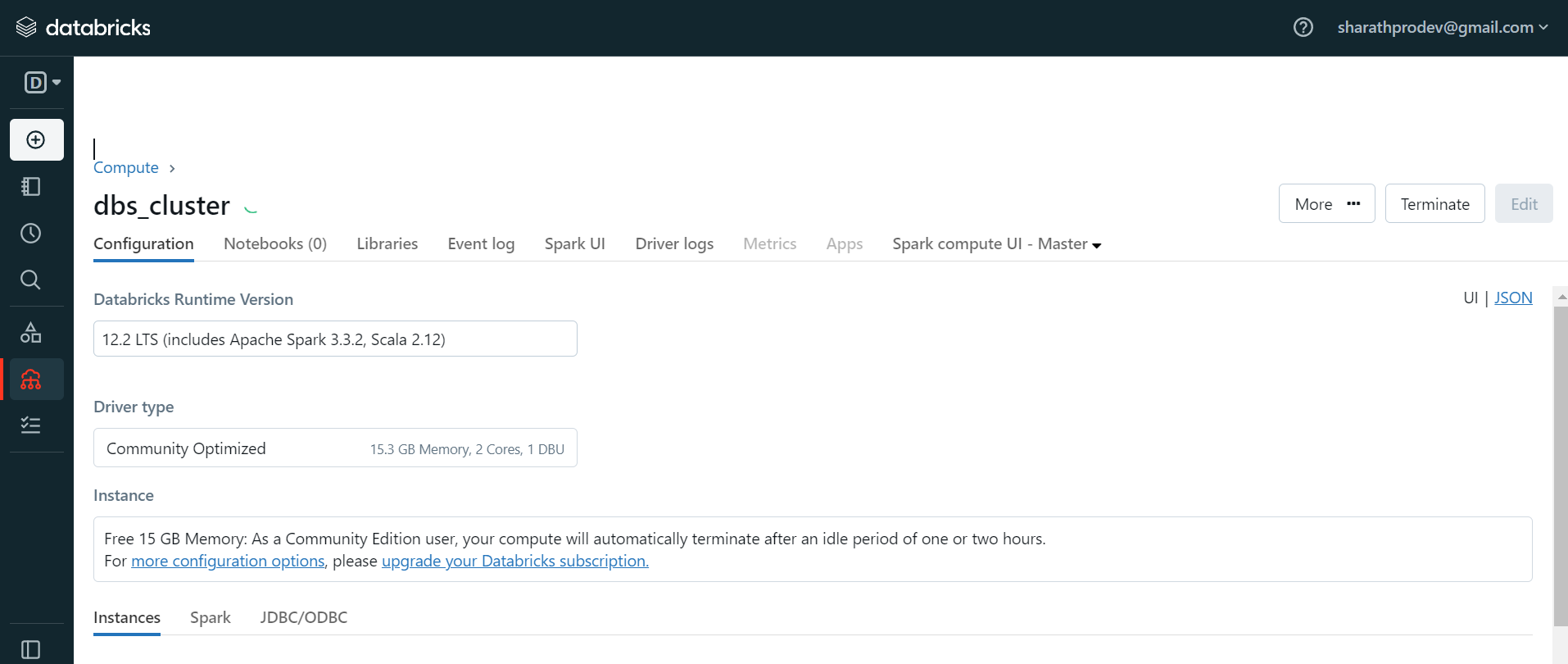
