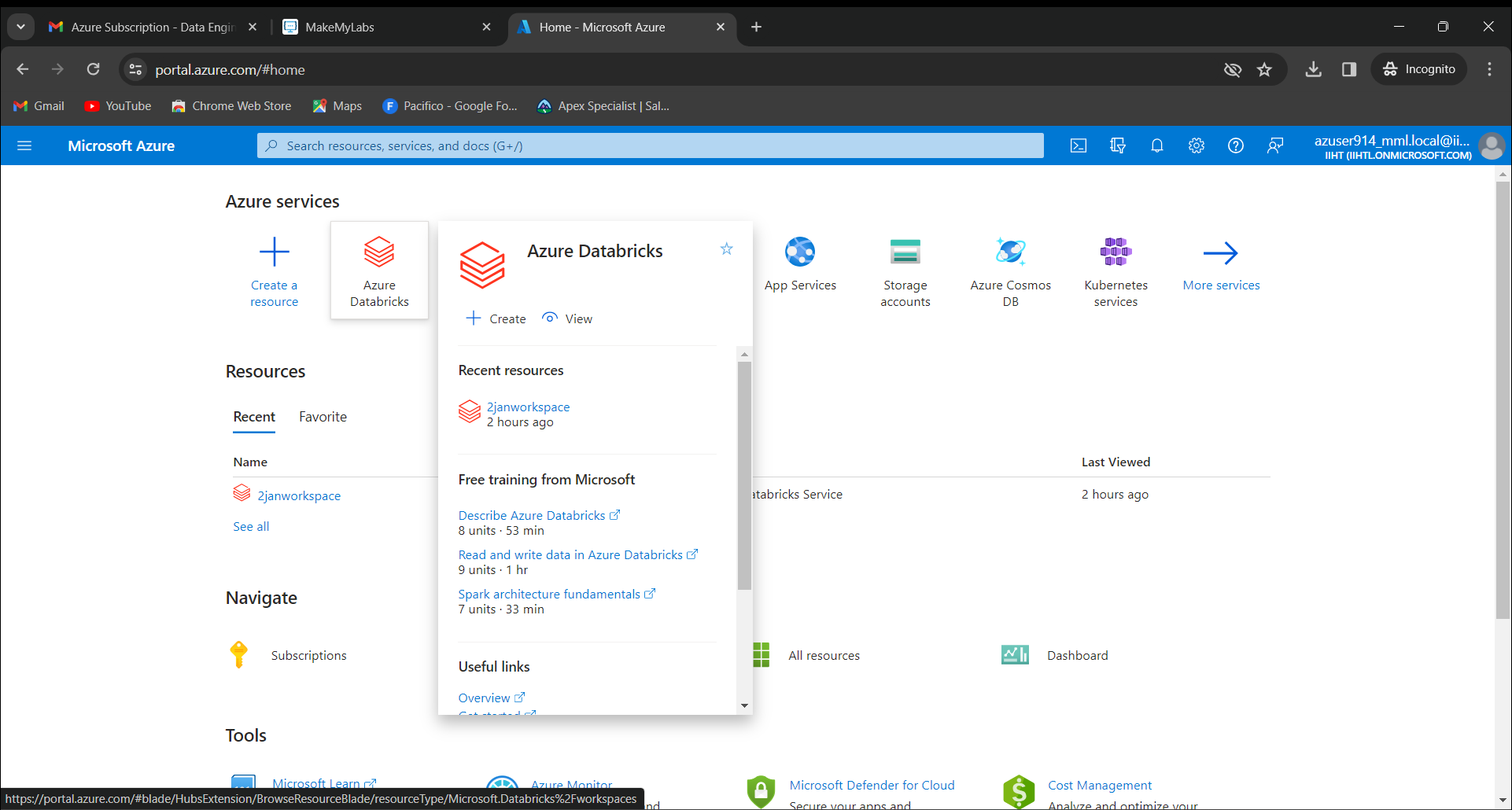
02/01/2024

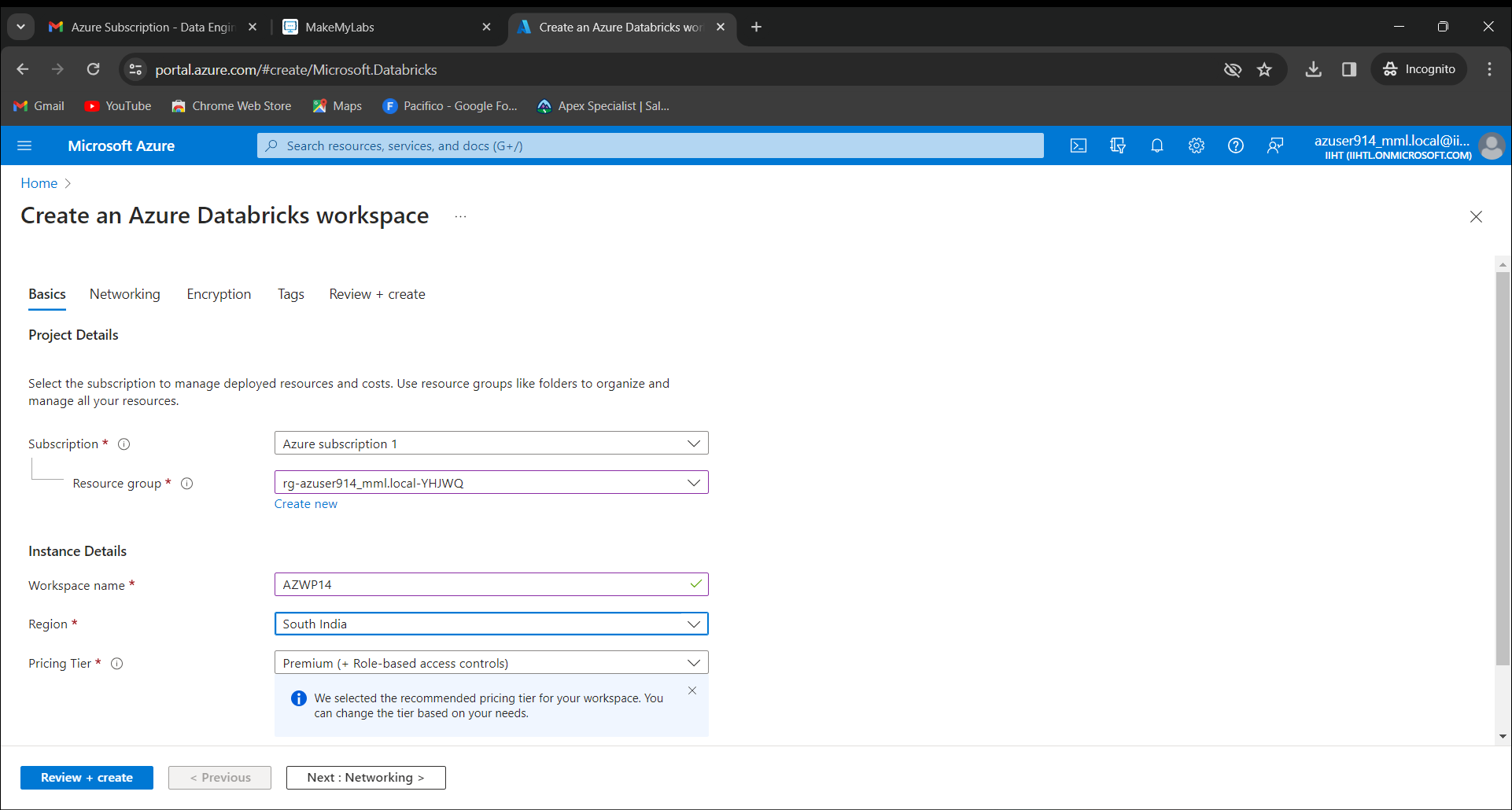
