based
- Scheduled
- Tumbling Window

In this workshop we will be using a scheduled trigger: <https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger>

In the ADF UI a pipeline can be triggered manually, or a created from the selections as seen in the image below:



We will be creating a scheduled trigger 'COVID-DATA_TRIGGER' to run on a daily basis to execute the 'COVID_DATA' pipeline.

COV_19 Analytics with Azure Data Explorer (ADX)

The JSON file for this ADF trigger is in the /download folder “COVID-DATA_TRIGGER.json”.

- From the COVID_DATA Pipeline canvas, select ‘Trigger’, then ‘New/Edit’ or ‘+ New’, depending upon the access path. Complete the blade as shown in the image below:

New trigger

Name *

Description

Type *

Start Date (UTC) *

Recurrence *
Every Day(s)

Advanced recurrence options

Execute at these times
Hours (UTC)
Minutes (UTC)
Schedule execution times (UTC)
20:00

End *
 No End On Date

End On (UTC) *

Annotations

Activated * Yes No

- Select ‘OK’
- Triggers can then be seen in the ‘Triggers’ blade:

Microsoft Azure | Data Factory > demoadffactory

Connections Triggers

To execute a pipeline set the trigger. Triggers represent a unit of processing that determines when a pipeline execution needs to be kicked off.

+ New

Showing 1 - 1 of 1 items

NAME ↑↓	TYPE ↑↓	STATUS ↑↓
COVID_DATA_TRIGGER	Schedule	Started

COV_19 Analytics with Azure Data Explorer (ADX)

Activity 4 – Security & Access

During the previous activities, and to complete the ADF pipeline build, we need to grant access to Azure services (ADX, ADF, ADLS). This is achieved via Azure Active Directory objects (App Registration, Service Principal, Managed Identity).

Terminology Explained

To assist with understanding the security principals within this workshop, *the document “Azure data Factory-Security.pdf” can be found in the /downloads folder.*

Application Registration

To delegate Identity and Access Management functions to Azure AD, an application must be registered with an Azure AD tenant. Refer to “Azure data Factory-Security.pdf” in /downloads.

<https://docs.microsoft.com/en-us/azure/active-directory/develop/howto-create-service-principal-portal>

Managed Identities

<https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview>

Service Principal

To access resources that are secured by an Azure AD tenant, the entity that requires access must be represented by a [security principal](#). This is true for both users (user principal) and applications (service principal).

<https://docs.microsoft.com/en-us/azure/active-directory/develop/app-objects-and-service-principals#:~:text=A%20service%20principal%20must%20be%20created%20in%20each,Web%20application%2FAPI%20also%20has%20a%20service%20principal%20>

ADF – Security Configuration (Linked Service)

Httplinkedservice

This Linked Service, being a public hosted server, uses anonymous authentication. The configuration can be seen here:

COV_19 Analytics with Azure Data Explorer (ADX)

Edit linked service (HTTP)

Name *
Httplinkedservice

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Base URL *
https://pandemicdatalake.blob.core.windows.net

Server Certificate Validation
 Enable Disable

Authentication type *
Anonymous

Annotations

Adlslinkedservice

This Linked Service uses the ADF generated Managed Identity to authenticate to ADLS.
Configuration is shown here:

Edit linked service (Azure Data Lake Storage Gen2)

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to test connection. Please make sure your self-hosted integration runtime is higher than version 4.0 if connecting via self-hosted integration runtime.

Name *
adlslinkedservice

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Authentication method
Managed Identity

Account selection method
 From Azure subscription Enter manually

URL *
https://demodatalakeadls.dfs.core.windows.net

Managed identity name: **demoadfdatafactory**
Managed identity object ID: **2ca7705e-9e33-4735-9f19-0ef48b5c7978**
Grant Data Factory service managed identity access to your Azure Data Lake Storage Gen2.
[Learn more](#)

Test connection
 To linked service To file path

Annotations

demoadxsink

This Service uses the Service Principal **datademosecurity** to authenticate, as shown here:

COV_19 Analytics with Azure Data Explorer (ADX)

Edit linked service (Azure Data Explorer (Kusto))

Name *
demoadxsink

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Account selection method
 From Azure subscription Enter manually

Endpoint *
https://demoadxclusteruk.uksouth.kusto.windows.net

Tenant *
72f\$

Service principal ID *
5e5ac...

Service principal key Azure Key Vault
Service principal key *

Database *
covid

Edit

Annotations

ADLS – Security Configuration

We will grant access to ADF (via **demoadfdatafactory**) and ADX (via **datademosecurity**). Security is managed via the ‘Access control (IAM)’ option off the **demodatalakeadls** blade as shown below:

Dashboard >

demodatalakeadls | Access control (IAM)
Storage account

Search (Ctrl+ /) + Add Download role assignments (preview) Edit columns Refresh Remove Got feedback?

Overview Activity log Access control (IAM) Data transfer Events Storage Explorer (preview)

Check access Role assignments Deny assignments Classic administrators Roles

Access control
Review the level of access a user, group, service principal, or managed identity has to this resource. [Learn more](#)

Find Search by name or email address

Add a role assignment
Grant access to resources at this scope by assigning one to a user, group, service principal, or managed identity.
Add [Learn more](#)

View role assignments
View the users, groups, service principals and managed identities that have role assignments granting them access at this scope.
View [Learn more](#)

Permissions [that need to be] granted are shown here:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the 'Access control (IAM)' blade for the 'demodatalakeadls' storage account. The 'Role assignments' tab is selected. It displays 6 role assignments across 4 items (2 Service Principals, 2 Managed Identities). The table includes columns for Name, Type, Role, Scope, and Group by. A red box highlights the 'Storage Blob Data Contributor' section, which lists two entries: 'datademosecurity' (App) and 'demoadfdatafactory' (Data Factory). Both have 'Storage Blob Data Contributor' assigned at 'This resource' scope.

Name	Type	Role	Scope	Group by
CSEORreader	App	Reader	All scopes	Management group
demoadfdatafactory Reader	App	Storage Blob Data Contributor	This resource	Role (selected)
datademosecurity	App	Storage Blob Data Contributor	This resource	Storage Blob Data Contributor
demoadfdatafactory	Data Factory	Storage Blob Data Contributor	This resource	Storage Blob Data Contributor

ADX – Security Configuration (cluster)

We will grant access to ADF (via **demoadfdatafactory**) and ADX (via **datademosecurity**). Security is managed via the ‘Access control (IAM)’ option off the **demoadxclusteruk** blade as shown below:

The screenshot shows the 'Access control (IAM)' blade for the 'demoadxclusteruk' Azure AD Data Explorer Cluster. The 'Check access' tab is selected. On the left, the 'Activity log' and 'Access control (IAM)' options are highlighted with a red box. On the right, there are two cards: 'Add a role assignment' (with a checked checkbox icon) and 'View role assignments' (with a user icon).

Permissions granted ('Owner' is used as an example, 'Contributor' could also be selected) at the ADX cluster level are shown here:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the 'Access control (IAM)' blade for the 'demoadxclusteruk' cluster. The 'Role assignments' tab is selected. A red box highlights the 'Owner' section, which lists two entries: 'datademosecurity' (App, Owner) and 'demoadffactory' (Data Factory, Owner). Below this is the 'Reader' section, which lists 'CSEORreader' (App, Reader) and 'Geneva Analytics Reader' (App, Reader).

Name	Type	Role
datademosecurity	App	Owner
demoadffactory	Data Factory	Owner
CSEORreader	App	Reader
Geneva Analytics Reader	App	Reader

Note AD access may be required for Azure Data Explorer Dashboard access. This should be granted 'Reader' role.