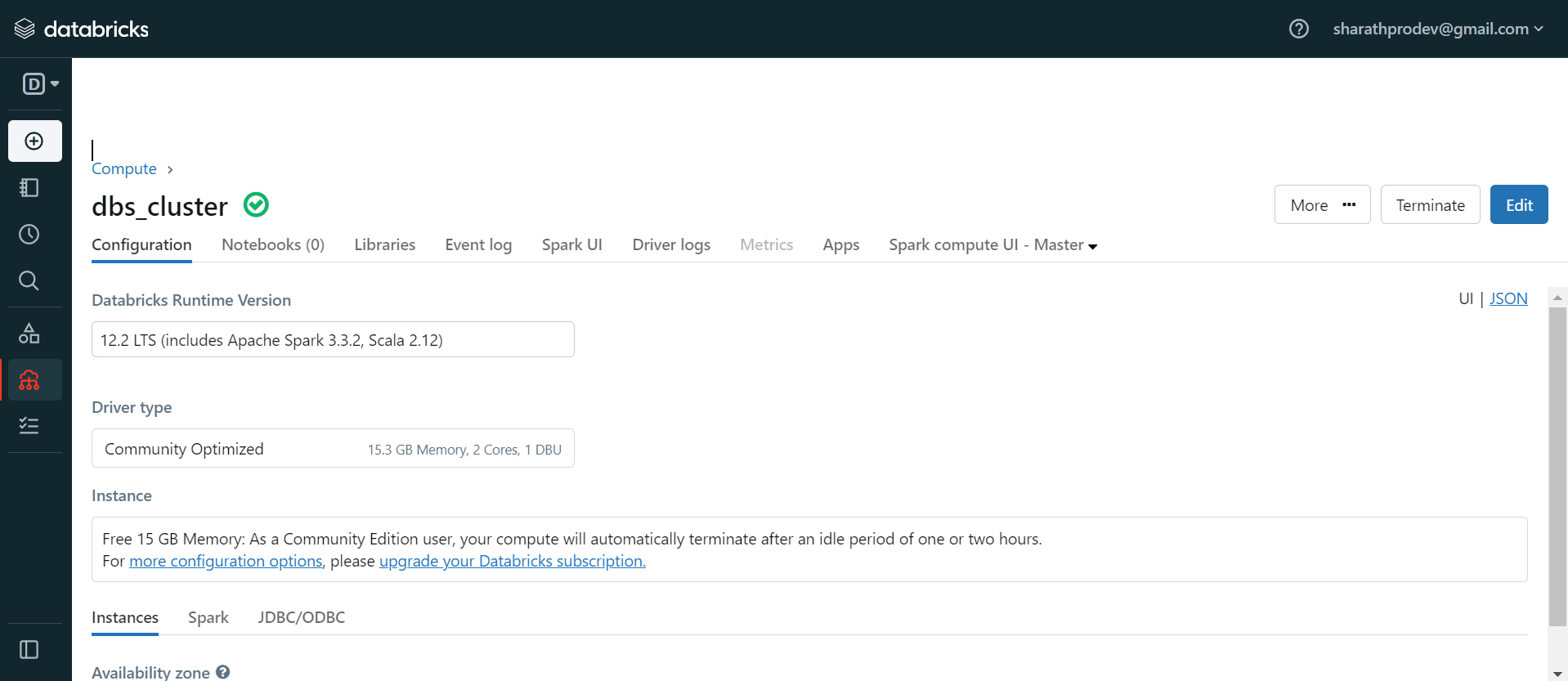