Chakradhar Bhogapurapu, BIBA

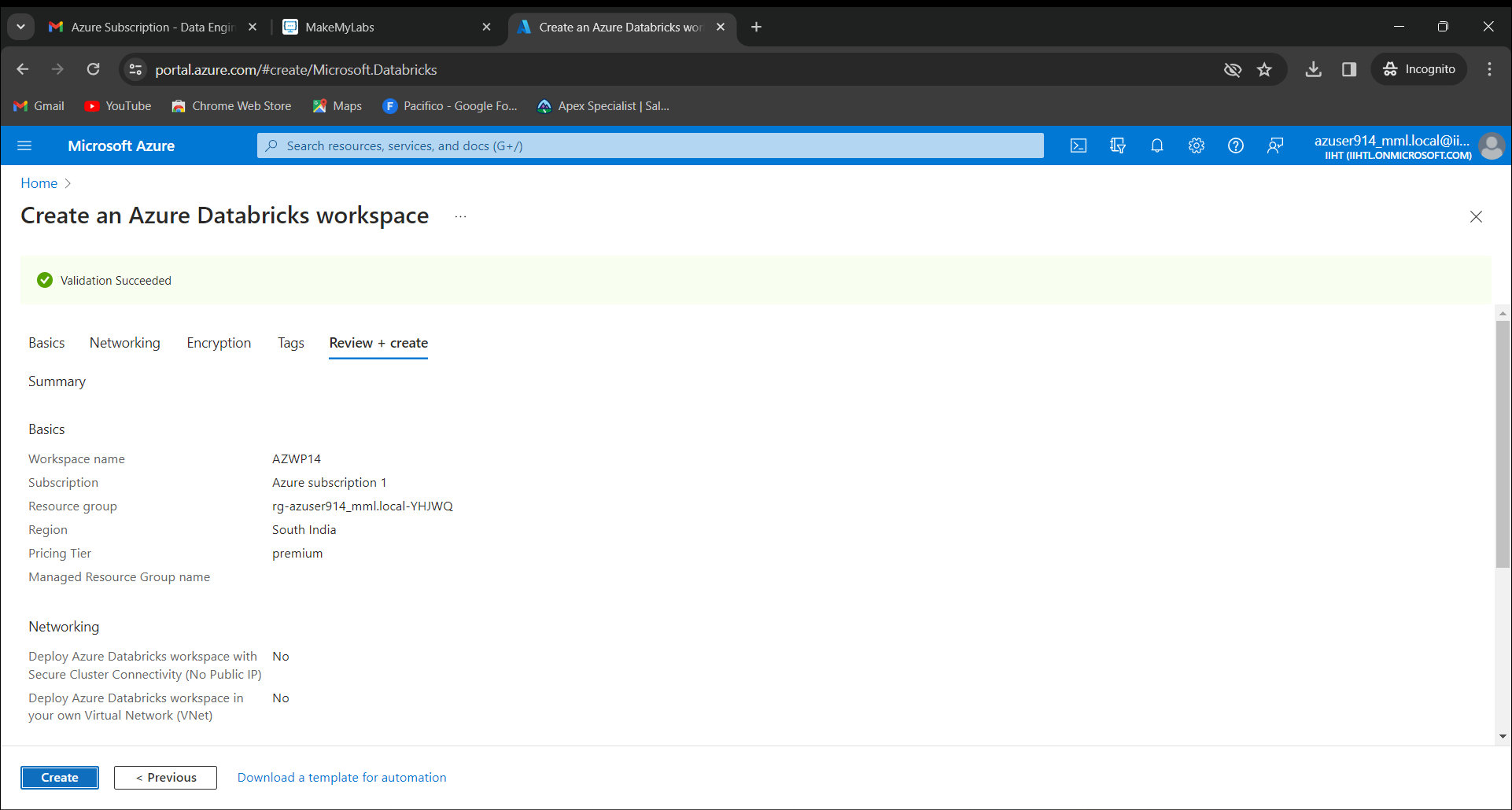
Azure Databricks

