

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

DESCRIPTION Use the data analytics stack to build a data pipeline using Data Factory, Databricks and Synapse. Problem Statement: As a Data Engineer, you've been asked to access the services that can help with ETL of data in the cloud data storage to enable analytics through Synapse. In this POC, we will be collecting the data from SQL Database using ADF and the transformed data will be the source for databricks to run complex transformations and once data is analysed using Databricks, it is synced into synapse analytics data warehouse as historical dataset for enabling various analytics.

Domain: Analytics

- Create a Resource Group.
- Create a Storage account.
- Create an Azure SQL Database.
- Create a data factory.
- Configure Databricks cluster
- Create Synapse analytics Data Warehouse

The screenshot shows the 'Create a resource group' wizard in the Azure portal. The 'Basics' tab is selected. The 'Project details' section shows a subscription dropdown set to 'Azure Pass - Sponsorship' and a resource group dropdown set to 'project1'. The 'Resource details' section shows a region dropdown set to '(US) East US'. At the bottom, there are buttons for 'Review + create', '< Previous', and 'Next : Tags >'.

Create a resource group

Basics Tags Review + create

Resource group - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. [Learn more](#)

Project details

Subscription * Azure Pass - Sponsorship

Resource group * project1

Resource details

Region * (US) East US

Review + create < Previous Next : Tags >

The screenshot shows the 'Create a storage account' wizard in the Azure portal. The 'Basics' tab is selected. The 'Project details' section shows a subscription dropdown set to 'Azure Pass - Sponsorship' and a resource group dropdown set to 'project'. At the bottom, there are buttons for 'Review + create', '< Previous', and 'Next : Advanced >'.

Create a storage account

Basics Advanced Networking Data protection Encryption 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 about Azure storage accounts](#)

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * Azure Pass - Sponsorship

Resource group * project

Review + create < Previous Next : Advanced >

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

Microsoft Azure Search resources, services, and docs (G+)

Home > SQL databases > Create SQL Database

Database name * project

Server * (new) project97 (East US) Create new

Want to use SQL elastic pool? No

Workload environment Development

Compute + storage * Basic 1 GB storage

Review + create Next : Networking >

Feedback

Service and compute tier

Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. [Learn more](#)

Service tier Basic (For less demanding workloads) Compare service tiers

DTUs [Compare DTU options](#)

5 (Basic)

Data max size (GB) 1

Cost summary

Cost per DTU (in INR)	70.57
DTUs selected	x 5
ESTIMATED COST / MONTH	352.85 INR

Apply

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the 'Create SQL Database' wizard in the Microsoft Azure portal. On the left, under 'Basics', the subscription is set to 'Azure Pass - Sponsorship', the resource group to 'project', the region to 'East US', and the database name to 'project'. The 'Cost summary' panel on the right indicates an estimated monthly cost of 352.85 INR, calculated from 70.57 DTUs selected at 5x. A note at the bottom states: 'By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms](#).^D'.

The screenshot shows the 'Create SQL Database' wizard in the Microsoft Azure portal. Under 'Security', the identity is set to 'Not enabled', service principal (preview) to 'Off', transparent data encryption to 'Service-managed key selected', advanced data security to 'Not now', and both Sql Ledger(Database) and Digest Storage to 'Disabled'. Under 'Additional settings', the use existing data is 'Sample' and the collation is 'SQL_Latin1_General_CI_AS'. A note at the bottom states: 'We'd love your feedback! →'.

The screenshot shows the 'Deployment' overview page for the database 'Microsoft.SQLDatabase.newDatabaseNewServer_64edfb5d4b18481d948b8'. The status bar at the top indicates 'Deployment is in progress'. The main area shows deployment details: Deployment name: Microsoft.SQLDatabase.newDatabaseNewServer..., Start time: 7/27/2022, 10:00:36 AM, Subscription: Azure Pass - Sponsorship, Correlation ID: b3742218-f7c9-4730-8a41-f0c165990b91, Resource group: project. A table shows the operation details for the resource 'project97': Resource (project97), Type (Microsoft.Sql/servers), Status (Accepted), and Operation details (link). The right sidebar includes links to Microsoft Defender for Cloud, free tutorials, and expert support.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

This screenshot shows the Microsoft Azure portal interface for a SQL database named 'project'. The top navigation bar includes 'Microsoft Azure', a search bar, and user information 'kiran.alvaa@outlook.com DEFAULT DIRECTORY'. The main content area displays the database's properties under the 'Essentials' section. Key details include:

- Resource group: project
- Status: Online
- Location: East US
- Subscription: Azure Pass - Sponsorship
- Subscription ID: 2a3a0c67-30a1-47b8-8ccb-43ed2b5cf86c
- Pricing tier: Basic
- Server name: project97.database.windows.net
- Elastic pool: No elastic pool
- Connection strings: Show database connection strings
- Earliest restore point: No restore point available
- Tags: Click here to add tags

The left sidebar lists various Azure services: Overview, Activity log, Tags, Diagnose and solve problems, Getting started, Query editor (preview), Power Platform (Power BI, Power Apps, Power Automate), and Settings (Compute + storage, Connection strings). A summary chart indicates 1.07% used space.

This screenshot shows the Microsoft Azure portal interface for creating a new Azure Databricks workspace. The top navigation bar includes 'Microsoft Azure', a search bar, and user information 'kiran.alvaa@outlook.com DEFAULT DIRECTORY'. The main content area is titled 'Create an Azure Databricks workspace' and is divided into several sections:

- Basics**: The currently selected tab, showing the subscription as 'Azure Pass - Sponsorship' and a dropdown for 'Resource group' with 'Create new' option.
- Networking**: Shows fields for 'Workspace name' (Enter name for Databricks workspace) and 'Region' (East US).
- Advanced**: Not visible in the screenshot.
- Tags**: Not visible in the screenshot.
- Review + create**: A blue button at the bottom left.
- < Previous** and **Next : Networking >**: Navigation buttons at the bottom.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the 'Project Details' step of the Azure Databricks workspace creation wizard. It includes fields for Subscription (Azure Pass - Sponsorship), Resource group (project), Workspace name (projectworkspace), Region (East US), and Pricing Tier (Trial (Premium - 14-Days Free DBUs)). Buttons at the bottom include 'Review + create', '< Previous', and 'Next : Networking >'.

The screenshot shows the 'Networking' step of the wizard. It displays validation results ('Validation Succeeded'), workspace details (Subscription: Azure Pass - Sponsorship, Resource group: project, Region: East US, Pricing Tier: trial), and networking options (Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP) set to No, Deploy Azure Databricks workspace in your own Virtual Network (VNet) set to No). Buttons at the bottom include 'Create', '< Previous', and 'Download a template for automation'.

The screenshot shows the 'Overview' page for the 'project_projectworkspace'. It displays deployment status ('Deployment is in progress'), deployment details (Deployment name: project_projectworkspace, Subscription: Azure Pass - Sponsorship, Resource group: project), and deployment logs ('Start time: 7/27/2022, 10:08:07 AM, Correlation ID: 3570bf9e-a4bc-41c4-8379-65e1220b4e68'). A message encourages feedback. A table for deployment details shows 'No results.'.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure Project workspace Overview page. At the top right, a message box indicates "Deployment succeeded" for the deployment "project_projectworkspace" in resource group "project". Below the message, there are two buttons: "Go to resource" and "Pin to dashboard". The main content area displays deployment details: Deployment name: project_projectworkspace, Subscription: Azure Pass - Sponsorship, Resource group: project, Start time: 7/27/2022, 10:08:07 AM, Correlation ID: 3570bf9e-a4bc-41c4-8379-65e1228b4e68. A "Deployment details" button with a download icon is also present. To the right, there are promotional cards for "Cost Management" and "Microsoft Defender for Cloud".

The screenshot shows the "Create a cluster" wizard in the Microsoft Azure Databricks interface. The title bar says "Clusters / New Compute". The main form is titled "New Cluster" with a "Create Cluster" button. It includes fields for "Cluster name" (set to "projectcluster"), "Cluster mode" (set to "Single Node"), "Databricks runtime version" (set to "7.3 LTS (Scala 2.12, Spark 3.0.1)"), and "Node type" (set to "Standard_DS3_v2"). Other options like "Use Photon Acceleration" and "Autopilot options" are shown. A "DBU / hour: 0.75" value is displayed next to the node type. A "Don't show again" link is at the bottom.

The screenshot shows the "Create a cluster" wizard in the Microsoft Azure Databricks interface, with the title bar "Clusters / projectcluster". The cluster has been completed, indicated by a green box labeled "Completed". The configuration page shows the cluster named "projectcluster" with a green status indicator. It lists tabs for Configuration, Notebooks (0), Libraries, Event log, Spark UI, Driver logs, Metrics, Apps, and Spark cluster UI - Master. The "Configuration" tab is selected. The cluster details include "Policy: Unrestricted", "Cluster mode: Single Node", "Databricks Runtime Version: 7.3 LTS (includes Apache Spark 3.0.1, Scala 2.12)", and "Node type: Standard_DS3_v2". The "DBU / hour: 0.75" value is also present. A "Next step" section with a "Import data" button is visible at the bottom.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

Microsoft Azure Search resources, services, and docs (G+) kiran.alvaa@outlook.com DEFAULT DIRECTORY

Home > Azure Synapse Analytics >

Create Synapse workspace

Subscription * Azure Pass - Sponsorship

Resource group * project Create new

Managed resource group Enter managed resource group name

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * projectsynapse97

Region * East US

Select Data Lake Storage Gen2 * From subscription Manually via URL

Account name * padlsgen2 Create new

File system name * (New) mycontainer Create now

Review + create < Previous Next: Security >

<https://portal.azure.com/#>

Microsoft Azure Search resources, services, and docs (G+) kiran.alvaa@outlook.com DEFAULT DIRECTORY

Home >

Microsoft.Azure.SynapseAnalytics-20220727101821 | Overview

Deployment

Search (Ctrl+F) <> Delete Cancel Redeploy Refresh

We'd love your feedback! →

■■■ Deployment is in progress

Deployment name: Microsoft.Azure.SynapseAnalytics-20220727101... Start time: 7/27/2022, 10:27:34 AM
Subscription: Azure Pass - Sponsorship Correlation ID: 00a3330c-0620-404b-91da-2595855250bc Download

Resource group: project

Deployment details (Download)

Resource	Type	Status	Operation details
No results.			