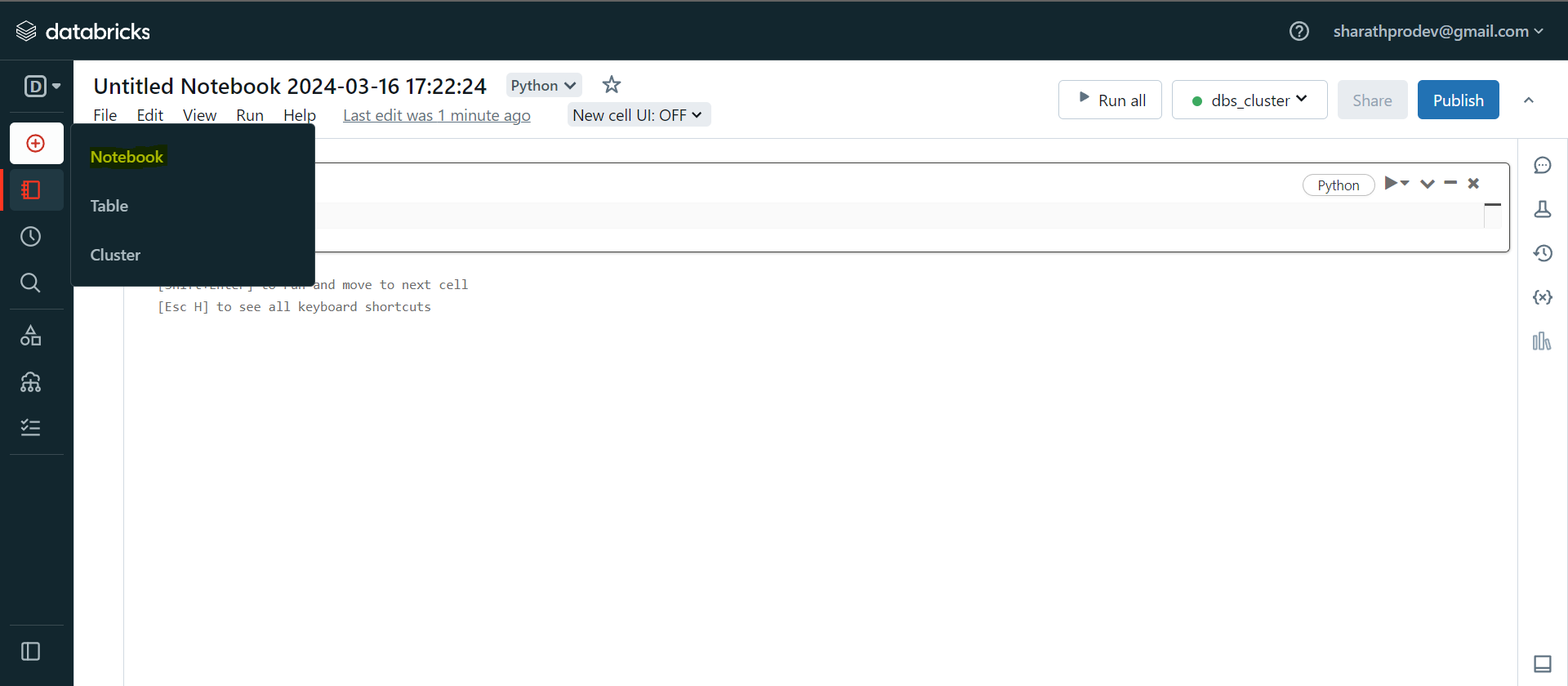
Cluster is getting created



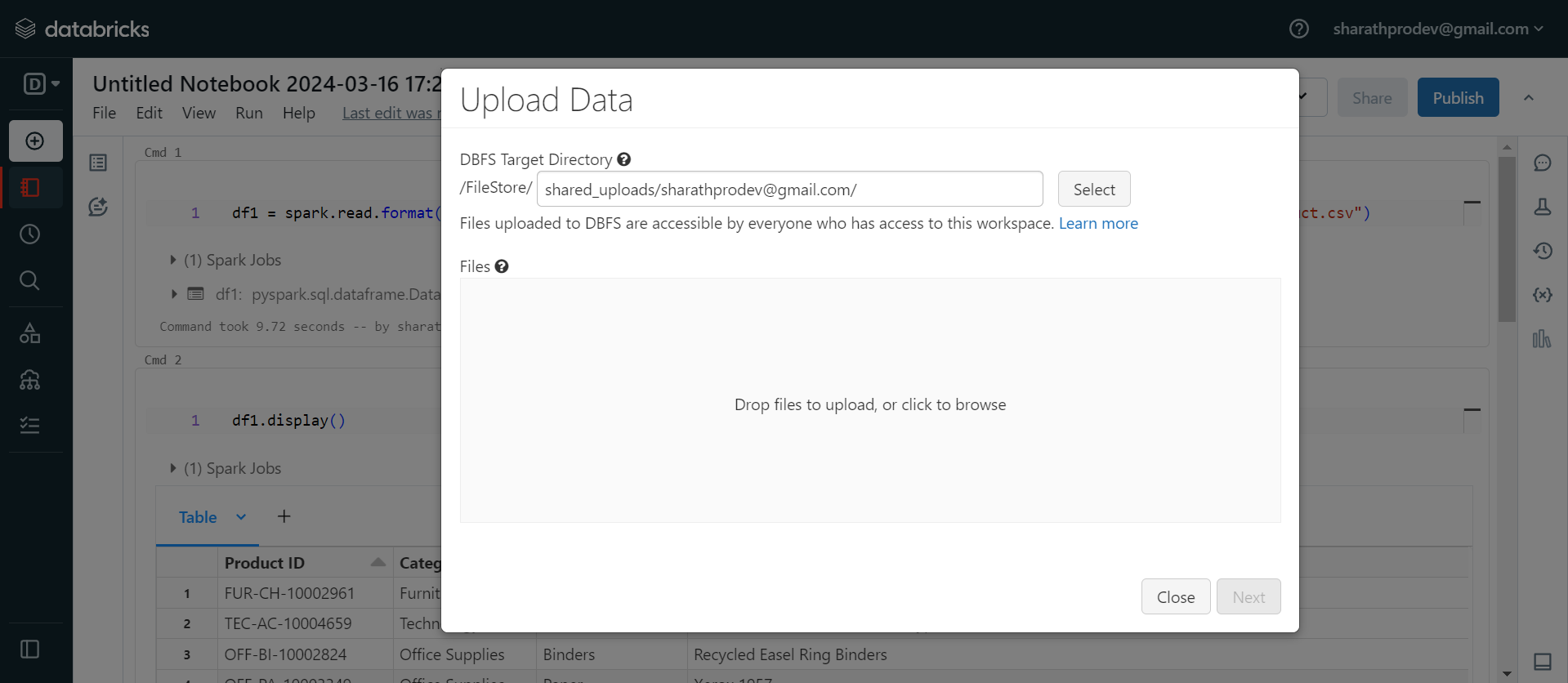
