Taller Práctico: Creación de Tableros de Monitoreo en Databricks

Objetivo: Diseñar y construir un conjunto de dashboards en Databricks para monitorear la salud operacional, la calidad de los datos y los resultados de negocio del pipeline Medallion del dataset Olist.

1. La Filosofía del Monitoreo: ¿Por Qué se Crean Tableros?

Un pipeline de datos en producción sin un monitoreo adecuado opera como una caja negra, lo que compromete su fiabilidad. Los tableros de monitoreo proporcionan la visibilidad necesaria para responder preguntas críticas para diferentes audiencias:

1. Para los Ingenieros de Datos (Salud del Pipeline): ¿Se completó el job? ¿Cuánto tardó? ¿Falló alguna tarea? ¿Cuántos registros se procesaron en cada capa? Este tablero ayuda a asegurar la fiabilidad operacional. 🏆
2. Para los Analistas y Gobernanza de Datos (Calidad de Datos): ¿Cuántos errores se detectaron? ¿Cuáles son los tipos de error más comunes? ¿La calidad de los datos de origen está mejorando o empeorando con el tiempo? Este tablero construye la confianza en los datos. 🛡️
3. Para los Usuarios de Negocio (Resultados y KPIs): ¿Cuáles son los ingresos del último mes? ¿Qué categorías de productos son las más vendidas? ¿Qué estados generan más ventas? Este tablero demuestra el valor de negocio del pipeline. 📈

Para lograr esto, se creará un nuevo notebook dedicado exclusivamente a las consultas que alimentarán los tableros.

2. Notebook de Consultas para Tableros

Se creará un nuevo notebook llamado 05_Dashboard_Queries. Mantener estas consultas separadas de la lógica del pipeline (ETL) es una buena práctica, ya que permite modificar los tableros sin afectar el proceso de transformación de datos.

Notebook 5: 05_Dashboard_Queries

Propósito: Contener todas las consultas SQL que servirán como fuente para las visualizaciones de los tableros.

```
-- Celda 1: KPI - Salud del Pipeline: Conteo de Registros por Capa
-- Esta consulta proporciona una vista rápida del volumen de datos que fluye a través de cada etapa.
-- 🤔 Punto de Análisis: Si el número de registros en Silver es drásticamente menor que en Bronze, podría indicar un filtrado de calidad de datos exitoso o un problema en el JOIN de la
```

capa de transformación.
USE CATALOG sesion_5;

SELECT '1. Bronze (orders)' AS capa, COUNT(*) AS total_registros FROM bronze.orders_bronze
UNION ALL
SELECT '2. Silver (fact_orders)' AS capa, COUNT(*) FROM silver.fact_orders
UNION ALL
SELECT '3. Gold (monthly_sales)' AS capa, COUNT(*) FROM
gold.monthly_sales_by_category_gold;
```sql
-- Celda 2: KPI - Calidad de Datos: Estado de los Pedidos en la Capa Silver
-- Permite entender la distribución de los estados de los pedidos que pasaron las validaciones.
-- 🤔 Punto de Análisis: Un alto número de pedidos 'canceled' o 'unavailable' puede indicar tanto un problema de negocio (ej. falta de stock) como un problema de datos en el origen.
SELECT
  order_status,
  COUNT(DISTINCT order_id) AS numero_de_pedidos
FROM silver.fact_orders
GROUP BY order_status
ORDER BY numero_de_pedidos DESC;
```sql
-- Celda 3: KPI - Negocio: Ingresos Totales por Año
-- Una vista de alto nivel del rendimiento del negocio a lo largo del tiempo.
SELECT
  anio,
  SUM(ingresos_totales) AS ingresos_anuales
FROM gold.monthly_sales_by_category_gold
GROUP BY anio
ORDER BY anio;
```sql
-- Celda 4: KPI - Negocio: Top 10 Categorías de Productos por Ingresos
-- Ayuda al negocio a identificar qué categorías son las más importantes.
SELECT
  categoria_producto,
  SUM(ingresos_totales) AS ingresos_por_categoria
FROM gold.monthly_sales_by_category_gold
WHERE categoria_producto IS NOT NULL
GROUP BY categoria_producto
ORDER BY ingresos_por_categoria DESC
LIMIT 10;
```sql
-- Celda 5: KPI - Negocio: Evolución Mensual de Ingresos
-- Permite analizar tendencias y estacionalidad en las ventas.

```
SELECT
  MAKE_DATE(anio, mes, 1) AS fecha_mes,
  SUM(ingresos_totales) AS ingresos_mensuales
FROM gold.monthly_sales_by_category_gold
GROUP BY anio, mes
ORDER BY anio, mes;
```

3. Guía para la Creación de los Tableros (Flujo de Trabajo Actualizado)

Instrucciones para los Participantes:

El flujo de trabajo en Databricks Dashboards se centra en la creación de "datasets" (conjuntos de datos) a partir de consultas SQL, que luego se utilizan para construir las visualizaciones.