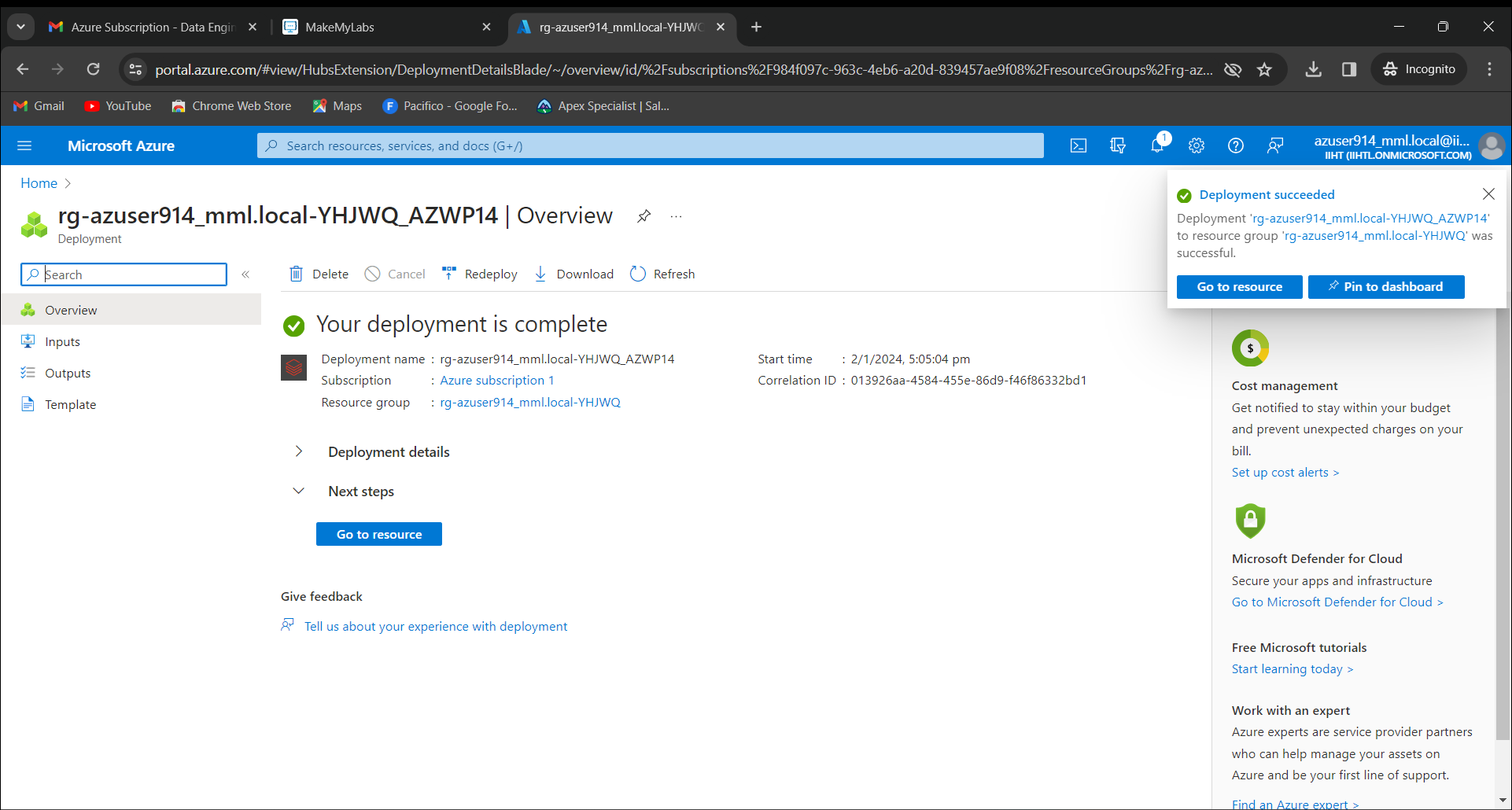
Create your own workspace in Azure Databricks