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure portal interface. At the top, there's a navigation bar with icons for back, forward, search, and user information. Below it is the main dashboard for a resource group named "Microsoft.Azure.SynapseAnalytics-20220727101821". The dashboard has a sidebar with links like Overview, Inputs, Outputs, and Template. The main content area displays a message: "Your deployment is complete" with a green checkmark icon. It provides deployment details: Deployment name: Microsoft.Azure.SynapseAnalytics-20220727101821..., Start time: 7/27/2022, 10:27:34 AM, Subscription: Azure Pass - Sponsorship, and Resource group: project. There are buttons for "Go to resource group" and "Deployment details (Download)". To the right, there's a sidebar with sections for Cost Management, Microsoft Defender for Cloud, and Free Microsoft tutorials.

The screenshot shows the Microsoft Azure portal interface for a SQL database named "project97". The left sidebar lists various management options: Settings, Azure Active Directory, SQL databases, SQL elastic pools, DTU quota, Properties, Locks, Data management, Backups, Deleted databases, Failover groups, Import/Export history, and security. The "Networking" option is selected. The main content area shows the "Networking" settings for the database. It includes sections for Firewall rules (allowing certain public internet IP addresses) and Exceptions (allowing Azure services and resources to access the server). A table lists a single rule named "allowall" with Start IPv4 address 0.0.0.0 and End IPv4 address 255.255.255.255. There are "Save" and "Discard" buttons at the bottom.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure portal interface for creating a new dedicated SQL pool. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information (kiran.alvaa@outlook.com, DEFAULT DIRECTORY). The current page is 'Azure Synapse Analytics > projectsynapse97 | SQL pools > New dedicated SQL pool'. The main content area has tabs: 'Basics' (selected), 'Additional settings', 'Tags', and 'Review + create'. Under 'Product details', it shows 'Azure Synapse Analytics dedicated SQL pool by Microsoft' with an 'Est. Cost Per Hour' of '108.79 INR' and a link to 'View pricing details'. A 'Terms' section contains legal text about agreeing to terms and conditions. Below that, under 'Basics', the 'Dedicated SQL pool name' is set to 'projectpool' and the 'Performance level' is 'DW100c'. At the bottom are buttons for 'Create', '< Previous', and 'Download a template for automation'.

The screenshot shows the Microsoft Azure portal interface for creating a new Synapse workspace. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information (kiran.alvaa@outlook.com, DEFAULT DIRECTORY). The current page is 'Home > Azure Synapse Analytics > Create Synapse workspace'. The main content area has tabs: 'Basics' (selected), 'Security' (selected), 'Networking', 'Tags', and 'Review + create'. Under 'Security', it says 'Configure security options for your workspace.' and 'Authentication'. It asks to choose an authentication method: 'Use both local and Azure Active Directory (Azure AD) authentication' (radio button selected) or 'Use only Azure Active Directory (Azure AD) authentication'. Below this, there are fields for 'SQL Server admin login *' (sqladminuser), 'SQL Password' (*****), and 'Confirm password' (*****). A note about 'System assigned managed identity permission' is present, with a link to learn more. At the bottom are buttons for 'Review + create', '< Previous', and 'Next: Networking >'.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows two consecutive pages from the Microsoft Azure portal.

Create Data Factory Page:

- Header: Microsoft Azure, Search resources, services, and docs (G+), kirin.alvaa@outlook.com, DEFAULT DIRECTORY.
- Breadcrumbs: Home > Data factories > Create Data Factory.
- Validation Status: Validation Passed.
- Step Progress: Basics, Git configuration, Networking, Advanced, Tags, **Review + create**.
- TERMS: A legal agreement section.
- Basics: Subscription (Azure Pass - Sponsorship).
- Action Buttons: Create (blue), < Previous, Next, Download a template for automation.

Microsoft.DataFactory-20220728091607 | Overview Page:

- Page Title: projectdf97 (Data factory (V2)).
- Left sidebar: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Networking, Managed identities, Properties, Locks, Setting started, Quick start.
- Essentials (JSON):
 - Resource group (move) : project
 - Status : Succeeded
 - Location : East US
 - Subscription (move) : Azure Pass - Sponsorship
 - Subscription ID : 2a3a8c67-30a1-47b8-8bcb-43ed2b5cf86c
 - Type : Data factory (V2)
 - Getting started : Quick start
- Getting started:
 - Open Azure Data Factory Studio: Start authoring and monitoring your data pipelines and data flows. [Open](#)
 - Read documentation: Learn how to be productive quickly. Explore concepts, tutorials, and samples. [Learn more](#)

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

PIPELINE 1 – COPY MOVIES.CSV FROM ADLS TO SQL TABLE

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (pipeline1, pipeline2), 'Datasets', 'Data flows' (dataflow1), and 'Power Query'. The main workspace displays a 'Copy data' activity named 'Copy data1'. The 'Source' tab is selected, showing the configuration for reading data from 'movies'. The 'File path type' is set to 'File path in dataset', and the 'Sink' tab is visible at the bottom.

The screenshot shows the Microsoft Azure Data Factory Studio interface. The 'Sink' tab is selected for the 'Copy data' activity 'Copy data1'. The 'Sink dataset' is set to 'AzureSqlTable1'. Other sink options like 'Write behavior' (Insert, Upsert, Stored procedure), 'Bulk insert table lock' (Yes, No), and 'Table option' (None, Auto create table) are also visible.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

```
CREATE TABLE [moviesdata]
(
    Film [varchar](200) NULL,
    Genre [varchar](200) NULL,
    LeadStudio [varchar](200) NULL,
    Audiencescore [varchar](200) NULL,
    Profitability [varchar](200) NULL,
    RottenTomatoes [varchar](200) NULL,
    WorldwideGross [varchar](200) NULL
)

SELECT * FROM moviesdata
```

movies Container

Search (Ctrl+F)

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease

Overview

Authentication method: Access key (Switch to Azure AD User Account)

Location: movies

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
movies.csv	7/28/2022, 11:12:29 ...	Hot (inferred)		Block blob	5.02 KiB	Available

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in.

Data Factory Validate all Publish all

Factory Resources

Pipelines

- pipeline1
- pipeline2
- pipeline3

Datasets

- AzureSqlTable1
- AzureSynapseAnalyticsTable1
- DelimitedText2
- fromdatabricks
- JoinedTable
- movies
- moviestable

Data flows

- dataflow1

Activities

Copy data

Copy data1

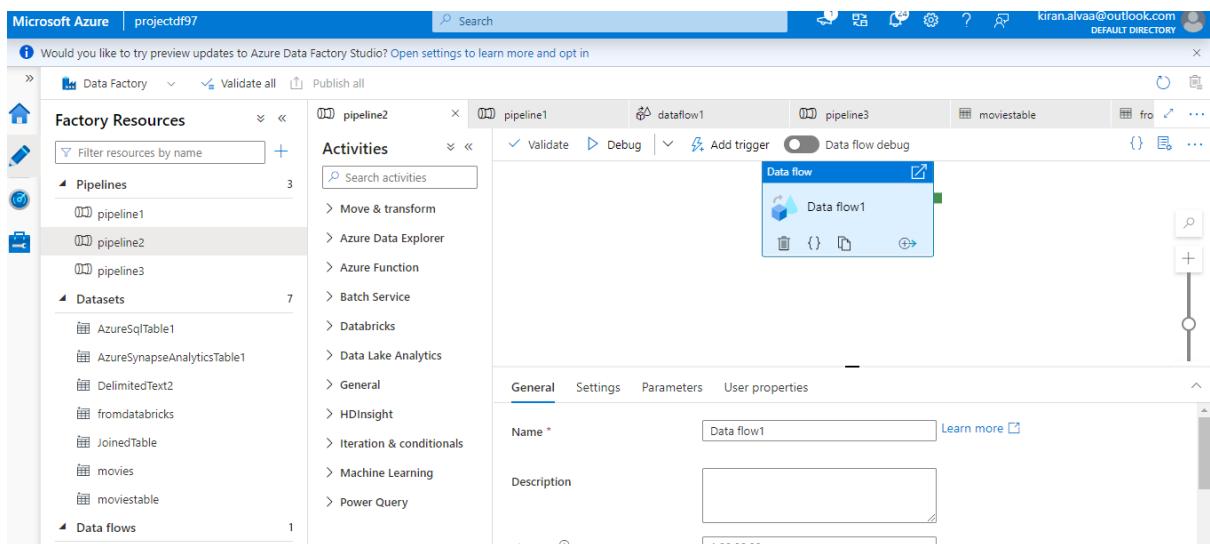
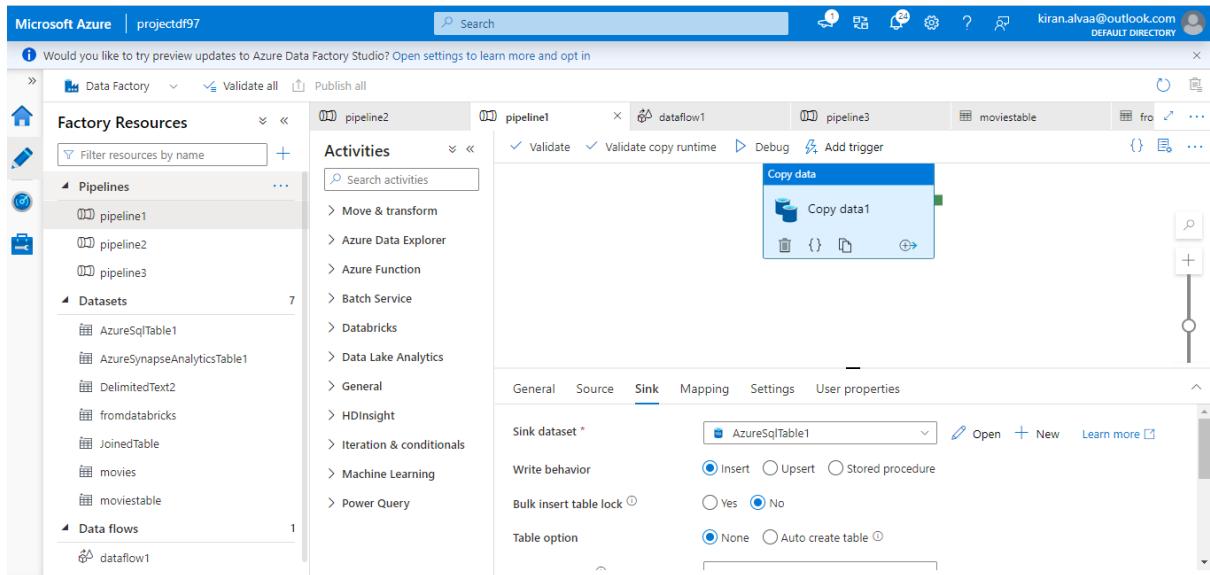
General Source Sink Mapping Settings User properties

