

# Introducción a Databricks y el ecosistema Spark

Databricks

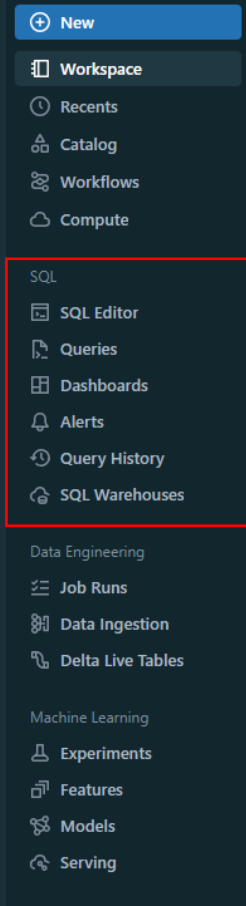
Data Analytics



IMMUNE  
TECHNOLOGY INSTITUTE



# Databricks – Data Analytics



Databricks ofrece un ecosistema unificado para el análisis de datos, permitiendo a los usuarios consultar datos ingestados a través de SQL y una interfaz gráfica intuitiva.

La plataforma combina la facilidad de SQL con las herramientas avanzadas de análisis para un procesamiento de datos eficaz, habilitando desde la exploración de datos hasta la creación de Dashboards y Alertas.

# Consulta de Datos con SQL

Utilizando SQL en Databricks, los usuarios pueden ejecutar consultas sobre los datos ingresados de manera eficiente.

La sección 'SQL Editor' dentro del menú de opciones permite escribir y ejecutar consultas SQL directamente, facilitando la manipulación y análisis de grandes volúmenes de datos.

Para consultas que se repitan con frecuencia podemos guardarlas para recuperarlas posteriormente a través de la opción "Queries"

The screenshot displays the Databricks SQL Editor interface. On the left, a 'Catalog' sidebar shows a tree view with 'hive\_metastore', 'ml', 'samples', and 'system'. The main editor area contains a SQL query: 

```
1 SELECT
2   o_orderdate AS Date,
3   o_orderpriority AS Priority,
4   sum(o_totalprice) AS 'Total Price'
5 FROM
6   `samples`.`tpch`.`orders`
7 WHERE
8   o_orderdate > '1994-01-01'
9   AND o_orderdate < '1994-01-31'
10 GROUP BY
11   1,
12   2
13 ORDER BY
14   1,
15   2
```

 Below the query, the 'Raw results' tab is active, showing a table with 9 rows and 4 columns: 'Date', 'Priority', 'Total Price', and an unnamed column with values like '96444609.82'. The status bar at the bottom indicates '12 s 751 ms | 145 rows returned' and 'Refreshed a minute ago'.

|   | Date       | Priority        | Total Price |
|---|------------|-----------------|-------------|
| 1 | 1994-01-02 | 1-URGENT        | 96444609.82 |
| 2 | 1994-01-02 | 2-HIGH          | 93497904.94 |
| 3 | 1994-01-02 | 3-MEDIUM        | 88800085.02 |
| 4 | 1994-01-02 | 4-NOT SPECIFIED | 97955477.98 |
| 5 | 1994-01-02 | 5-LOW           | 98015661.37 |
| 6 | 1994-01-03 | 1-URGENT        | 92534508.96 |
| 7 | 1994-01-03 | 2-HIGH          | 92286715.43 |
| 8 | 1994-01-03 | 3-MEDIUM        | 93521575.91 |
| 9 | 1994-01-03 | 4-NOT SPECIFIED | 97569521.46 |

# SQL Warehouses

Los SQL Warehouses representan una evolución en la gestión y análisis de datos en la nube.

Diseñados específicamente para consultas SQL de alto rendimiento, estos clusters están optimizados para cargar, transformar y analizar grandes volúmenes de datos con una eficiencia excepcional

Se diferencian de los clusters de procesamiento de datos en las siguientes características:

- **Optimización para Consultas SQL:** Mediante una cola de ejecución en paralelo y Photon por defecto habilitado para la ejecución de consultas
- **Escalabilidad Automática:** Escalado automático para ajustarse a la carga de trabajo
- **Gestión de Costes:** Más predecible basado en el rendimiento de las consultas.
- **Experiencia de Usuario y Herramientas de BI:** Están diseñados para ser integrados con herramientas de BI. Con los clusters normales, la experiencia no es tan directa ni optimizada como con SQL Warehouses

# SQL Warehouses en Azure

Databricks ofrece una variedad de tamaños de clusters, desde pequeños para tareas ligeras de desarrollo hasta grandes para el procesamiento de datos a escala de petabytes.

Esto permite a los usuarios escalar sus recursos según las necesidades específicas de cada proyecto.

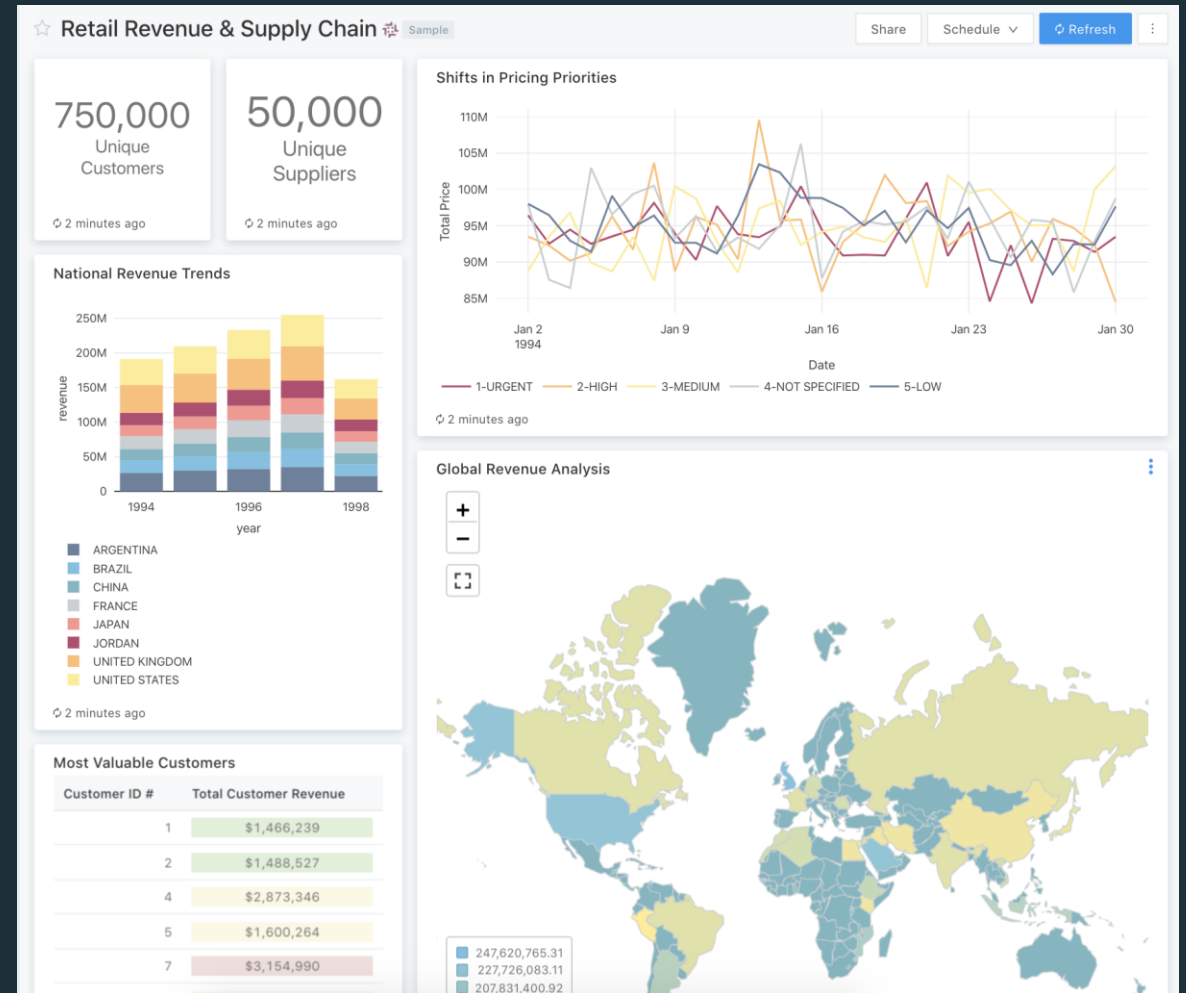
| Cluster size | Instance type for driver (applies only to pro and classic SQL warehouses) | Worker count           |
|--------------|---|------------------------|
| 2X-Small     | Standard_E8ds_v4  | 1 x Standard_E8ds_v4   |
| X-Small      | Standard_E8ds_v4  | 2 x Standard_E8ds_v4   |
| Small        | Standard_E16ds_v4   | 4 x Standard_E8ds_v4   |
| Medium       | Standard_E32ds_v4   | 8 x Standard_E8ds_v4   |
| Large        | Standard_E32ds_v4   | 16 x Standard_E8ds_v4  |
| X-Large      | Standard_E64ds_v4   | 32 x Standard_E8ds_v4  |
| 2X-Large     | Standard_E64ds_v4   | 64 x Standard_E8ds_v4  |
| 3X-Large     | Standard_E64ds_v4   | 128 x Standard_E8ds_v4 |
| 4X-Large     | Standard_E64ds_v4   | 256 x Standard_E8ds_v4 |

# Creación de Dashboards y Visualizaciones Interactivas

Databricks facilita la creación de dashboards y visualizaciones interactivas, permitiendo a los usuarios transformar sus datos en insights accionables.

La plataforma soporta una amplia gama de gráficos y visualizaciones personalizables.

Permite compartir los resultados de manera efectiva con stakeholders, potenciando la toma de decisiones basada en datos



# Query History

El Query History es una herramienta vital en Databricks, ya que permite a los usuarios rastrear y analizar el rendimiento de sus consultas SQL.

Esto no solo ayuda en la optimización de consultas sino también en la auditoría y el cumplimiento de políticas de gobernanza de datos.

The screenshot displays the Databricks Query History interface. On the left is a sidebar with navigation options: SQL, Create, SQL Editor, Queries, Dashboards, Alerts, Data, SQL Endpoints, and Query History (highlighted). Below these are Partner Connect, Help, Settings, a user profile (CUJ lucas.cerdan+aws@da...), and Menu options.

The main panel shows a table of query history with columns for Query, SQL Endpoint, and Status. The selected query is highlighted in blue.

| Query   | SQL Endpoint            | Status    |
|---|-------------------------|-----------|
| Listing columns 'catalog : hive_metastore, schemaPattern ...  | Data Engineering end... | 2022-01-1 |
| SHOW VIEWS IN 'default'                                       | Data Engineering end... | 2022-01-1 |
| Listing tables 'catalog : hive_metastore, schemaPattern : ... | Data Engineering end... | 2022-01-1 |
| SHOW DATABASES  | Data Engineering end... | 2022-01-1 |
| select c_custkey, c_name, sum(l_extendedprice * (1 - l_dis... | Data Engineering end... | 2022-01-0 |
| USE samples.tpch  | Data Engineering end... | 2022-01-0 |
| show tables in 'tpch'   | Data Engineering end... | 2022-01-0 |
| use catalog 'samples'   | Data Engineering end... | 2022-01-0 |
| show tables in 'tpch'   | Data Engineering end... | 2022-01-0 |
| use catalog 'samples'   | Data Engineering end... | 2022-01-0 |
| show tables in 'default'                                      | Data Engineering end... | 2022-01-0 |
| use catalog 'samples'   | Data Engineering end... | 2022-01-0 |
| show tables in 'default'                                      | Data Engineering end... | 2022-01-0 |
| use catalog 'samples'   | Data Engineering end... | 2022-01-0 |
| show databases  | Data Engineering end... | 2022-01-0 |
| use catalog 'samples'   | Data Engineering end... | 2022-01-0 |
| Listing columns 'catalog : hive_metastore, schemaPattern ...  | Data Engineering end... | 2022-01-0 |
| SHOW VIEWS IN 'default'                                       | Data Engineering end... | 2022-01-0 |
| Listing tables 'catalog : hive_metastore, schemaPattern : ... | Data Engineering end... | 2022-01-0 |
| -- TPC-H Q1 select l_returnflag, l_linestatus, sum(l_quant... | Data Engineering end... | 2022-01-0 |
| -- TPC-H Q22 select cntrycode, count(*) as numcust, sum(c...  | Data Engineering end... | 2022-01-0 |
| -- TPC-H Q21 select s_name, count(*) as numwait from sampl... | Data Engineering end... | 2022-01-0 |
| -- TPC-H Q20 select s_name, s_address from samples.tpch.su... | Data Engineering end... | 2022-01-0 |

The detailed view of the selected query (ID: 598b82bf-510f-438e-8da3-37fa6bbf9a2c) shows the SQL code and execution details.

```
-- TPC-H Q1
select
  l_returnflag,
  l_linestatus,
  sum(l_quantity) as sum_qty,
  sum(l_extendedprice) as sum_base_price,
  sum(l_extendedprice * (1 - l_discount)) as sum_dis
  sum(l_extendedprice * (1 - l_discount)) * (1 + l_ta
  avg(l_quantity) as avg_qty,
  avg(l_extendedprice) as avg_price,
  avg(l_discount) as avg_disc,
  count(*) as count_order
from
  samples.tpch.lineitem
where
```

**Duration**

| Category                         | Time    | Percentage |
|----------------------------------|---------|------------|
| Total                            | 25.45 s |            |
| Optimizing query & pruning files | 8.86 s  | 35%        |
| Execution                        | 16.34 s | 64%        |
| Result fetching                  | 257 ms  | 1%         |

**Overview**

SQL Endpoint: Data Engineering endpoint

**Execution summary**

Start - End time: 12:01:59.799 - 12:02:25.596

Task total time: 1.44 m

Rows returned: 4 (29,999,795 read)

Files read: 6

Partitions read: 0

Bytes read: 252.37 MB

Bytes spilled to disk: 0 bytes

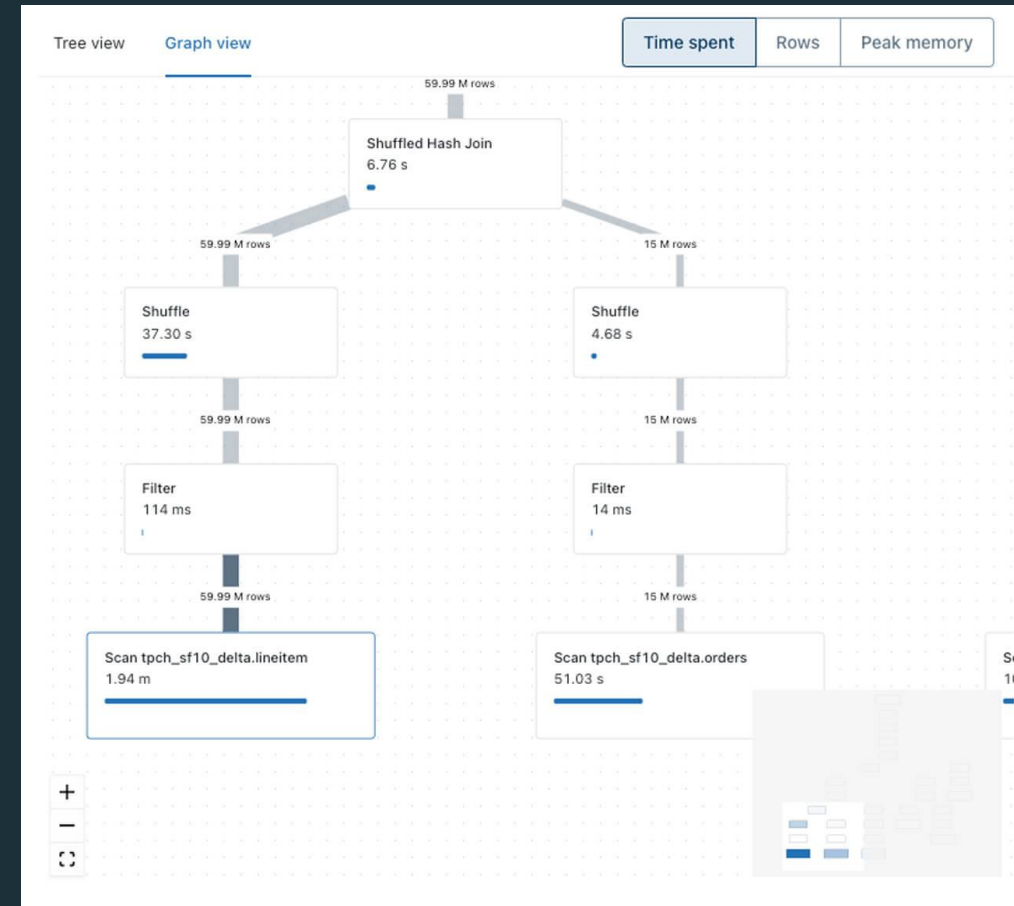
[View full execution details](#)



# Query History

El Query History es una herramienta vital en Databricks, ya que permite a los usuarios rastrear y analizar el rendimiento de sus consultas SQL.

Esto no solo ayuda en la optimización de consultas sino también en la auditoría y el cumplimiento de políticas de gobernanza de datos.



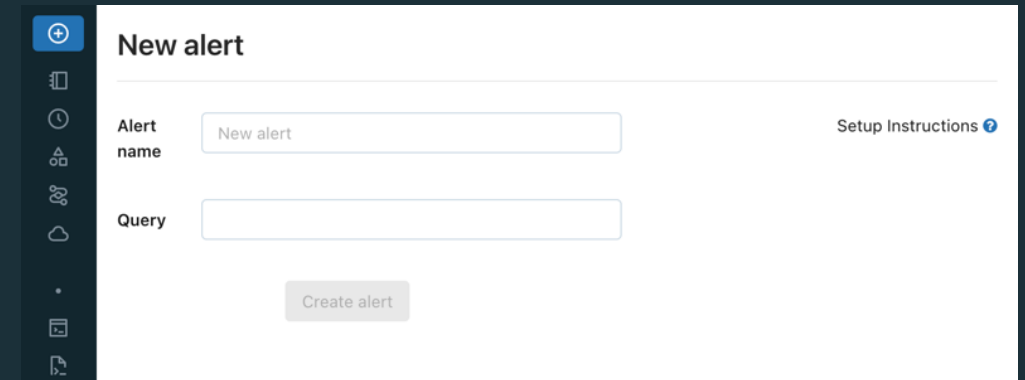


# Alertas

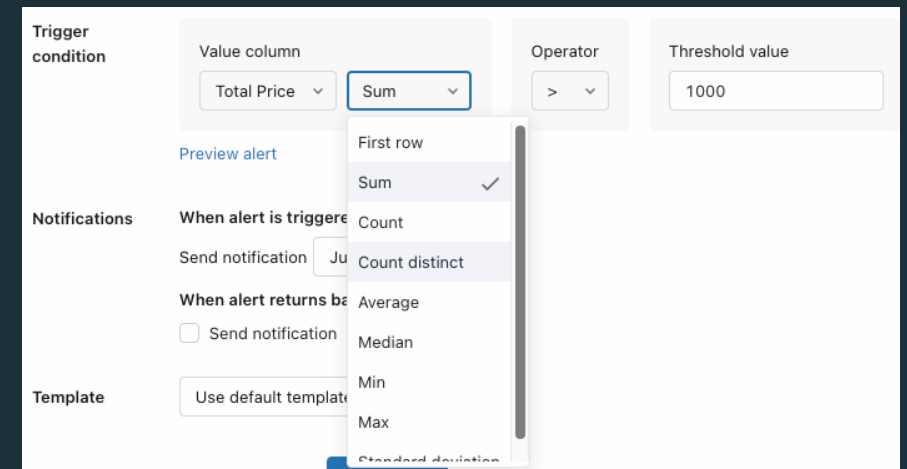
Permiten monitorizar resultados de consultas y recibir notificaciones cuando se cumplen condiciones específicas

Basadas en umbrales, como cuando un valor supera un límite o se detecta una anomalía en los datos.

Se ejecutan a intervalos regulares y envían notificaciones por correo electrónico o integraciones como Slack o Teams entre otros



The screenshot shows the 'New alert' configuration page in Databricks. On the left is a dark sidebar with navigation icons. The main area has a title 'New alert' and a 'Setup Instructions' link. Below the title are two input fields: 'Alert name' (containing 'New alert') and 'Query'. At the bottom right is a 'Create alert' button.



This screenshot shows the detailed configuration for an alert. It includes sections for 'Trigger condition', 'Notifications', and 'Template'. The 'Trigger condition' section has dropdowns for 'Value column' (set to 'Total Price'), 'Operator' (set to '>'), and 'Threshold value' (set to '1000'). A 'Preview alert' link is also present. The 'Notifications' section has a 'When alert is triggered' dropdown (set to 'Sum') and a 'Send notification' checkbox. The 'Template' section has a 'Use default template' button. A dropdown menu is open, showing options: 'First row', 'Sum' (selected with a checkmark), 'Count', 'Count distinct', 'Average', 'Median', 'Min', 'Max', and 'Standard deviation'.

# Integración con Herramientas Externas de BI y Visualización

Databricks se integra a la perfección con herramientas externas de BI y visualización como Tableau, Power BI y Looker, ampliando las capacidades de análisis y reporte de los usuarios.

Esta integración permite a las organizaciones aprovechar sus inversiones existentes en herramientas de BI, facilitando la visualización avanzada de datos y la creación de dashboards interactivos