ADX – Security Configuration (database)

Permissions granted at the ADX database level are shown here:

The screenshot shows the 'Permissions' blade for the 'covid' database. A red box highlights the 'Permissions' section in the left sidebar. The main area shows a table of permissions. A second red box highlights the 'Ingestor' section, which lists 'demoadffactory' (App, Database Ingestor) and 'datademosecurity' (App, Database Ingestor).

Name	Type	Role
Neil .	User	Database Admin
demoadffactory	App	Database Ingestor
datademosecurity	App	Database Ingestor

Note AD access may be required for Azure Data Explorer Dashboard access. This should be granted 'Database Viewer' role.

Further Reading

Azure documentation around the security aspects of the services used within this workshop can be found in the following links:

ADF – [Security Considerations](#)

COV_19 Analytics with Azure Data Explorer (ADX)

<https://docs.microsoft.com/en-us/azure/data-factory/data-movement-security-considerations#:~:text=Data%20Factory%20management%20resources%20are%20built%20on%20Azure,grouping%20of%20activities%20that%20together%20perform%20a%20task.>

ADLS – [Access Control](https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control)

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

ADX - [Roles](https://docs.microsoft.com/en-us/azure/data-explorer/kusto/management/security-roles)

<https://docs.microsoft.com/en-us/azure/data-explorer/kusto/management/security-roles>

ADX – [Managed Identity](https://docs.microsoft.com/en-us/azure/data-explorer/managed-identities?tabs=portal)

<https://docs.microsoft.com/en-us/azure/data-explorer/managed-identities?tabs=portal>

ADX Dashboard – Sharing

It is possible to share your ADX Dashboard. This is achieved via the ‘Share’:

The screenshot shows the Azure Data Explorer interface with a dashboard titled "D_DATA". The top navigation bar includes "Edit", "Share", "Save as copy", and "Export to JSON". A red box highlights the "Share" dropdown menu, which contains "Copy link" and "Manage permissions". Below the dashboard title, there is a date range selector showing "Date : 2020-06-30" and a location selector showing "Country : United Kingdom". The main content area displays a pie chart titled "Top 20 Countries - Infections (date sele...)" showing the distribution of infections by country.

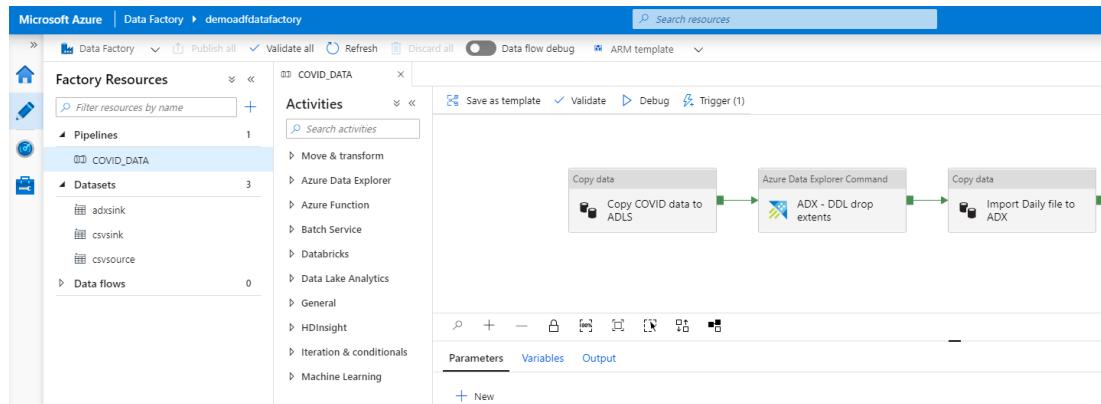
Selecting ‘Manage Permissions’ allows edit or read access to be granted to a given Active Directory resource. Use ‘Add new member’ and ‘Permission’ to grant dashboard access, as shown below:

The screenshot shows the "Dashboard permissions" settings page. It features a "Members (14)" section with a search bar and a table listing members with their names, emails, types, and current permissions. A red box highlights the "Add new members" button and the "Permission" dropdown for the first row, which is set to "Can view". Another red box highlights the "Can edit" permission dropdown for the "Owners (1)" row, which currently shows "User".

Remember ADX database [Reader permission will also need to be granted. See ‘[ADX – Security Configuration \(database\)](#)’ above.

Activity 5 – ADF Pipeline

We now have a completed ADF Pipeline '**COVID_DATA**':



The scheduled trigger '**COVID_DATA_TRIGGER**' will run the ADF '**COVID_DATA**' landing COVID-19 files to ADLS in a date orientated hierarchy (landingzone/YYYY/MM/DD/). Repeat executions of the trigger(pipeline) will [on consecutive days] produce the same results but into landingzone/YYYY/MM/**DD + 1** where + 1 is the next/consecutive date, as shown in the image below [date shown = 6th July 2020]:

NAME	LAST MODIFIED
01	07/07/2020, 21:00:39
02	07/07/2020, 21:00:34
03	07/07/2020, 21:00:46
04	07/07/2020, 21:01:05
05	07/07/2020, 21:00:43
06	07/07/2020, 21:00:55
07	07/07/2020, 21:01:47
08	08/07/2020, 21:00:35
09	09/07/2020, 21:01:06

Start/activate the ADF trigger

To test the ADF pipeline we need to either (a) manually trigger it or (b) run the pipeline through the ADF monitor.

For (a):

COV_19 Analytics with Azure Data Explorer (ADX)

1. Select the **COVID_DATA** pipeline
2. Select 'Trigger' / 'Trigger now' option

For (b):

The screenshot shows the Microsoft Azure Data Factory Pipeline runs page. The left sidebar has options: Dashboards, Pipeline runs (which is selected and highlighted with a red box), Trigger runs, Integration runtimes, and Alerts & metrics. The main area is titled 'Pipeline runs' with a subtitle 'Showing 1 - 1 items'. It lists one run for the 'COVID_DATA' pipeline. The run details are: PIPELINE NAME: COVID_DATA, RUN START: 7/9/20, 9:00:01 PM, DURATION: 00:01:59, TRIGGERED BY: COVID_DATA_TRIGGER, STATUS: Succeeded (green checkmark). There are three icons next to the run entry: a blue 'Rerun' icon, a blue 'Cancel' icon, and a blue 'Edit columns' icon. A red box highlights the 'Pipeline runs' link in the sidebar and the 'Rerun' icon.

