ASHWIN: Research and cloud support

* Compare Azure and AWS to determine the best fit for our use case( 1 week)
* Study in brief the different resources available in the platform and in detail the resources that would be used for this project (By tuesday week)
* Test deploy the required resources (Data lake, PostrgreSQL DB, Data Factory) and learn its functionalities
* Create dummy scripts and trigger them using data factory peipelines and assess their effectiveness in data retrieval and storage
* Research about ways to connect the instance of PostgreSQL DB to the front-end web application and test it
* **NEEDED Snapshot of data, the raw data from different apis, and what the Visualization team aims to use as a tool for the visualization**

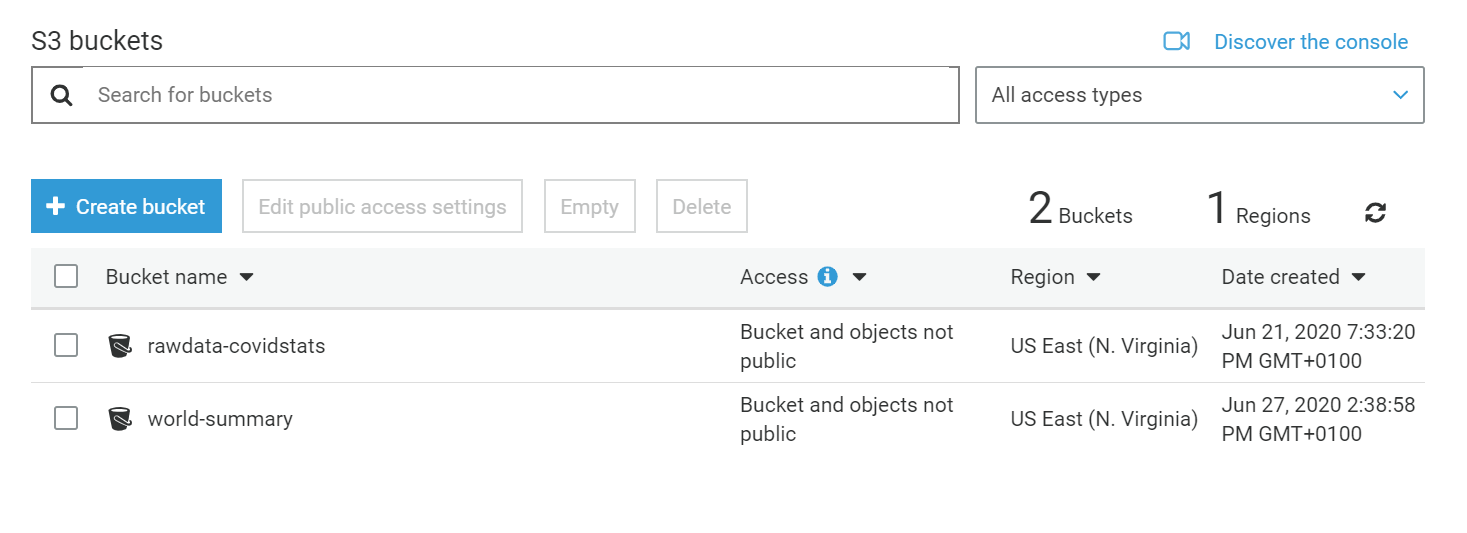
The work on integrating our project with a cloud platform consisted of 3 broad tasks. In the first stage, we compared between AWS and Azure, the different resources they provide and pricing to deploy the required resources on Azure. Once that was done, data was transported to the data lake and an instance of MySQL server was hosted on Azure that was connected to train the chatbot. Finally different datasets and pipelines were created on Azure Data Factory for the data in transit between different resources. These pipelines were triggered to completely automate this process.

Considering our project needs, for cloud we have used Azure; wherein we could create our own database and store the data in multiple tables. Azure pipelines were very useful to pick the data from a local file and push them onto the cloud.