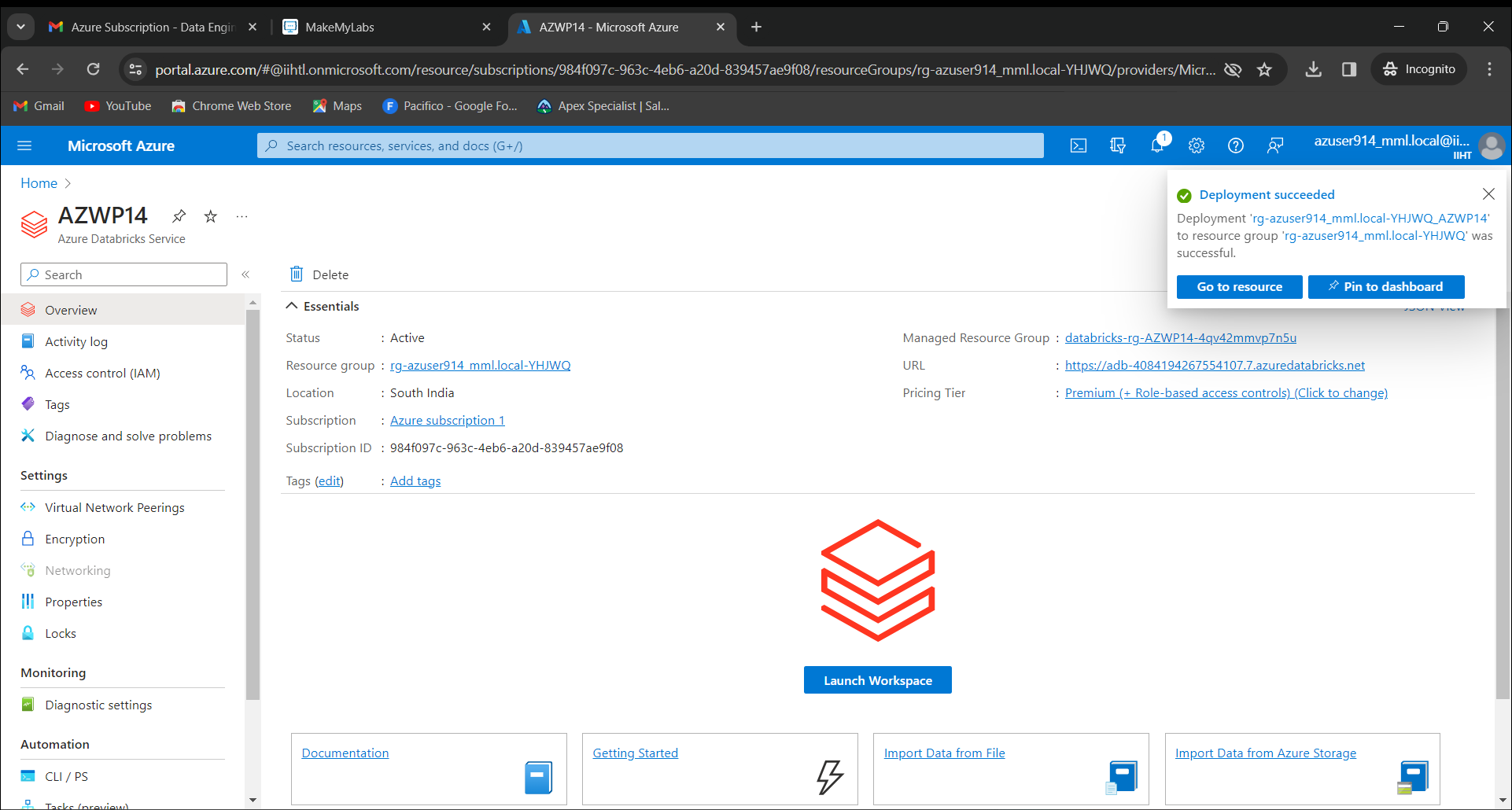


It will ask you the details and now click review+create