3. Select the 'Pipeline runs'
4. Select the **COVID_DATA** pipeline
5. The 'Rerun' icon will appear, select this to execute the pipeline

Monitor ADF Pipeline execution

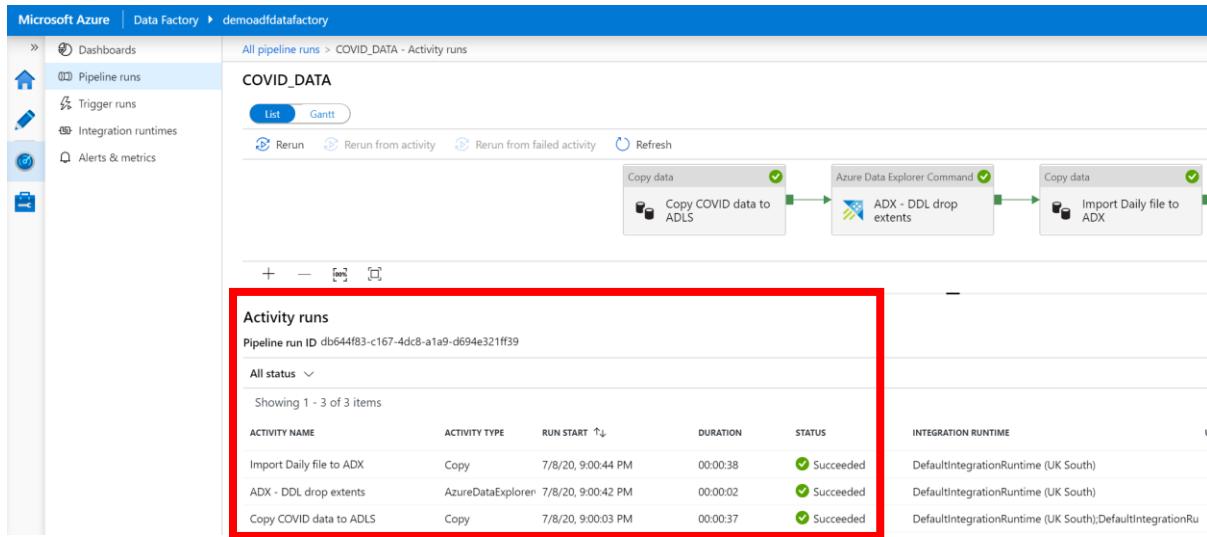
We can monitor the progress (and historical executions) and status of the ADF pipeline:

1. Select **Pipeline runs**
2. Select the date range you want to monitor:

The screenshot shows the Microsoft Azure Data Factory Pipeline runs page. The left sidebar has options: Dashboards, Pipeline runs (selected and highlighted with a red box), Trigger runs, Integration runtimes, and Alerts & metrics. The main area is titled 'Pipeline runs' with a subtitle 'Choose an option'. It shows a calendar for June 2020 with the 29th selected. Below the calendar are 'Start time' and 'End time' fields set to 06/29/2020 and 07/05/2020 respectively, both at 10:00:00 PM. At the bottom are 'OK' and 'Cancel' buttons. A red box highlights the 'Pipeline runs' link in the sidebar and the date range selection area.

COV_19 Analytics with Azure Data Explorer (ADX)

3. ADF will now show the historical pipeline runs. Selecting the required pipeline run, or the currently executing pipeline, will open a view displaying the individual pipeline activities:



Familiarise yourself with monitoring, and reviewing, ADF pipeline activity:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually>

Activity 6 – Analytics via KQL

We can now use KQL to query data ingested [via ADF] to ADX.

ADX KQL queries

We can use the Azure Portal ADX Query for this activity or the URL:

<https://dataexplorer.azure.com/>

For the Azure Portal:

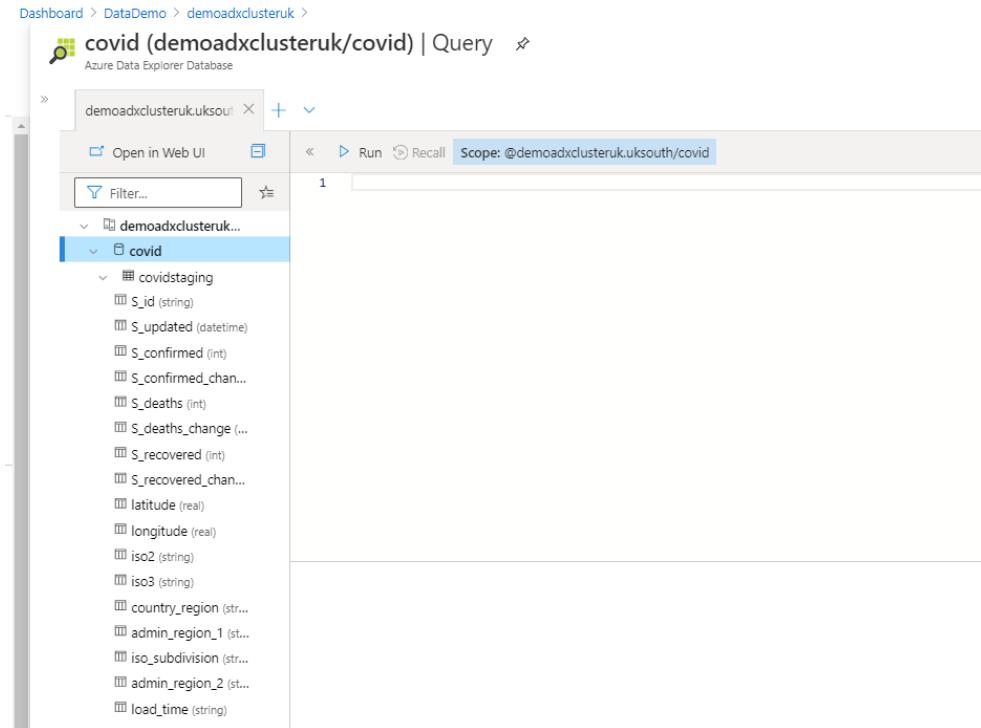
1. Select your resource group
2. Select the ADX cluster
3. Select **Databases**, select the **covid** database:

The screenshot shows the Azure Portal interface for an Azure Data Explorer (ADX) cluster named "demoadxclusteruk". The "Databases" blade is open, displaying a list of databases. The "covid" database is selected, as indicated by a red box around its row in the table. The top navigation bar includes a "Query" button, which is also highlighted with a red box. On the left side, there is a sidebar with various navigation links, and the "Databases" link is highlighted with a red box.

DATABASE	SIZE
DATABASE	
covid	5.92 MB

COV_19 Analytics with Azure Data Explorer (ADX)

This will then open an ADX KQL query window:



Basic Analytics

So, lets run some basic analytics. To facilitate with this section the following links will help understand KQL:

- Pluralsight: <https://www.pluralsight.com/courses/kusto-query-language-kql-from-scratch>
- Pluralsight: <https://www.pluralsight.com/courses/microsoft-azure-data-explorer-advanced-query-capabilities>
- <https://docs.microsoft.com/en-us/sharepoint/dev/general-development/keyword-query-language-kql-syntax-reference>

All KQL used in this section can be found in the /downloads folder – “KQL – Query tool.kql”

With the database name highlighted (as per image above) type the following KQL and select **Run**:

```
covidstaging
| where country_region == 'Worldwide'
| where S_confirmed > 0
```

This will show you the data for ‘Worldwide’ geography, on a daily basis with a positive *S_confirmed* count. Adding in the **Project** will permit specific columns to be selected:

```
covidstaging
```

COV_19 Analytics with Azure Data Explorer (ADX)

```
| where country_region == 'Worldwide'  
| where S_confirmed > 0  
| project S_updated, S_confirmed, S_confirmed_change
```

The following command will select all data for ‘New Zealand’ geography:

```
covidstaging  
| where country_region == 'New Zealand'
```

Play with the data, selecting specific dates or geographies to become familiar with its schema and structure. The following command will show the **covidstaging** schema:

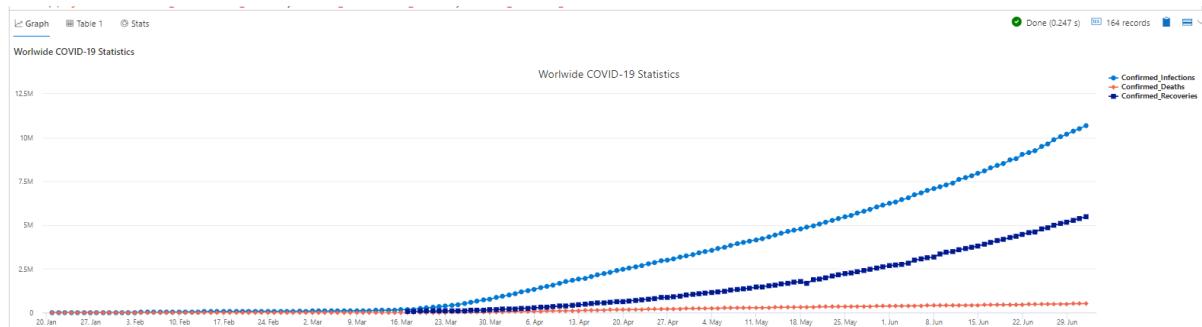
```
.show table covidstaging
```

We need KQL queries that are more sophisticated and and/or visual. The following takes the ‘Worldwide’ geography data and visualises it as a line chart. **Render** activates KQL query graphical visualisations:

(<https://docs.microsoft.com/en-us/azure/data-explorer/kusto/query/renderoperator?pivots=azuredatexplorer>)

```
// queries - Linechart - worldwide, all dates  
covidstaging  
| where country_region == 'Worldwide'  
| project S_updated, S_confirmed, S_deaths, S_recovered  
| render linechart with (title = "Worldwide COVID-19 Statistics")  
| project-rename Confirmed_Infections = S_confirmed, Confirmed_Recoveries =  
S_recovered, Confirmed_Deaths = S_deaths
```

Notice how **project-rename** has been used to tidy the column names displayed:



Toggle the ‘Graph’, ‘Table’ and ‘Stats’ tab views to see the underlying data and ADX performance statistics:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the Azure Data Explorer interface. At the top, there is a code editor with the following KQL query:

```
43 // queries - stacked area chart - country, latest date
44 let maxDateTable = covidstaging | where country_region == 'U
```

Below the code editor, there are three tabs: "Graph" (selected), "Table 1", and "Stats". The "Graph" tab displays a stacked area chart titled "Worldwide COVID-19 Statistics, Confirmed/ Recovered/ Deaths". The chart shows data for the top 20 countries, with the y-axis ranging from 0 to 4M. The legend indicates three series: "Confirmed_Recoveries" (blue), "Confirmed_Infections" (orange), and "Confirmed_Deaths" (black). The chart shows a significant decline in infections over time, with the United States having the highest number of infections.

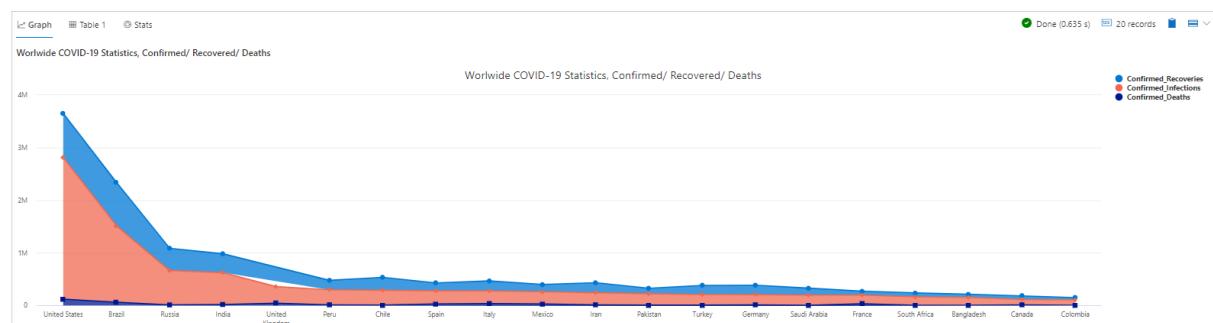
Increasing the KQL sophistication will give greater sophistication to the analytics. Introducing the `let` command allows a temporary table to be built to determine the maximum date:

```
let maxDateTable = covidstaging | where country_region == 'Worldwide' | order by S_updated desc | take 1 | project S_updated;
```

If we join this temporary table to our main KQL, for `country_region` data, we can take the `top 20` confirmed cases count:

```
// queries - stacked area chart - country, latest date

let maxDateTable = covidstaging | where country_region == 'Worldwide' | order by S_updated desc | take 1 | project S_updated;
covidstaging
| join kind= inner (maxDateTable) on $left.S_updated == $right.S_updated
| where country_region != 'Worldwide'
| where admin_region_1 == ''
| project country_region, S_recovered, S_confirmed, S_deaths
| order by S_confirmed desc
| render stackedareachart with ( title = "Worldwide COVID-19 Statistics, Confirmed/ Recovered/ Deaths" , xcolumn = country_region)
| project-rename Confirmed_Deaths = S_deaths, Confirmed_Infections = S_confirmed,
Confirmed_Recoveries = S_recovered
| take 20
```

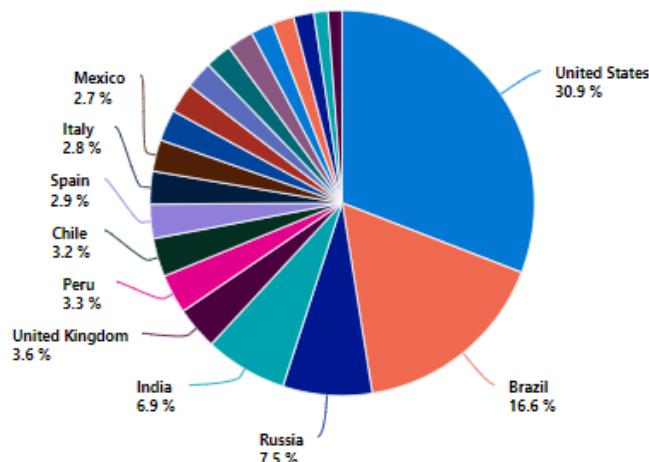


Using visualisations can have a large impact on the understanding of the data. For example, making some minor changes to the KQL above and utilising a pie chart gives a clearer picture of infections by `country_region`:

COV_19 Analytics with Azure Data Explorer (ADX)

```
// queries - piechart - country, latest date, confirmed
let maxDateTable = covidstaging | where country_region == 'Worldwide' | order by S_updated
desc | take 1 | project S_updated;
covidstaging
| join kind= inner (maxDateTable) on $left.S_updated == $right.S_updated
| where country_region != 'Worldwide'
| where admin_region_1 == ''
| project country_region, S_confirmed
| order by S_confirmed desc
| render piechart with ( title = "Worldwide COVID-19 Statistics, Confirmed Infections" ,
xcolumn = country_region, legend = hidden)
| project-rename Confirmed_Infections = S_confirmed
| take 20
```