Source dataset: movies

File path type: File path in dataset

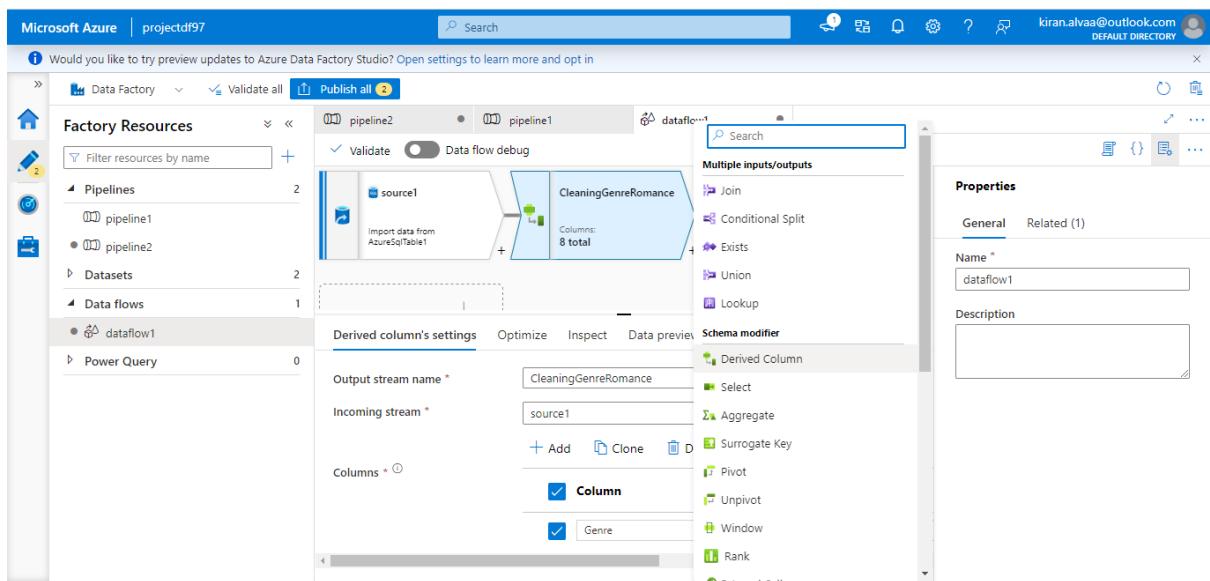
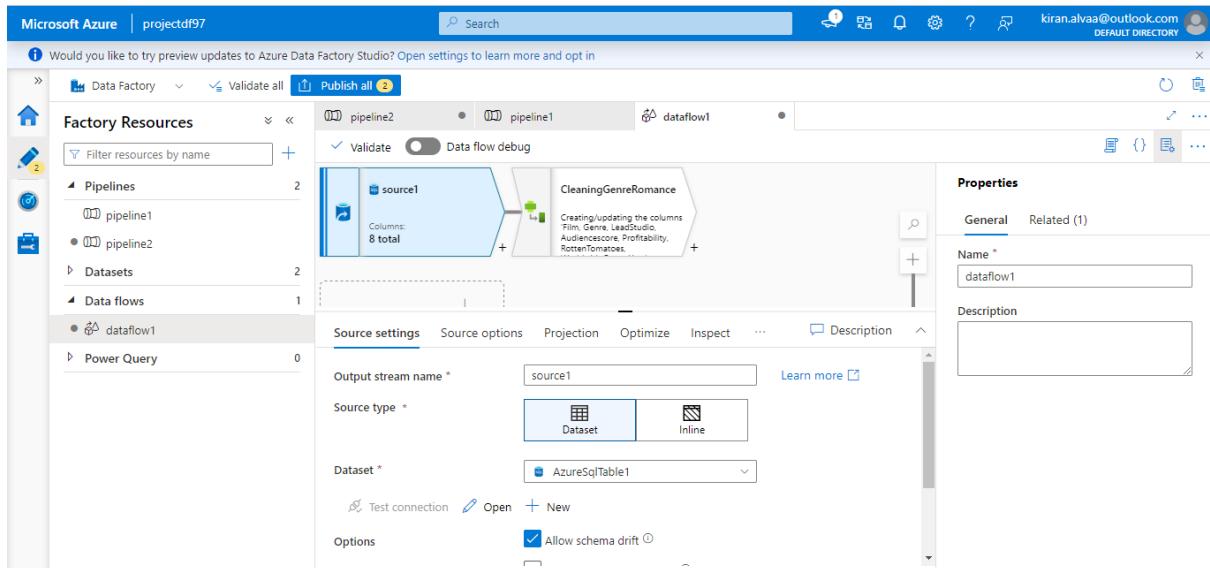
Filter by last modified Start time (UTC) End time (UTC)

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



1. Create a data stream named CleaningGenreRomance and perform data cleansing on the Genre column using Derived Column and case expression. (While collecting data it was observed that some genres have spelling mistakes like romance, Romence for Romance, comedy, Comdy for Comedy.)

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the 'Dataflow expression builder' interface in Microsoft Azure. The top navigation bar includes 'Microsoft Azure', 'projectdf97', a search bar, and user information 'kiran.alvaa@outlook.com DEFAULT DIRECTORY'. The main area is titled 'Dataflow expression builder' and shows a derived column named 'CleaningGenreRomance'. The 'Column name' is set to 'Genre'. The 'Expression' field contains the following code:

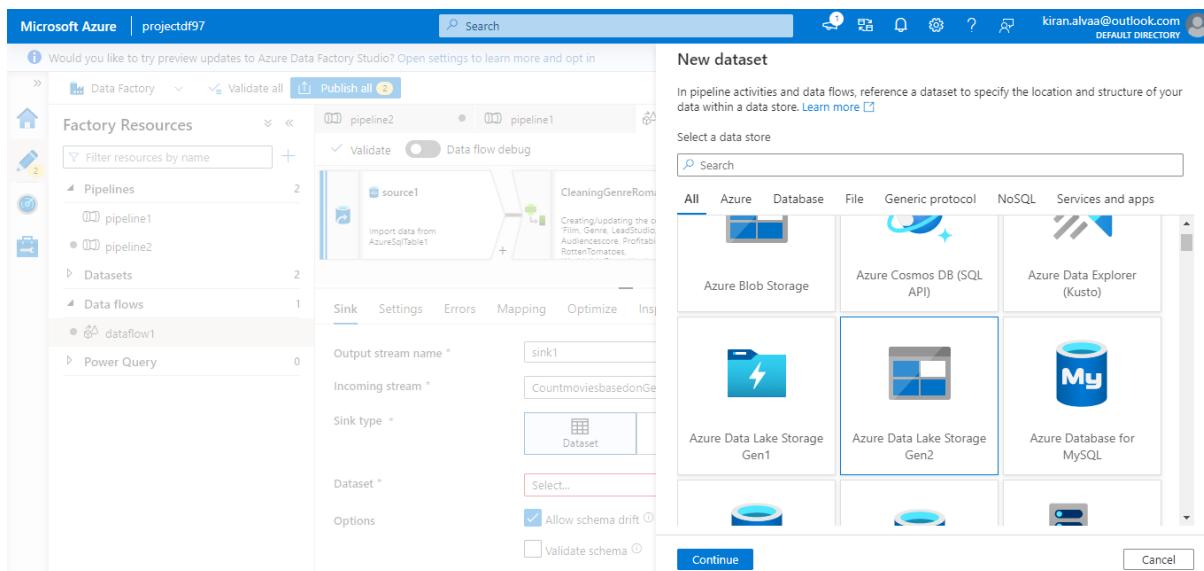
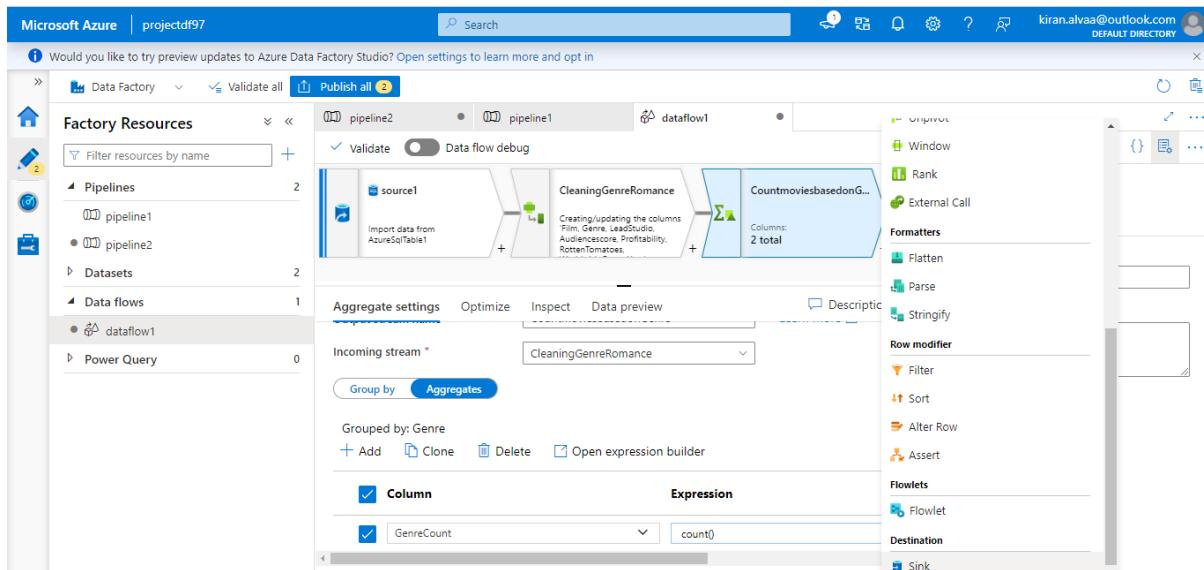
```
lower(iif(lower(Genre)='romence','Romance',iif(lower(Genre)='comdy','Comedy',Genre)))
```

