**Name: Shaktim Nepal**

**Projectt: Movies Analytics using Azure Data Factory**

**Problem Statement**

Use Azure Data Factory's visual authoring experience to create a pipeline that copies movie data stored to Azure Data Lake Storage Gen2 and then executes a Mapping Data Flow to transform and write the data to a Azure SQL Database.

## **Ingesting data into Azure Data Lake Storage Gen2**

* Create a new azure data lake container with name as **azure-project**
* Upload the data into this container

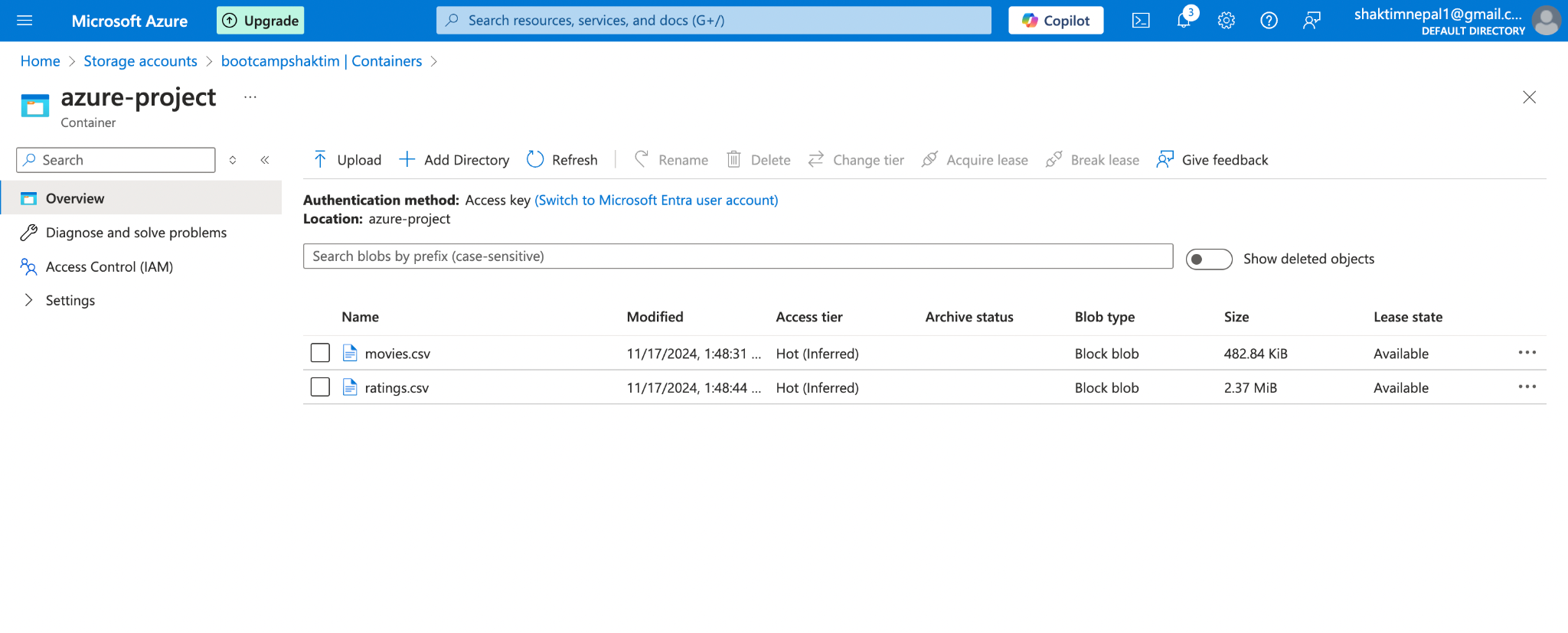
**Create an ETL pipeline from data lake to sql db with following transformations.**

* List down all unique genres and their number of occurrences for all those movies where rating is greater than 5. (I did greater than or equal to 5)
* Write down these genres and count combinations into 2 column tables in azure sql db.
  + Schema = **Test**
  + Table Name = **generes**