Worldwide COVID-19 Statistics, Confirmed Infections



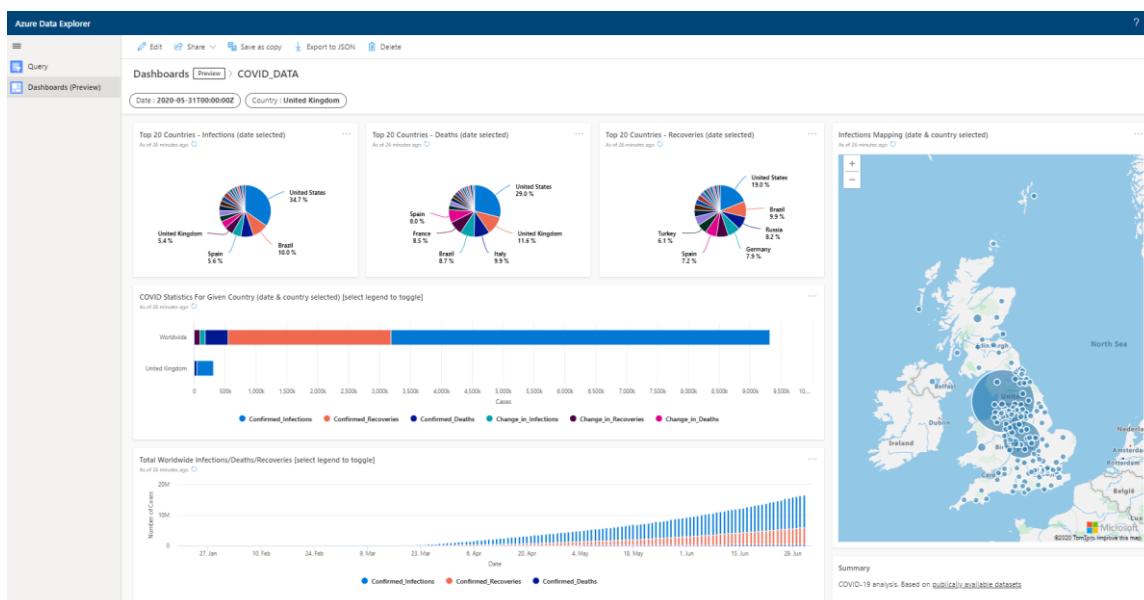
Activity 7 – Analytics via ADX Dashboards

Using the ADX query tool is a great way to perform “what if” analysis over your data, but it assumes a knowledge of KQL is required. If this is not the case, or you need to build a sophisticated dashboarding [with secure sharing capabilities] experience then the Azure Data Explorer Dashboards can be utilised.

An explanation of (the new) Azure Data Explorer Dashboards can be found here:

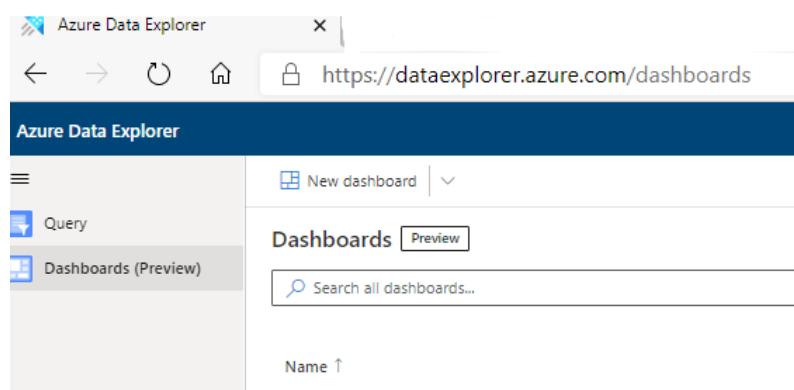
<https://docs.microsoft.com/en-us/azure/data-explorer/azure-data-explorer-dashboards#:~:text=%20Visualize%20data%20with%20Azure%20Data%20Explorer%20dashboards,support%20a...%205%20Next%20Steps.%20%20More%20>

We will now build the following Azure Data Explorer Dashboard:



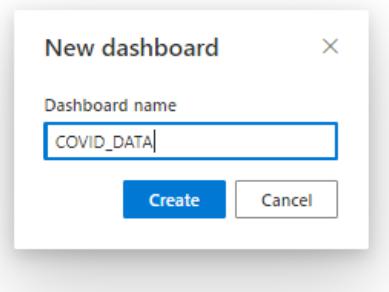
Create Dashboard

Type the ADX Dashboard URL: <https://dataexplorer.azure.com/dashboards>



Select ‘+ New’ and enter the Dashboard name ‘COVID_DATA’:

COV_19 Analytics with Azure Data Explorer (ADX)



Data sources

- Select '**Data sources**':

- Select '**+ New data source**'
- Enter the ADX Cluster URI (see section [Activity 2](#))
- Select '**Connect**'
- Select '**covid**' in the listbox
- Enter the Data source name: '**covidstaging**'
- Click '**Apply**'

This will create the data source:

Create Parameters

We will now create the Dashboard parameters that will drive the Dashboard visualisations:

An explanation of ADX Dashboard parameters is here: <https://docs.microsoft.com/en-us/azure/data-explorer/dashboard-parameters>

Date Parameter

The **Date** parameter allows the dashboard to be context sensitive to the selected date value.

COV_19 Analytics with Azure Data Explorer (ADX)

- Select '**Parameters**'
- Select '**+ New parameter**'
- Enter the fields as shown in the image below. The '**edit query**' KQL is shown beneath also:

The screenshot shows the 'Edit parameter' dialog in the Azure Data Explorer interface. The 'Parameter display name' is 'Date'. The 'Parameter type' is 'Single selection'. The 'Variable name' is 'dateselected'. The 'Data type' is 'String'. The 'Pin as dashboard filter' checkbox is checked. The 'Source' section shows 'Query' selected. The 'Data source' dropdown is set to 'covidstaging'. A red box highlights the 'Query results' section, which contains a link 'Edit query' and a list of dates: S_updated, 2020-07-02, 2020-07-01, 2020-06-30, 2020-06-29, 2020-06-28. Below this, it says 'Showing 5 out of 154 records'. The 'Value' dropdown is set to 'S_updated'. The 'Display name' dropdown is also set to 'S_updated'. There are checkboxes for 'Select all' and 'Add empty "Select all" value'. The 'Default value' is '2020-06-30'. At the bottom are 'Done' and 'Cancel' buttons.

- Select '**edit query**' and enter the KQL below:

```
covidstaging
| where country_region == 'United Kingdom'
| distinct S_updated
| order by S_updated desc
| project format_datetime(S_updated, 'yyyy-MM-dd')
```

- Select 'Done' to save the KQL date query into the date parameter definition:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the Azure Data Explorer interface. At the top, there's a search bar with the text "covidstaging" and a dropdown arrow, followed by a "Run" button. Below the search bar is a code editor containing the following Kusto query:

```
1 | covidstaging
2 | where country_region == 'United Kingdom'
3 | distinct s_updated
4 | order by s_updated desc
5 | project format_datetime(s_updated, 'yyyy-MM-dd')
```

Below the code editor is a "Results" table with a single column "S.updated". The table lists dates from 2020-07-02 down to 2020-06-15. A red box highlights the "Done" button at the bottom of the table.

- Select 'Done' to then save the **Date** parameter itself. You should now have created the **Date** parameter:

The screenshot shows the Power BI Parameters pane. At the top, there's a toolbar with icons for help, search, settings, and a user profile. Below the toolbar is a header with the word "Parameters" and a close button. Underneath the header is a "New parameter" button. The main area displays two parameters:

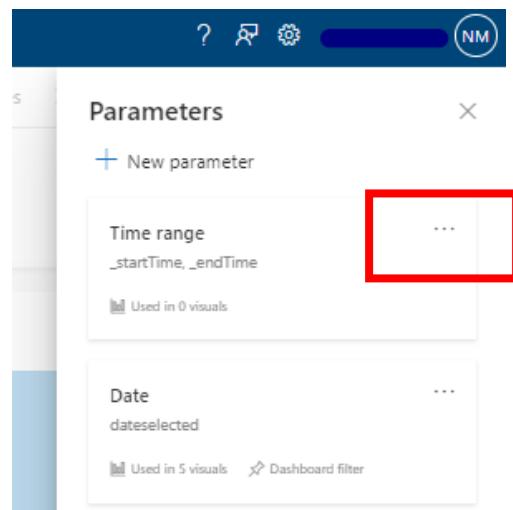
- Time range**: `_startTime, _endTime`. This parameter is not highlighted with a red box.
- Date**: `dateselected`. This parameter is highlighted with a large red box. Below it, it says "Used in 5 visuals" and "Dashboard filter".