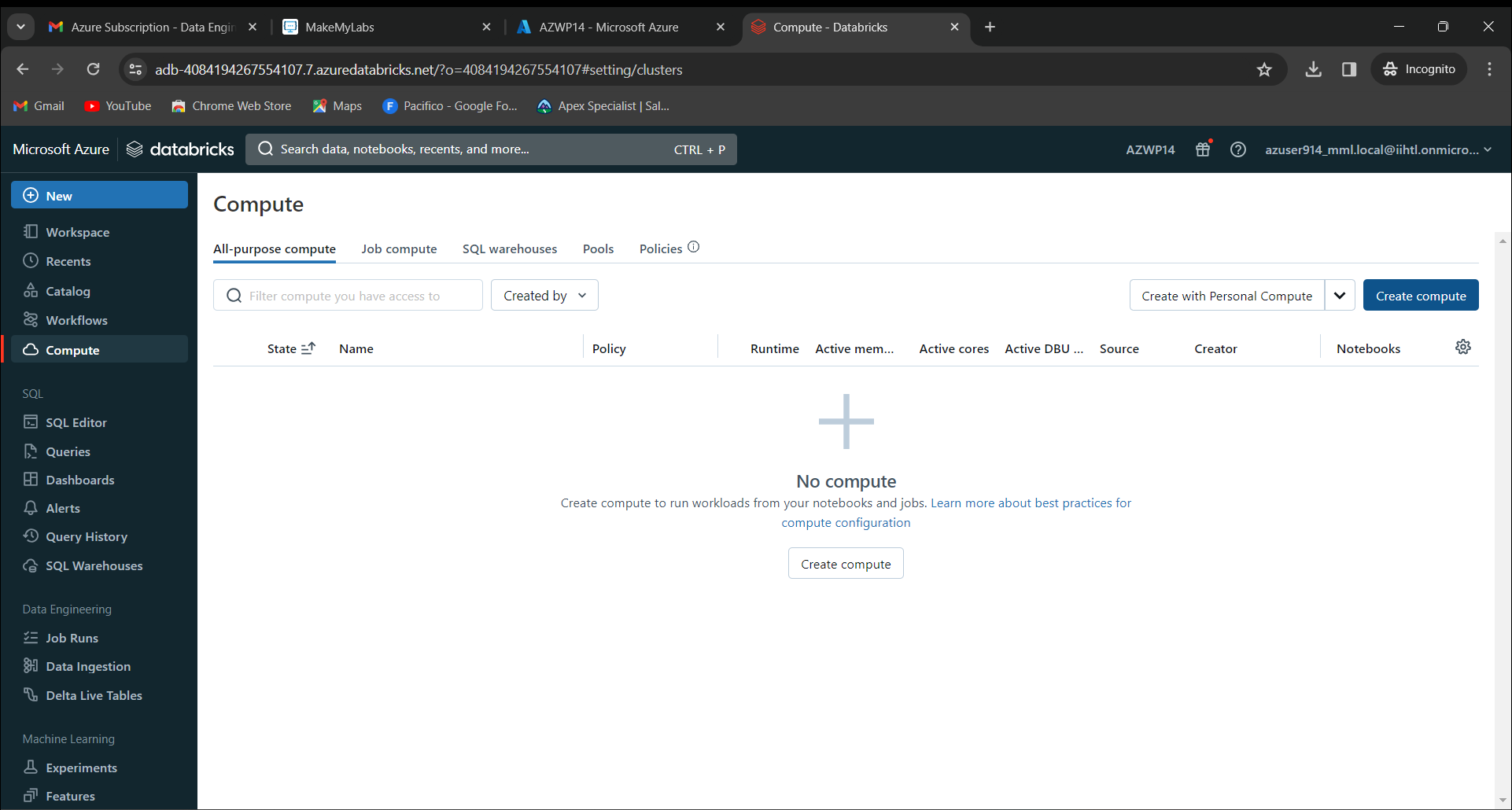


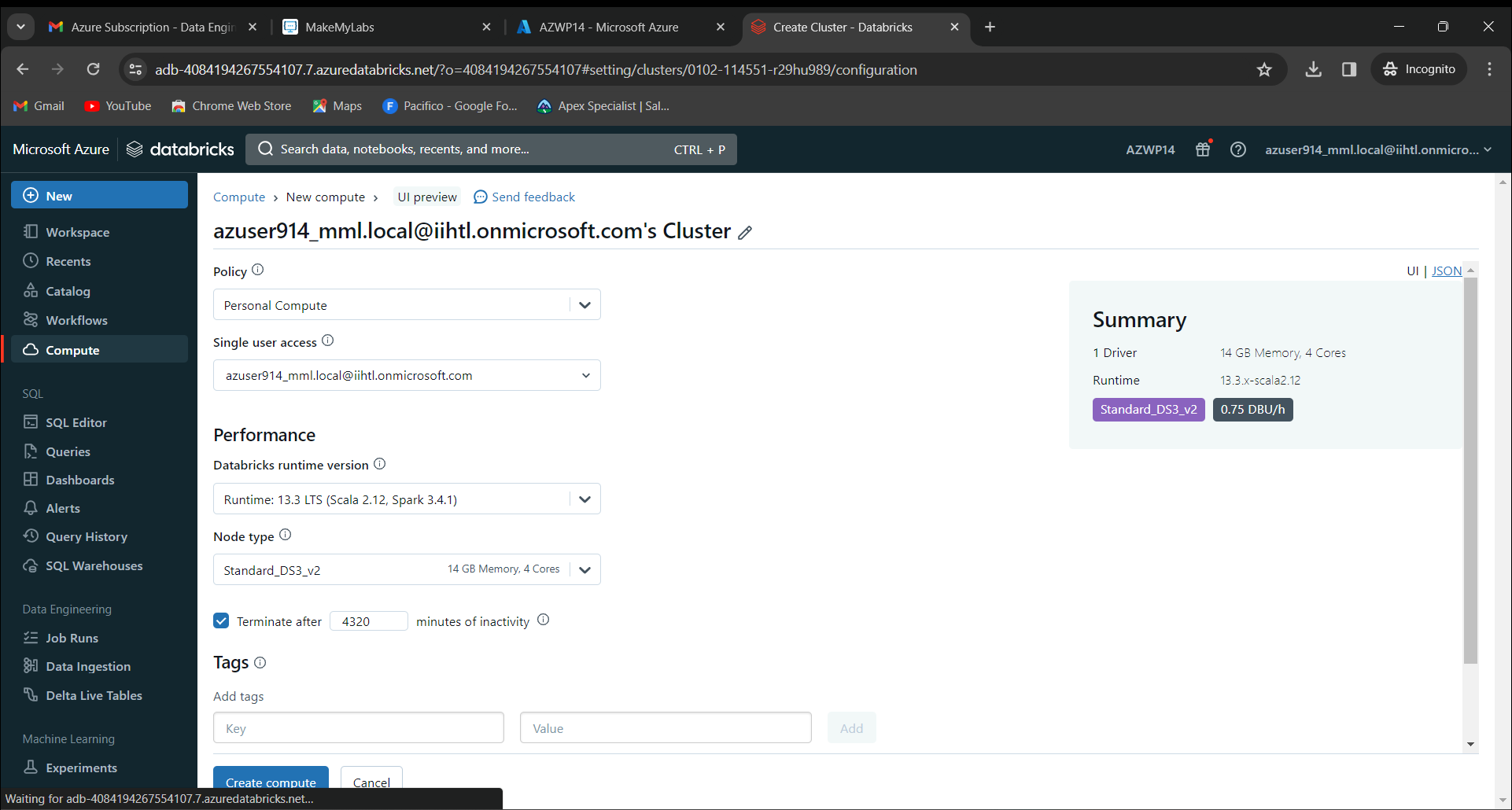
Now click create