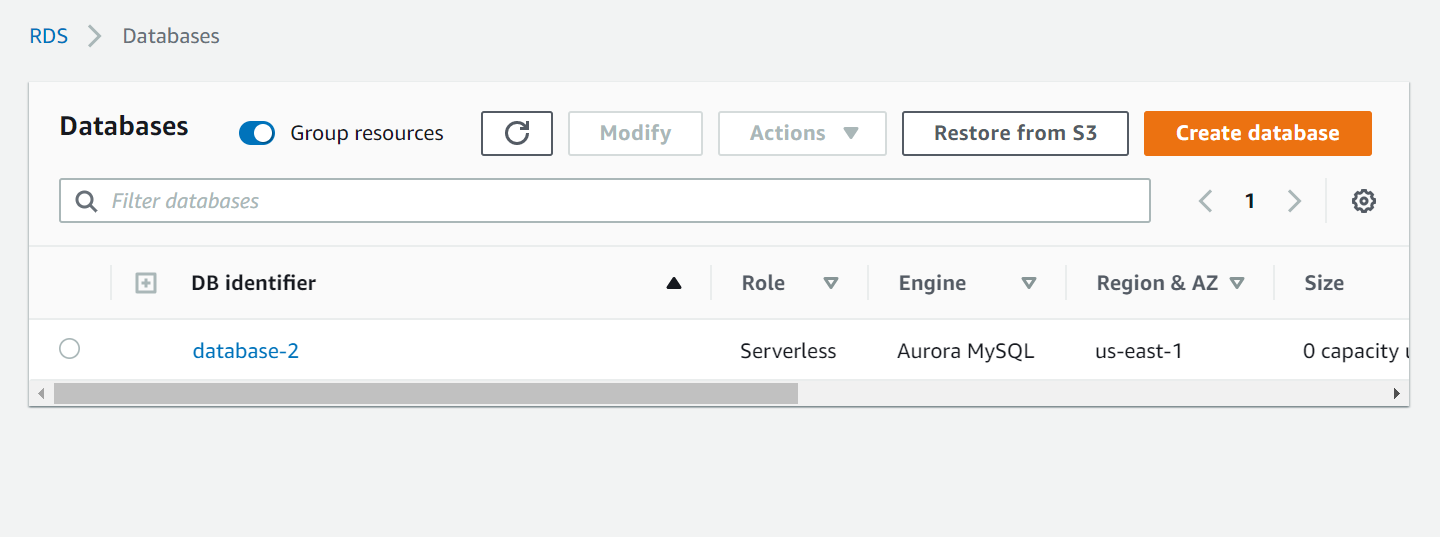
**Before use of azure we tried creating pipelines in AWS:**

We have AWS Student accounts set up and have limited access to AWS resources(Resource usage restrictions and a $100 credit limit).

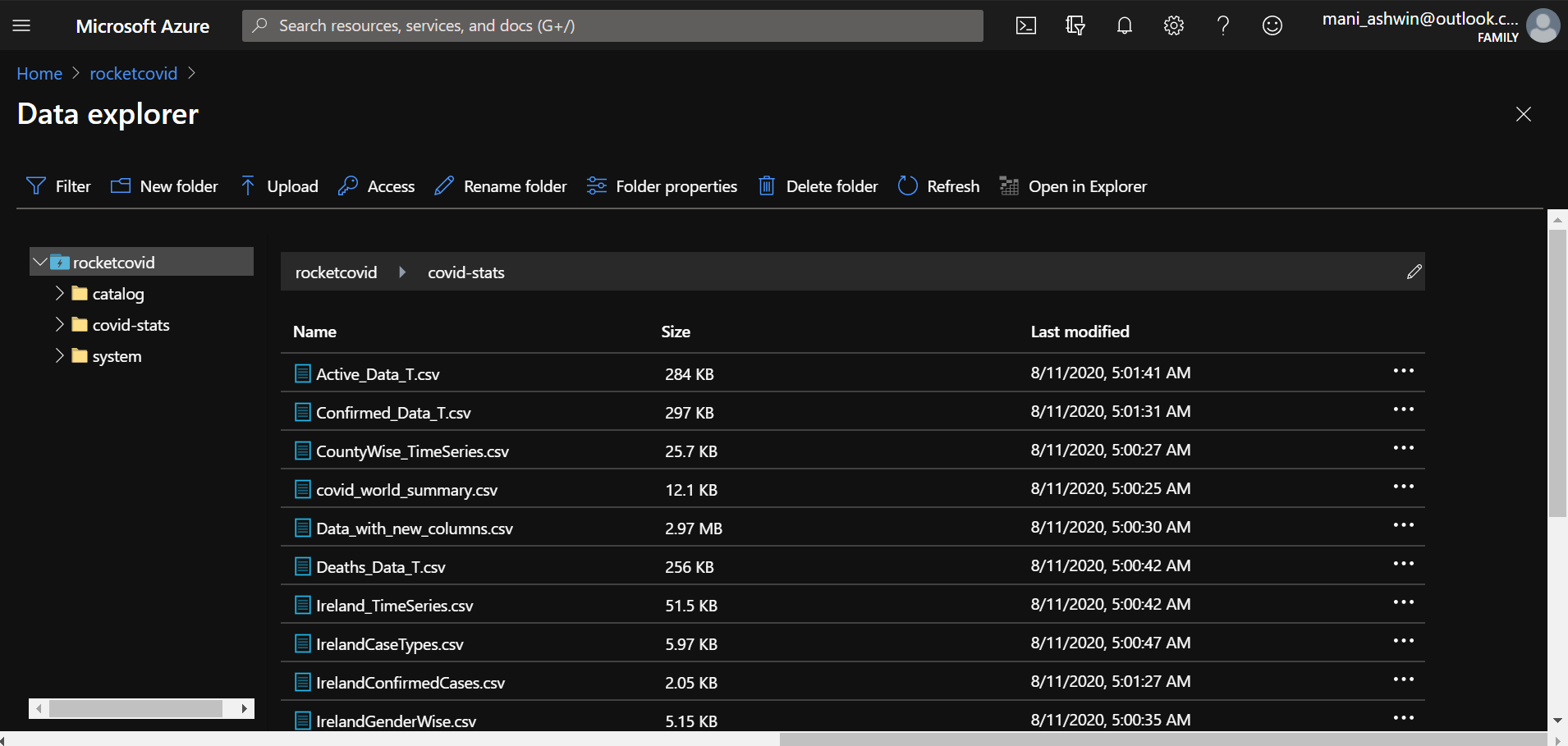
Created S3 buckets that contains data from different APIs