Cluster is now Created



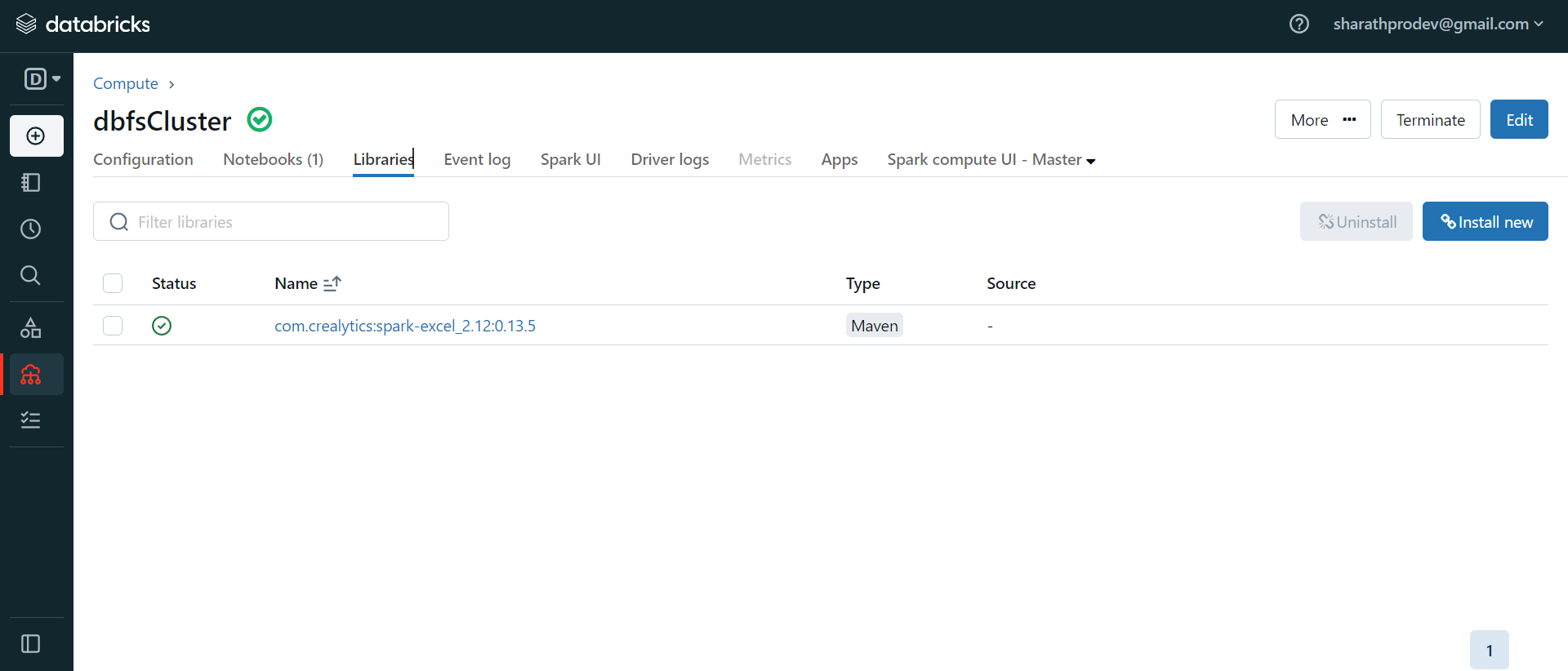
After Creating Cluster, Created Notebook