Paso A: Creación de los Datasets del Tablero

1. Navegar a Dashboards: En el menú de la izquierda, se debe seleccionar Dashboards.
2. Crear un Nuevo Tablero: Se debe hacer clic en Create Dashboard y nombrarlo Monitoreo de Negocio Olist.
3. Entrar al Modo de Edición: Dentro del nuevo tablero, se debe seleccionar la pestaña Data.
4. Crear el Primer Dataset:
    - Hacer clic en Create from SQL.
    - En el editor que aparece, pegar el código de la Celda 1 del notebook 05_Dashboard_Queries.
    - Hacer clic en Run.
    - En la parte superior, renombrar este dataset de "Untitled Dataset" a Conteo_por_Capa.
5. Repetir para todas las Consultas:
    - Hacer clic en el botón + Add data source y seleccionar Create from SQL de nuevo.
    - Pegar el código de la Celda 2 y renombrar el dataset a Estado_Pedidos.
    - Repetir este proceso para las celdas 3, 4 y 5, creando los datasets Ingresos_Anuales, Top_Categorias y Evolucion_Mensual respectivamente.

Al final de este paso, se tendrán 5 datasets listos para usar en la pestaña "Data" del tablero.

Paso B: Construcción de las Visualizaciones en el Canvas

1. Volver al Canvas: Hacer clic en la pestaña Canvas en la parte superior.
2. Añadir el Primer Widget: Hacer clic en el botón Add a visualization en el canvas.
3. Configurar el Widget:
    - En el panel derecho, seleccionar el dataset Conteo_por_Capa.
    - Elegir el tipo de visualización: Bar (Gráfico de Barras).
    - Arrastrar capa al eje X y total_registros al eje Y.
    - Cambiar el título del widget a "Registros Procesados por Capa".

4. Repetir para los demás Widgets: Añadir nuevas visualizaciones y conectarlas a los datasets correspondientes para construir los dos tableros.

Tablero 1: Salud del Pipeline y Calidad de Datos

- Widget 1: Volumen de Datos por Capa (Ya creado)
- Widget 2: Distribución de Estados de Pedidos
  - Dataset: Estado_Pedidos
  - Visualización: Gráfico de Torta (Pie Chart)
  - Título: "Estado de los Pedidos en la Capa Silver"

🏆 Desafío Práctico: Añadir un nuevo widget a este tablero. Se debe escribir una consulta en el notebook 05 que cuente el número de pedidos por año de compra (order_purchase_year) en la tabla silver.fact_orders. Luego, crear un nuevo dataset en el tablero con esta consulta y visualizarlo como un gráfico de líneas para observar el crecimiento del volumen de pedidos a lo largo de los años.

Tablero 2: KPIs de Negocio

- Widget 1: Ingresos Anuales
  - Dataset: Ingresos_Anuales
  - Visualización: Gráfico de Barras
  - Título: "Ingresos Totales por Año"
- Widget 2: Top 10 Categorías por Ingresos
  - Dataset: Top_Categorias
  - Visualización: Gráfico de Barras Horizontales
  - Título: "Top 10 Categorías de Productos por Ingresos"
- Widget 3: Evolución Mensual de Ingresos
  - Dataset: Evolucion_Mensual
  - Visualización: Gráfico de Líneas (Line Chart)
  - Título: "Ingresos Mensuales a lo Largo del Tiempo"

🏆 Desafío Práctico: Crear un nuevo KPI para el tablero de negocio. Se debe escribir una consulta en el notebook 05 que encuentre el Top 5 de estados (customer_state) con más clientes distintos. Crear un dataset con ella y visualizarlo como un mapa de Brasil o como un gráfico de barras.

Resultado Final: Al completar estos pasos, los participantes tendrán dos tableros funcionales que monitorean todo el pipeline, basados en los datos reales de Olist que han procesado.