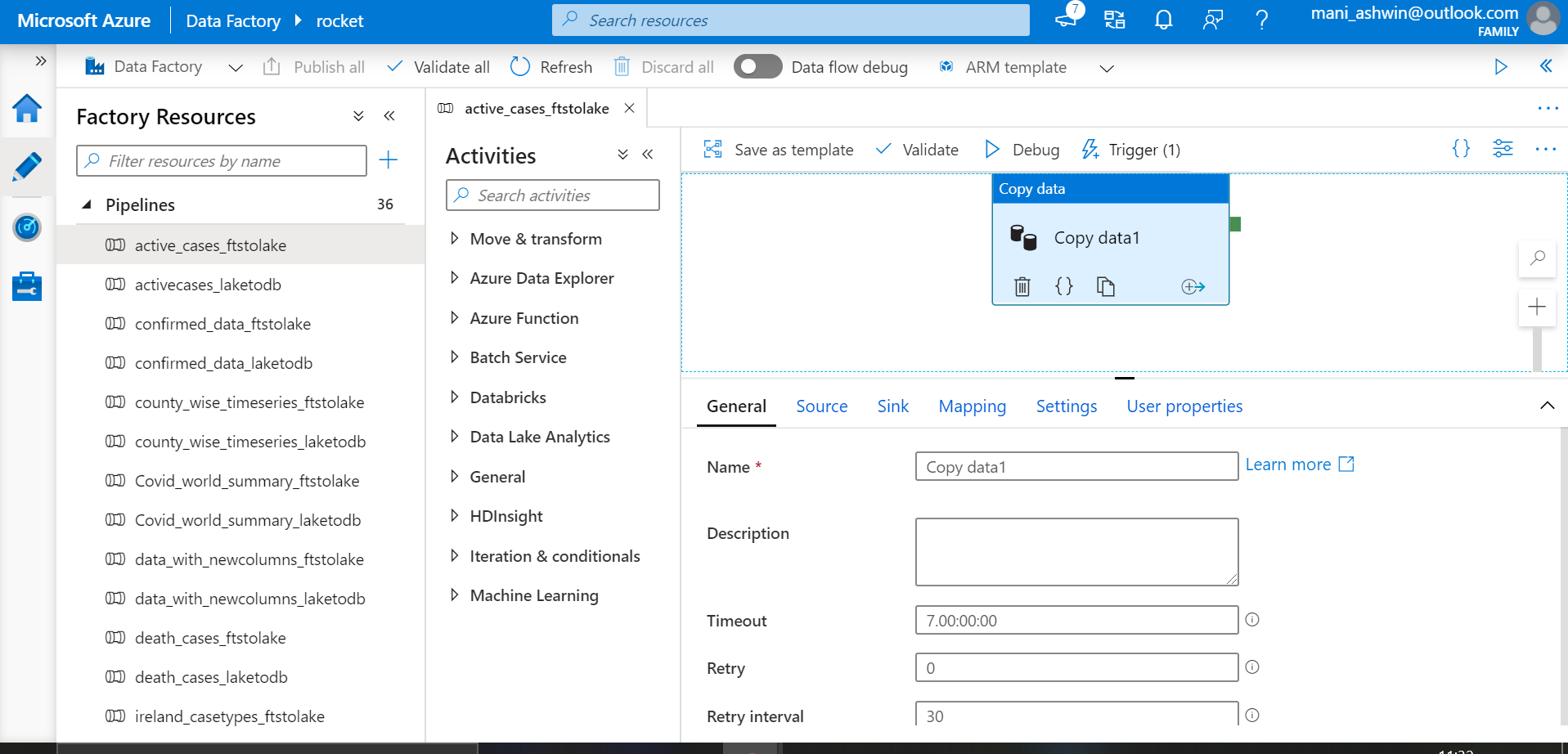
Created an instance of Amazon RDS with an Aurora Serverless API that uses a MySQL engine for processing queries



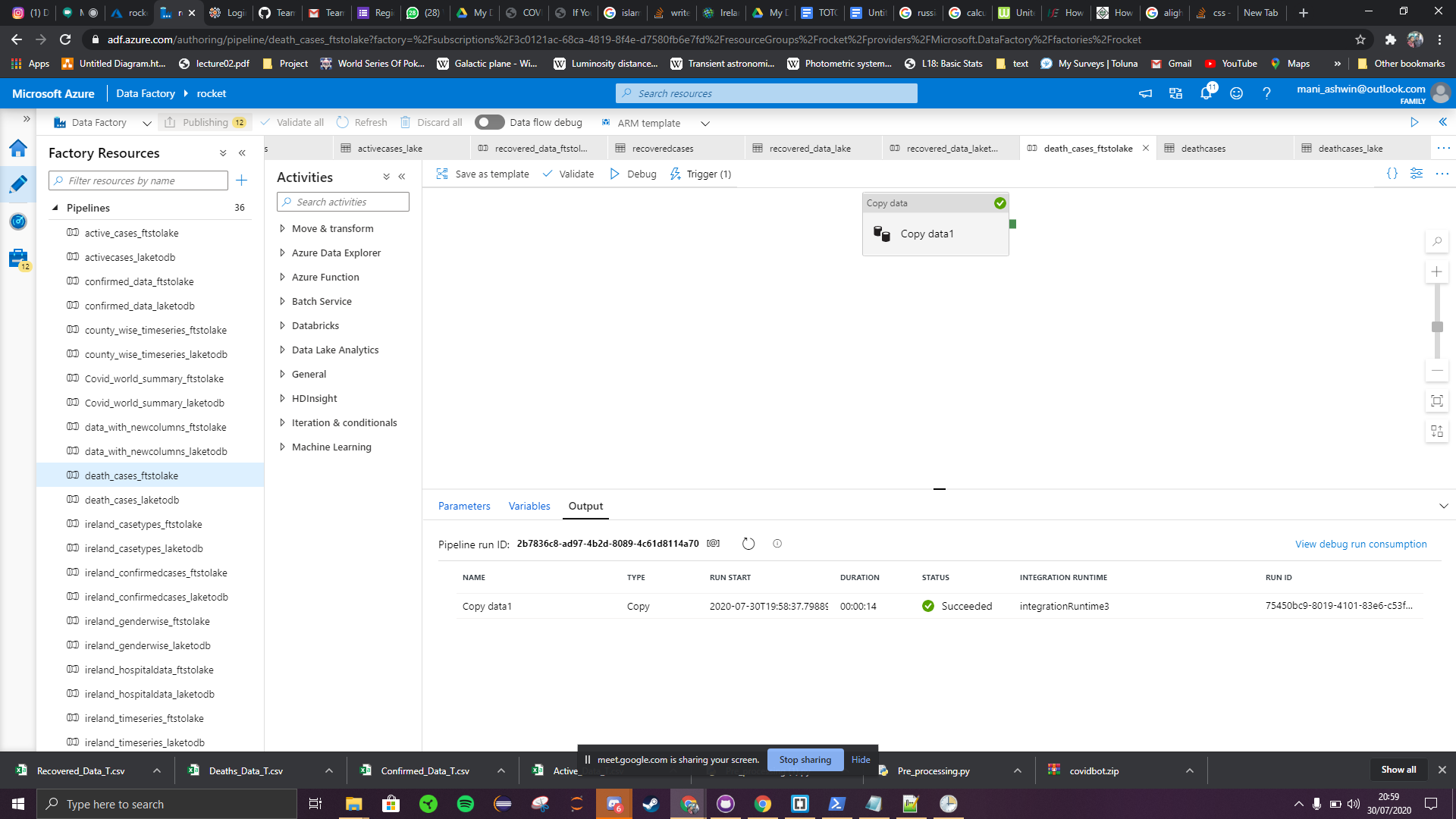
**Creation of data lake and pipelines in azure:**

****

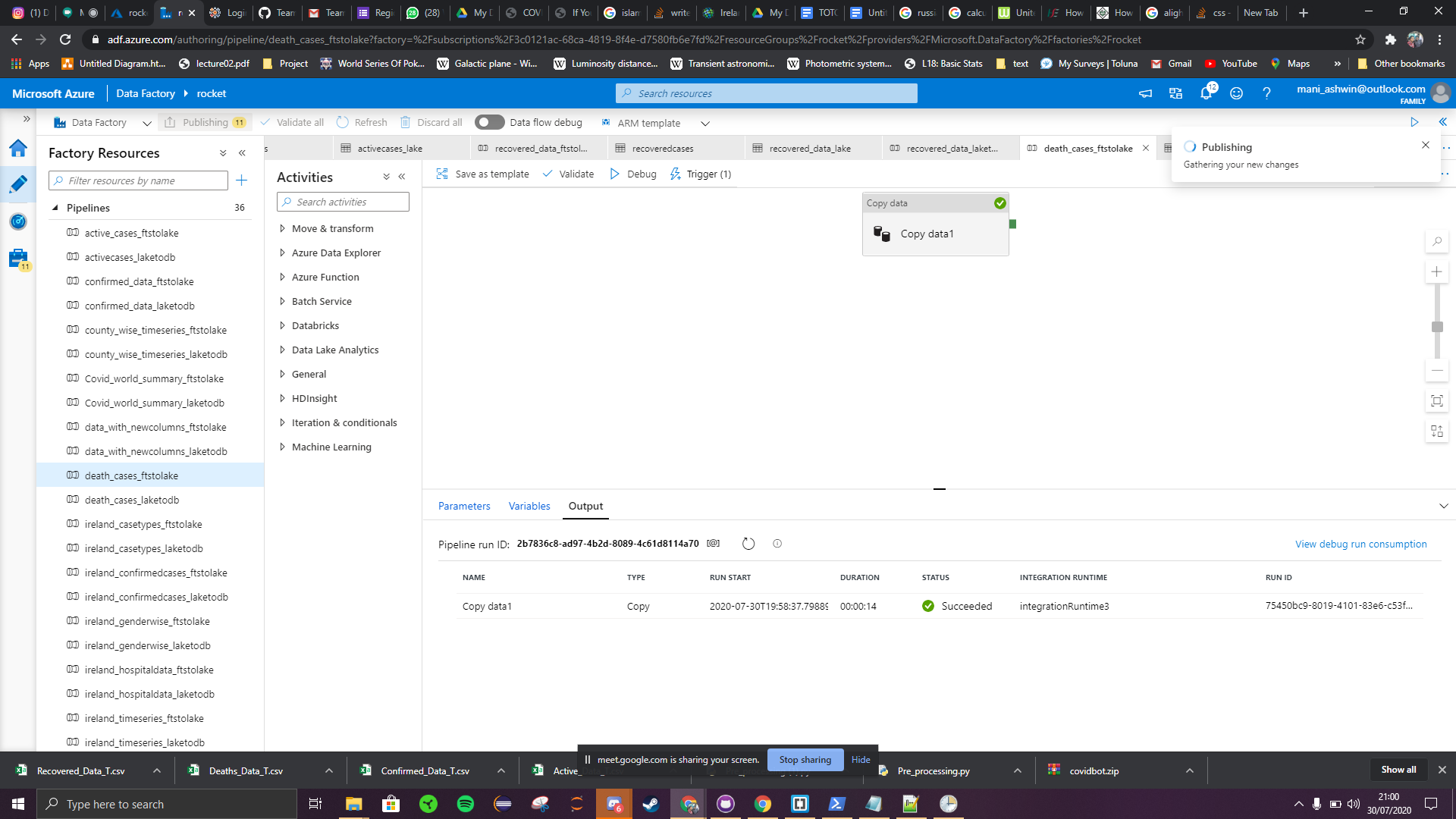
Data lake is like a storage space on cloud for all formats of data and mimics the structure of our file system with folders and files



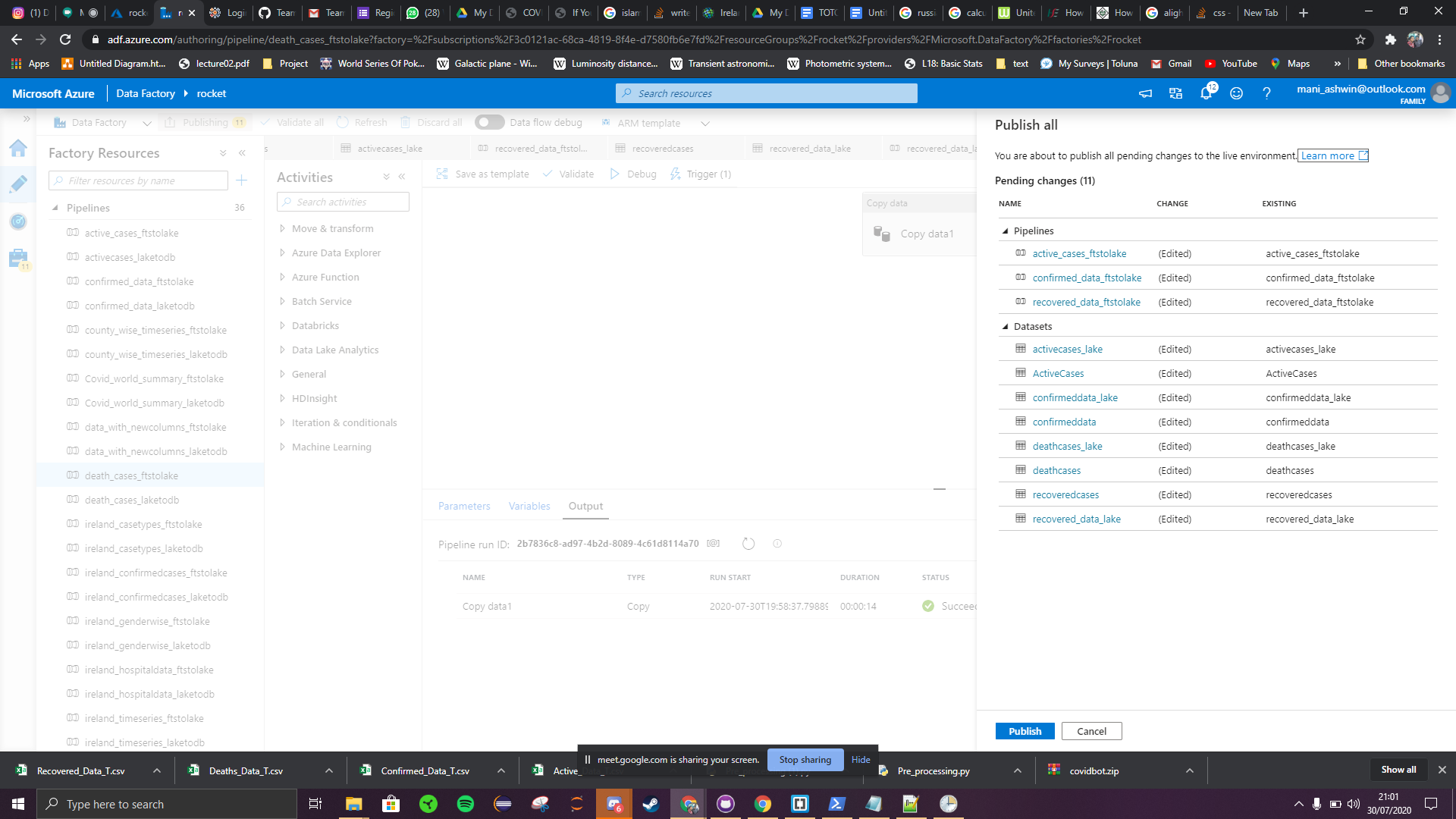
36 pipelines were created in total that connected 57 datasets. Each pipelines connects data from a source to a sink. It also has parameters for mapping and an option to add user properties



Once a pipeline is built, it can be debugged to check if it is successfully deployed



If it undergoes a successful debug, it can be published and saved. This is repeated for all pipelines



After the pipelines are created, triggers are created and mapped to all the pipelines and the triggers are published to save the entity as one