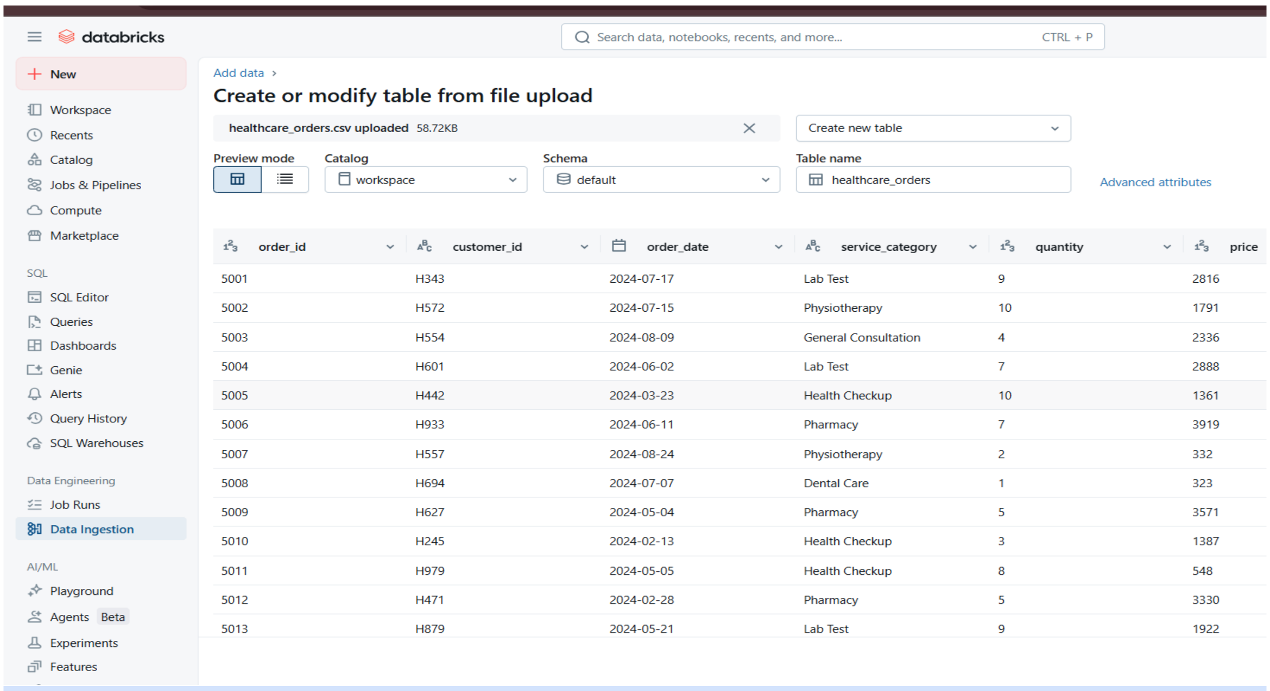
LAB 4

Name : Sahanaa M S

USN : 1RVU23CSE394

1) UPLOAD healthcare_orders.csv TO WORKSPACE:



Add data ›

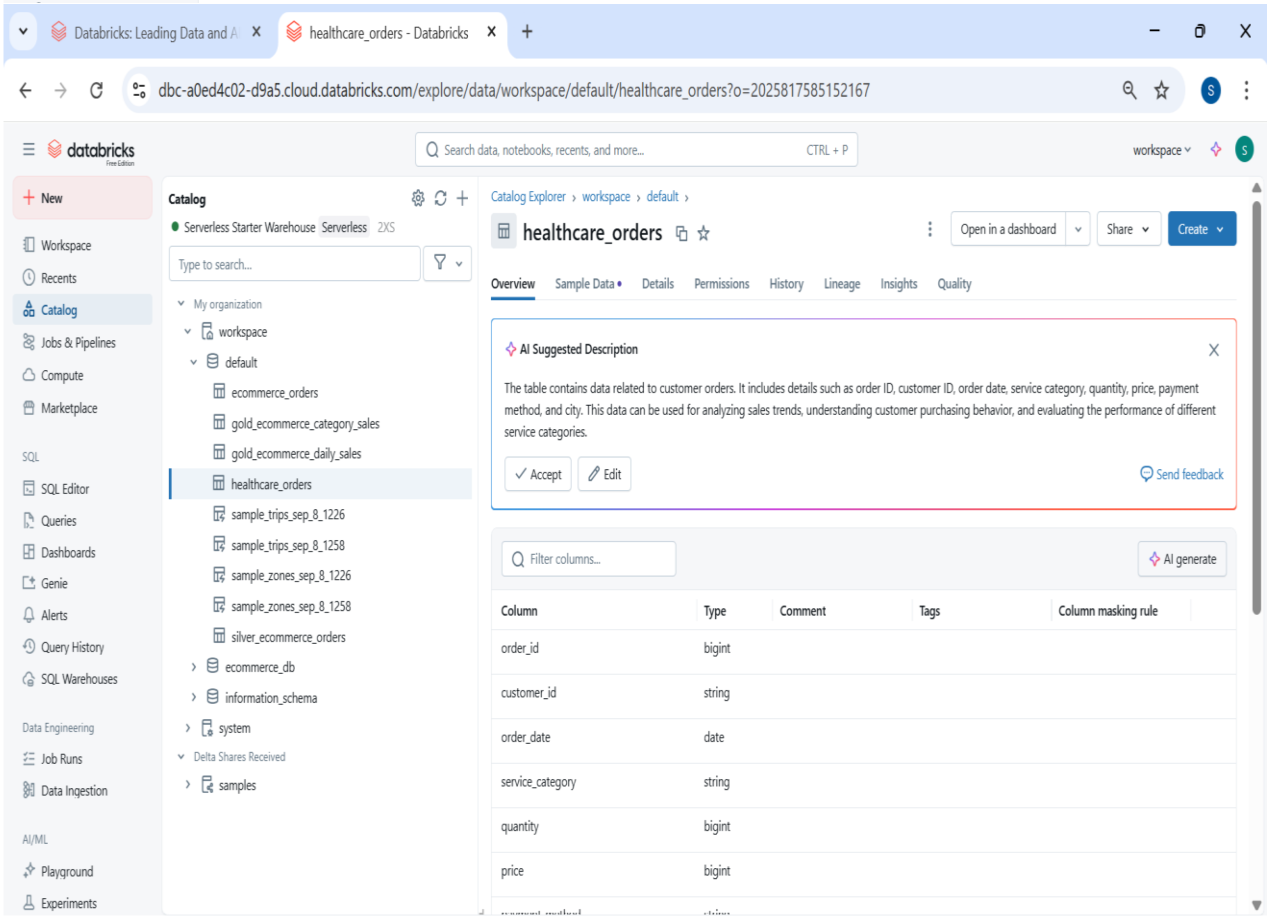
Create or modify table from file upload

healthcare_orders.csv uploaded 58.72KB

Create new table

Preview mode | Catalog: workspace | Schema: default | Table name: healthcare_orders | Advanced attributes

order_id	customer_id	order_date	service_category	quantity	price
5001	H343	2024-07-17	Lab Test	9	2816
5002	H572	2024-07-15	Physiotherapy	10	1791
5003	H554	2024-08-09	General Consultation	4	2336
5004	H601	2024-06-02	Lab Test	7	2888
5005	H442	2024-03-23	Health Checkup	10	1361
5006	H933	2024-06-11	Pharmacy	7	3919
5007	H557	2024-08-24	Physiotherapy	2	332
5008	H694	2024-07-07	Dental Care	1	323
5009	H627	2024-05-04	Pharmacy	5	3571
5010	H245	2024-02-13	Health Checkup	3	1387
5011	H979	2024-05-05	Health Checkup	8	548
5012	H471	2024-02-28	Pharmacy	5	3330
5013	H879	2024-05-21	Lab Test	9	1922



dbc-a0ed4c02-d9a5.cloud.databricks.com/explore/data/workspace/default/healthcare_orders?o=2025817585152167

Catalog Explorer › workspace › default ›

healthcare_orders

Overview | Sample Data | Details | Permissions | History | Lineage | Insights | Quality

AI Suggested Description

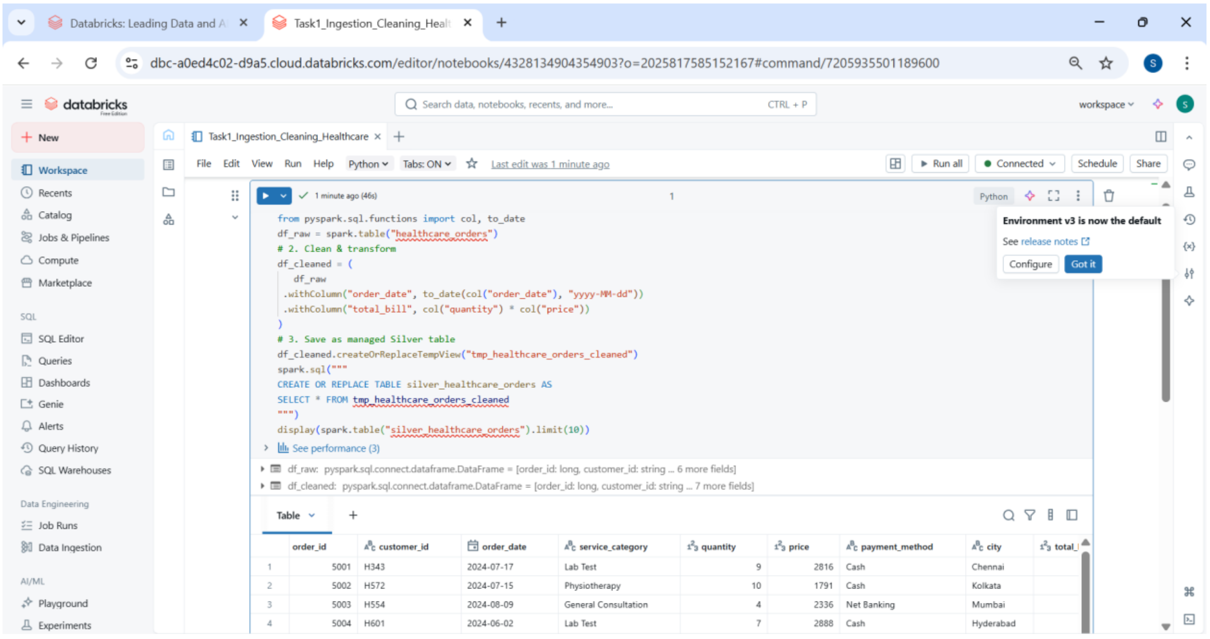
The table contains data related to customer orders. It includes details such as order ID, customer ID, order date, service category, quantity, price, payment method, and city. This data can be used for analyzing sales trends, understanding customer purchasing behavior, and evaluating the performance of different service categories.

Accept | Edit | Send feedback

Catalog: My organization › workspace › default: ecommerce_orders, gold_ecommerce_category_sales, gold_ecommerce_daily_sales, healthcare_orders, sample_trips_sep_8_1226, sample_trips_sep_8_1258, sample_zones_sep_8_1226, sample_zones_sep_8_1258, silver_ecommerce_orders; ecommerce_db; information_schema; system; Delta Shares Received: samples

Column	Type	Comment	Tags	Column masking rule
order_id	bigint			
customer_id	string			
order_date	date			
service_category	string			
quantity	bigint			
price	bigint			

2) TASK1_INGESTION_CLEANING_HEALTHCARE:



dbc-a0ed4c02-d9a5.cloud.databricks.com/editor/notebooks/4328134904354903?o=2025817585152167#command/7205935501189600

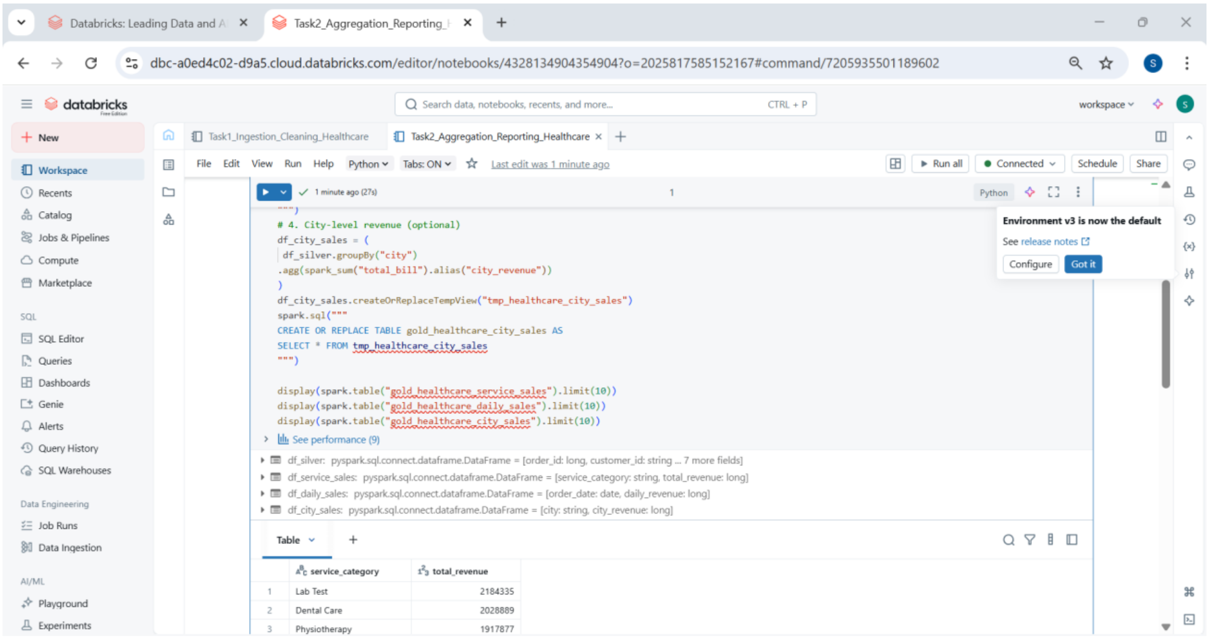
Task1_Ingestion_Cleaning_Healthcare

```python
from pyspark.sql.functions import col, to_date
df_raw = spark.table("healthcare_orders")
# 2. Clean & transform
df_cleaned = (
    df_raw
  .withColumn("order_date", to_date(col("order_date"), "yyyy-MM-dd"))
  .withColumn("total_bill", col("quantity") * col("price"))
)
# 3. Save as managed Silver table
df_cleaned.createOrReplaceTempView("tmp_healthcare_orders_cleaned")
spark.sql("""
CREATE OR REPLACE TABLE silver_healthcare_orders AS
SELECT * FROM tmp_healthcare_orders_cleaned
""")
display(spark.table("silver_healthcare_orders").limit(10))
```

```
df_raw:  pyspark.sql.connect.dataframe.DataFrame = [order_id: long, customer_id: string ... 6 more fields]
df_cleaned:  pyspark.sql.connect.dataframe.DataFrame = [order_id: long, customer_id: string ... 7 more fields]
```

	order_id	customer_id	order_date	service_category	quantity	price	payment_method	city	total_
1	5001	H343	2024-07-17	Lab Test	9	2816	Cash	Chennai	
2	5002	H572	2024-07-15	Physiotherapy	10	1791	Cash	Kolkata	
3	5003	H554	2024-08-09	General Consultation	4	2336	Net Banking	Mumbai	
4	5004	H601	2024-06-02	Lab Test	7	2888	Cash	Hyderabad	

3) TASK2_AGGREGATION_REPORTING_HEALTHCARE:



dbc-a0ed4c02-d9a5.cloud.databricks.com/editor/notebooks/4328134904354904?o=2025817585152167#command/7205935501189602

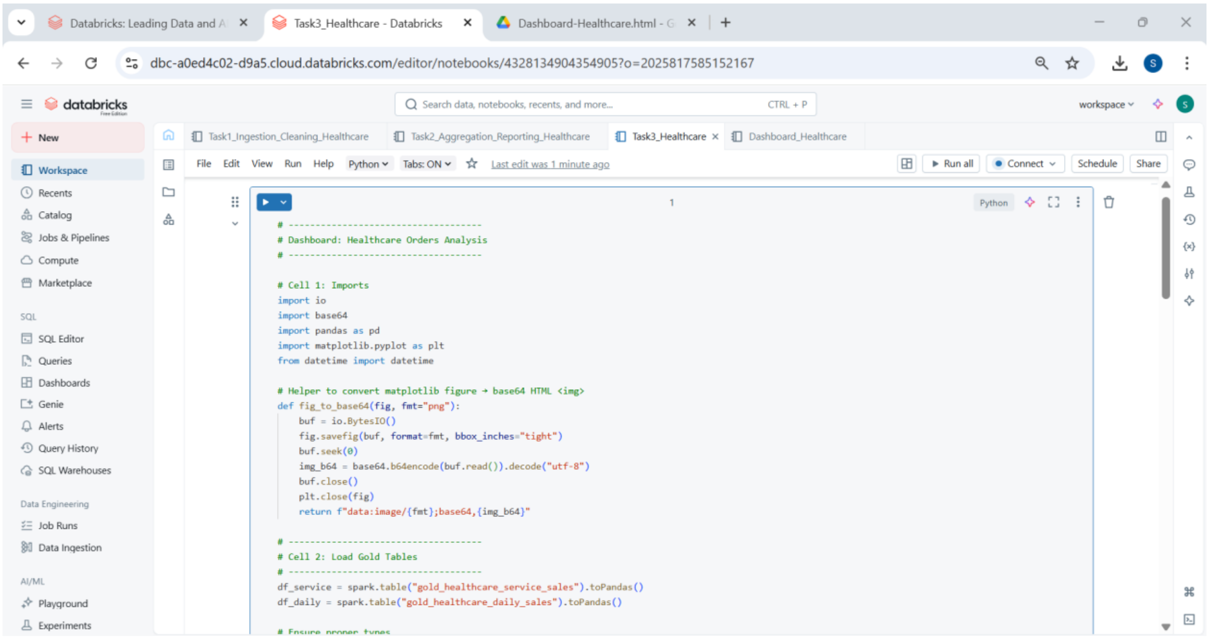
Task2_Aggregation_Reporting_Healthcare

```python
""")
# 4. City-level revenue (optional)
df_city_sales = (
 df_silver.groupBy("city")
.agg(spark_sum("total_bill").alias("city_revenue"))
)
df_city_sales.createOrReplaceTempView("tmp_healthcare_city_sales")
spark.sql("""
CREATE OR REPLACE TABLE gold_healthcare_city_sales AS
SELECT * FROM tmp_healthcare_city_sales
""")

display(spark.table("gold_healthcare_service_sales").limit(10))
display(spark.table("gold_healthcare_daily_sales").limit(10))
display(spark.table("gold_healthcare_city_sales").limit(10))
```

```
df_silver:  pyspark.sql.connect.dataframe.DataFrame = [order_id: long, customer_id: string ... 7 more fields]
df_service_sales:  pyspark.sql.connect.dataframe.DataFrame = [service_category: string, total_revenue: long]
df_daily_sales:  pyspark.sql.connect.dataframe.DataFrame = [order_date: date, daily_revenue: long]
df_city_sales:  pyspark.sql.connect.dataframe.DataFrame = [city: string, city_revenue: long]
```

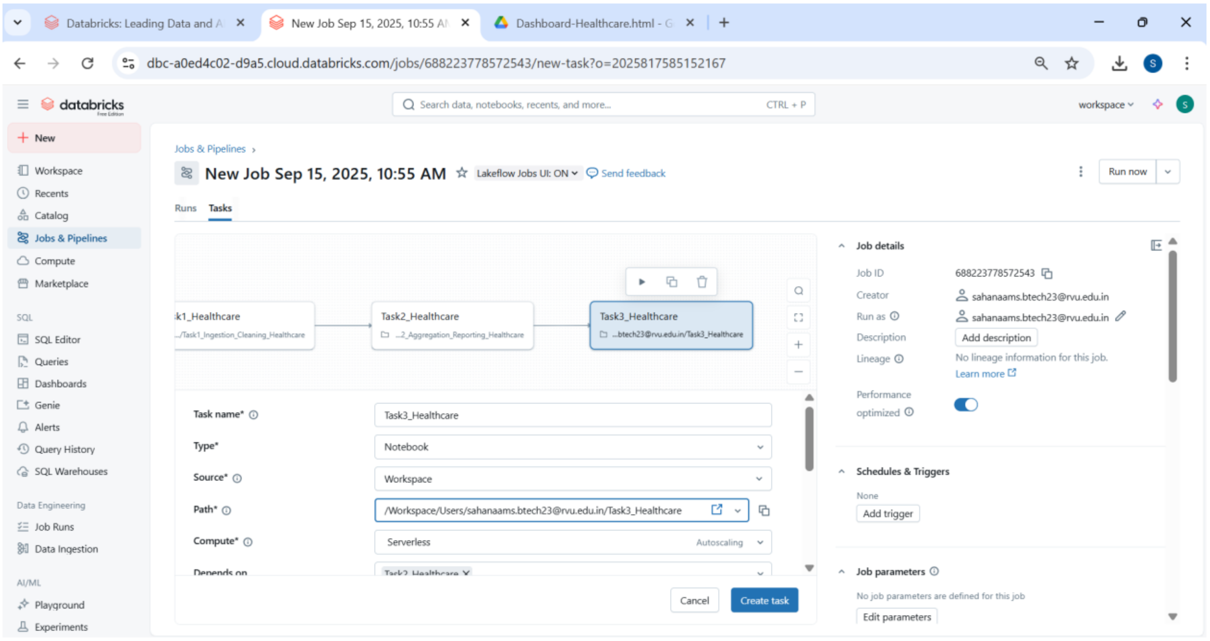
	service_category	total_revenue
1	Lab Test	2184335
2	Dental Care	2028889
3	Physiotherapy	1917877

4) TASK3_HEALTHCARE AND DASHBOARD_HEALTHCARE :


dbc-a0ed4c02-d9a5.cloud.databricks.com/editor/notebooks/4328134904354905?o=2025817585152167

```python
# ----------------------------------
# Dashboard: Healthcare Orders Analysis
# ----------------------------------

# Cell 1: Imports
import io
import base64
import pandas as pd
import matplotlib.pyplot as plt
from datetime import datetime

# Helper to convert matplotlib figure → base64 HTML <img>
def fig_to_base64(fig, fmt="png"):
    buf = io.BytesIO()
    fig.savefig(buf, format=fmt, bbox_inches="tight")
    buf.seek(0)
    img_b64 = base64.b64encode(buf.read()).decode("utf-8")
    buf.close()
    plt.close(fig)
    return f"data:image/{fmt};base64,{img_b64}"

# ----------------------------------
# Cell 2: Load Gold Tables
# ----------------------------------
df_service = spark.table("gold_healthcare_service_sales").toPandas()
df_daily = spark.table("gold_healthcare_daily_sales").toPandas()

# Ensure proper types
```

5) JOB AND PIPELINES – CREATE JOB:


dbc-a0ed4c02-d9a5.cloud.databricks.com/jobs/688223778572543/new-task?o=2025817585152167

Jobs & Pipelines ›

New Job Sep 15, 2025, 10:55 AM

Task name*	Task3_Healthcare
Type*	Notebook
Source*	Workspace
Path*	/Workspace/Users/sahanaams.btech23@rvu.edu.in/Task3_Healthcare
Compute*	Serverless Autoscaling
Depends on	Task2_Healthcare

Job details

Job ID	688223778572543
Creator	sahanaams.btech23@rvu.edu.in
Run as	sahanaams.btech23@rvu.edu.in
Description	Add description
Lineage	No lineage information for this job. Learn more
Performance optimized	●

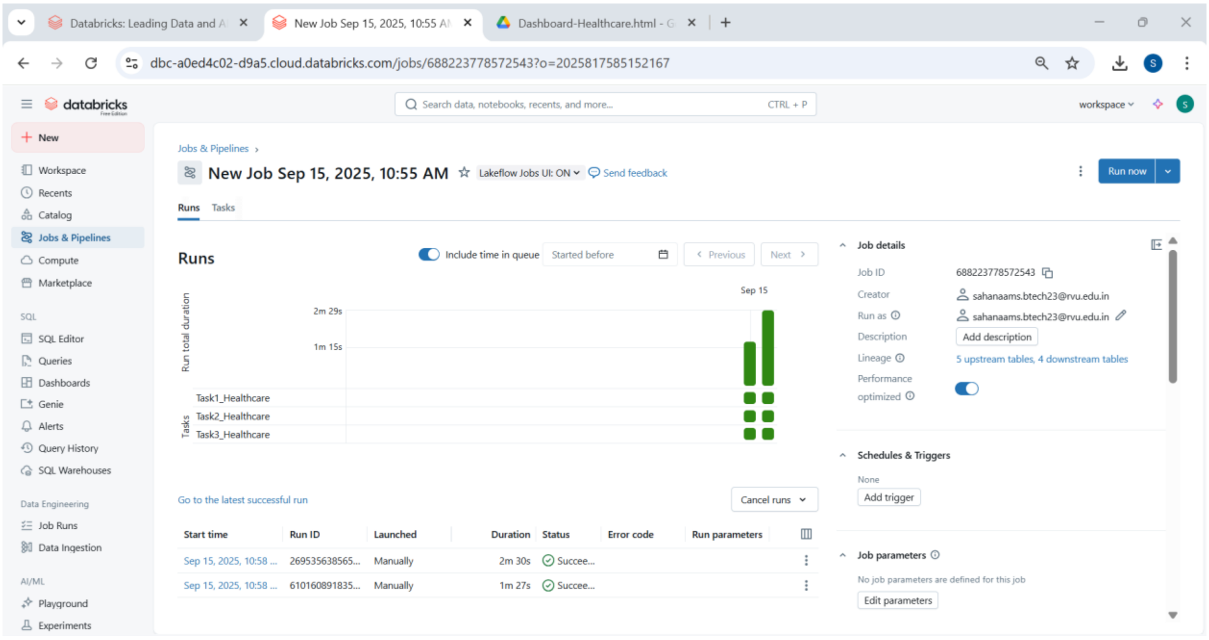
Schedules & Triggers

None

Job parameters

No job parameters are defined for this job

6) JOB AND PIPELINES – SUCCESSFUL JOB CREATED:



New Job Sep 15, 2025, 10:55 AM

Runs

Start time	Run ID	Launched	Duration	Status	Error code	Run parameters
Sep 15, 2025, 10:58 ...	269535638565...	Manually	2m 30s	Succee...		
Sep 15, 2025, 10:58 ...	610160891835...	Manually	1m 27s	Succee...		

Job details

Job ID	688223778572543
Creator	sahanaams.btech23@rvu.edu.in
Run as	sahanaams.btech23@rvu.edu.in
Description	Add description
Lineage	5 upstream tables, 4 downstream tables
Performance optimized	ON

Schedules & Triggers

None

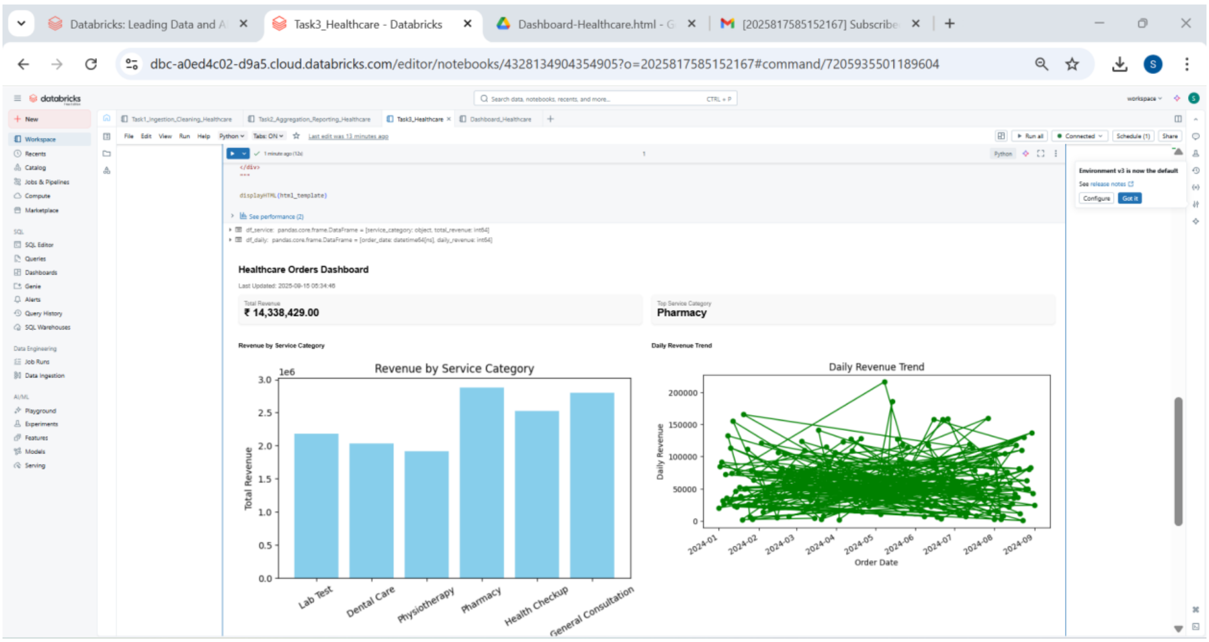
Add trigger

Job parameters

No job parameters are defined for this job

Edit parameters

7) DASHBOARD REPORT:


Databricks: Leading Data and A
Task3_Healthcare - Databricks
Dashboard-Healthcare.html - G
[2025817585152167] Subscribe
dbc-a0ed4c02-d9a5.cloud.databricks.com/editor/notebooks/4328134904354905?o=2025817585152167#command/7205935501189604
Task1_Ingestion_Cleaning_Healthcare
Task2_Aggregation_Reporting_Healthcare
Task3_Healthcare
Dashboard_Healthcare
File
Edit
View
Run
Help
Python
Tabs: ON
Last edit was 13 minutes ago
Run all
Connected
Schedule (1)
Share
1 minute ago (12s)
Environment v3 is now the default
See release notes
Configure
Got it
Revenue by Service Category
1e6
Total Revenue
Lab Test
Dental Care
Physiotherapy
Pharmacy
Health Checkup
General Consultation
Daily Revenue Trend
Daily Revenue
Order Date
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09

```
</div>
"""

displayHTML(html_template)
```

See performance (2)

```
df_service:  pandas.core.frame.DataFrame = [service_category: object, total_revenue: int64]
df_daily:  pandas.core.frame.DataFrame = [order_date: datetime64[ns], daily_revenue: int64]
```

Healthcare Orders Dashboard

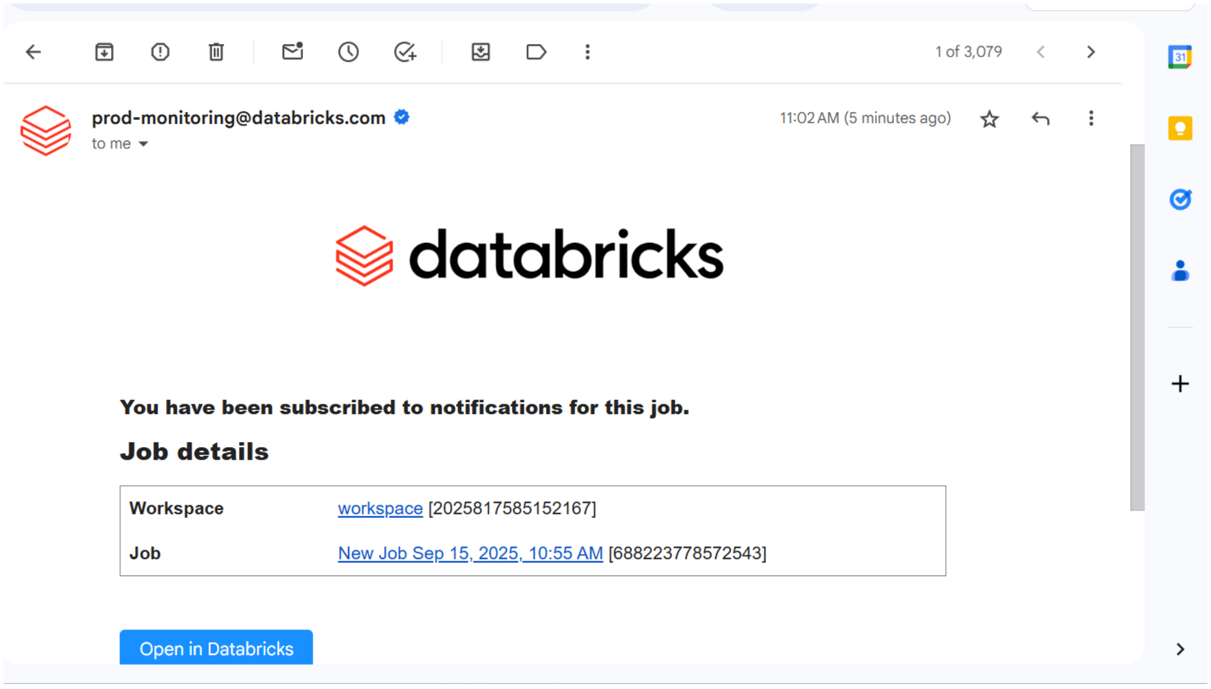
Last Updated: 2025-09-15 05:34:46

Total Revenue	Top Service Category
₹ 14,338,429.00	Pharmacy

Revenue by Service Category

Daily Revenue Trend

8) EMAIL FROM DATABRICKS FOR SUCCESSFUL JOB CREATION:



prod-monitoring@databricks.com
to me

11:02 AM (5 minutes ago)

databricks

You have been subscribed to notifications for this job.

Job details

Workspace	workspace [2025817585152167]
Job	New Job Sep 15, 2025, 10:55 AM [688223778572543]