Import all Data, under file 🡪 Upload



Added below library to load xlsx data file



**Senior Data Engineer Task**

**Scenario: E-commerce Sales Data Processing with Databricks**

You've been assigned to design and implement a data processing system using Databricks for an e-commerce platform. This platform generates a lot of sales data, including details about orders, products, customers, and transactions. Your goal is to use Databricks to create a scalable, efficient, and reliable data engineering solution. This solution should process and analyze the data to provide valuable insights for business stakeholders.

**Source Datasets:**

Order.json:

|  |
| --- |
| **Orders Table** |
| Row ID |
| **Order ID** |
| Order Date |
| Ship Date |
| Ship Mode |
| **Customer ID** |
| **Product ID** |
| Quantity |
| Price |
| Discount |
| Profit |

Customer.xlsx:

|  |
| --- |
| **Customers table** |
| Customer ID |
| Customer Name |
| email |
| phone |
| address |
| Segment |
| Country |
| City |
| State |
| Postal Code |
| Region |

Product.csv:

|  |
| --- |
| **Product table** |
| **Product ID** |
| Category |
| Sub-Category |
| Product Name |
| State |
| Price per product |

**Data Transformation and Processing:**

Your task is to process the raw sales data using **Databricks notebooks** and **PySpark**. You need to clean up the data and transform it into structured formats suitable for analysis. Specifically, you should **create a master table** and **perform aggregations** based on the requirements provided.

**Note**: Follow a **test-driven development (TDD) approach**. Write appropriate test cases to ensure the correctness of the aggregations for the given scenarios.

**Task**

1. **Create raw tables for each source dataset**
2. **Create an enriched table for customers and products**
3. **Create an enriched table which has**
   1. **order information**
      1. **Profit rounded to 2 decimal places**
   2. **Customer name and country**
   3. **Product category and sub category**
4. **Create an aggregate table that shows profit by**
   1. **Year**
   2. **Product Category**
   3. **Product Sub Category**
   4. **Customer**
5. **Using SQL output the following aggregates**
   1. **Profit by Year**
   2. **Profit by Year + Product Category**
   3. **Profit by Customer**
   4. **Profit by Customer + Year**

**Notes:**

* Ensure you understand the task requirements thoroughly before starting.
* Pay attention to specific details and expectations outlined in the task descriptions.
* Use a test-driven development approach to validate the correctness of your implementations.
* Write comprehensive test cases to cover different scenarios and edge cases.
* Ensure your solution handles data quality issues and implements robust error-handling mechanisms.
* Document your code and assumptions clearly to make it understandable for others.
* Consider performance implications and optimize your code for efficiency and scalability.