Time Range Parameter

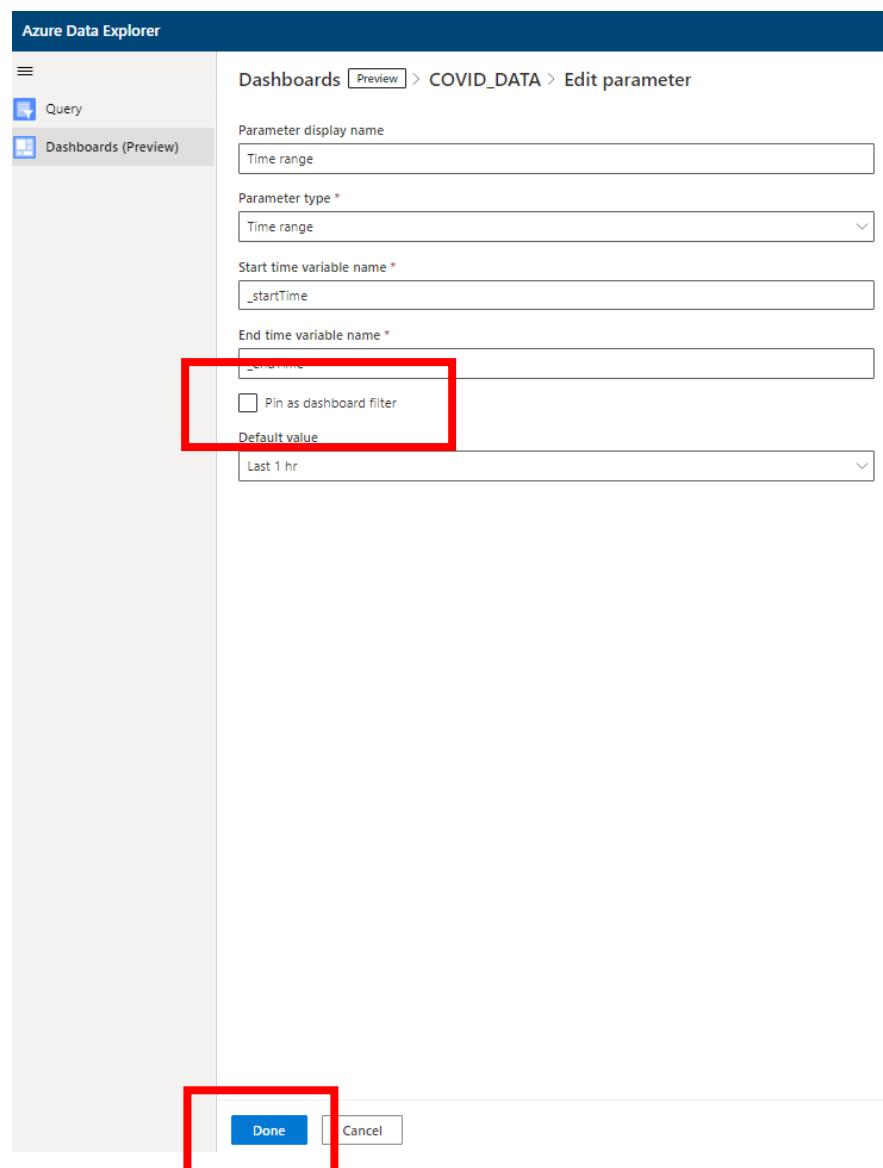
This parameter may appear as a default parameter on your dashboard. It is not required for the COVID-19 data (or dashboard) as the data is date (day, 24hr) based. We will, therefore, disable this option from the dashboard:

COV_19 Analytics with Azure Data Explorer (ADX)

- Next to the ‘Time range’ parameter definition, select ‘...’ and ‘edit’ :



- Unselect the ‘Pin as dashboard filter’ option
- Select ‘Done’ to save:



COV_19 Analytics with Azure Data Explorer (ADX)

Country Parameter

The **Country** parameter allows the dashboard to be context sensitive to the selected country value.

- Select '**Parameters**'
- Select '**+ New parameter**'
- Enter the fields as shown in the image below. The '**edit query**' KQL is shown beneath:

The screenshot shows the 'Edit parameter' dialog in the Azure Data Explorer interface. The 'Parameter display name' is set to 'Country'. The 'Parameter type' is 'Single selection'. The 'Variable name' is 'countryselected'. The 'Data type' is 'String'. The 'Pin as dashboard filter' checkbox is checked. Under 'Source', the 'Query' option is selected, and the 'Data source' dropdown is set to 'covidstaging'. The 'Query results' section shows a list of countries starting with 'Afghanistan', with 'Andorra' at the bottom. A red box highlights the 'Edit query' link next to the 'Query results' label. Below this, the 'Value' is 'country_region', 'Display name' is 'country_region', and 'Select all' is checked. The 'Default value' is 'United Kingdom'. At the bottom are 'Done' and 'Cancel' buttons.

- Select '**edit query**' and enter the KQL below:

COV_19 Analytics with Azure Data Explorer (ADX)

```
covidstaging
| where country_region != 'Worldwide'
| distinct country_region
| order by country_region asc
```

- Select 'Done' to save the KQL date query into the date parameter definition:

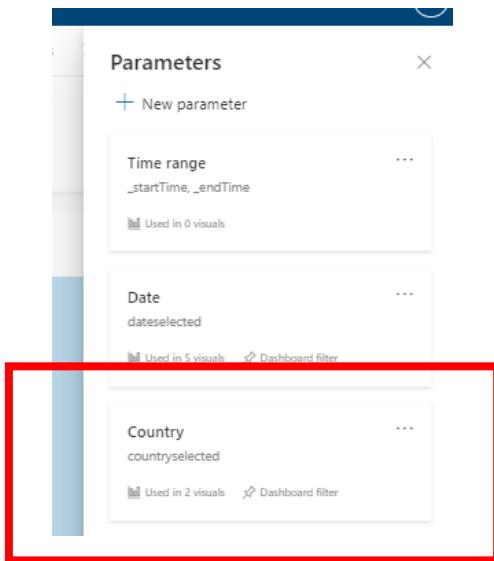
The screenshot shows the 'Add a query' dialog in Azure Data Explorer. At the top, there is a code editor with the following KQL query:

```
covidstaging
| where country_region != 'Worldwide'
| distinct country_region
| order by country_region asc
```

Below the code editor is a 'Results' pane showing a list of country names under the 'country_region' column. The list includes: Afghanistan, Albania, Algeria, American Samoa, Andorra, Angola, Anguilla, Antigua and Bar..., Argentina, Armenia, Aruba, Australia, Austria, Azerbaijan, Bahamas, Bahrain, and Bangladesh. The 'Done' button at the bottom of the results pane is highlighted with a red box.

