Prerana Annapantula

Azure Data Catalog

1. What is Azure Data Catalog?

Metadata management tool to help organizations find and manage large amounts of data. Users can discover, understand, and consume data sources.

1. What are the Challenges faced by Producers and Consumers of data?

|  |  |
| --- | --- |
| Producers | Consumers |
| Annotating data sources with descriptive metadata is often a lost effort. Client applications typically ignore descriptions that are stored in the data source. | Users may not know that a data source exists until contact. No central location where the data sources are registered. |
| Creating documentation for data sources is often a lost effort. Keeping documentation in sync with data sources is an ongoing responsibility, and users might lack trust in documentation that’s perceived as being out of date. | Users do not know the location of the data source. They cannot connect to the data by using a client application/connection string or path. |
| Creating and maintaining documentation for data sources is complex and time-consuming. Making that documentation readily available to everyone who uses the data source can be even more so. | Data source’s documentation – they cannot understand the intended uses of the data. Data sources and documentation might live in a variety of places and be consumed through a variety of experiences. |
| Restricting access to data sources and ensuring that data consumers know how to request access is an ongoing challenge. | If users have questions about an information asset, they may not be able to find the expert or team that’s responsible. There is no explicit connection between data and those with expert perspectives on its use. |
|  | Unless users understand the process for requesting access to the data source, discovering the data source and its documentation still does not help them access the data. |

1. List the Capabilities of the Azure Data Catalog with flow diagram.

Discover 🡪 Understand 🡪 Consume 🡪 Contribute

Discover by searching, browsing, and filtering.

Understand the metadata, who the experts may be, and what the content is.

Consume your data with your tools the way you want to.

Contribute to the data catalog by tagging, documenting, and publishing.

1. What is Registration?

Registration is the process of extracting the metadata from the data source and copying it into the Data Catalog service. It is the first step to making the data source discoverable via Data Catalog.

1. What are the different types of information extracted in Registration?

Structural Metadata – physical data about the objects

Descriptive Metadata – tags and identity of experts

Preview – a sample of the data

Data Profiles – aggregate statistics about the data (more about the table/schema itself)

1. What are the ways you can discover the data in Azure Data Catalog?

Searching – similar to using keywords to find what you’re looking for

Filtering – selecting specific characteristics to limit/complement the search – experts/tags/object type/source type

1. Provide an example of Syntactical search?

Basic Search: “Sales data”

Property Scoping: “Name: finance”

Boolean Operators: “Finance NOT Marketing”

Grouping with Parenthesis: “Name: finance AND (tags:F10 NOT tags:F1)”

Comparison operators: “Modifiedtime < “03/22/2019””

1. What are annotations and give examples.

Annotations are people’s own points of view/understanding of specific data sources.

Friendly names – given at a data asset level to make the object name more understandable

Description – given at data asset and attribute/column levels to describe their perspective on the use of the data asset

Tags: given at data asset and attribute/column levels; User tags – Categorizes data assets or attributes; Glossary tags – categorize data assets

Experts – given at data asset level to allow contact with expert users or groups with experts who discovered the data sources but users cannot understand the complete data

Request access – users want to have access to a certain data source but don’t have permissions

Documentation – rich text information with links and pictures that aren’t providing the same information as tags.

1. Can we annotate multiple assets? If so, how?

Yes – we can annotate some selected assets in a single operation. For multiple tables/views only the columns that all the selected tables have in common will be displayed in the portal.

1. What is preview and Data Profile? Why do we need them?

Preview is seeing a sample of the data asset. It is necessary to get an understanding about how the data looks and works. Since a preview gives a sample, the user can figure out what the data asset is about/used for, with the multiple rows of data that they’re shown.

Data profile is the aggregate statistics about the data asset. It essentially provides some overall information such as mean, median, maximum, of the data asset. It’ll group it in a way to allow the user to get an understanding of the data itself. Also there’s a column data profile and object data profile to help get an understanding of either the specific columns or the data as a whole.

Azure Data Factory

1. What is Azure Data Factory?

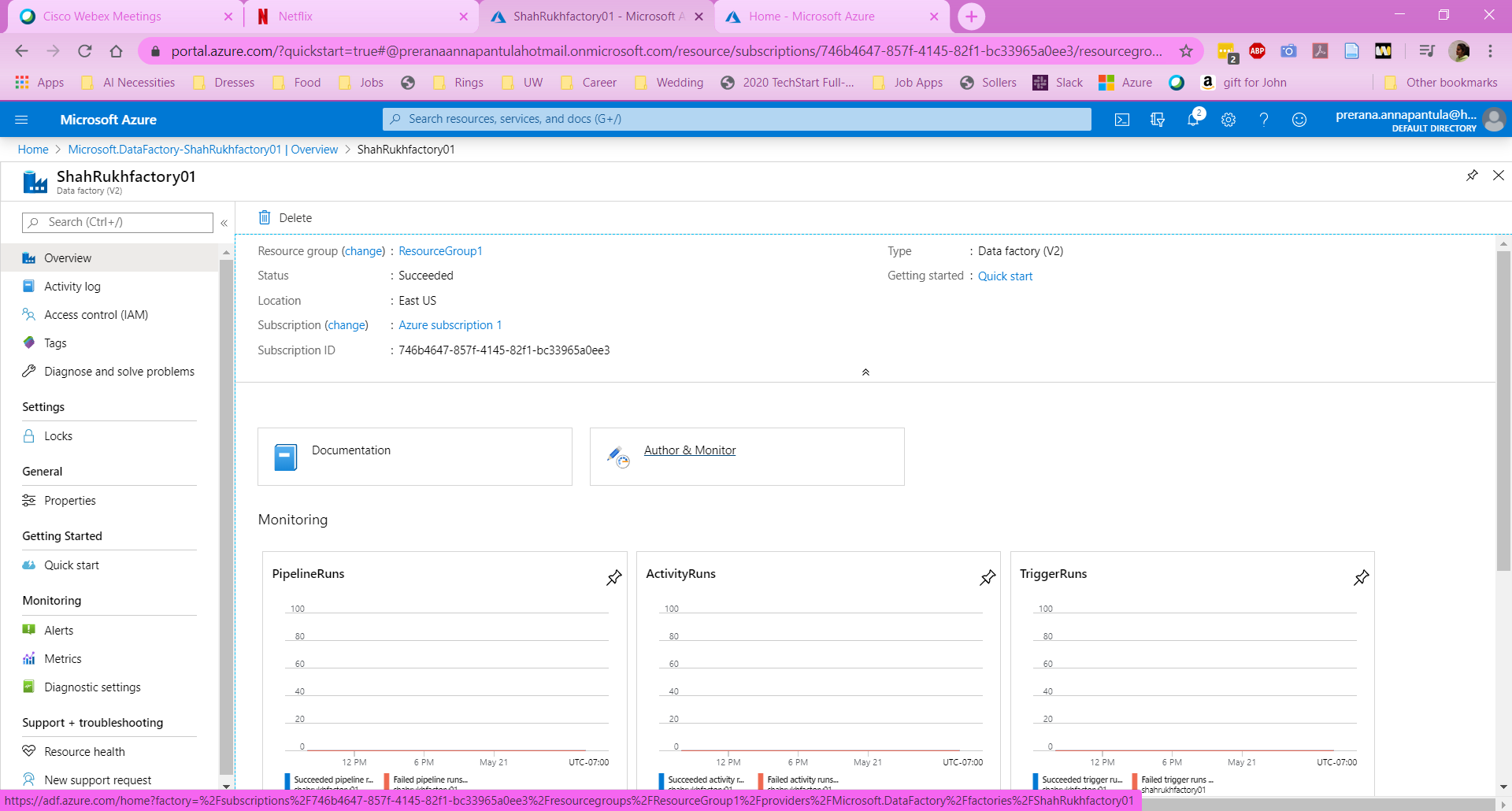
Azure Data factory is the cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.

1. How transactional processing is different from Analytical processing?

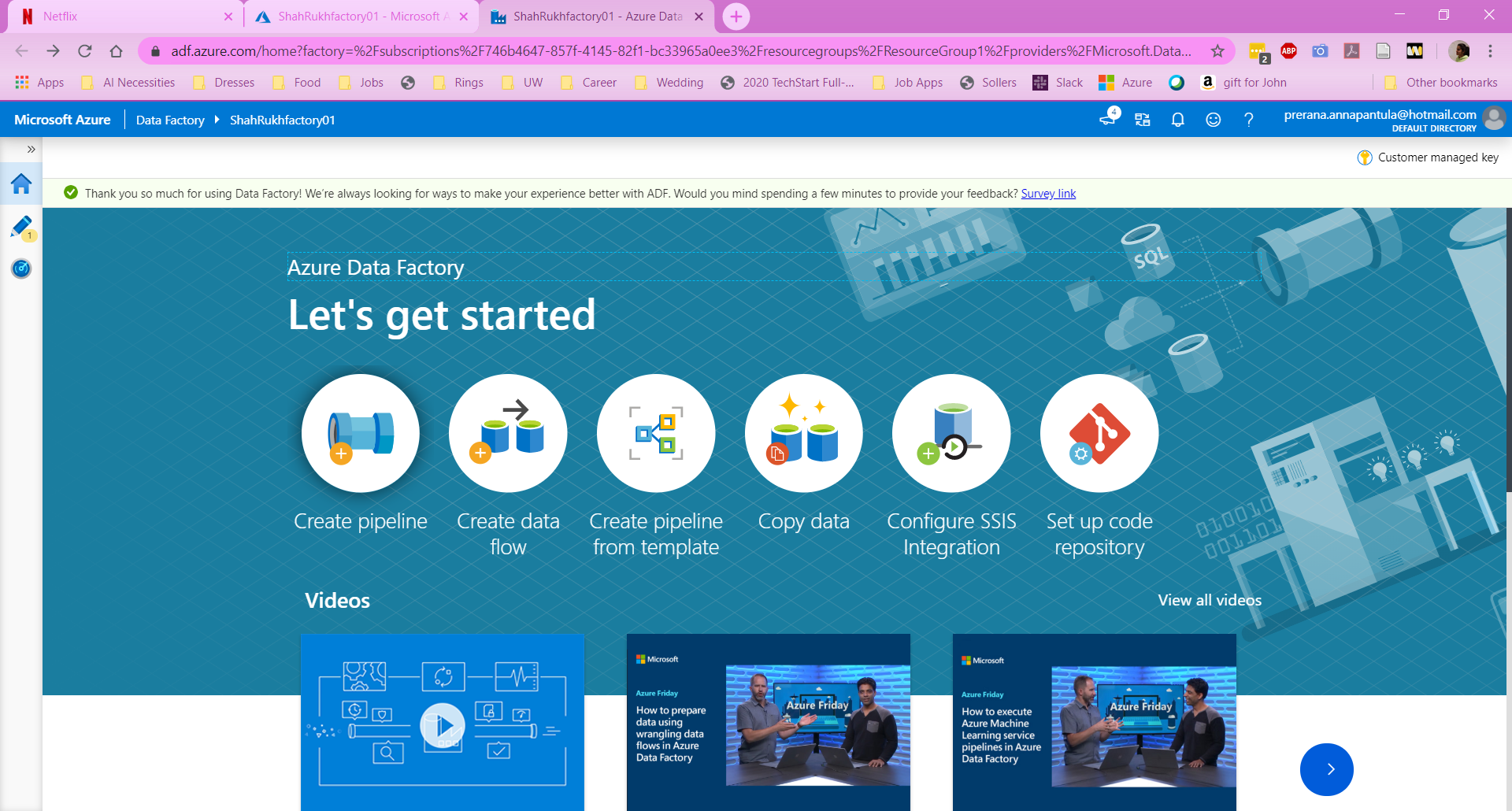
Transactional processing is based on manipulating specific rows of data. Analytical processing is based on mainly just analyzing large groups of data.

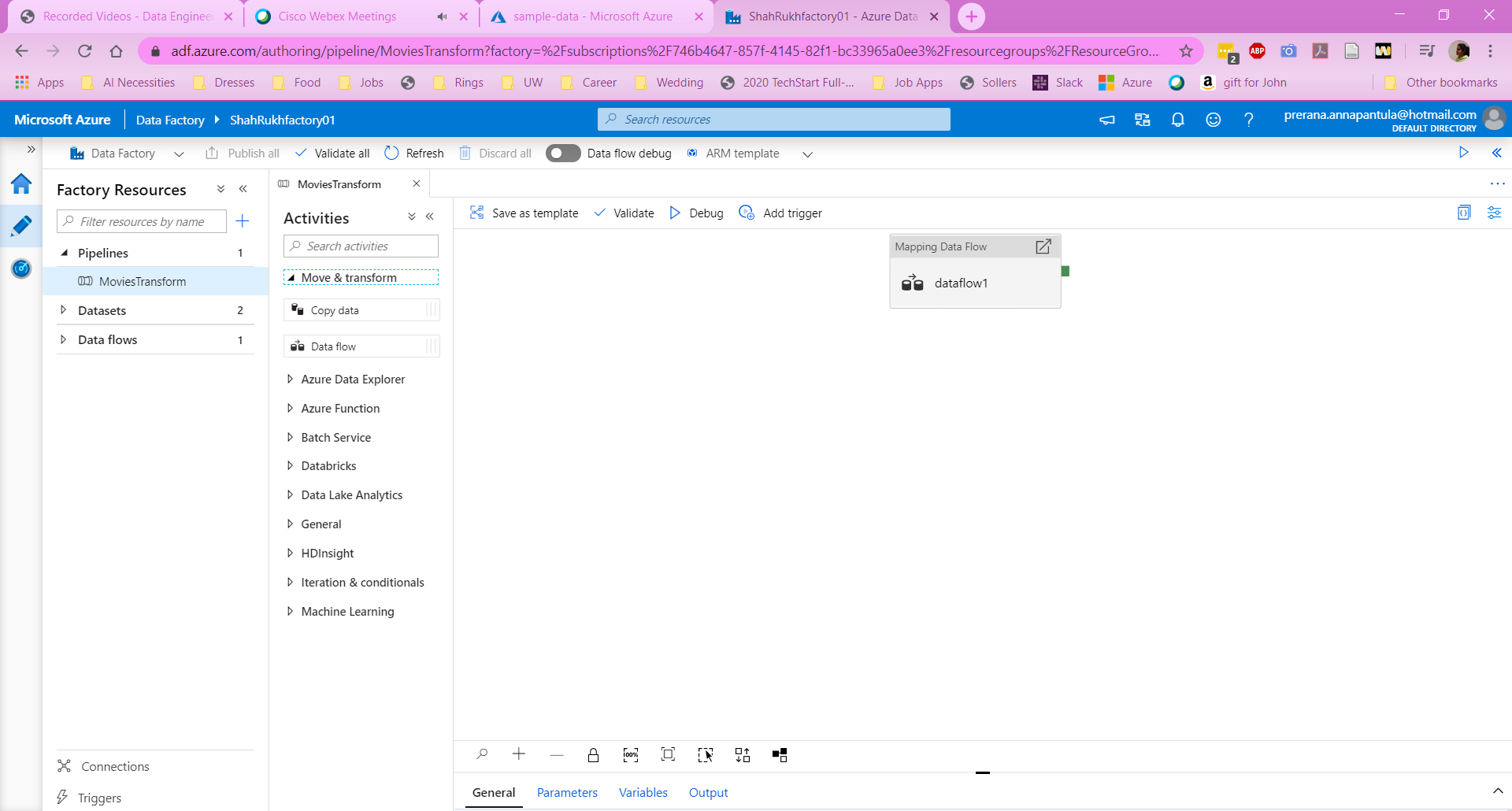
1. Please create a Pipeline from the scratch to move and transform the data as we did for MoviesDB file. Please use the Filter and Aggregate transformation. Please document each and every step and elaborate.