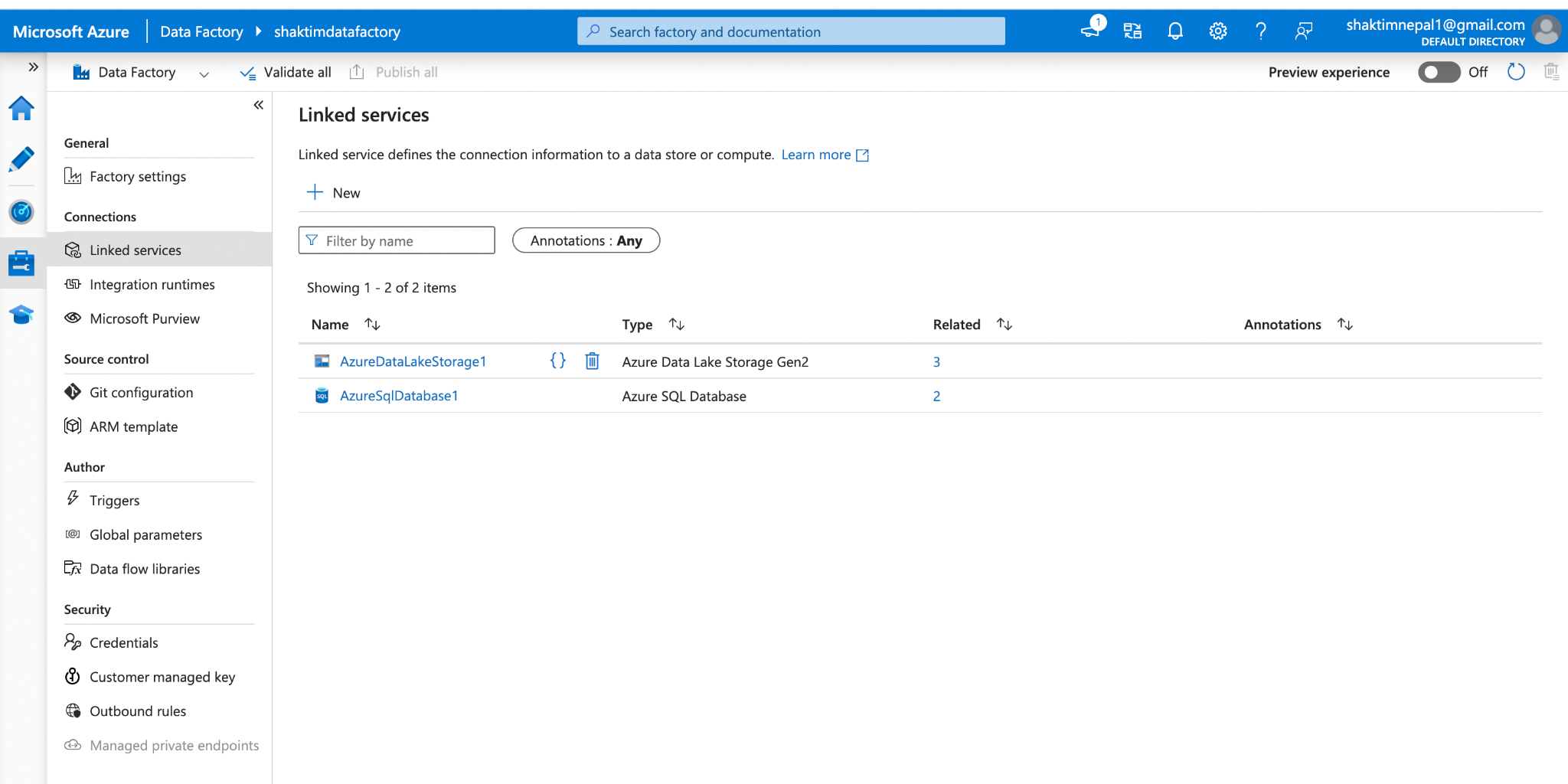
**Submissions**

You are supposed to submit the screenshots of all the activities like

