PIPELINE AND ACTIVITIES

ForEach Activity

Control Flow vs Data Flow

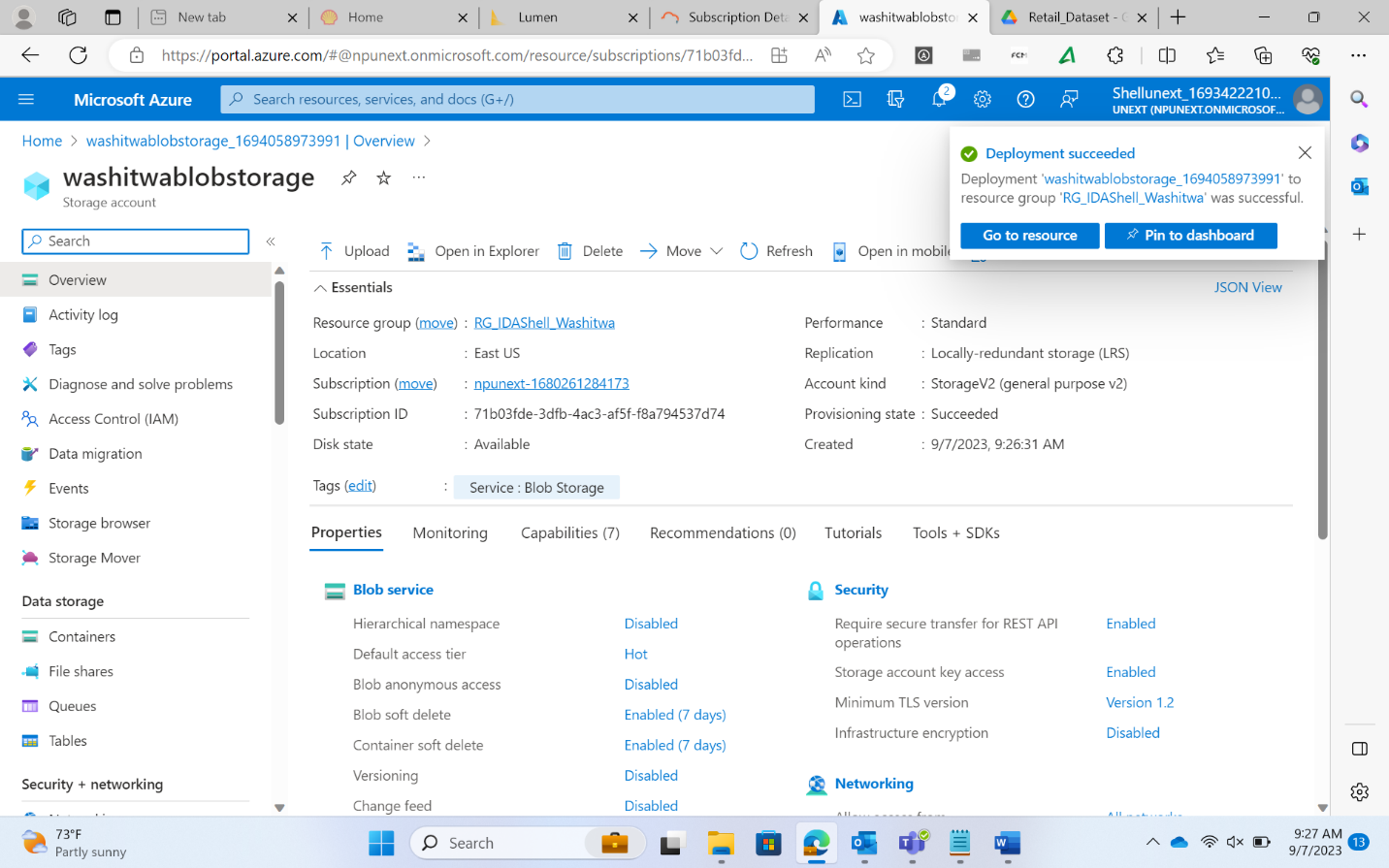
Data flow