The 'Save' button is visible at the top right. Below the expression editor, there are sections for 'Expression elements' (listing 'All', 'Functions', 'Input schema', 'Parameters', 'Cached lookup', and 'Data flow library') and 'Expression values' (listing 'Film', 'Genre', and 'LeadStudio'). At the bottom, there are 'Data preview', 'Save and finish', 'Cancel', and 'Clear contents' buttons.

2. Create a data stream named CountMoviesBasedOnGenre that can calculate number of films for each genre and store it as a separate dataset in ADLS under folder name “solution/genreCount”

The screenshot shows the 'Data Factory Studio' interface in Microsoft Azure. The top navigation bar includes 'Microsoft Azure', 'projectdf97', a search bar, and user information 'kiran.alvaa@outlook.com DEFAULT DIRECTORY'. The left sidebar shows 'Factory Resources' with 'Pipelines' (pipeline1, pipeline2), 'Datasets' (2), 'Data flows' (1: dataflow1), and 'Power Query' (0). The main workspace displays a data flow pipeline named 'dataflow1'. The pipeline consists of three main stages: 'source1' (Import data from AzureSqlTable1), 'CleaningGenreRomance' (Creating/updating the columns Film, Genre, LeadStudio, Audiencescore, Profitability, RottenTomatoes, ...), and 'CountmoviesbasedonGenre' (Aggregate settings: Grouped by: Genre, Output stream name: CountmoviesbasedonGenre, Incoming stream: CleaningGenreRomance). The 'Properties' panel on the right shows the 'Name' as 'dataflow1'. At the bottom, there are buttons for 'Group by', 'Aggregates', 'Add', 'Clone', 'Delete', and 'Open expression builder'.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in

Factory Resources > Pipelines > pipeline2 > Sink

Select format

Choose the format type of your data

Avro	CSV	JSON
DelimitedText		
ORC	Parquet	Binary

Output stream name: sink1
Incoming stream: CountmoviesbasedonGenre
Sink type: Dataset
Dataset: Select...
Options: Allow schema drift (checked), Validate schema (unchecked)

Continue Back Cancel

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in

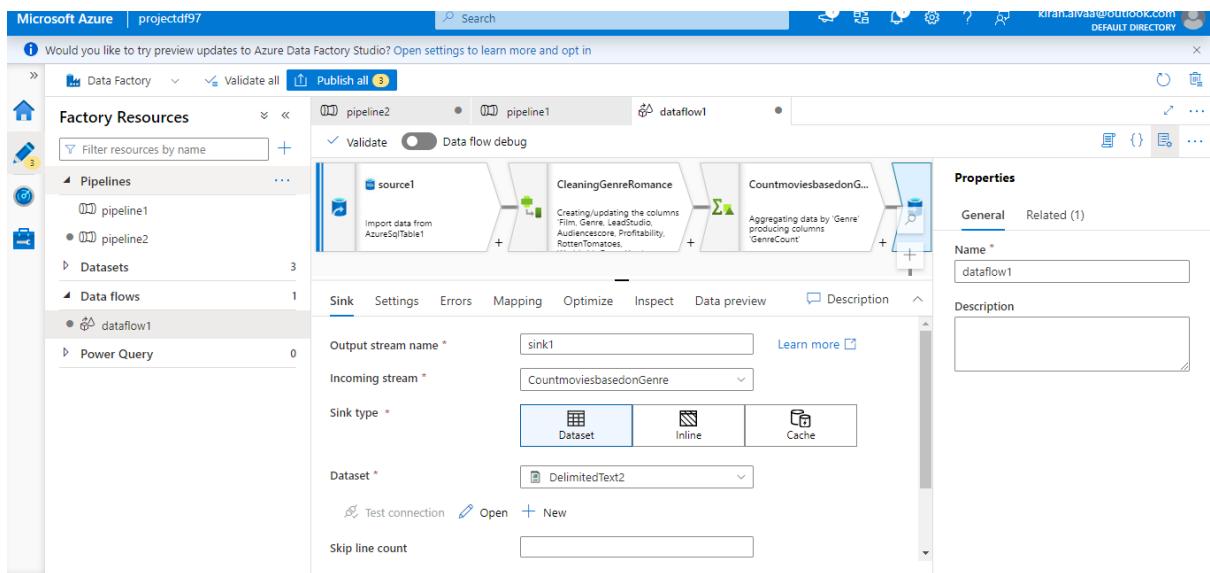
Factory Resources > Pipelines > pipeline2 > Sink

Set properties

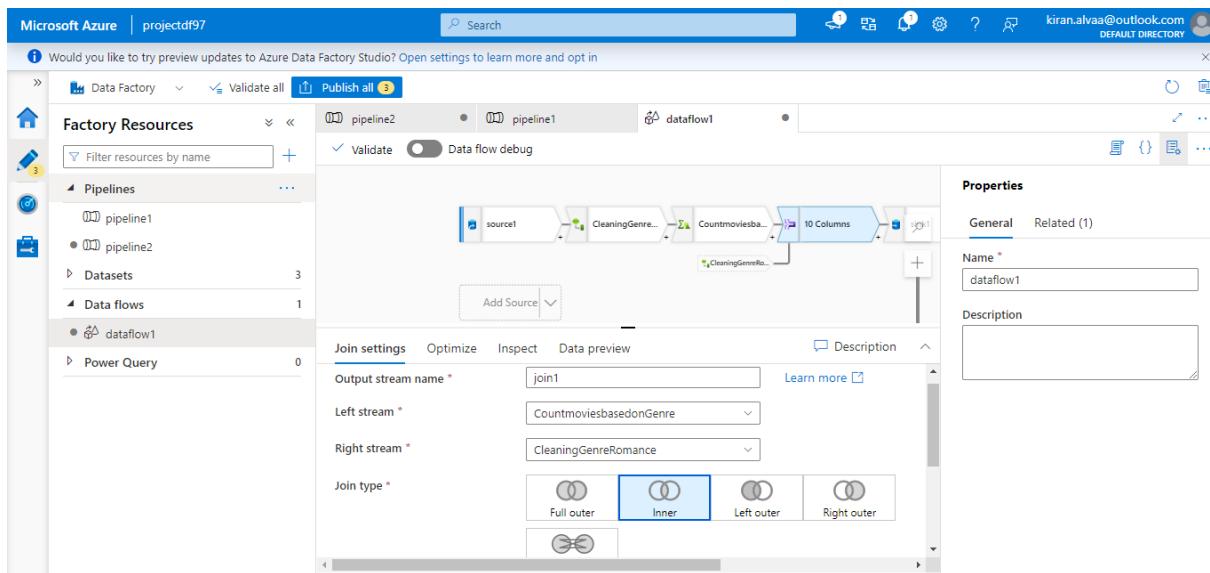
Name: DelimitedText2
Linked service: AzureDataLakeStorage1
File path: mycontainer / solution/genreCount / File
First row as header:
Import schema: From connection/store (radio button selected), From sample file, None
Advanced