- Select 'Done' to then save the **Country** parameter itself. You should now have created the **Country** parameter:

COV_19 Analytics with Azure Data Explorer (ADX)



COV_19 Analytics with Azure Data Explorer (ADX)

Dashboard Visualisations

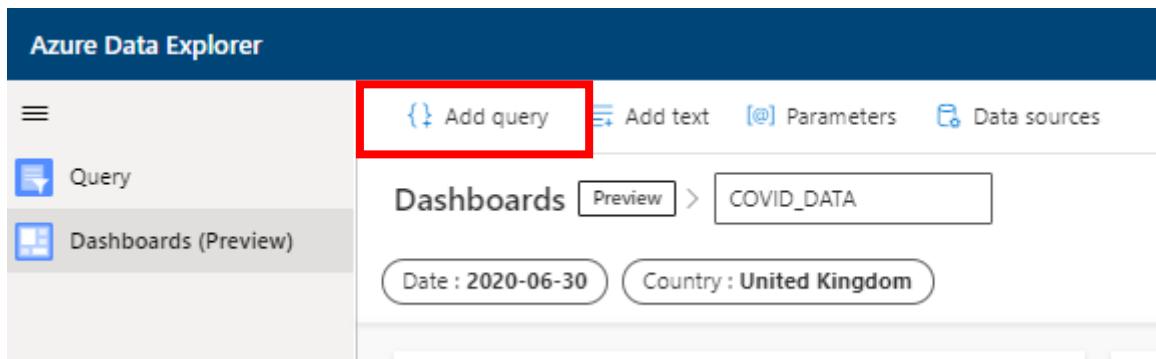
We can now add report visualisations to the dashboard.

The following KQL queries will be created. **All KQL can be found in the /downloads container in this repository. It is saved under the corresponding name.**

- a) Top 20 Countries – Infections (date selected) = **KQL_dashboard_1.kql**
- b) Top 20 Countries – Deaths (date selected) = **KQL_dashboard_2.kql**
- c) Top 20 Countries – Recoveries (date selected) = **KQL_dashboard_3.kql**
- d) COVID Statistics (date & country selected) = **KQL_dashboard_4.kql**
- e) Total Worldwide Infections/Deaths/Recoveries = **KQL_dashboard_5.kql**
- f) Infections Mapping (date & country selected) = **KQL_dashboard_6.kql**

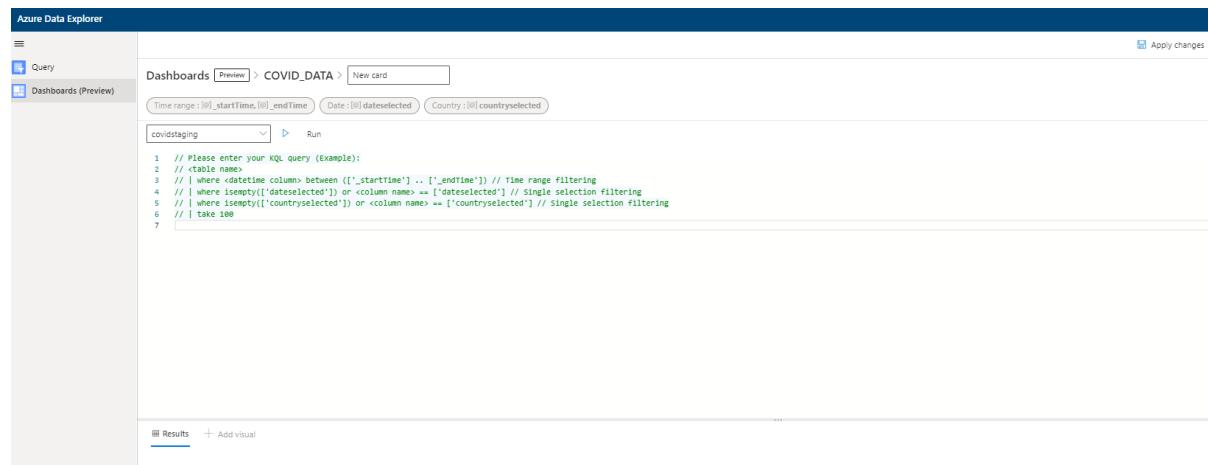
Create a Dashboard Visualisation

- Select ‘{} Add Query’



The screenshot shows the Azure Data Explorer interface. At the top, there's a dark blue header with the title 'Azure Data Explorer'. Below it is a navigation bar with icons for 'Query' (selected), 'Dashboards (Preview)', 'Add query' (highlighted with a red box), 'Add text', 'Parameters', and 'Data sources'. The main area is titled 'Dashboards' with a 'COVID_DATA' card. Below the card are two input fields: 'Date : 2020-06-30' and 'Country : United Kingdom'.

- Insert the KQL into the query panel as per the list above



The screenshot shows the Azure Data Explorer interface with the 'Query' tab selected. The main area contains a code editor with the following KQL script:

```
covidd staging
1 // Please enter your KQL query (Example):
2 // Table Name:
3 // | where <datetime column> between ('[_starttime'] .. [_endtime']) // Time range filtering
4 // | where isempty(['dateSelected']) or <column name> == ['dateSelected'] // Single selection filtering
5 // | where isempty(['countrySelected']) or <column name> == ['countrySelected'] // Single selection filtering
6 // | take 100
7
```

Below the code editor, there are tabs for 'Results' and '+ Add visual'.

- Select ‘Run’
- Once the results are displayed from the query in ‘Results’ tab, select ‘+ Add visual’:

COV_19 Analytics with Azure Data Explorer (ADX)

The screenshot shows the Azure Data Explorer interface. On the left, there's a sidebar with 'Query' and 'Dashboards (Preview)' options. The main area has a breadcrumb navigation bar: Dashboards > COVID_DATA > New card. Below it are filter buttons for Time range, Date, and Country. A search bar contains 'covidstaging' and a red box highlights the 'Run' button. The code editor shows a KQL query:

```

1 // Please enter your kql query (Example):
2 // <table name>
3 // | where <datetime column> between (['_startTime'] .. ['_endTime']) // Time range filtering
4 // | where isempty(['dateselected']) or <column name> == ['dateselected'] // Single selection filtering
5 // | where isempty(['countryselected']) or <column name> == ['countryselected'] // Single selection filtering
6 // | take 100
7
8 covidstaging
9 | where country_region != 'Worldwide'
10 | where admin_region_1 != ''
11 | where s_updated == dateselected
12 | project country_region, s_confirmed
13 | order by s_confirmed desc
14 | project-rename Confirmed_infections = s_confirmed
15 | take 20

```

Below the code editor is a table titled 'Results' with two columns: 'country_region' and 'Confirmed_infections'. It lists three rows: United States (2,588,017), Brazil (1,402,041), and Russia (547,849). A red box highlights this table.

- Enter the visualisation card title in the panel, and,
- Select the required visual under ‘Visual Formatting’. In this case [KQL_dashboard_1.kql] this is a pie chart:

The screenshot shows the Azure Data Explorer interface with a visual card titled 'COVID_DATA > Top 20 Countries - Infec...'. The visual card panel includes a 'Visual formatting' section with a 'Chart type' dropdown set to 'Pie Chart', highlighted by a red box. The main results panel shows a pie chart titled 'Top 20 Countries - Infections (date selected)'. The chart displays the percentage distribution of infections across several countries. Labels on the chart include: United States (30.5 %), Brazil (16.5 %), India (6.7 %), Russia (3.6 %), Peru (3.4 %), United Kingdom (3.7 %), and others. A red box highlights the pie chart.

- Select ‘Apply Changes’ to save the visualisation card. Resize and position the visualisation on the dashboard as needed.

COV_19 Analytics with Azure Data Explorer (ADX)

Remaining Dashboard Visualisations

Repeat the above steps [in section '**Create a Dashboard Visualisation**'] for the remaining visualisations:

- a) Top 20 Countries – Infections (date selected) = **KQL_dashboard_1.kql** **created**
- b) Top 20 Countries – Deaths (date selected) = **KQL_dashboard_2.kql**
- c) Top 20 Countries – Recoveries (date selected) = **KQL_dashboard_3.kql**
- d) COVID Statistics (date & country selected) = **KQL_dashboard_4.kql**
- e) Total Worldwide Infections/Deaths/Recoveries = **KQL_dashboard_5.kql**
- f) Infections Mapping (date & country selected) = **KQL_dashboard_6.kql**

Specific visualisation properties are shown as follows:

- b) Top 20 Countries – Deaths (date selected)

Pie chart – as per (a):

The screenshot shows the 'Visual formatting' section of a visualization settings panel. At the top, there are buttons for 'Apply changes' (highlighted with a red box), 'Discard changes', 'Share', and 'Refresh'. Below these are sections for 'Reset', 'Visual formatting', and 'Collapse all'. A dropdown menu labeled 'Chart type' is open, showing 'Pie Chart' as the selected option (also highlighted with a red box). The rest of the panel is mostly empty.

- c) Top 20 Countries – Recoveries (date selected)

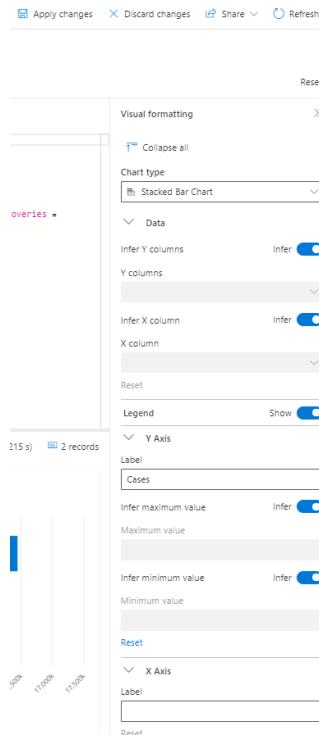
Pie chart – as per (a):

This screenshot is identical to the one above, showing the 'Visual formatting' section of a visualization settings panel. The 'Apply changes' button at the top is highlighted with a red box. The 'Chart type' dropdown is open, showing 'Pie Chart' as the selected option, which is also highlighted with a red box. The rest of the panel is mostly empty.

COV_19 Analytics with Azure Data Explorer (ADX)

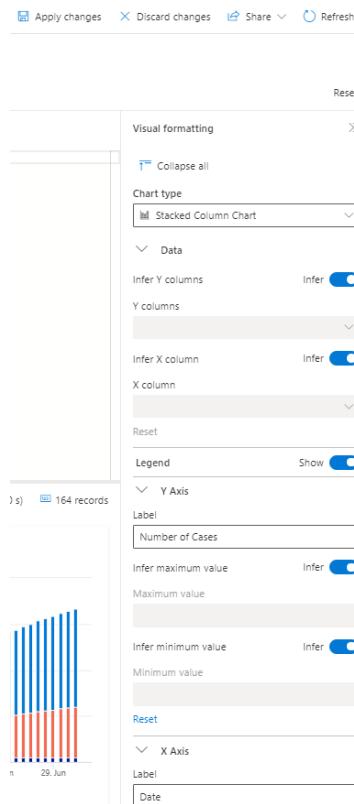
d) COVID Statistics (date & country selected)

Stacked Bar Chart:



e) Total Worldwide Infections/Deaths/Recoveries

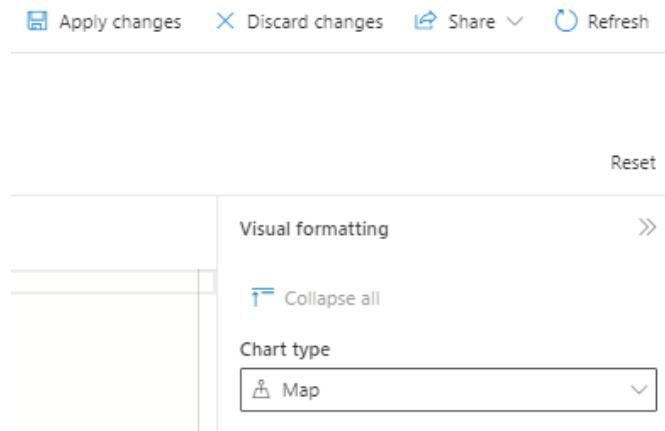
Stacked Column Chart:



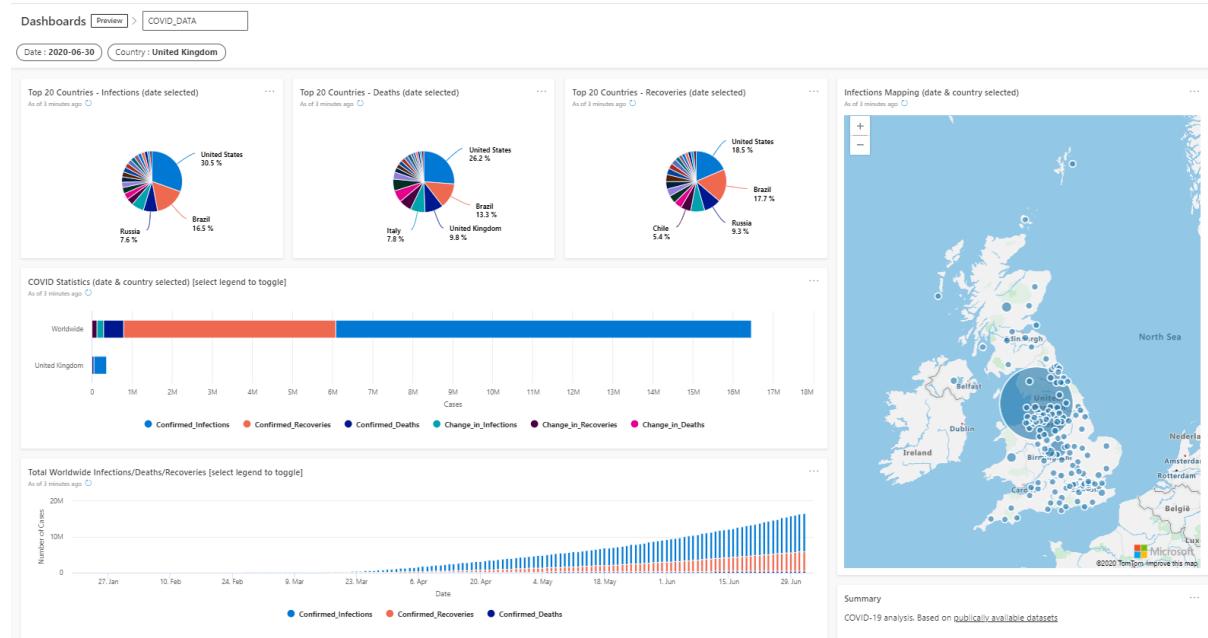
COV_19 Analytics with Azure Data Explorer (ADX)

f) Infections Mapping (date & country selected)

Map:



The completed dashboard should look like the graphic below. If not, check the KQL behind each visualisation card definition and the visualisation properties (were appropriate):



Activity 8 – Wrap-up

You have completed Workshop 1. Now try to enhance the activities as follows:

- Try different file formats (JSON or JSON Lines) or larger volumes of files
- Experiment with KQL to perform richer queries and visualisations
- Expand your reporting & visualisations to include Power BI

Clean-up

Remember to stop both the ADX cluster and the ADF trigger to prevent Azure charges being incurred unnecessarily.