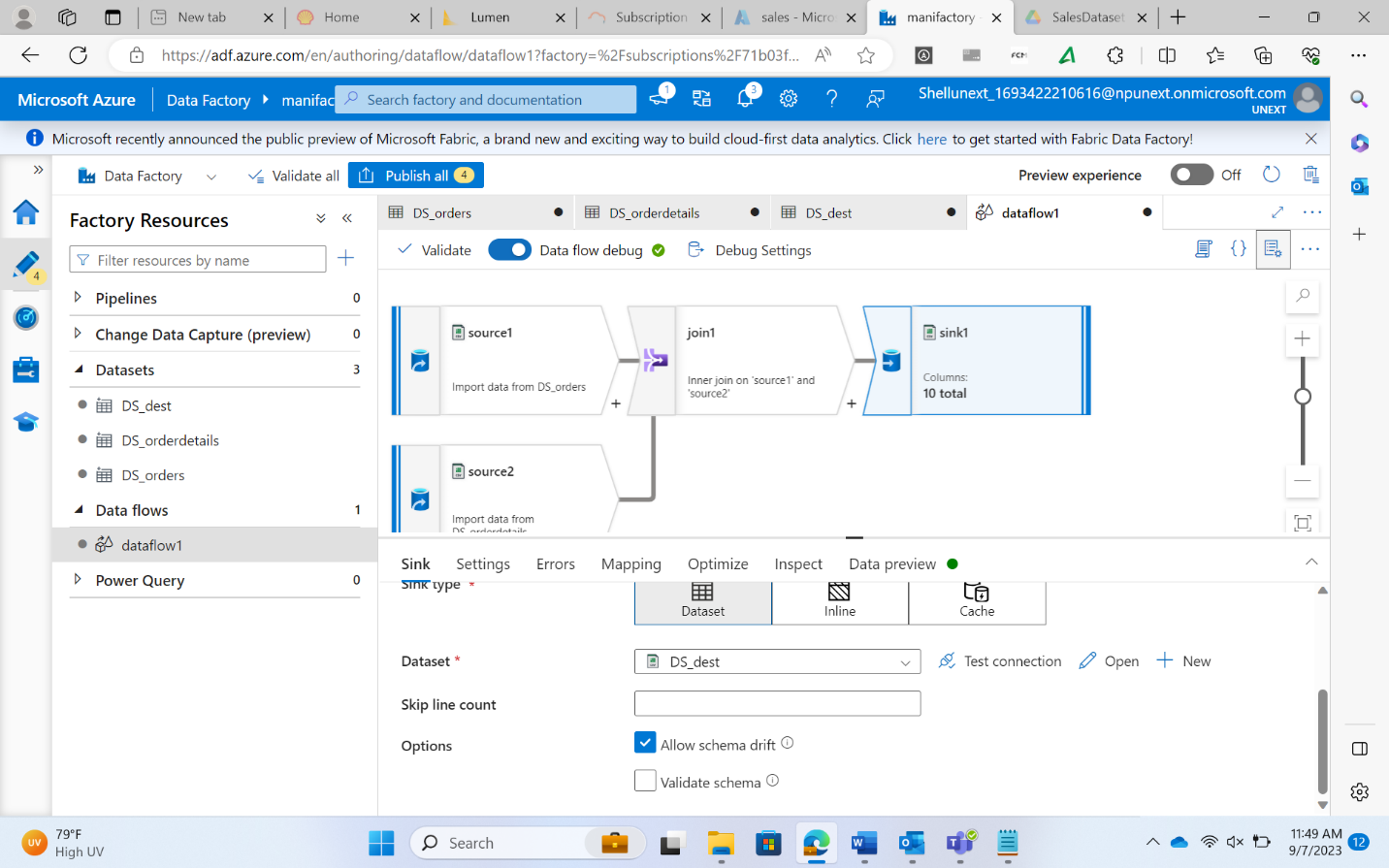
* Spark is used for transformations
* Databricks uses Spark engine in the backend for data transformations

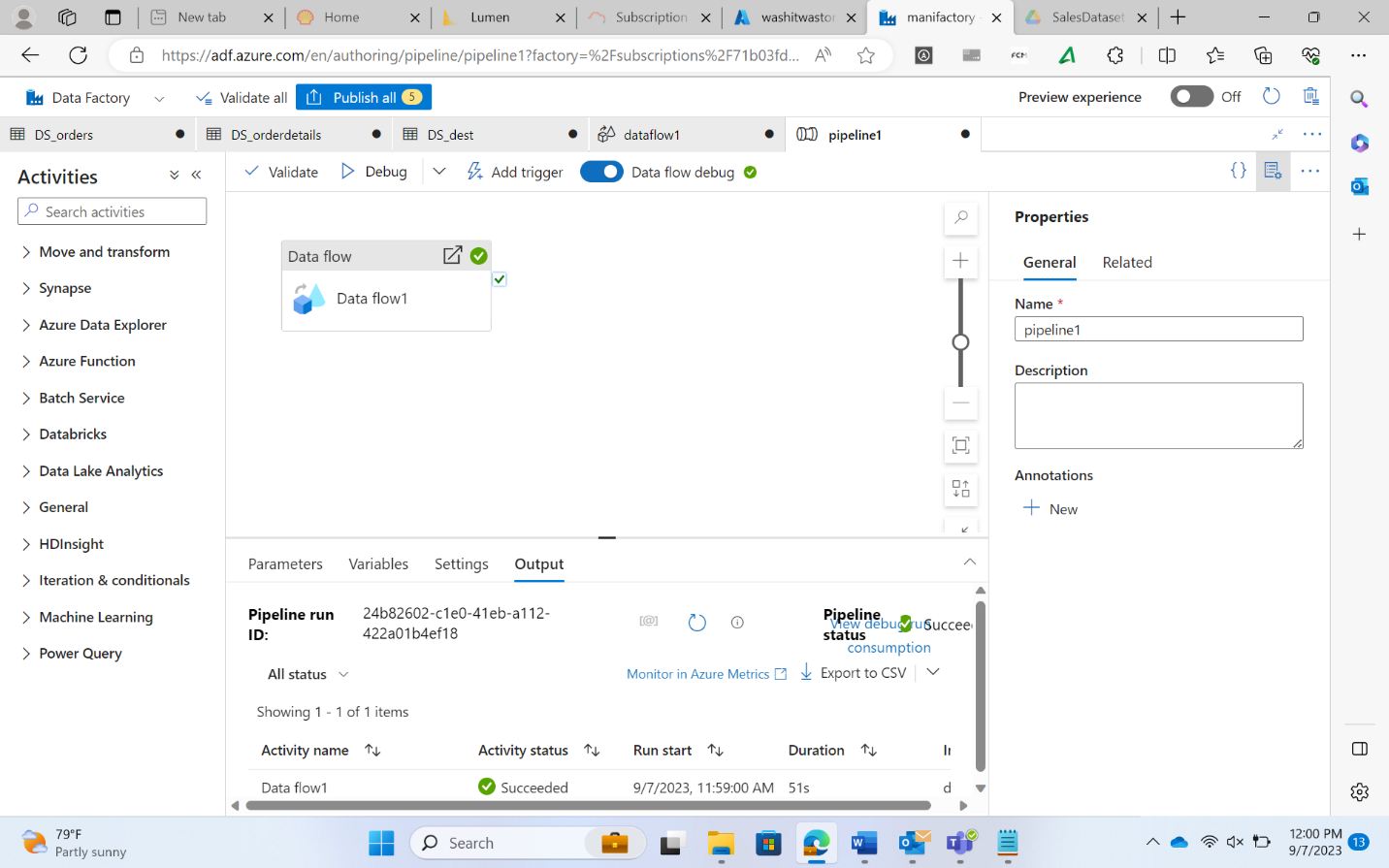
\*Azure Data Factory Studio

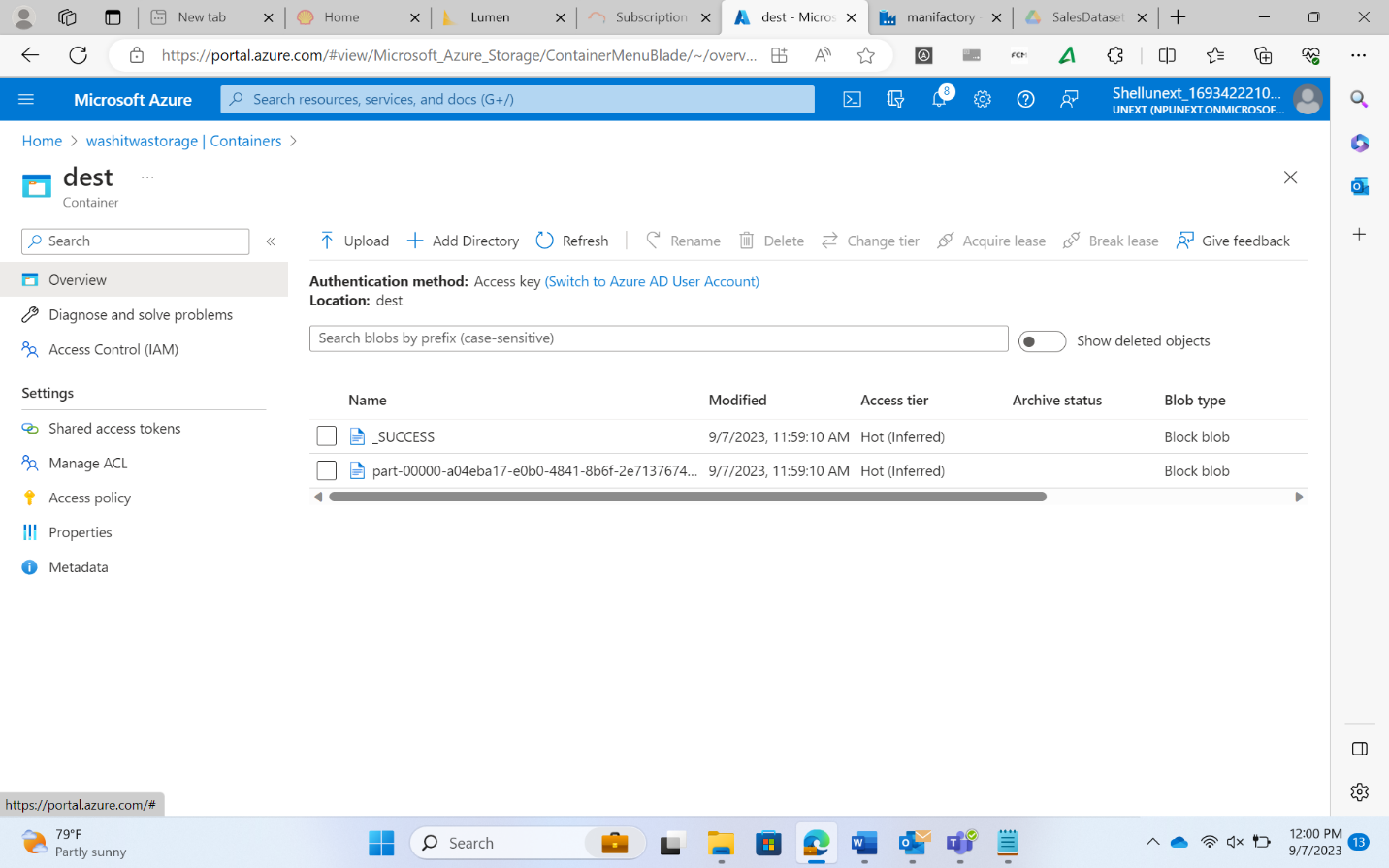
Task

* Requirement
  + Create 2 storage accounts: 1 Blob Storage, 1 ADLS Gen2
  + Create a Data Factory
  + Use Data Flow activity to carry out transformations for the csv files present
* Greater(tointeger(Amount), 1000)
* Read about ADF expressions









Derived filters

\*\*explore transformations

Surrogate key

* Substitute
* To make a column unique
* Example: if the same order is placed in the same day
* The surrogate keys are always unique

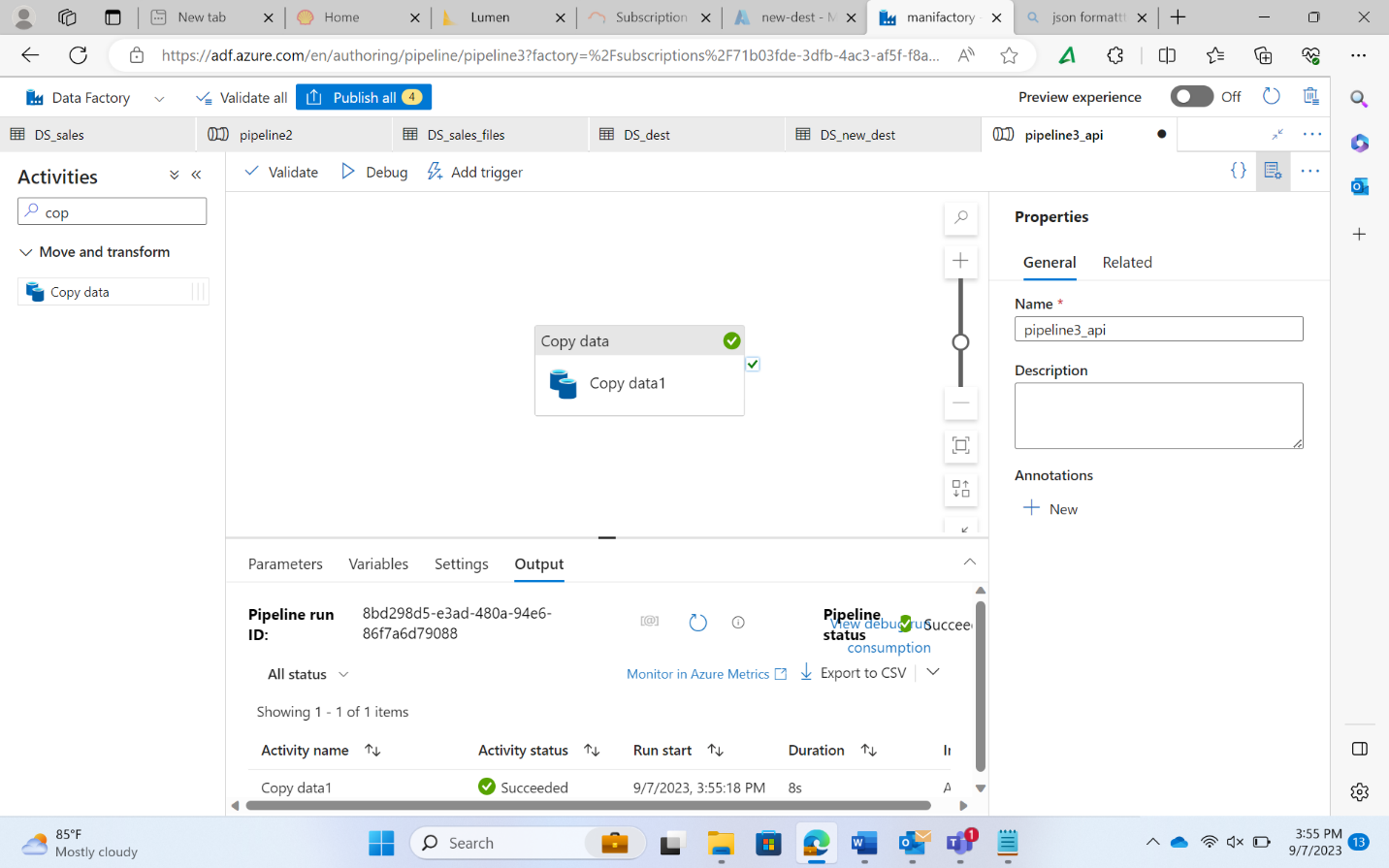
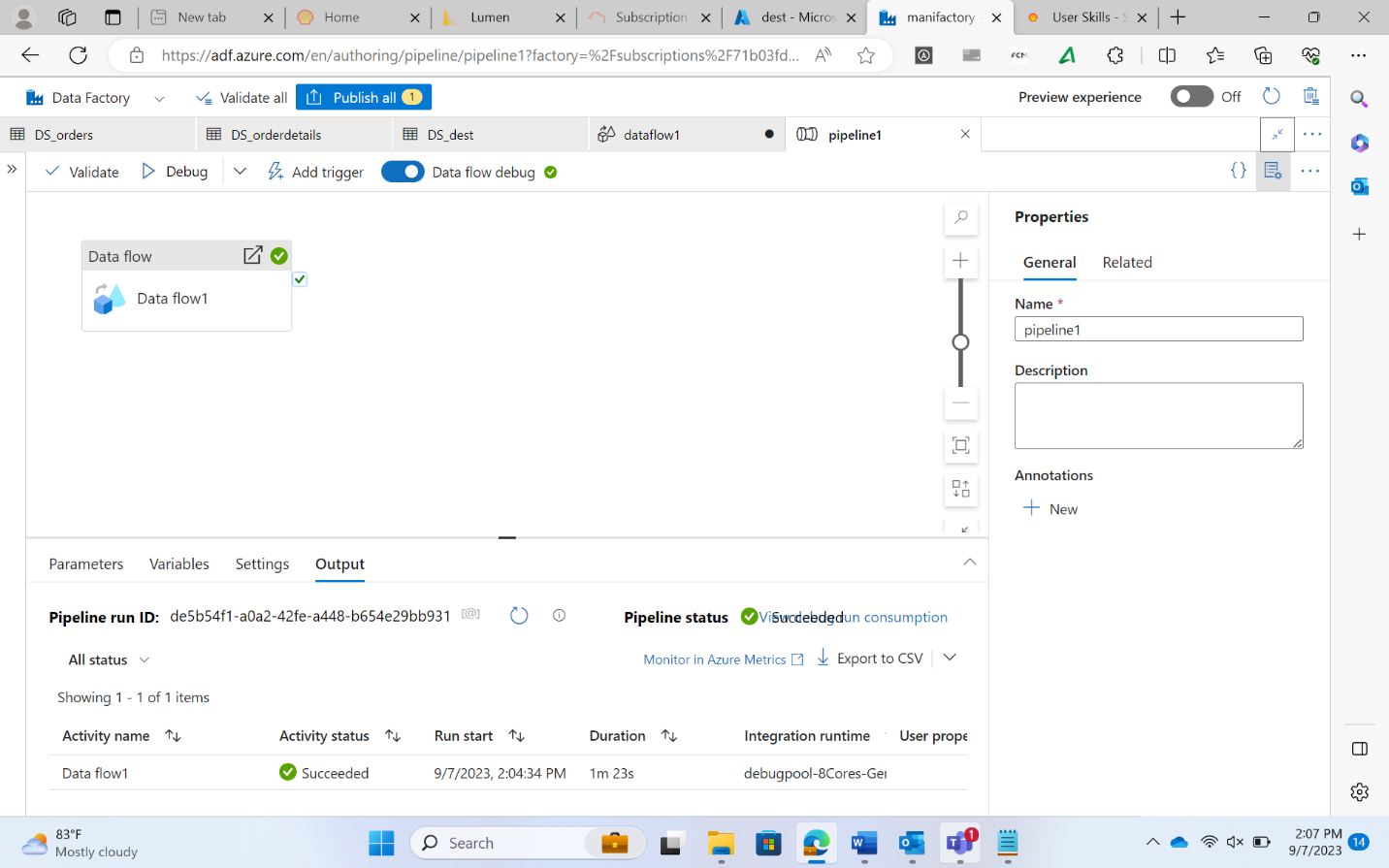
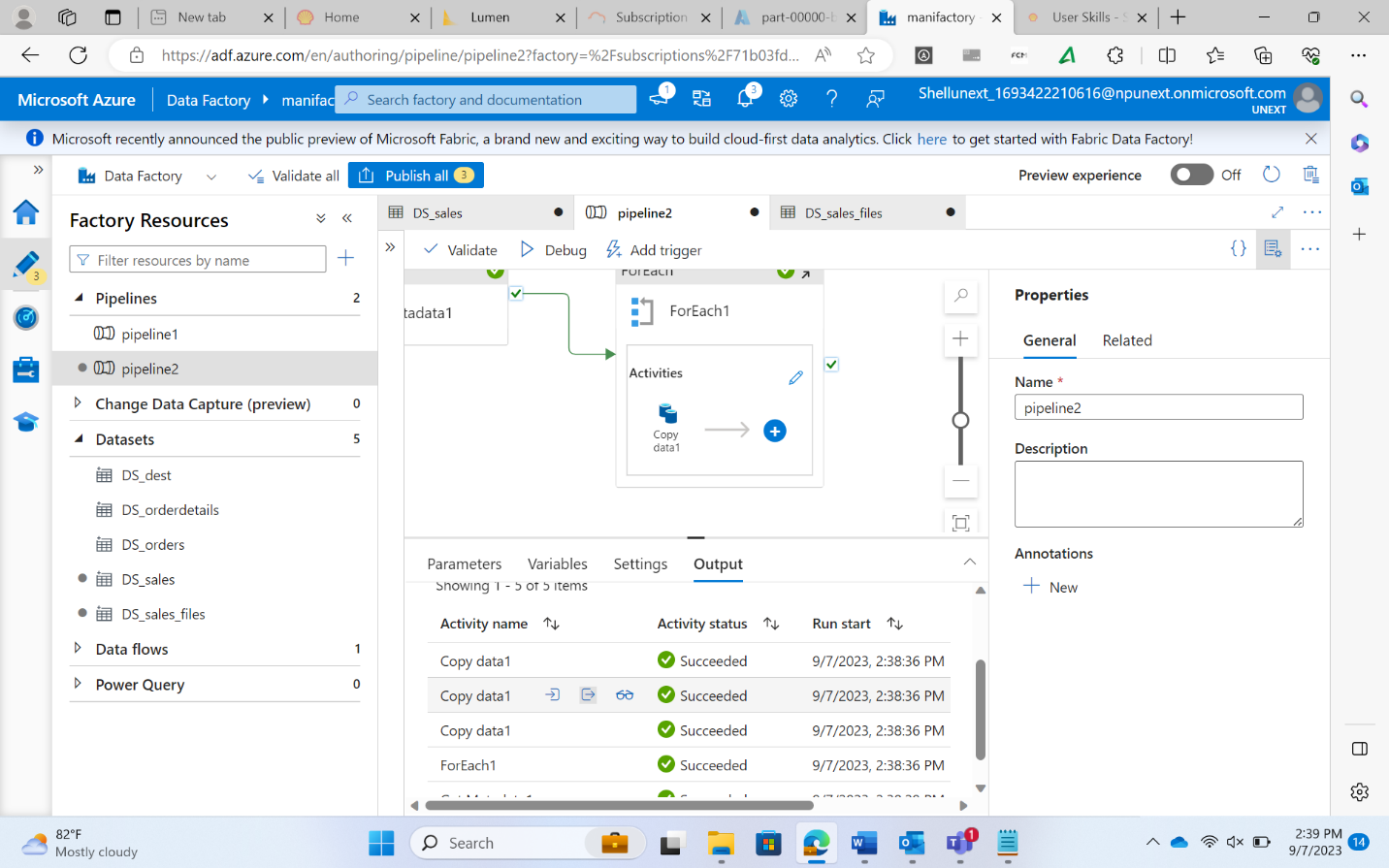
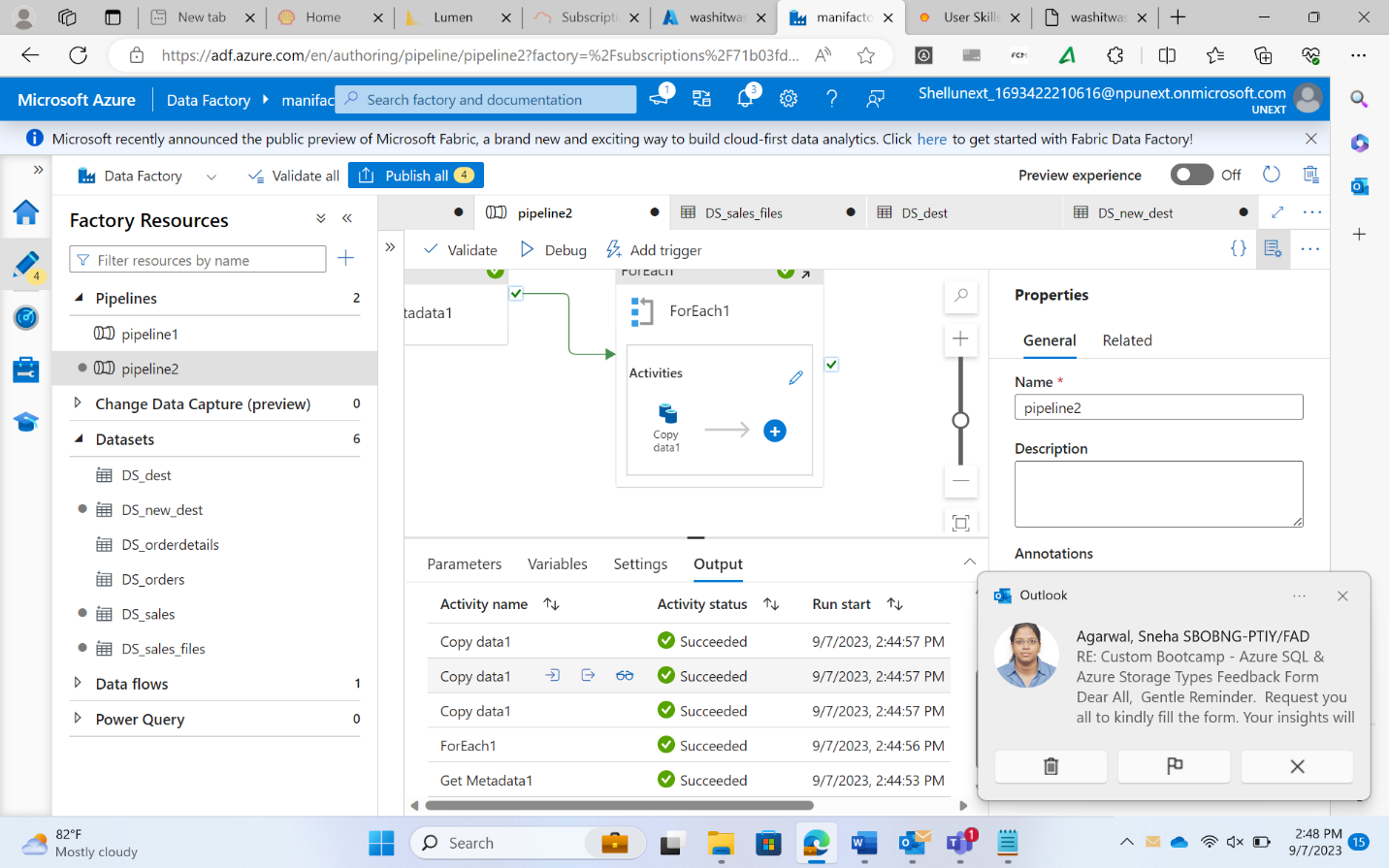
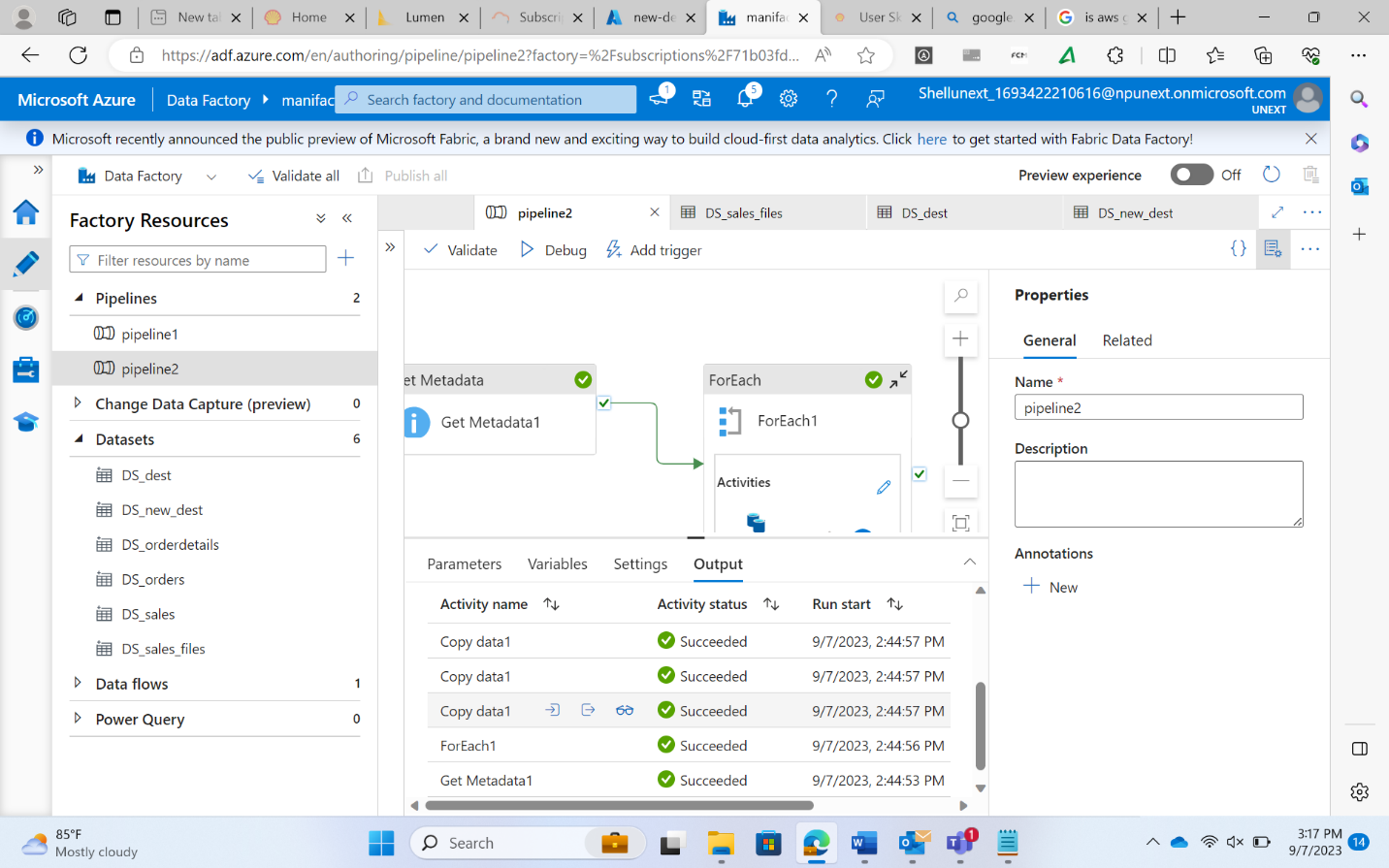
Conditional split vs filter

Task

* Create a new pipeline
* Use ForEach activity to read the file names

API

* Use Web Activity
* Create new LS for HTTP



CDC (change data capture)

* CDC takes care of the following things
  + Inserts
  + Updates
  + Deletes

Azure Resource Manager Template (ARM Template)

* It maintains a record of all the services used and published on Azure
* It is used to recover and go back to the same state if all the resources from the portal is removed