OK Back Cancel

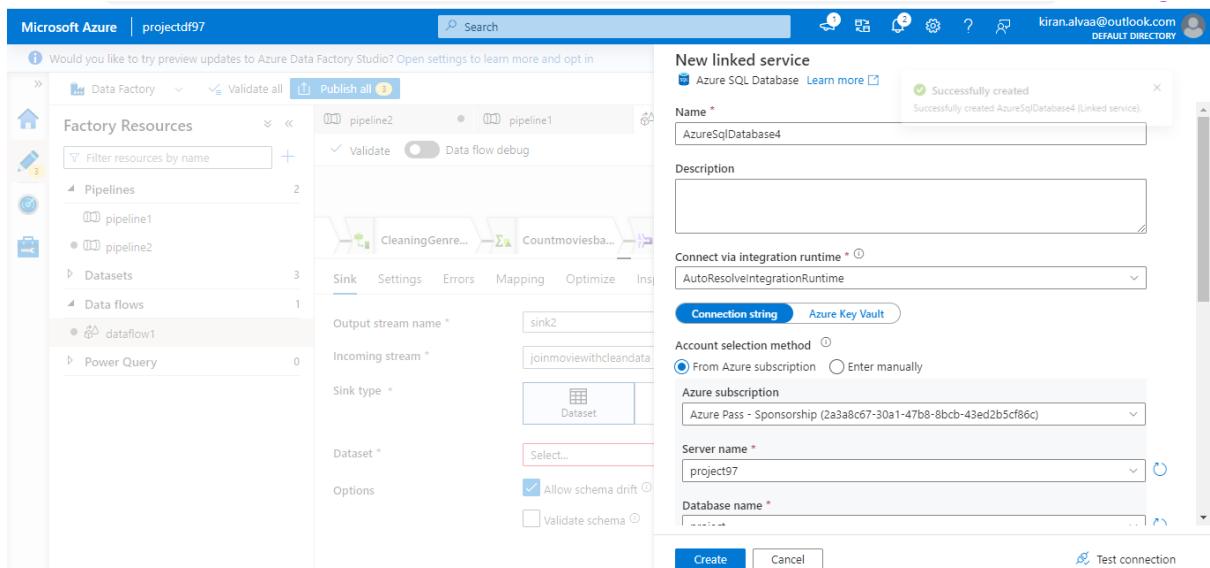
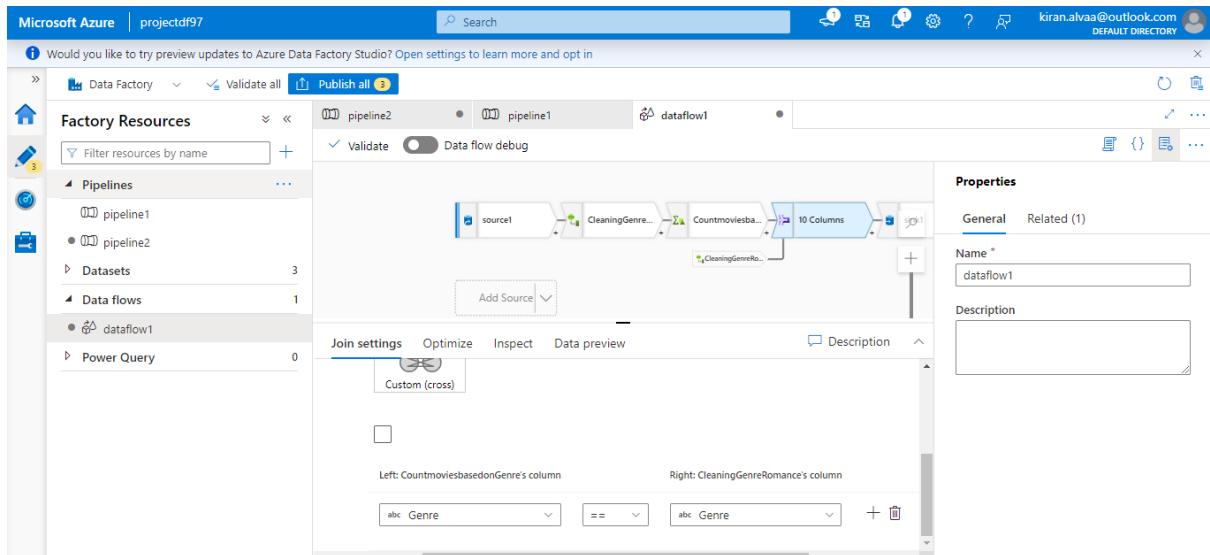
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



3. Create a new stream named JoinMovieCountWithCleanData. Perform join operation on CountMoviesBasedOnGenre with CleaningGenreRomance stream and store the same in the Azure SQL Database.



Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines (pipeline1, pipeline2), Datasets, Data flows (dataflow1), and Power Query. The main workspace displays a data flow pipeline with various components like CleaningGenre, joinmovewithcleandata, Countmoviesbase, and CleaningGenreRo. A 'Sink' configuration panel is open on the right, showing settings for an output stream named 'sink2' derived from an incoming stream 'joinmovewithcleandata'. The sink type is set to 'Dataset', and the dataset is 'Select...'. Under 'Options', the 'Allow schema drift' checkbox is checked. The 'Schema and table name' section specifies 'dbo.movies'. At the bottom of the dialog are 'OK', 'Back', and 'Cancel' buttons.

This screenshot shows the 'Notifications' pane in the Azure Data Factory Studio after a successful publish. It lists three notifications: 1) 'Publishing completed' (Successfully published a minute ago), 2) 'Successfully created' (Successfully created AzureSqlDatabase4 (Linked service) 9 minutes ago), and 3) 'Successfully created' (Successfully created AzureDataLakeStorage2 (Linked service) 24 minutes ago). A 'Close' button is at the bottom of the notifications pane.

The screenshot shows the 'Pipeline run' configuration dialog. It includes a warning message: 'Trigger pipeline now using last published configuration.' Below it is a 'Parameters' section with a table:

Name	Type	Value
No records found		

Below the parameters is a 'General' settings section with fields for 'Data flow' (set to 'dataflow1'), 'Run on (Azure IR)' (set to 'AutoResolveIntegrationRuntime'), and 'Compute size' (set to 'Small'). At the bottom are 'OK' and 'Cancel' buttons.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, a sidebar lists navigation options: Dashboards, Runs, Pipeline runs (selected), Trigger runs, Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main area displays 'All pipeline runs > pipeline2 - Activity runs'. It shows a list of activity runs, with one entry for 'Data flow1' listed under 'Activity type'. To the right, a 'Notifications' panel is open, showing four successful events: 'Succeeded' (ran pipeline2), 'Publishing completed' (published pipeline2), 'Successfully created' (AzureSqlDatabase4 linked service), and 'Successfully created' (AzureDataLakeStorage2 linked service). A 'Close' button is at the bottom of the notifications panel.

The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar shows a tree structure: Home > padlsgen2 | Containers > mycontainer (Container selected). Under 'mycontainer', there are sections for Overview, Diagnose and solve problems, Access Control (IAM), Settings, Shared access tokens, Manage ACL, Access policy, Properties, and Metadata. The main area displays a list of blobs in the 'mycontainer' container. The table has columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The blobs listed are: '_committed_0680...', '_started_0680644...', '_SUCCESS', 'part-00000-tid-86...', 'part-00013-tid-86...', 'part-00040-tid-86...', 'part-00043-tid-86...', 'part-00127-tid-86...', 'part-00178-tid-86...', and 'part-00199-tid-86...'. All blobs are of type 'Block blob' and are currently 'Available'. The 'Modified' column shows dates from August 1, 2022, to August 2, 2022.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
_committed_0680...	8/1/2022, 9:07:04 AM	Hot (Inferred)		Block blob	642 B	Available
_started_0680644...	8/1/2022, 9:07:04 AM	Hot (Inferred)		Block blob	0 B	Available
_SUCCESS	8/1/2022, 9:07:04 AM	Hot (Inferred)		Block blob	0 B	Available
part-00000-tid-86...	8/1/2022, 9:07:04 AM	Hot (Inferred)		Block blob	0 B	Available
part-00013-tid-86...	8/1/2022, 9:06:58 AM	Hot (Inferred)		Block blob	90 B	Available
part-00040-tid-86...	8/1/2022, 9:07:04 AM	Hot (Inferred)		Block blob	58 B	Available
part-00043-tid-86...	8/1/2022, 9:06:58 AM	Hot (Inferred)		Block blob	255 B	Available
part-00127-tid-86...	8/1/2022, 9:07:00 AM	Hot (Inferred)		Block blob	862 B	Available
part-00178-tid-86...	8/1/2022, 9:07:01 AM	Hot (Inferred)		Block blob	1.03 KiB	Available
part-00199-tid-86...	8/1/2022, 9:07:01 AM	Hot (Inferred)		Block blob	2.82 KiB	Available

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

1. Get the clean data from Azure SQL DB. Create an activity that can copy the data from SQLDB to ADLS Gen2.

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (pipeline1, pipeline2, pipeline3), 'Datasets', 'Data flows' (dataflow1), and 'Power Query'. The main workspace displays the 'Activities' section with a 'Copy data' activity named 'sqltabletoADLSGen2'. The 'Properties' pane on the right shows the pipeline is named 'pipeline3' with a description and annotations. The 'Source' tab of the activity configuration is selected, showing 'Source dataset' set to 'JoinedTable' and 'Use query' set to 'Table'.

The screenshot shows the 'Preview data' interface for the 'JoinedTable' dataset. It displays a table of movie data from the 'dbo.movies' object in the 'AzureSqlDatabase4' linked service. The columns are: ID, Genre, GenreCount, Film, LeadStudio, Audiencescore, Profitability, RottenTomatoes, and WorldWideGross. The data includes entries for movies like 'The Curious Case of Benjamin Button', 'Killers', 'WALL-E', 'Tangled', 'Gnomeo and Juliet', 'Gnomeo and Juliet', and 'Water For Elephants'.