* Creation of data lake container

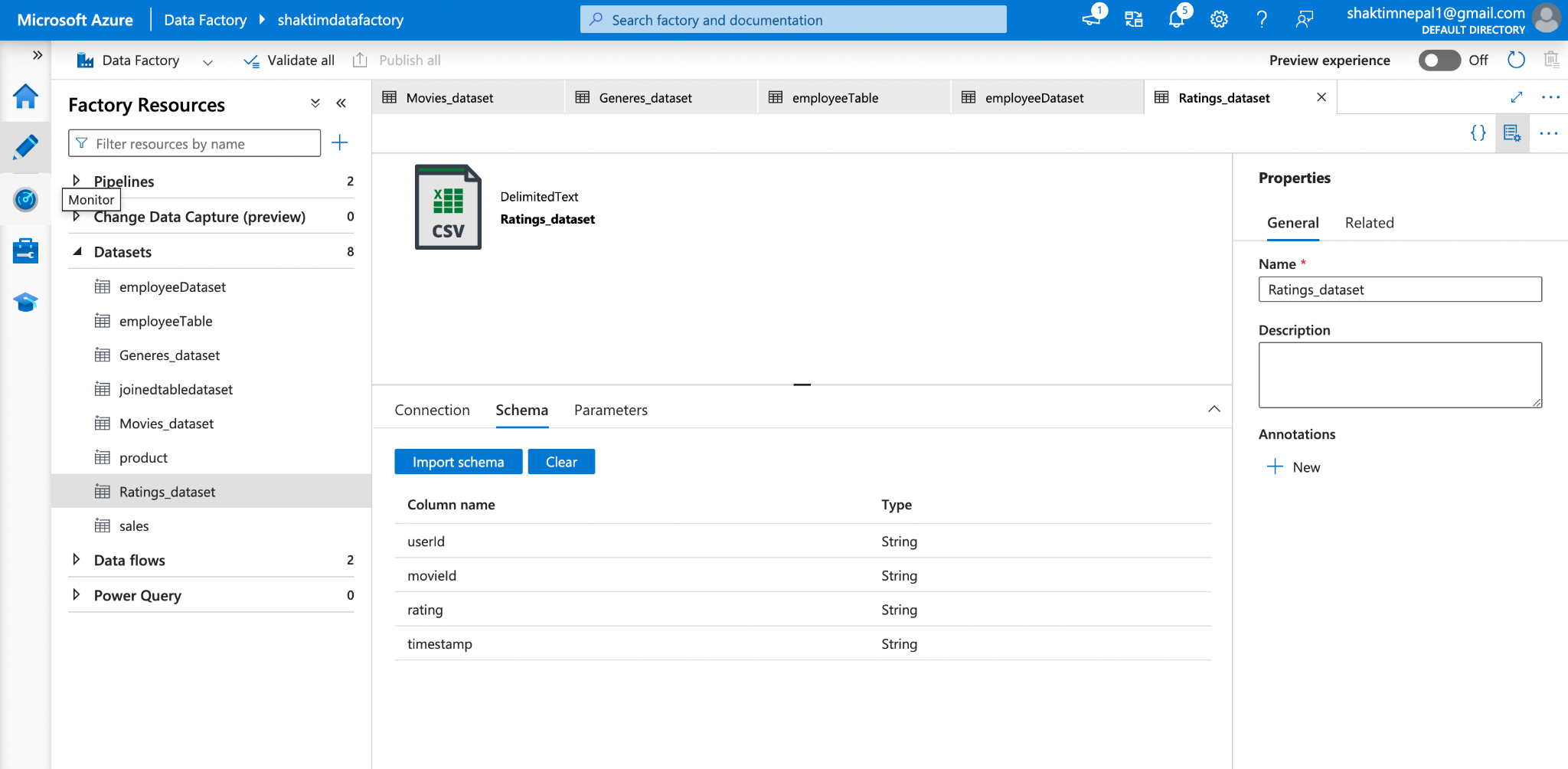


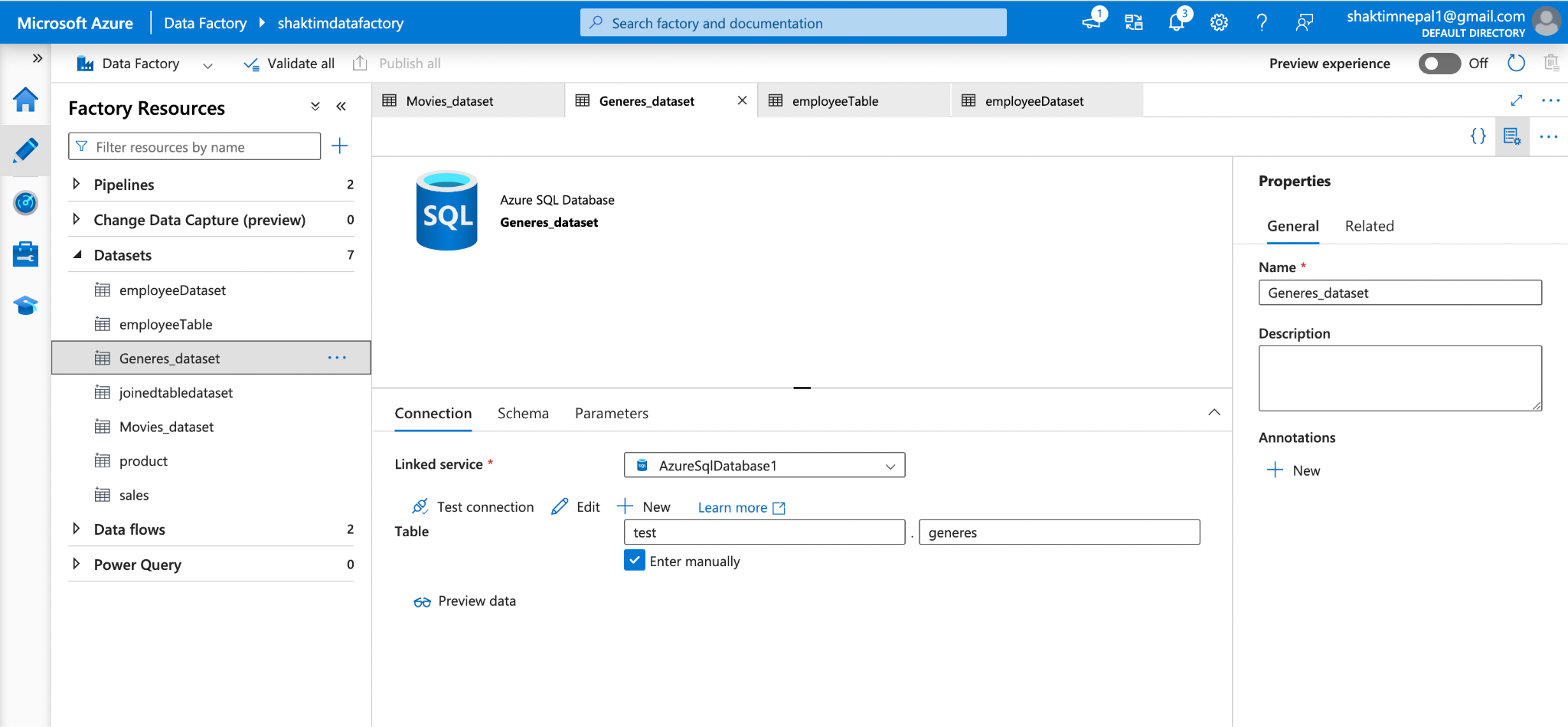
* Creation of 2 linked services