Now our deployment is successfully done.

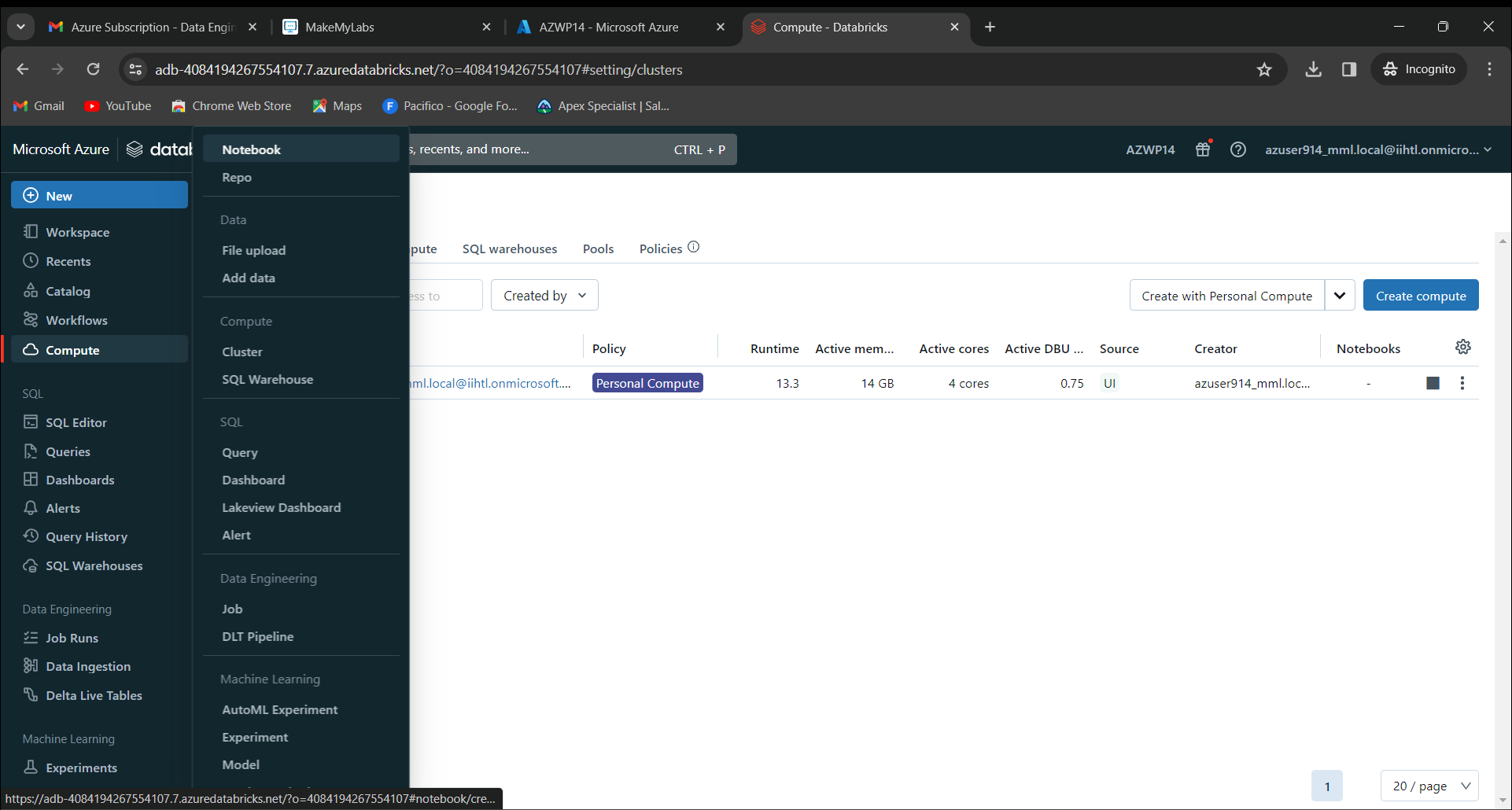
Click on launch work space. 