ID	Genre	GenreCount	Film	LeadStudio	Audiencescore	Profitability	RottenTomatoes	WorldWideGross
1	fantasy	1	The Curious Case of Benjamin Button	Warner Bros.	81	1.78394375	73	\$285
2	action	1	Killers	Lionsgate	45	1.245333333	11	\$93.4
3	animation	4	WALL-E	Disney	89	2.896019067	96	\$521
4	animation	4	Tangled	Disney	88	1.365692308	89	\$355
5	animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56	\$193
6	animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56	\$193
7	drama	13	Water For Elephants	20th Century Fox	72	3.081421053	60	\$117

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

Search

All Azure Database File Generic protocol NoSQL Services and apps

Azure Blob Storage	Azure Cosmos DB (MongoDB API)	Azure Cosmos DB (SQL API)
Azure Data Explorer (Kusto)	Azure Data Lake Storage Gen1	Azure Data Lake Storage Gen2

Continue Cancel

Select format

Choose the format type of your data

Avro	Binary	DelimitedText
JSON	ORC	Parquet

Set properties

Name: moviestable

Linked service: AzureDataLakeStorage1

File path: moviestable / Directory / File

First row as header:

Import schema: From connection/store From sample file None

> Advanced

OK Back Cancel

2. Create an activity that can use Azure Databricks to read the data from the ADLS Gen2 and perform rank operation on the Genre column. Ensure this activity gets activated only after the data is stored in ADLS from SQL DB. The result of Databricks must be stored in the ADLS.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines (pipeline1, pipeline2, pipeline3), Datasets, Data flows, and Power Query. In the center, a pipeline named 'pipeline2' is selected. A 'Copy data' activity is being configured, pointing from 'sqltabletoADLS' to 'Databricks'. On the right, a 'New linked service' dialog is open for 'Azure Databricks'. It shows the 'Databrick Workspace URL' as <https://adb-8934044172288797.17.azuredatabricks.net>. The 'Authentication type' is set to 'Managed service identity'. The 'Workspace resource ID' is listed as `/subscriptions/2a3a8c67-30a1-47b8-8bcb-43ed2b5cf86c/resourceGroups/project/providers`. Below this, it notes a managed identity name: `projectdf97` and object ID: `7e4c38ff-c85a-4e54-9c50-cc073b228197`. The 'Existing cluster ID' dropdown is set to 'Add workspace and access token to list options'. Other sections include 'Annotations', 'Parameters', and 'Advanced'.

The screenshot shows the 'Access control (IAM)' page for the 'projectworkspace' Databricks workspace. The left sidebar includes 'Overview', 'Activity log', 'Access control (IAM)', 'Tags', 'Settings' (Virtual Network Peering, Encryption, Networking, Properties, Locks), 'Automation' (Tasks, Export template), and 'Support + troubleshooting'. The main area displays 'Check access' and 'My access' sections. Under 'Check access', there are tabs for 'Role assignments', 'Roles', 'Deny assignments', and 'Classic administrators'. The 'My access' section shows a 'View my access' button. To the right, there are three panels: 'Grant access to this resource' (with a 'Add role assignment' button), 'View access to this resource' (with a 'View' button), and a 'Learn more' link for both. A search bar at the top is set to 'Search (Ctrl+/'.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

Microsoft Azure Search resources, services, and docs (G+/)

Home > Azure Databricks > projectworkspace | Access control (IAM) > Add role assignment ...

Got feedback?

Role Members * Review + assign

Selected role Contributor

Assign access to User, group, or service principal Managed identity

Members + Select members

Name	Object ID	Type
No members selected		

Description Optional

Review + assign Previous Next

Select managed identities

Got feedback?

Subscription * Azure Pass - Sponsorship

Managed identity Data factory (V2) (2)

Select ... Search by name

kirandfactory97 /subscriptions/2a3a8c67-30a1-47b8-8bcb-43ed2b5cf86c/resourceGroups/labdem...

Selected members: projectdf97 /subscriptions/2a3a8c67-30a1-47b8-8bcb-43ed2b5cf86c/resourceGroups/... Remove

Select Close

Microsoft Azure Search resources, services, and docs (G+/)

Home > Azure Databricks > projectworkspace | Access control (IAM) > Add role assignment ...

Got feedback?

Role Members * Review + assign

Selected role Contributor

Assign access to User, group, or service principal Managed identity

Members + Select members

Name	Object ID	Type
projectdf97	7e4c38ff-c85a-4e54-9c50-cc073b228197	Data factory (V2) <input type="radio"/>

Description Optional

Review + assign Previous Next

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the 'User Settings' page in the Microsoft Azure portal. The 'Access tokens' tab is selected. A modal window titled 'Generate new token' is open, prompting for a 'Comment' (set to 'project') and a 'Lifetime (days)' (set to 90). There are 'Cancel' and 'Generate' buttons at the bottom.

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines (pipeline1, pipeline2, pipeline3), Datasets, Data flows, and Power Query. In the center, under 'Activities', a 'Copy data' activity is selected, pointing to a 'sqltabletoADLS' destination. On the right, a 'New linked service' dialog is open for 'Azure Databricks'. It includes fields for 'Name' (AzureDatabricks1), 'Description', 'Connect via integration runtime' (AutoResolveIntegrationRuntime), 'Account selection method' (From Azure subscription), 'Azure subscription' (Azure Pass - Sponsorship), 'Databricks workspace' (projectworkspace), 'Select cluster' (Existing interactive cluster), and a 'Create' button. A 'Test connection' link is also present.

This screenshot shows the same 'New linked service' dialog for 'Azure Databricks', but with more detailed configuration. The 'Databricks workspace' is set to 'projectworkspace'. Under 'Select cluster', the 'Existing interactive cluster' option is selected. The 'Databrick Workspace URL' is set to 'https://adb-8934044172288797.17.azuredatabricks.net'. The 'Authentication type' is set to 'AccessToken', and the 'Access token' field contains a redacted token value. Other sections like 'Choose from existing clusters' and 'Annotations' are also visible.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The image consists of three vertically stacked screenshots of the Microsoft Azure Data Factory Studio interface.

Screenshot 1: Shows the 'Factory Resources' sidebar with 'Pipelines' expanded, containing 'pipeline1', 'pipeline2', and 'pipeline3'. The 'Activities' pane shows a 'Copy data' activity named 'sqltabletoADLS' selected. A 'Browse' dialog is open, showing a file tree under 'Root folder > Users > kiran.alvaa@outlook.com' with a single item named 'project'. Buttons for 'OK' and 'Cancel' are at the bottom right of the dialog.

Screenshot 2: Shows the 'Factory Resources' sidebar with 'Pipelines' expanded, containing 'pipeline1', 'pipeline2', and 'pipeline3'. The 'Activities' pane shows a 'Copy data' activity named 'sqltabletoADLSSGen2' connected to a 'Notebook' activity named 'Notebook1'. The 'Properties' pane on the right shows the 'General' tab for 'Notebook1' with 'Name' set to 'pipeline3' and 'Notebook path' set to '/Users/kiran.alvaa@outlook.com/project'. Buttons for 'Validate', 'Debug', and 'Add trigger' are at the top of the activities pane.

Screenshot 3: Shows the 'Factory Resources' sidebar with 'Datasets' expanded, containing 'AzureSqlTable1', 'DelimitedText2', 'JoinedTable', 'movies', and 'moviestable'. The 'Activities' pane shows a 'DelimitedText' activity named 'moviestable'. The 'Properties' pane on the right shows the 'General' tab for 'moviestable' with 'Name' set to 'moviestable'. The 'Connection' tab shows 'Linked service' set to 'AzureDataLakeStorage1' and 'File path' set to 'moviestable / Directory / [fromblob]'. The 'Schema' and 'Parameters' tabs are also visible.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the 'Access keys' section of the Azure Storage account settings for 'padlsgen2'. It displays two keys: 'key1' and 'key2'. Key1 was last rotated 5 days ago and its value is partially visible as '9XL1BGsKKW3hg3FdHkwuq27mrODUKfoyROT...'. Key2 was also last rotated 5 days ago and its value is partially visible as '9XL1BGsKKW3hg3FdHkwuq27mrODUKfoyROT...'. A 'Show' button is available for both keys.

The screenshot shows a Databricks notebook titled 'project' in Python. The code in the notebook is as follows:

```
# Use Access Key
spark.conf.set("fs.azure.account.auth.type", "padlsgen2.dfs.core.windows.net", "SharedKey")
spark.conf.set("fs.azure.account.key", "padlsgen2.dfs.core.windows.net", "9XL1BGsKKW3hg3FdHkwuq27mrODUKfoyROT...")

dataFrame1 = spark.read.option('header', True).option("inferSchema", True).csv("abfss://moviestable@padlsgen2.dfs.core.windows.net/fromblob/")
dataFrame1.registerTempTable("movies")

dfresult= spark.sql("select rank() over(order by Genre) as GenreRank, * from movies")
dfresult.write.option("header", True).csv("abfss://moviestable@padlsgen2.dfs.core.windows.net/fro mdatabricks1")
```

```
dataFrame1=
spark.read.option('header',True).option("inferSchema",True).csv("abfss://moviestable@padlsgen2.d
fs.core.windows.net/fromblob/")

dataFrame1.registerTempTable("movies")

dfresult= spark.sql("select rank() over(order by Genre) as GenreRank, * from movies")

dfresult.write.option("header",True).csv("abfss://moviestable@padlsgen2.dfs.core.windows.net/fro
mdatabricks1")
```

3. Create a final activity that will read the output of previous activity in ADLS and store the same in Synapse.

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines (pipeline1, pipeline2, pipeline3), Datasets (AzureSqlTable1, DelimitedText2, JoinedTable, movies, moviestable), Data flows (dataflow1), and Power Query. Pipeline 3 is selected. The main workspace displays a pipeline diagram with three activities: 'Copy data' (source: sqltabletoADLSgen2, sink: Notebook, named 'Notebook1'), 'Notebook' (represented by a red cube icon), and another 'Copy data' activity (source: Notebook, sink: adlsgen2tossynapse, named 'Copy data'). The 'Properties' panel on the right shows the pipeline is named 'pipeline3'. The 'Source' tab of the pipeline properties is selected, showing a dropdown menu for 'Source dataset' with options like 'Select...' and '+ New'.

This screenshot shows the 'New dataset' dialog box overlaid on the Data Factory Studio interface. The dialog title is 'New dataset' and it instructs the user to 'Select a data store'. It includes a search bar and a grid of data store icons categorized under 'All', 'Azure', 'Database', 'File', 'Generic protocol', 'NoSQL', and 'Services and apps'. The 'Azure' category is selected. Visible data stores include Azure Cosmos DB (SQL API), Azure Data Explorer (Kusto), Azure Data Lake Storage Gen1, Azure Data Lake Storage Gen2, Azure Database for MariaDB, Azure Database for MySQL, and others. At the bottom of the dialog are 'Continue' and 'Cancel' buttons.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines (pipeline1, pipeline2, pipeline3), Datasets (AzureSqlTable1, DelimitedText2, JoinedTable, movies, moviestable), Data flows (dataflow1), and Power Query. The main workspace displays a pipeline named 'pipeline2' with two steps: 'Copy data' and 'Notebook'. The 'Copy data' step has its 'Source' tab selected. To the right, a 'Browse' dialog box is open, showing a file tree under the root folder 'adfstagedcopytempdata'. The item 'moviestable' is highlighted. At the bottom of the dialog are 'OK' and 'Cancel' buttons.

The screenshot shows the 'Set properties' dialog box for the 'Copy data' step in pipeline2. The 'Source' tab is selected. The 'Name' field is set to 'fromdatabricks', the 'Linked service' is set to 'AzureDataLakeStorage1', and the 'File path' is set to 'moviestable / Directory / fromdatabricks'. The 'First row as header' checkbox is checked. The 'Import schema' section shows 'From connection/store' selected. At the bottom are 'OK', 'Back', and 'Cancel' buttons.

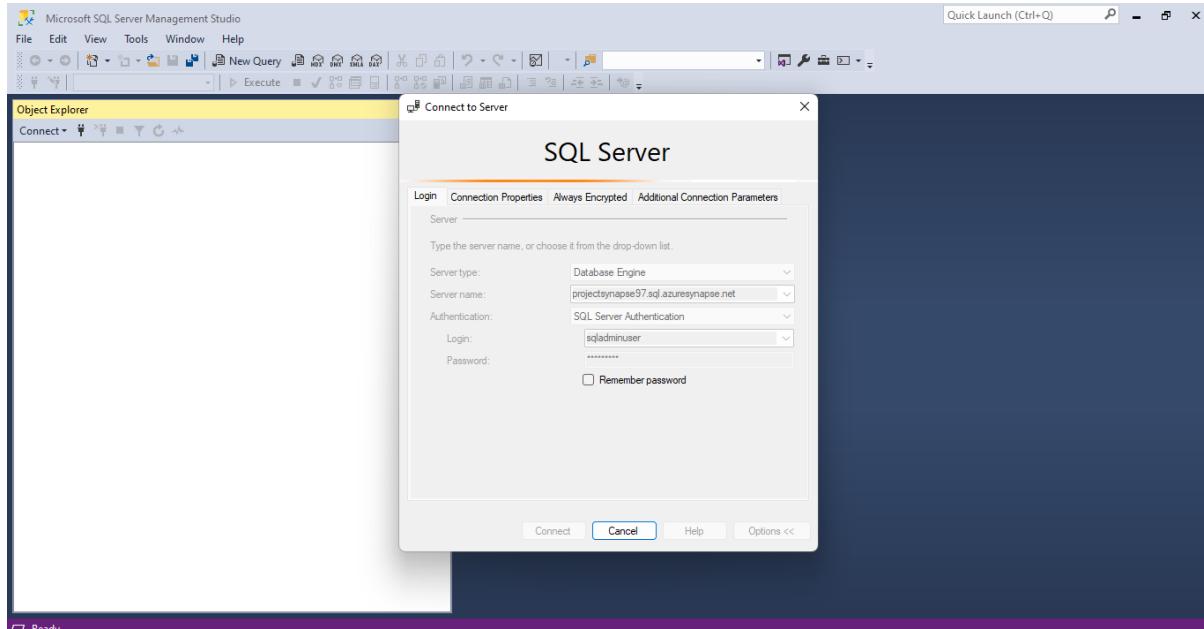
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (pipeline1, pipeline2, pipeline3), 'Datasets' (AzureSqlTable1, DelimitedText2, fromdatabricks, JoinedTable, movies, moviestable), 'Data flows' (dataflow1), and 'Power Query'. In the main workspace, pipeline2 is selected. A 'Copy data' activity is connected to a 'Notebook' activity. The 'Sink' tab is active, showing a 'Sink dataset' dropdown with 'Select...' button. To the right, a 'New dataset' dialog is open, titled 'New dataset'. It contains a search bar and tabs for 'All', 'Azure', 'Database', 'File', 'Generic protocol', 'NoSQL', and 'Services and apps'. Under 'Azure', 'Azure Synapse Analytics' is selected and highlighted with a blue border. Other options include 'Azure Search' and 'Azure Table Storage'. At the bottom of the dialog are 'Continue' and 'Cancel' buttons.

The screenshot shows the Microsoft Azure Data Factory Studio interface. The left sidebar shows 'Factory Resources' with pipelines pipeline1, pipeline2, and pipeline3, and datasets AzureSqlTable1, DelimitedText2, fromdatabricks, JoinedTable, movies, and moviestable. Pipeline pipeline3 is selected. In the main workspace, a 'Copy data' activity is connected to a 'Notebook' activity. The 'Sink' tab is active, showing a 'Sink dataset' dropdown with 'Select...' button. To the right, a 'New linked service' dialog is open, titled 'New linked service'. It shows 'Azure Synapse Analytics' selected as the provider. The 'Name' field is set to 'AzureSynapseAnalytics1'. The 'Description' field is empty. Under 'Connect via integration runtime', 'AutoResolveIntegrationRuntime' is selected. The 'Connection string' tab is active, showing 'Azure Key Vault'. The 'Account selection method' section has 'From Azure subscription' selected. The 'Azure subscription' dropdown shows 'Azure Pass - Sponsorship (2a3a8c67-30a1-47b8-8bcb-43ed2b5cf86c)'. The 'Server name' dropdown shows 'projectsynapse97 (Synapse workspace)'. The 'Database name' dropdown is empty. At the bottom are 'Create' and 'Cancel' buttons, and a 'Test connection' link.

The screenshot shows the Microsoft Azure Data Factory Studio interface. The left sidebar shows 'Factory Resources' with pipelines pipeline1, pipeline2, and pipeline3, and datasets AzureSqlTable1, DelimitedText2, fromdatabricks, JoinedTable, movies, and moviestable. Pipeline pipeline3 is selected. In the main workspace, a 'Copy data' activity is connected to a 'Notebook' activity. The 'Sink' tab is active, showing a 'Sink dataset' dropdown with 'Select...' button. To the right, a 'New linked service' dialog is open, titled 'New linked service'. It shows 'Azure Synapse Analytics' selected as the provider. The 'Server name' field is set to 'projectsynapse97 (Synapse workspace)'. The 'Database name' field is set to 'project'. The 'SQL pool' field is set to 'project'. The 'Authentication type' field is set to 'SQL authentication'. The 'User name' field is set to 'projectsunapse97'. The 'Password' tab is active, showing a masked password field. The 'Additional connection properties' section is collapsed. At the bottom are 'Create' and 'Cancel' buttons, and a 'Test connection' link.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



```
create table movies (
GenreRank int,
Genre varchar(200),
GenreCount int,
Film varchar(200),
LeadStudio varchar(200),
AudienceScore int,
Profitability varchar(200),
RottenTomatoes int,
WorldWideGross varchar(200),
```

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

[Year] int

)

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The title bar indicates the connection is to 'SQLQuery1.sql - projectsynapse97.sql.azuresynapse.net.project (sqladminuser (117)) - Microsoft SQL Server Management Studio'. The Object Explorer sidebar shows the database 'projectsynapse97.sql.azuresynapse.net (SQL Server 12.0.2000.8 - sqladminuser)' selected, with its databases, security, and integration services catalogs listed. The main pane displays a T-SQL script for creating a 'movies' table:

```
create table movies (
    GenreRank int,
    Genre varchar(200),
    GenreCount int,
    Film varchar(200),
    LeadActor varchar(200),
    AudienceScore varchar(200),
    Profitability int,
    WorldwideGross varchar(200),
    [Year] int
)
```

Below the script, the 'Messages' pane shows the command completed successfully with a completion time of 2022-08-01T10:32:56.0666184+05:00. The status bar at the bottom right shows the project name 'projectsynapse97.sql.azures...', user 'sqladminuser (117)', and duration '00:00:01 | 0 rows'.

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The title bar indicates the connection is to 'SQLQuery1.sql - projectsynapse97.sql.azuresynapse.net.project (sqladminuser (117))* - Microsoft SQL Server Management Studio'. The ribbon menu includes File, Edit, View, Query, Project, Tools, Window, Help. The toolbar has various icons for file operations like Open, Save, Print, and Database. The Object Explorer on the left shows the database structure: 'projectsynapse97.sql.azuresynapse.net (SQL Server 12.0.2000.8 - sqladminuser)' with Databases, System Databases, project, Security, and Integration Services Catalogs. The main pane displays a query window titled 'SQLQuery1.sql - pr.sqladminuser (117)*' containing the following T-SQL code:

```
create table movies (
    GenreRank int,
    Genre varchar(200),
    GenreCount int,
    Film varchar(200),
    LeadStudio varchar(200),
    AudienceScore int,
    Profitability varchar(200),
    RottenTomatoes int,
    WorldWideGross varchar(200),
    [Year] int

select * from movies
```

The results pane at the bottom shows the data from the 'movies' table with the following columns: GenreRank, Genre, GenreCount, Film, LeadStudio, AudienceScore, Profitability, RottenTomatoes, WorldWideGross, and Year. The results grid has 82% scroll position.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in.

Factory Resources

- Pipelines
 - pipeline1
 - pipeline2
 - pipeline3**
- Datasets
 - AzureSqlTable1
 - DelimitedText2
 - fromdatabricks
 - JoinedTable
 - movies
 - moviestable
- Data flows
 - dataflow1
- Power Query

Search

Validate all Publish all 3

pipeline1 dataflow1

Validate Validate copy runtime Debug

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

Search

All Azure Database File Generic protocol NoSQL Services and apps

Azure File Storage	Azure SQL Database	Azure SQL Database Managed Instance
Azure Search	Azure Synapse Analytics	Azure Table Storage

Continue Cancel

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in.