Step 1: Click on “Author & Monitor” to be able to create a pipeline within the Data Factory.

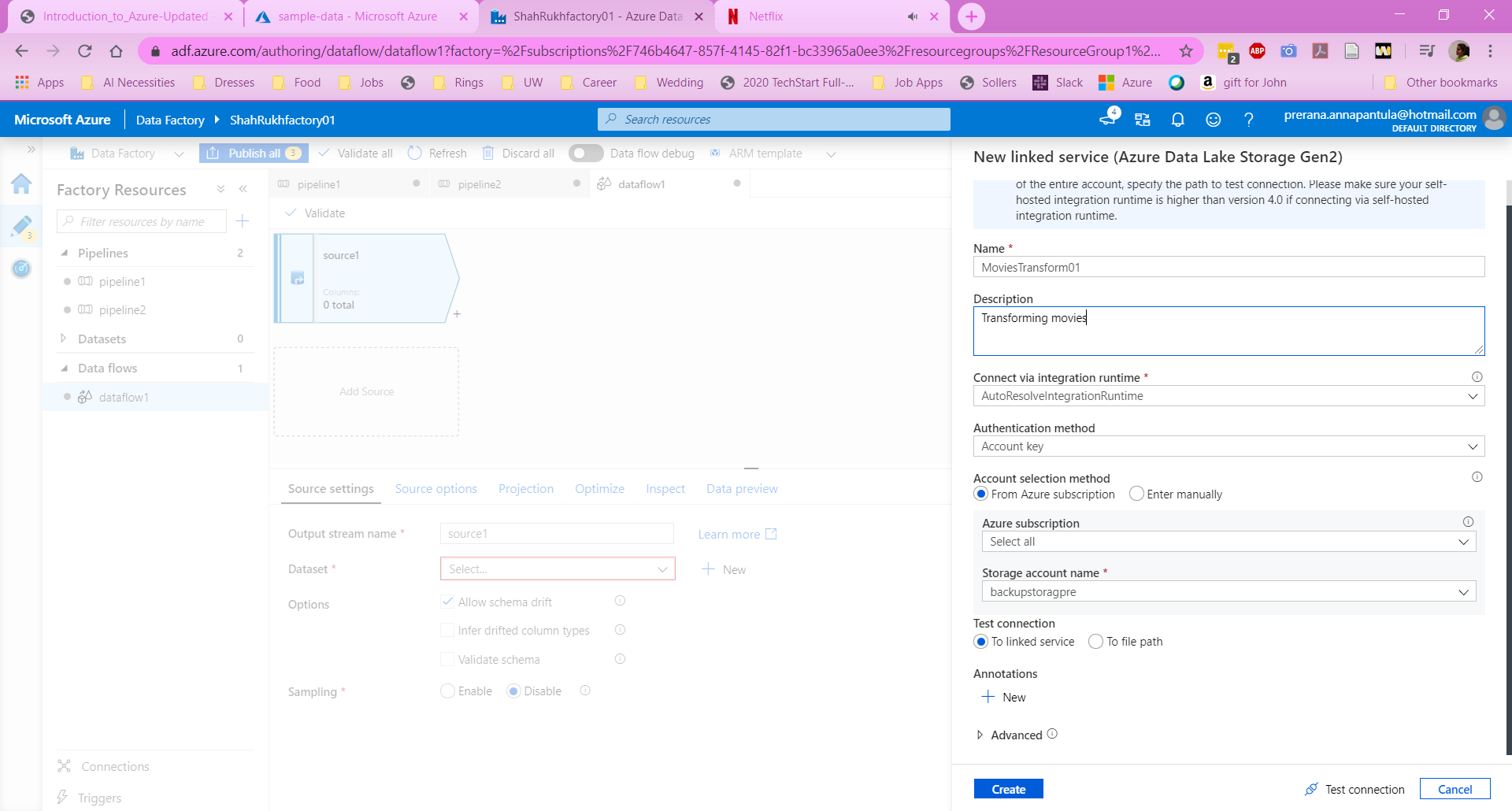


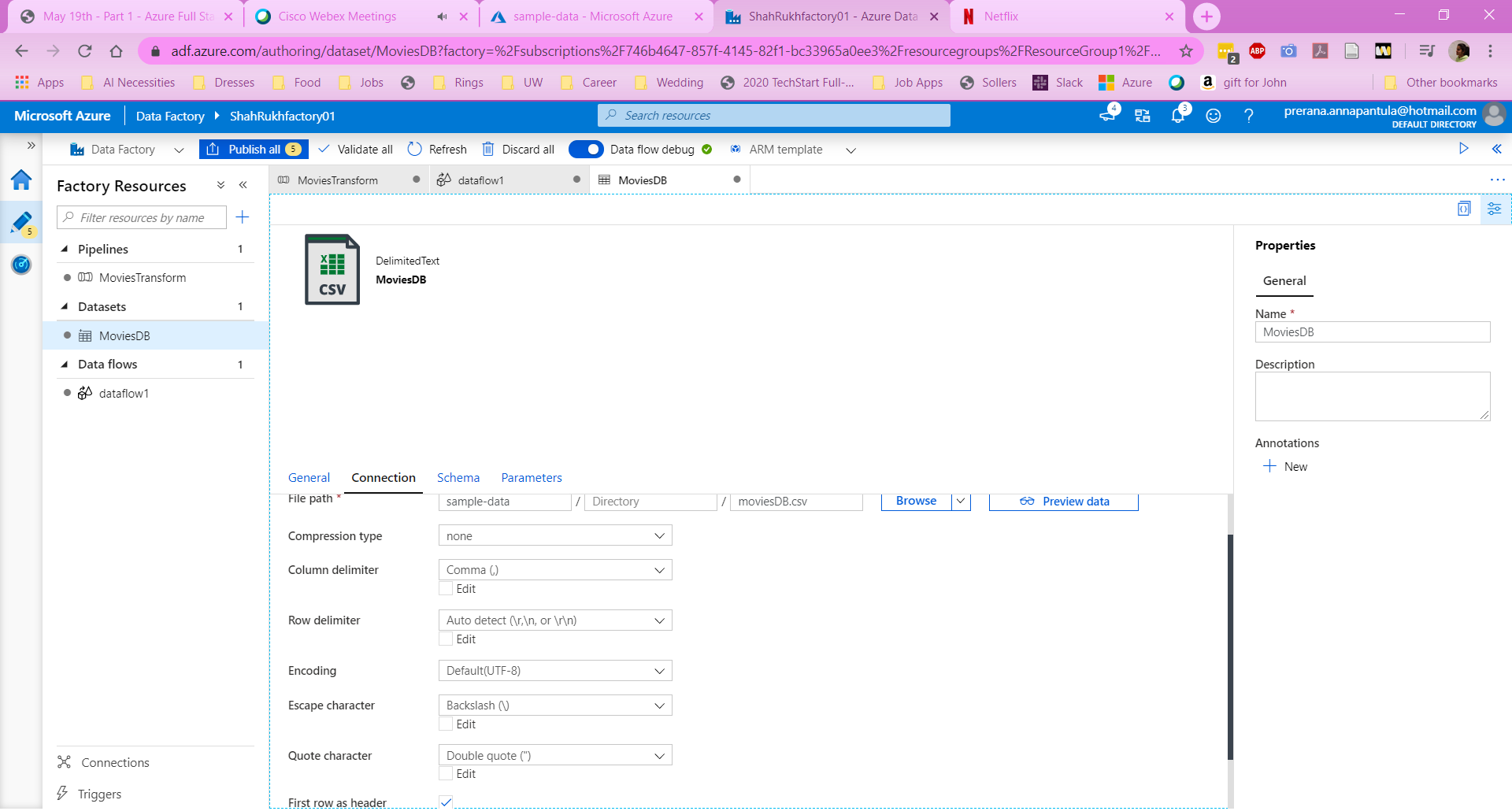
Step 2: Click on “Create pipeline” to create a new pipeline.

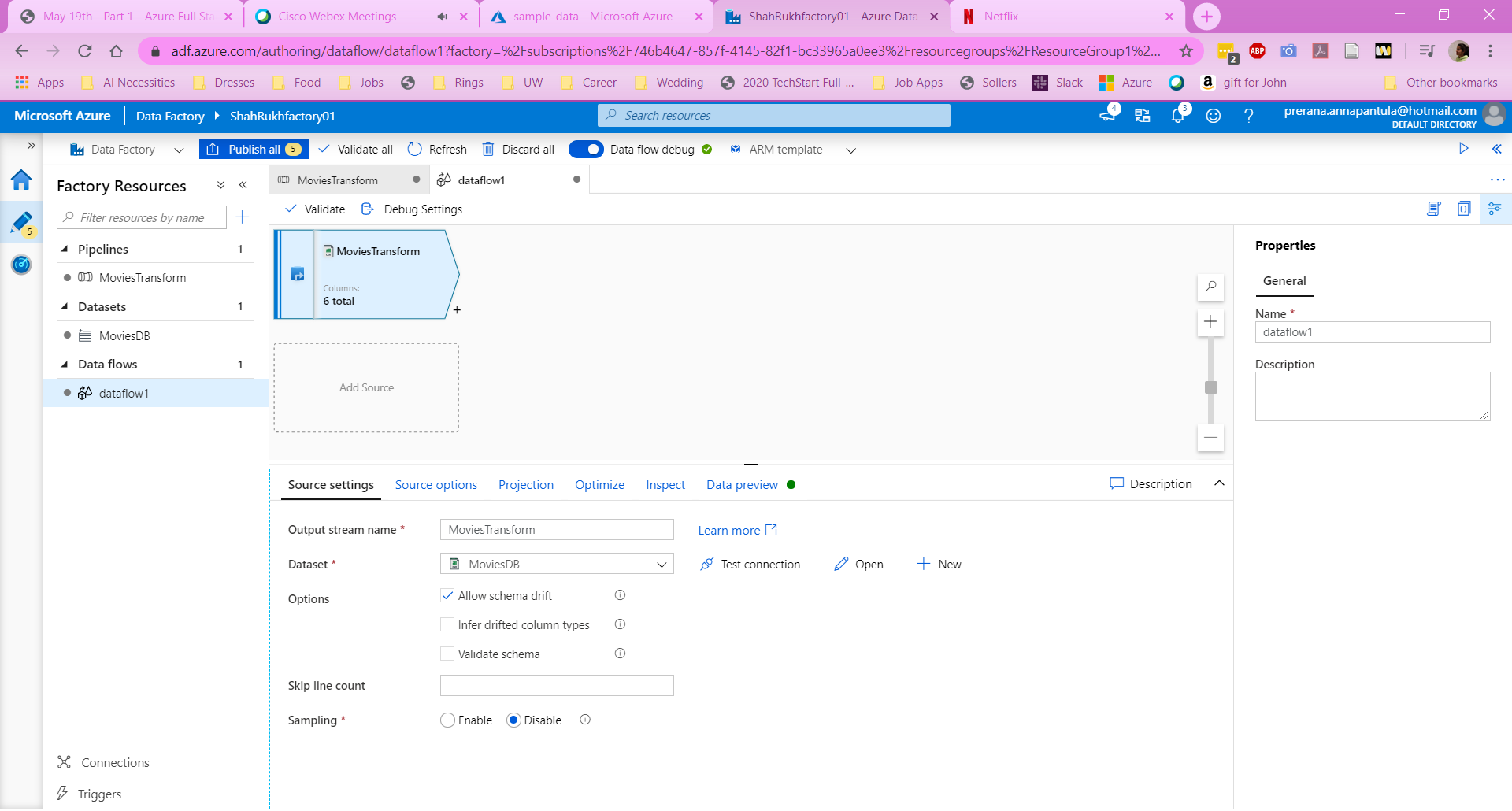


Step 3: Create a Pipeline by dragging “Data Flow” under “Move & Transform” into the white area. In the pop-up, select “Create new data flow” and then select “Mapping Data Flow.” This will create a new data flow within the pipeline. Click on Data flow debug to make sure that the data does flow throughout the making/usage of the pipeline.

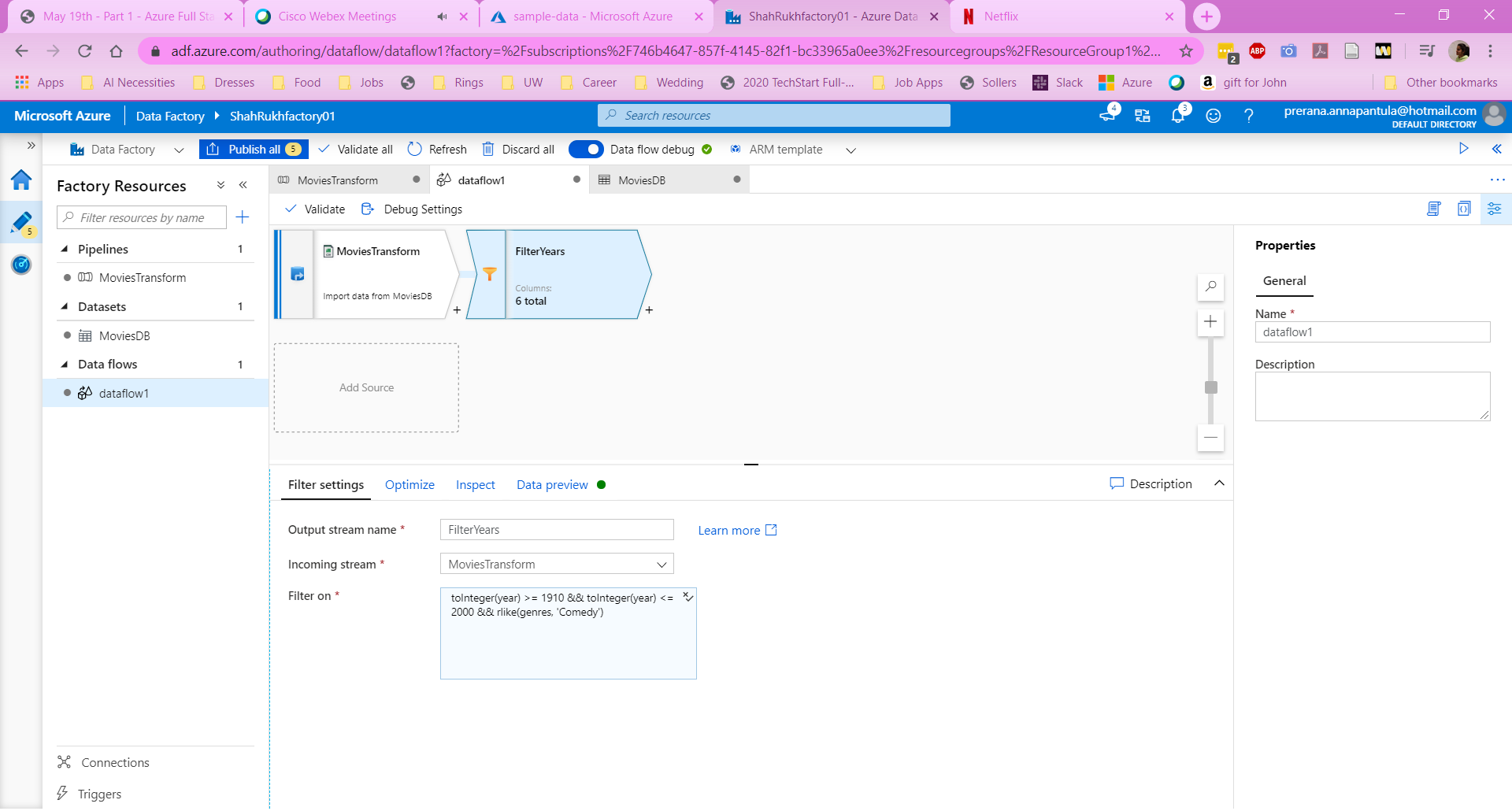
Step 4: Create a source as a “movement” for data to be accessed. Select the “+ New” button next to the Dataset box to create a new dataset for the pipeline. Make a new dataset by selecting Azure Data Lake Storage Gen 2 and then Delimiting text to be able to store and then access the data that is located within the Data Lake Storage. Create a linked service to be able to access the data correctly. To create a linked service, ensure that you select the runtime connection and authentication method required. Select which storage account is being accessed for the data and then test the connection. Once the linked service is created, input the file path and if needed, import the schema and make the file’s top row as the header.



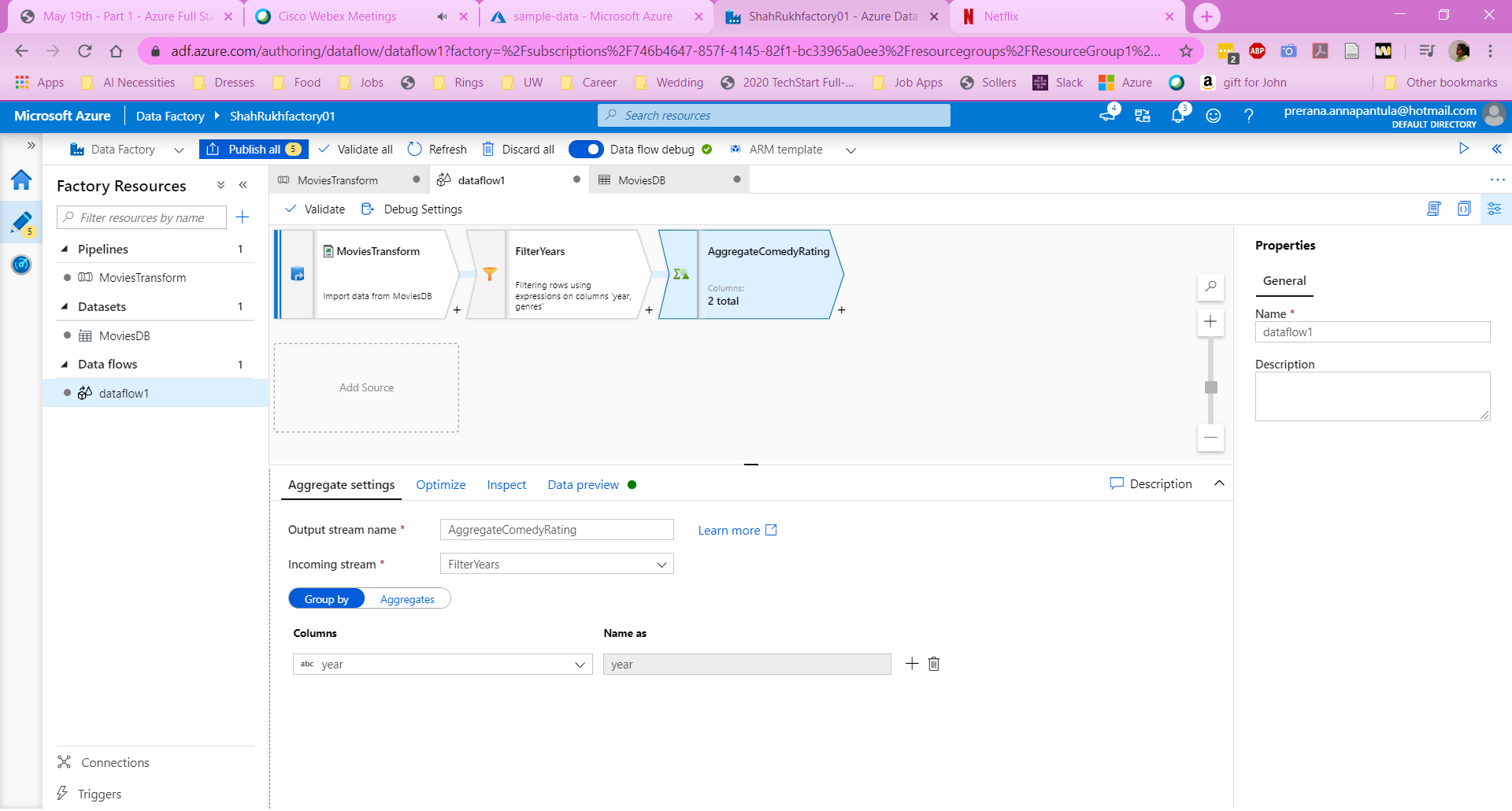




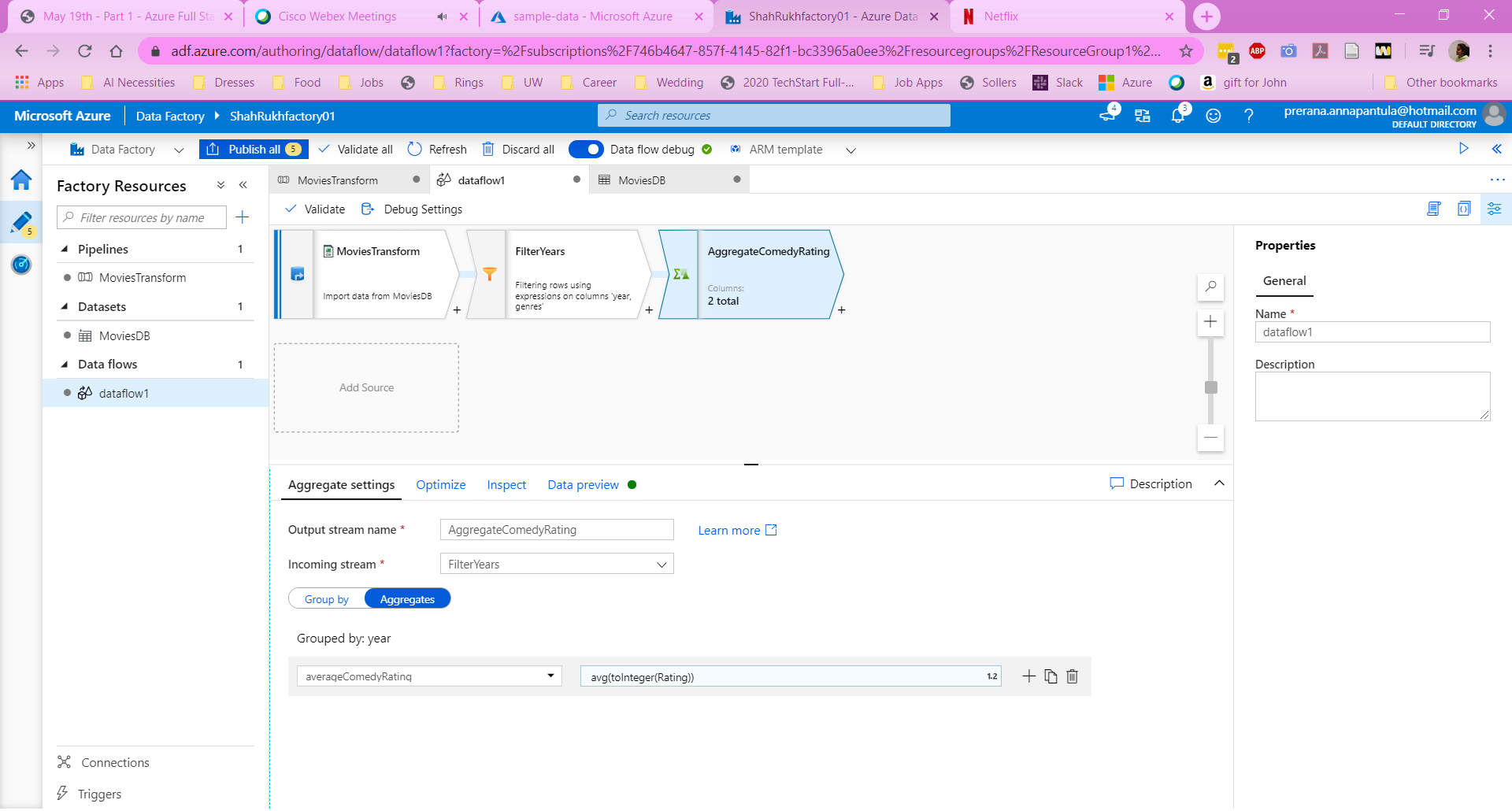
Step 5: Create a filter by choosing the “+” right next to the Movement activity to create a Transformation activity. To create a filter, choose the incoming stream and write, in the Azure syntax, what the filter should be based on. Select “Data Preview” and press “Refresh” to view the filtered data.



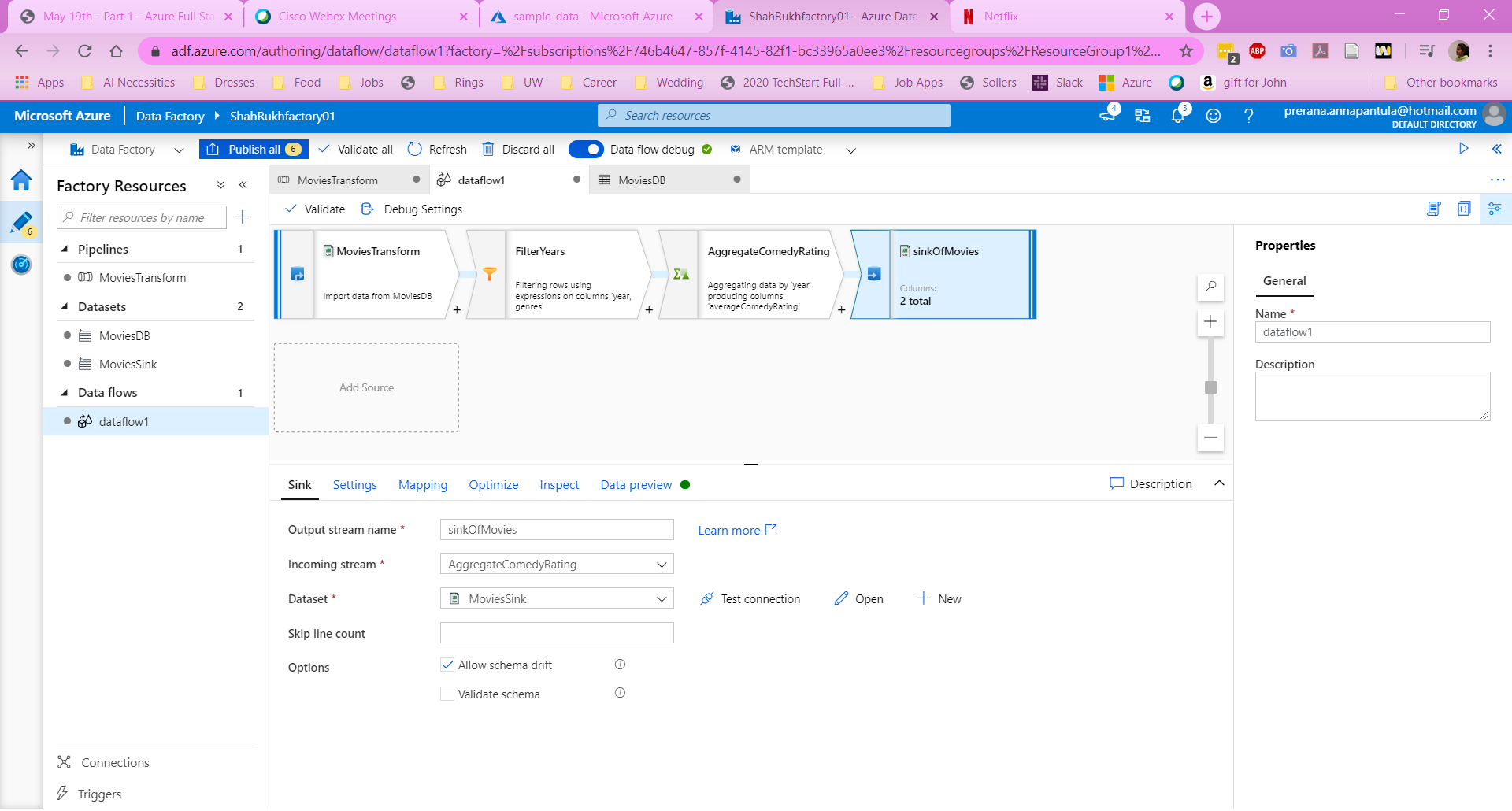
Step 6: Select the “+” next to the Filter to create another transformation activity, Aggregate. While in the “Group By” selection, choose the column by which to group the data and then the name for the grouped column.



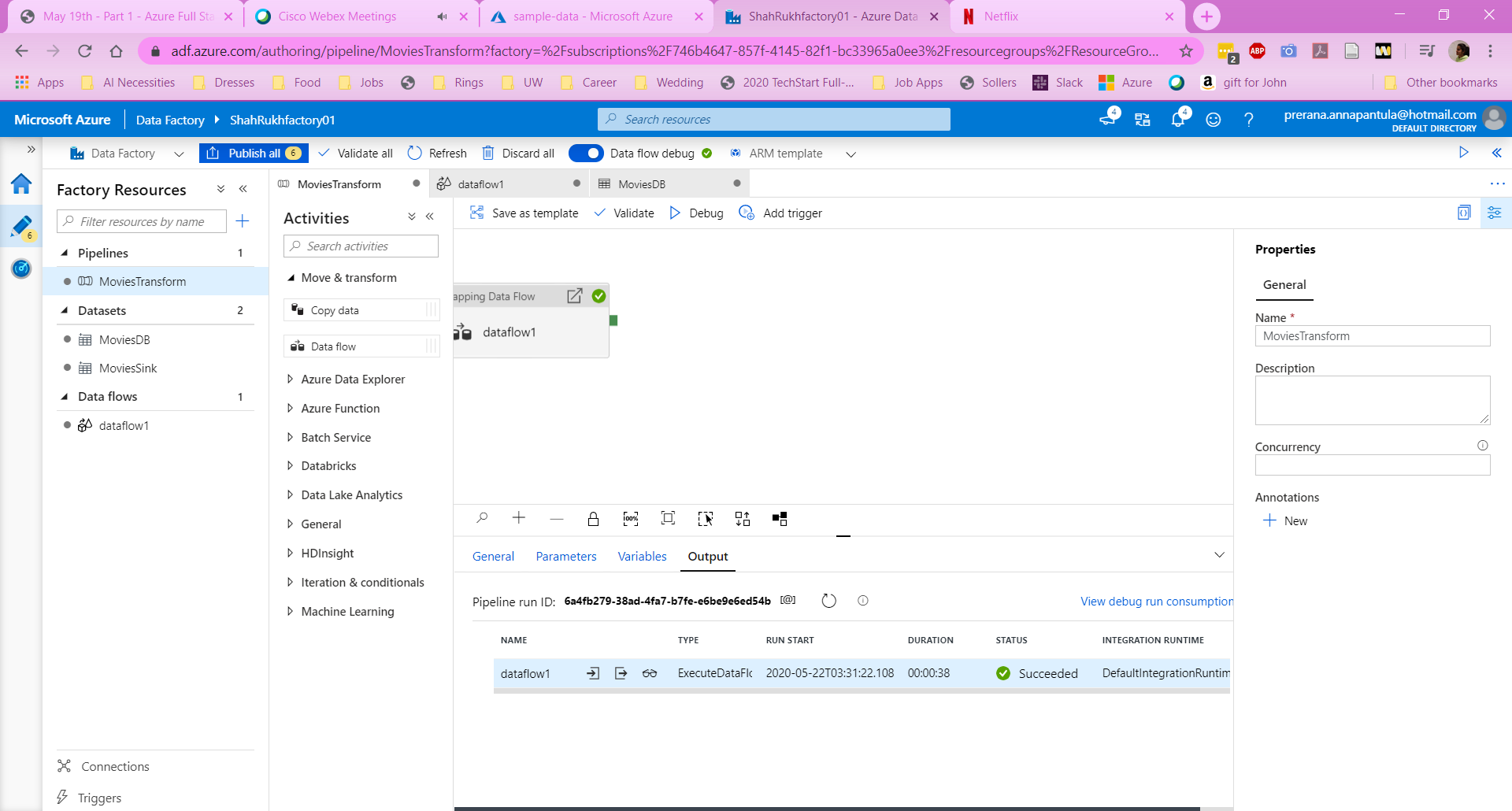
Step 7: Click on the “Aggregate” tab to be able to form the groups together even more. For this, select the name you’d like to give the new column and then choose how you want to aggregate the data. To aggregate the data, write the specifications in Azure syntax.



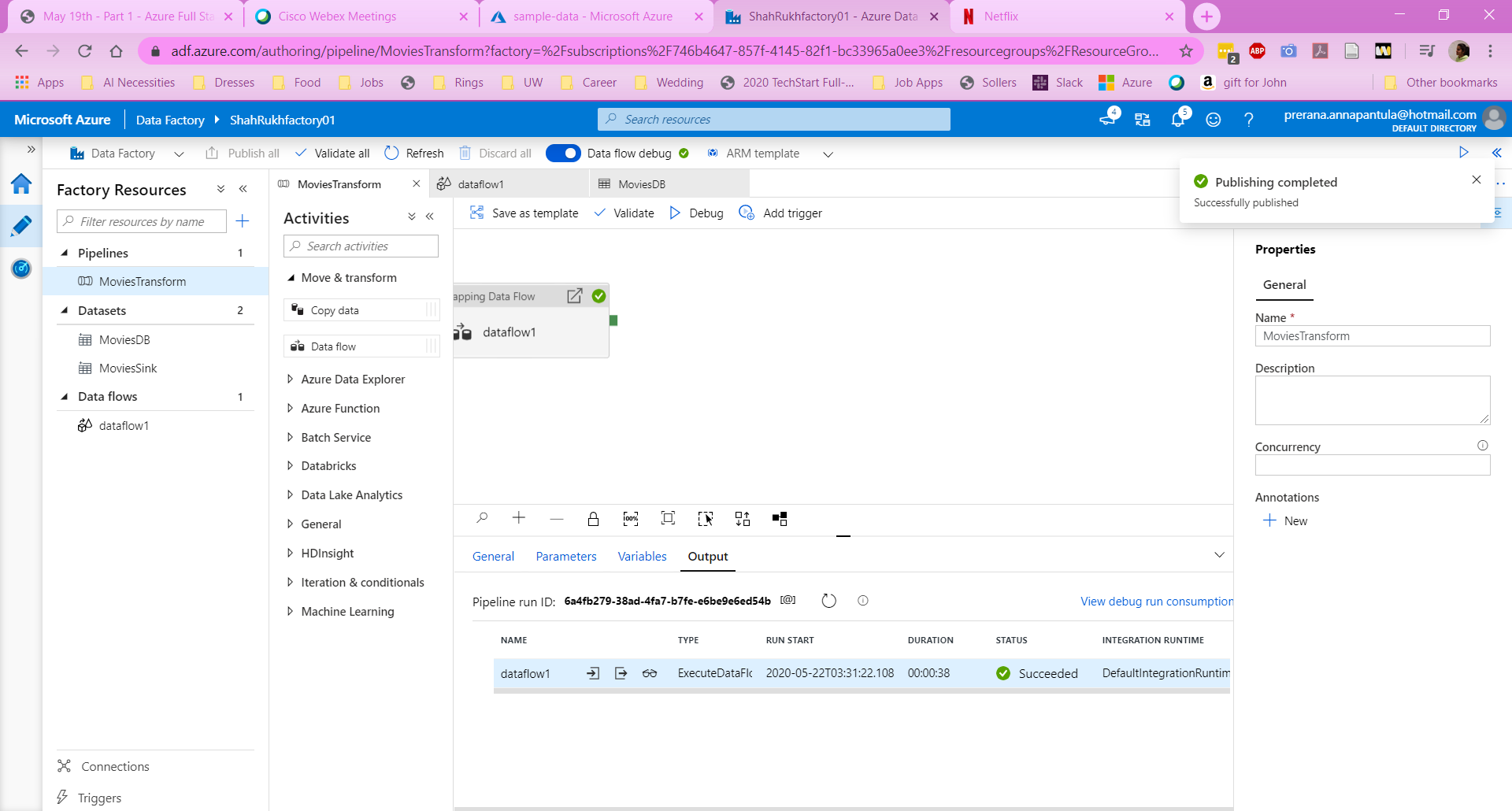
Step 8: Click the “+” button next to the aggregate activity to create a sink. A sink is used to represent the end of the pipeline. At this stage, the stream is outputted into a specific location. For this, create a new dataset by selecting Azure Data Lake Storage Gen 2 and then Delimiting text to be able to store and then access the data that is located within the Data Lake Storage. Choose the same Linked Service that’s used in the beginning of the pipeline and give a specific file path for the dataset output. Select ok and the pipeline is essentially finished.



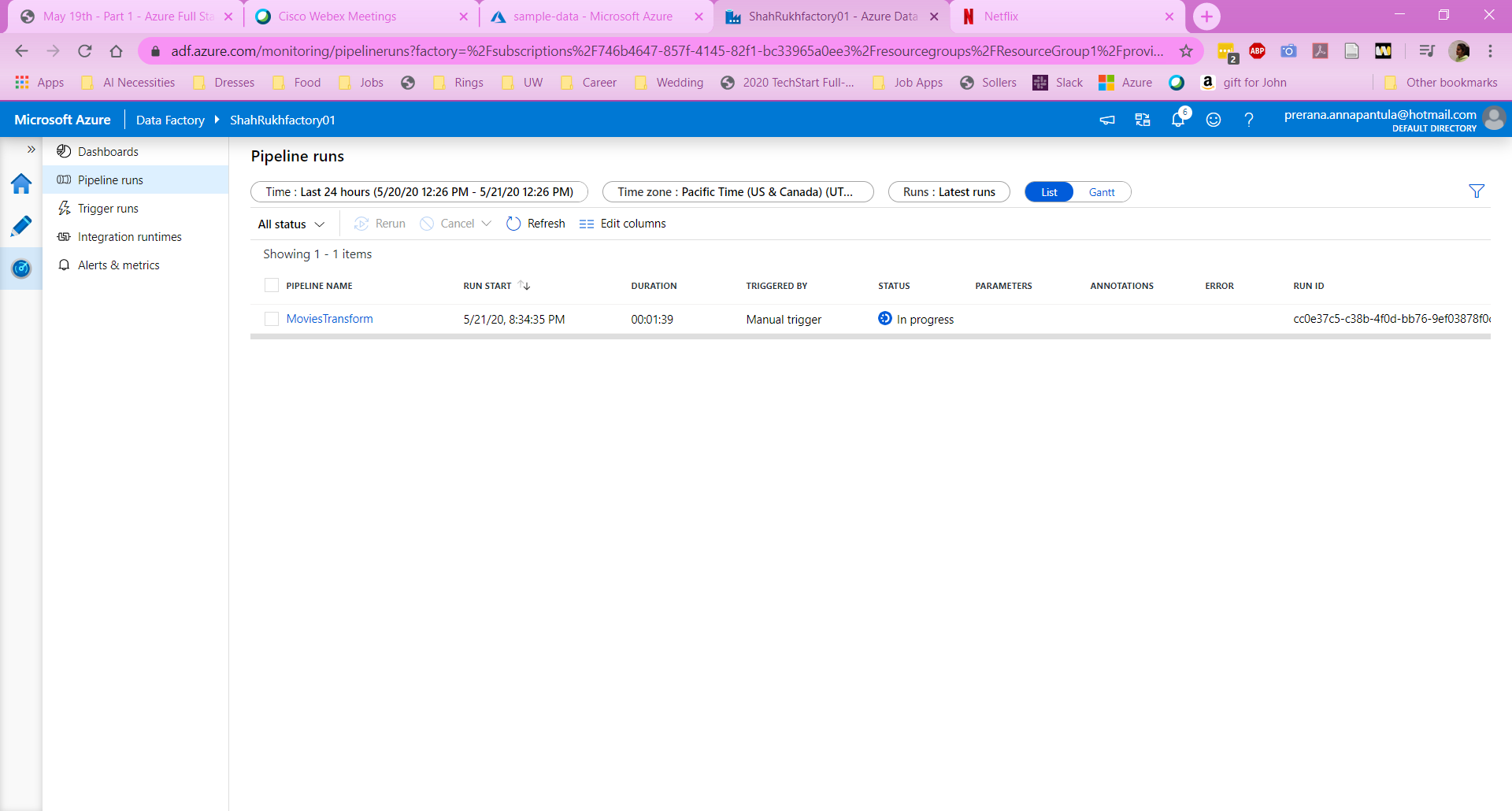
Step 9: Go back into the pipeline tab that was originally created and select debug to ensure that everything is working well. If the status shows as succeeded, then the pipeline will execute well.

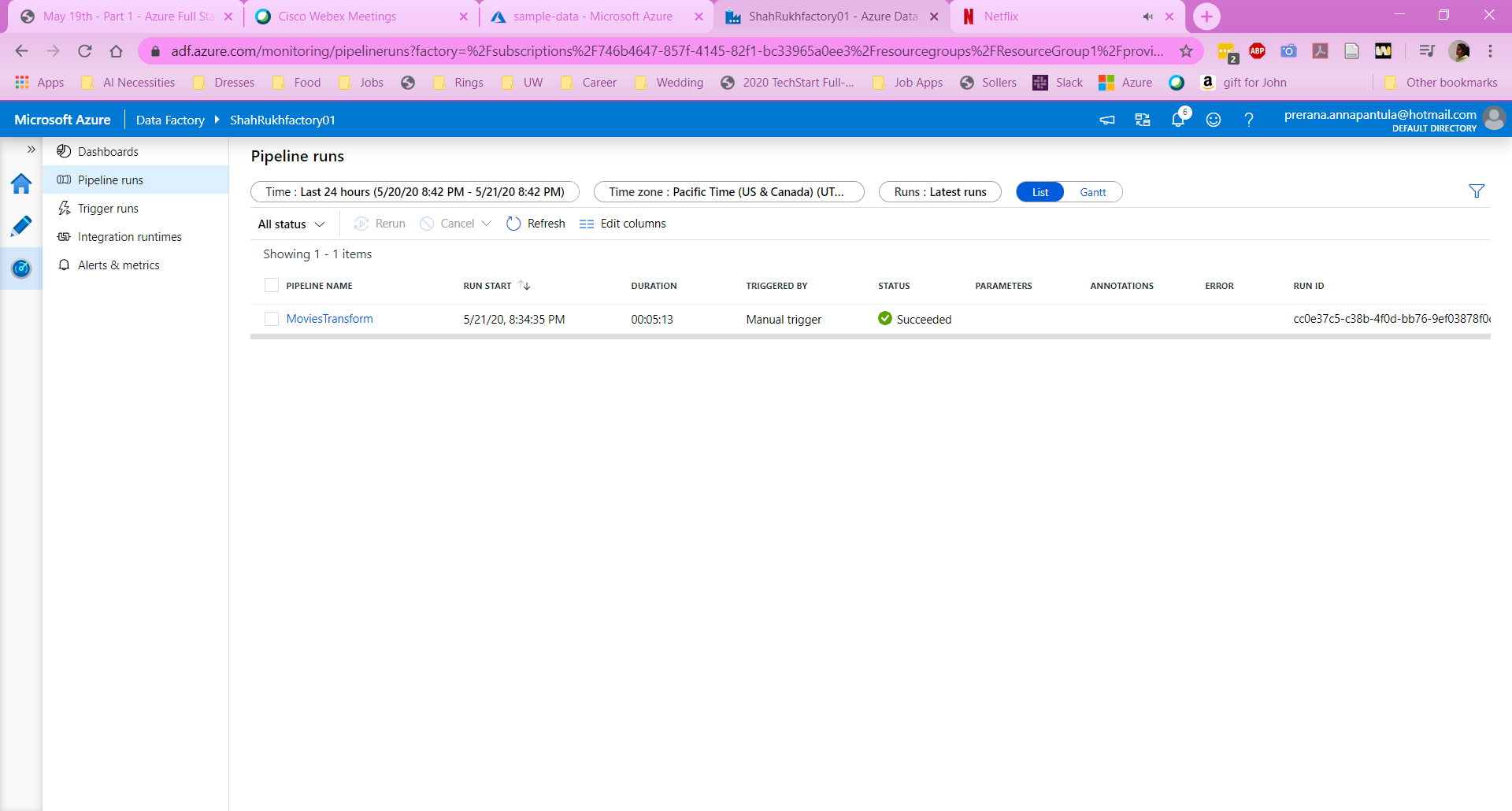


Step 10: Publish the pipeline by selecting “Publish All” and then “Publish” in the tab that opens up.



Step 11: Once published, press “Add trigger” and select “Trigger now” to execute the pipeline. The pipeline will take around 5-10 minutes to execute. Each run is given a specific run ID to identify it which means that each execution of the pipeline is essentially an instance of the specified pipeline. Once the pipeline is finished, it will show four different statuses, Succeeded, Failed, Skipped, and Completed. In this case, “Succeeded” is the best status.





* What is Azure Data Factory?  
  Azure Data Factory is Microsoft’s managed cloud service that's built for complex hybrid extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects.
* What are the Four important components of Azure Data Factory?  
  The four important components are pipeline, dataset, activities, and linked services.
* Write a detailed note about each of the components?  
  *Pipeline* – group of activities that overall perform a task. It allows to manage activities as a set instead having to manage each activity individually. They can be put together to operate in a sequence or operate by themselves parallelly.   
  *Dataset* – Data structure that is within the data stores that reference the data to use in activities – it can be used as an input or output.  
  *Activity* – Processing step in a pipeline. There are data movement activities, data transformation activities, and control activities.  
  *Linked Service* – Connection string to define the connection information that’s needed for the data factory to connect to an external resource. It is used to represent a data store or compute resource.
* What is the relationship between the different components of Azure Data Factory?  
  Activities consume and produce datasets. A pipeline is a group of activities. Linked services allow activities and datasets to consume resources. Activities require linked services to run and the datasets represent the data items that are stored in the linked services.
* What are triggers and describe the different types of Triggers?  
  There are scheduled triggers, tumbling window triggers, and event-based triggers. A scheduled trigger is a trigger that is scheduled to be kicked off at a specific time. With this trigger, pipeline executions can overlap. A tumbling window trigger is a trigger that is scheduled to run at specific intervals. With this trigger, executions cannot overlap. The pipeline execution will run during the specific time and stop running when triggered again. The event-based trigger is based on the execution completion of another pipeline run.