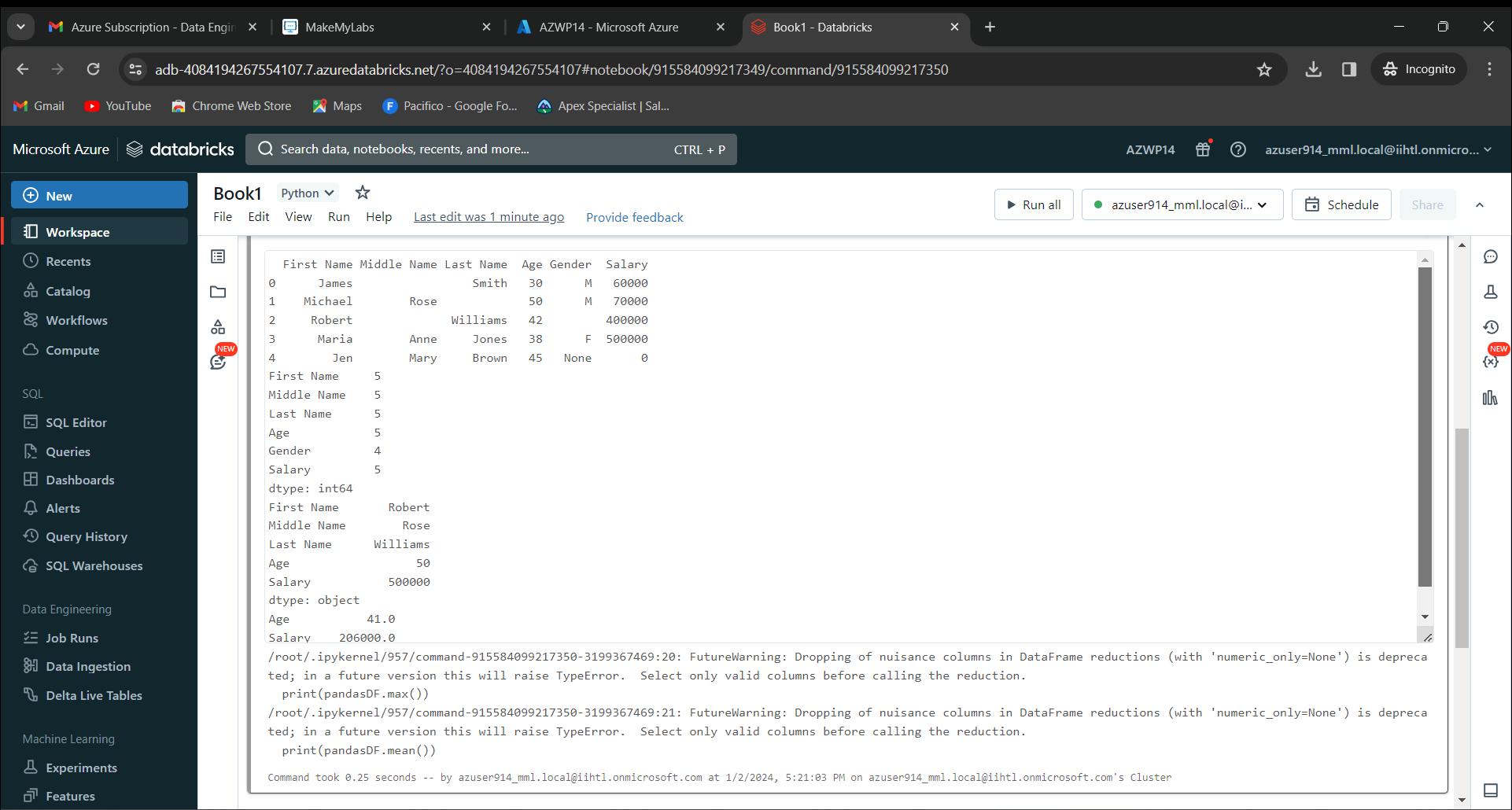
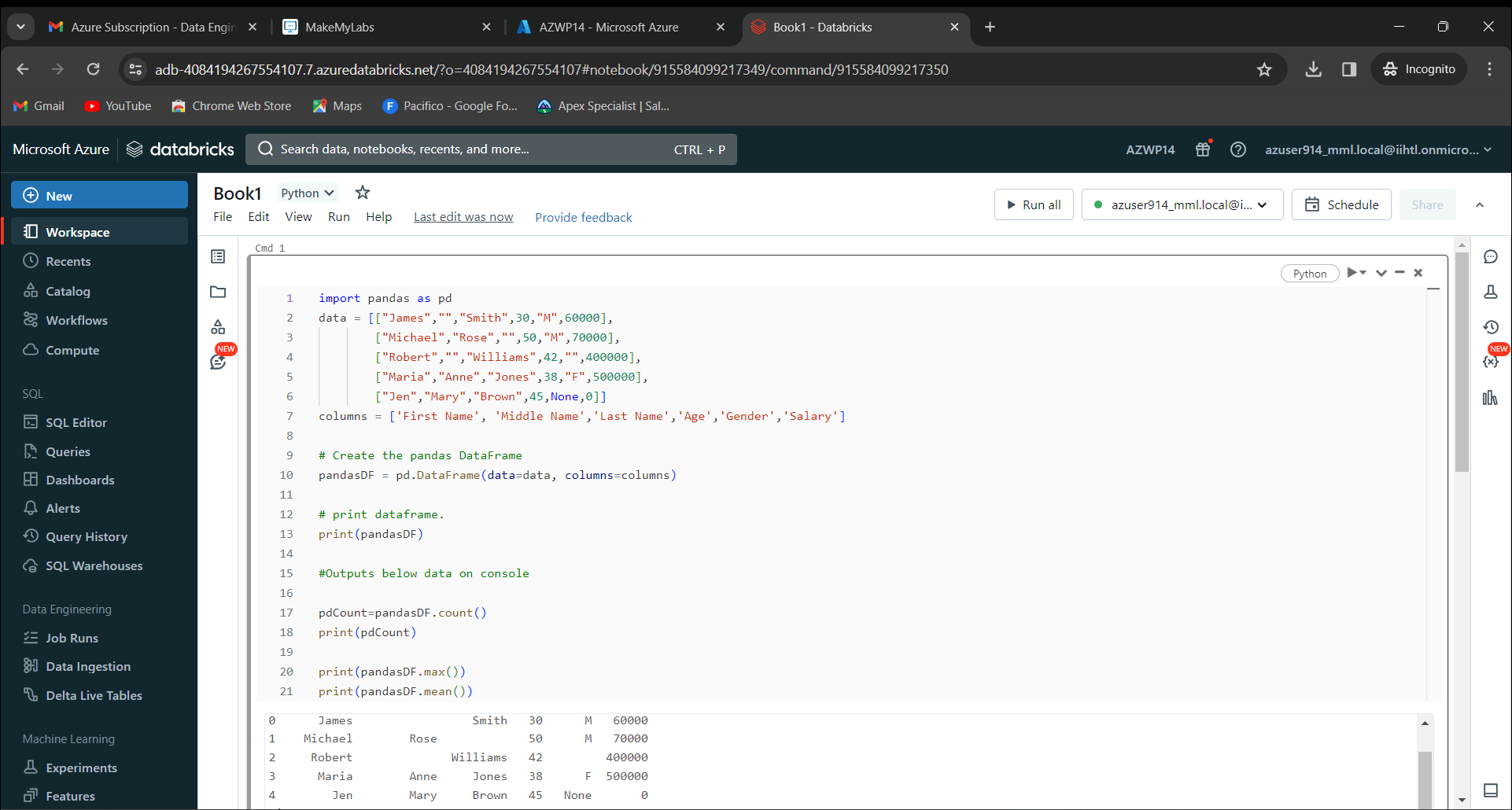
Create your own cluster





Now create a new notebook in your cluster.



Now run a program and we will see the output.

🡪Spark context is represented with sc

🡪Spark session is represented with spark

🡪RDD is an immutable distributed collection of data partitioned across nodes in your cluster with a low-level API.

🡪Dataframe is a distributed collection of rown under columns.

Delta tables

There are lot of ways to store the data in present world. Data can be stored in CSV files, JSON files, Parquet, etc. These are used to host the data. Then introduced the delta lakes in the industry that became really popular. Azure databricks helps in creating tables in various formats that includes delta also. Databricks supprts many platforms to consume the tables created in various platforms with external azure services, many of the require the table format to be of delta format.

Delta lake

Delta lake is the efficient storage layer that gives the foundation for storing the table in the databricks lakehouse. Delta lake is an open-source software extends parquet data files with a file-based transaction log for ACID and scala masterdata handling. A Delta table stores data as a directory of files on cloud object storage and registers table metadata to the metastore within a catalog and schema. Delta lake is the default storage format for all operations on databricks. All tables on databricks are delta tables unless they are specified. Databricks by default developed the delta lake protocol and continuous to actively contribute to the open-source project.

Project

Project statement will be provided for us that covers atleast 3 to 4 concepts that are covered from the training. Now we need to analize the project statement. We need to understand what are the basic concepts covered in the project statement and we need to provide the bried introduction for the project. Also we need to give the brief introduction about the tools that are used in the project. Now have a detailed plan on how to do the project and steps that how can we get the output and explain the steps that we did. According to that steps, make an ER diagram. For showing the workflow. To do this project, we need to have azure account subscription so that we can do project easily. Now we need to give the explanation on how the project works and overwiew of the execution process. List the resources and analyse the results that came as output. Make list of what technology and tools are used. At last, give a conclusion for the project.