Factory Resources

- Pipelines
 - pipeline1
 - pipeline2
 - pipeline3**
- Datasets
 - AzureSqlTable1
 - DelimitedText2
 - fromdatabricks
 - JoinedTable
 - movies
 - moviestable
- Data flows
 - dataflow1
- Power Query

Search

Validate all Publish all 3

pipeline1 dataflow1

Validate Validate copy runtime Debug

Edit linked service

Azure Synapse Analytics [Learn more](#)

Connection string Azure Key Vault

Account selection method

From Azure subscription Enter manually

Fully qualified domain name * projectsynapse97.sql.azuresynapse.net

Database name * project

Authentication type * SQL authentication

User name * sqldadminuser

Password [Azure Key Vault](#)

Additional connection properties

Apply Cancel Test connection

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in.

Factory Resources

- Pipelines
 - pipeline1
 - pipeline2
 - pipeline3**
- Datasets
 - AzureSqlTable1
 - DelimitedText2
 - fromdatabricks
 - JoinedTable
 - movies
 - moviestable
- Data flows
 - dataflow1
- Power Query

Search

Validate all Publish all 3

pipeline1 dataflow1

Validate Validate copy runtime Debug

Set properties

Name AzureSynapseAnalyticsTable1

Linked service * AzureSynapseAnalytics1

Table name dbo.movies

Import schema From connection/store None

Advanced

OK Back Cancel

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in

Data Factory Publishing (4)

Factory Resources

- Pipelines
 - pipeline1
 - pipeline2
 - pipeline3
- Datasets
 - AzureSqlTable1
 - AzureSynapseAnalyticsTable1
 - DelimitedText2
 - fromdatabricks
 - JoinedTable
 - movies
 - moviestable
- Data flows
 - dataflow1

Validate all Publishing (4)

dataflow1 pipeline3

Validate Validate copy runtime Debug

Copy data Notebook

sqltabletoADLSGen2 Notebook1

General Source Sink Mapping Settings

Source dataset * fromdatabricks

Open New Preview data Learn more

File path type File path in database

Wildcardmoviestable / fromdatabricks

Publish Cancel

Publish all

You are about to publish all pending changes to the live environment. [Learn more](#)

Pending changes (4)

Name	Change	Existing
Pipelines		-
pipeline3	(New)	-
Datasets		-
moviestable	(New)	-
fromdatabricks	(New)	-
AzureSynapseAnalyticsTab...	(New)	-

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in

All pipeline runs > pipeline3 - Activity runs

pipeline3

List Gantt

Rerun Rerun from activity Rerun from failed activity Refresh Update pipeline

Copy data Notebook Copy data

sqltabletoADLSGen2 Notebook1 adlsgen2tosynapse

+ - []

Activity runs

Pipeline run ID: 26732e1f-d6dc-47bb-83f9-73e0bb5355ce

All status

Showing 1 - 1 of 1 items

Microsoft Azure | projectdf97

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in

All pipeline runs > pipeline3 - Activity runs

pipeline3

List Gantt

Rerun Rerun from activity Rerun from failed activity Refresh Update pipeline

Copy data Notebook Copy data

sqltabletoADLSGen2 Notebook1 adlsgen2tosynapse

+ - []

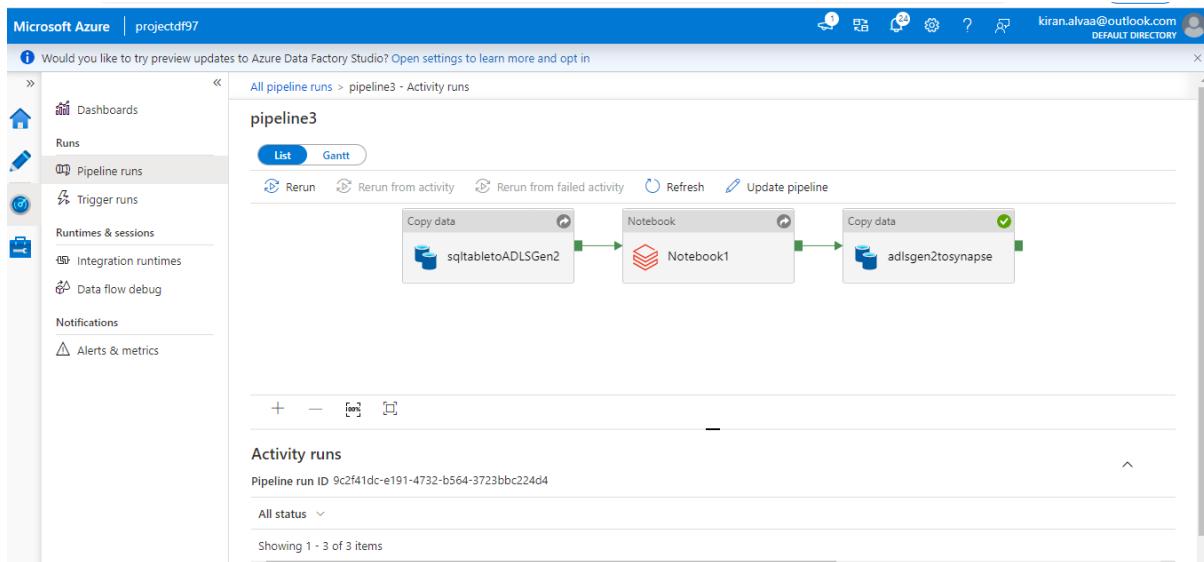
Activity runs

Pipeline run ID: 26732e1f-d6dc-47bb-83f9-73e0bb5355ce

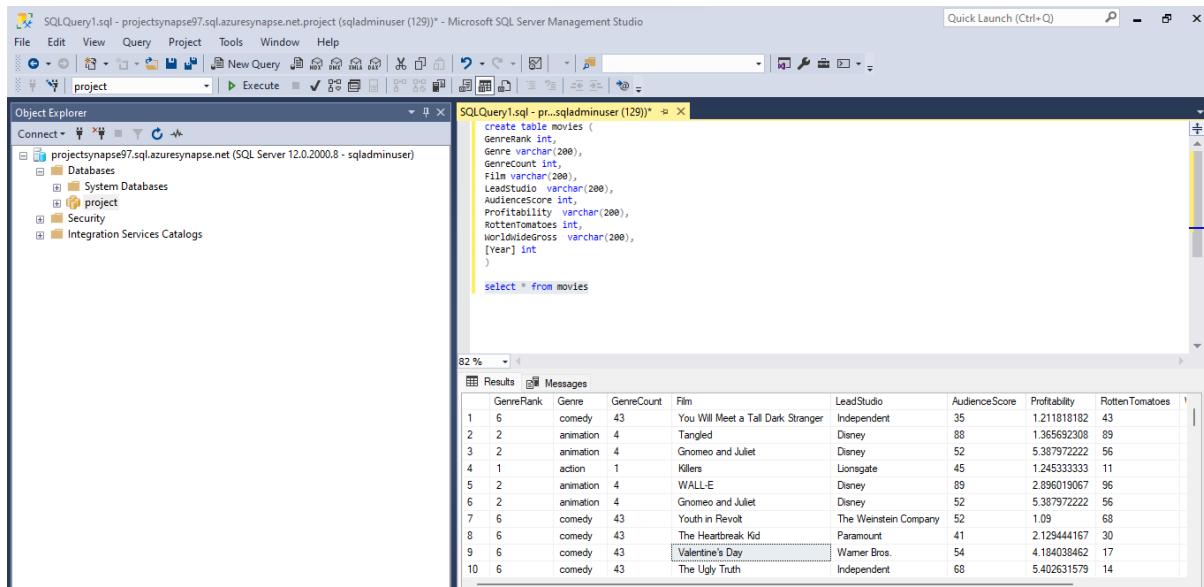
All status

Showing 1 - 2 of 2 items

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, the navigation pane is open with 'Pipeline runs' selected. The main area displays the 'pipeline3' pipeline. It consists of three activities connected sequentially: 'Copy data' (source: 'sqltabletoADLSGen2'), 'Notebook' (name: 'Notebook1'), and another 'Copy data' (sink: 'adlsgen2tosynapse'). The 'Notebook1' activity has a green checkmark indicating success. Below the pipeline, the 'Activity runs' section shows one completed run for the 'Notebook1' activity.



The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The Object Explorer on the left shows a database named 'projectsynapse97.sql.azuresynapse.net.project' containing a 'movies' table. A query window titled 'SQLQuery1.sql - projectsynapse97.sql.azuresynapse.net.project (sqladminuser (129))' is open, displaying the following SQL code:

```
GenreRank int,
Genre varchar(200),
GenreCount int,
Film varchar(200),
LeadStudio varchar(200),
AudienceScore int,
RottenTomatoes int,
WorldwideGross varchar(200),
[year] int
)
select * from movies
```

The results grid shows the following data for the first 10 rows:

	GenreRank	Genre	GenreCount	Film	LeadStudio	AudienceScore	Profitability	RottenTomatoes
1	6	comedy	43	You Will Meet a Tall Dark Stranger	Independent	35	1.21181812	43
2	2	animation	4	Tangled	Disney	88	1.355692308	89
3	2	animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56
4	1	action	1	Killers	Lionsgate	45	1.245333333	11
5	2	animation	4	WALL·E	Disney	89	2.996019067	96
6	2	animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56
7	6	comedy	43	Youth in Revolt	The Weinstein Company	52	1.09	68
8	6	comedy	43	The Heartbreak Kid	Paramount	41	2.125444167	30
9	6	comedy	43	Valentine's Day	Warner Bros.	54	4.184038462	17
10	6	comedy	43	The Ugly Truth	Independent	68	5.402631579	14