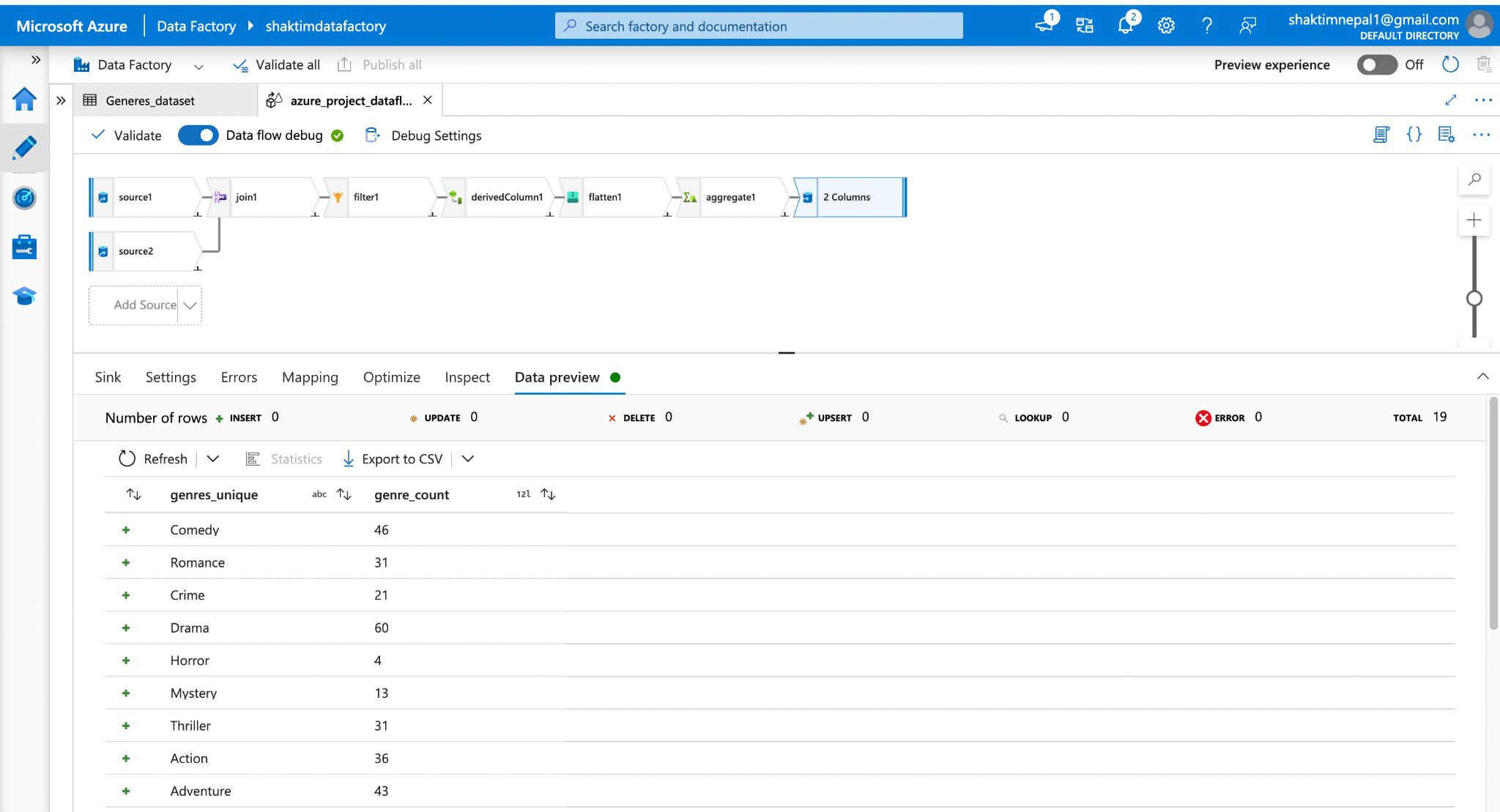
* Creation of datasets



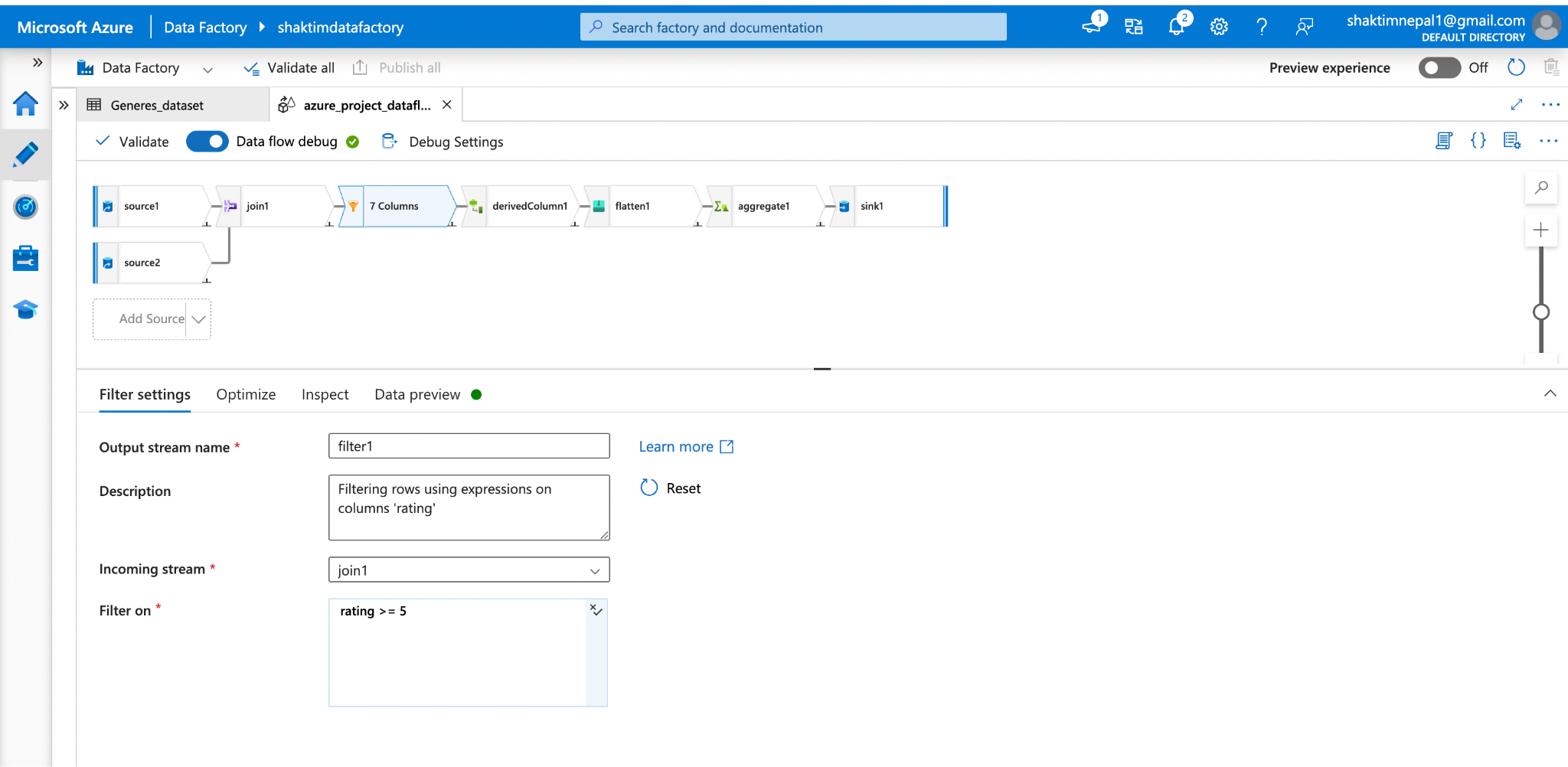




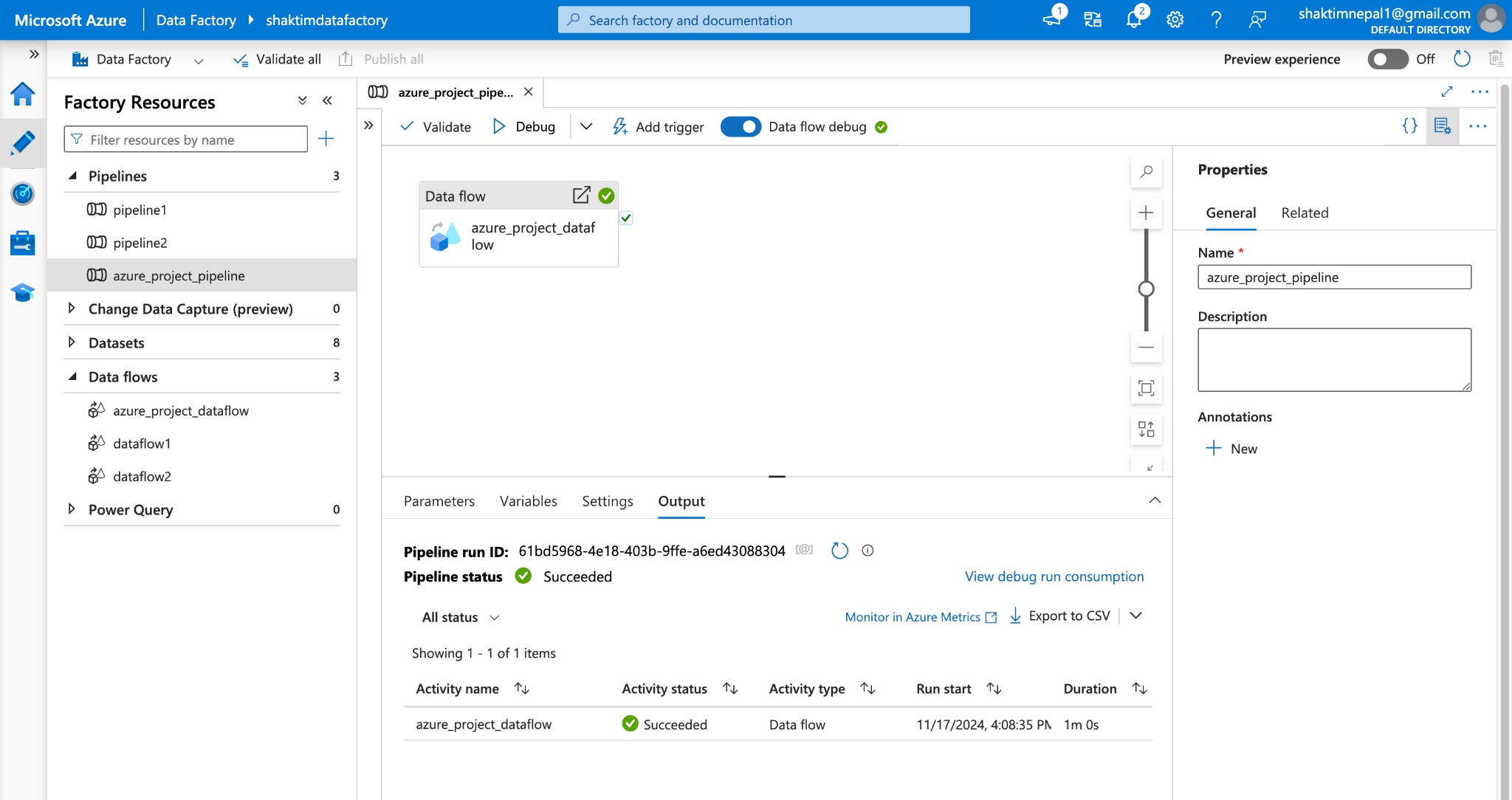
* Creation of dataflow



Note: For the filtering step, I used ratings greater than or equal to 5.0 instead of greater than 5 in order to get meaningful output because there were no movies with ratings greater than 5.0 in the ratings.csv data and we would not be able to validate our pipeline with that filter condition. And here is the screenshot of that filter expression implementation:

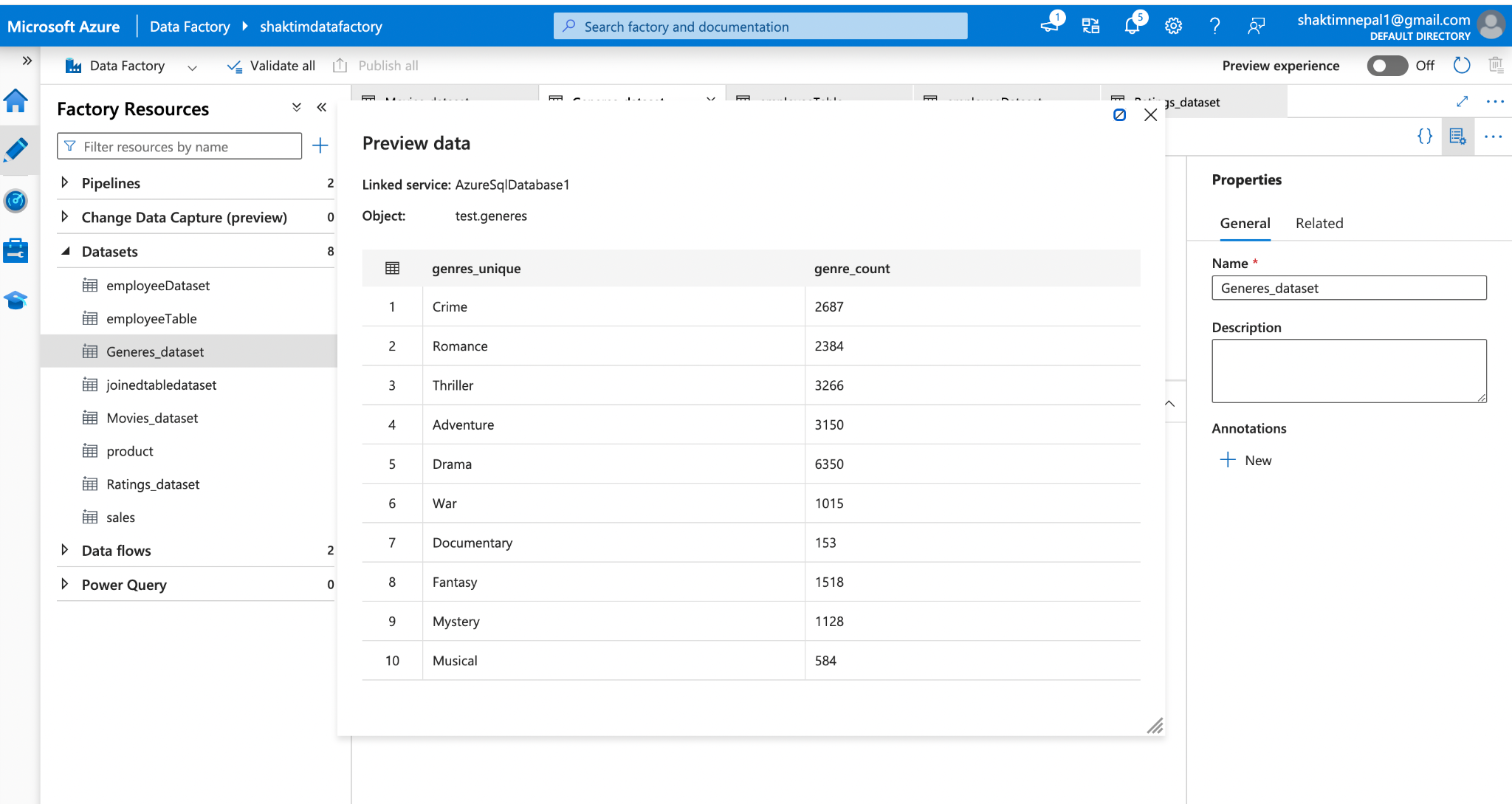


* Creation of pipeline



* Preview output of final sql db table

First validate if the dataset Generes\_dataset has been populated inside datasets in Azure Data Factory itself.



Then, preview data from test.generes table in bootcampdb sql database by running a simple query as follows:

