Universidad Panamericana
Maestría en Ciencia de Datos
Datos Masivos

Proyecto Final: *Pipeline Distribuido de Predicción para Iowa Liquor Sales en GCP*

Enrique Ulises Báez Gómez Tagle, Luis Alejandro Guillén Alvarez

3 de diciembre de 2025

Índice

1. Dataset utilizado

1.1. Fuente y descripción

El dataset utilizado proviene de **BigQuery Public Data** y contiene registros de ventas de licores en el estado de Iowa, Estados Unidos. Este conjunto de datos es mantenido por el Iowa Department of Commerce y está disponible públicamente para análisis.

- **Fuente:** BigQuery Public Data - `bigquery-public-data.iowa_liquor_sales.sales`
- **Tamaño:** 32,816,143 registros
- **Periodo:** 2012-01-03 a 2025-10-31 (13.8 años)
- **Características principales:**
  - `date`: Fecha de la transacción
  - `store_number`: Identificador de la tienda
  - `city`: Ciudad donde se realizó la venta
  - `category`: Categoría del producto
  - `item_number`: Identificador del producto
  - `sale_dollars`: Monto de la venta (variable objetivo)
  - `bottles_sold`: Cantidad de botellas vendidas
  - `volume_sold_liters`: Volumen vendido en litros

1.2. Cardinalidades y dimensiones

El dataset presenta alta cardinalidad en múltiples dimensiones, lo que lo hace ideal para procesamiento distribuido:

Cuadro 1: Cardinalidades del dataset Iowa Liquor Sales.

Dimensión	Valores Únicos
Tiendas	3,337
Ciudades	504
Productos	15,183
Categorías	185

1.3. Calidad de los datos

El análisis exploratorio reveló una excelente calidad de datos con mínimos valores faltantes:

Cuadro 2: Valores nulos por campo.

Campo	Valores Nulos	Porcentaje
sale_dollars	10	0.00003 %
category	16,974	0.052 %
city	84,575	0.258 %
Total	**101,559**	**0.31 %**

Calidad general: 99.69 % de datos completos, lo que indica un dataset de alta calidad para modelado predictivo.

1.4. Distribución de ventas

La distribución de la variable objetivo (`sale_dollars`) muestra las siguientes características:

Cuadro 3: Distribución de ventas en dólares.

Percentil	Valor (USD)
P50 (Mediana)	$78.66
P90	$269.88
P99	$1,185.60

1.5. Top 10 categorías por volumen de ventas

Las categorías más vendidas representan una parte significativa del volumen total de transacciones:

Cuadro 4: Top 10 categorías por ventas totales.

Rank	Categoría	Ventas Totales (USD)	Transacciones
1	1012100.0	$495,078,200	2,778,490
2	1031100.0	$441,329,100	2,988,622
3	1011200.0	$288,427,900	1,859,256
4	1081600.0	$219,643,200	1,360,017
5	1062400.0	$169,326,700	861,360
6	1022200.0	$152,794,300	668,286
7	1031080.0	$145,760,500	1,265,930
8	1022100.0	$143,383,100	849,580
9	1011400.0	$119,534,300	538,956
10	1011100.0	$117,536,600	1,213,606
Total Top 10		**$2,292,813,900**	**15,384,103**

1.6. Justificación de selección

Este dataset fue seleccionado por las siguientes razones:

1. **Volumen masivo:** Con más de 32 millones de registros, cumple ampliamente con el requisito de 32M registros y justifica el uso de procesamiento distribuido con PySpark en Dataproc.
2. **Datos temporales:** El rango de 13.8 años permite análisis de series temporales y patrones estacionales, ideal para feature engineering temporal.
3. **Alta dimensionalidad:** La combinación de 15K+ productos, 185 categorías, 3.3K tiendas y 504 ciudades proporciona un espacio de características rico para modelado predictivo.
4. **Calidad excepcional:** Con 99.69 % de datos completos, minimiza la necesidad de imputación compleja y permite enfocarse en transformaciones y modelado.
5. **Variable objetivo continua:** `sale_dollars` es una variable continua ideal para regresión lineal, permitiendo predecir montos de venta basados en características de productos, ubicación y temporalidad.
6. **Disponibilidad pública:** Al estar en BigQuery Public Data, facilita la reproducibilidad del proyecto y el acceso sin restricciones de licenciamiento.
7. **Relevancia práctica:** Los modelos predictivos de ventas tienen aplicaciones directas en optimización de inventario, planificación de demanda y estrategias de pricing.

2. Descripción de la arquitectura implementada

2.1. Diagrama de arquitectura

La arquitectura implementada sigue un patrón de medallion con dos capas (Bronze y Gold) sobre Google Cloud Platform, integrando servicios de almacenamiento, procesamiento distribuido y análisis de datos masivos.

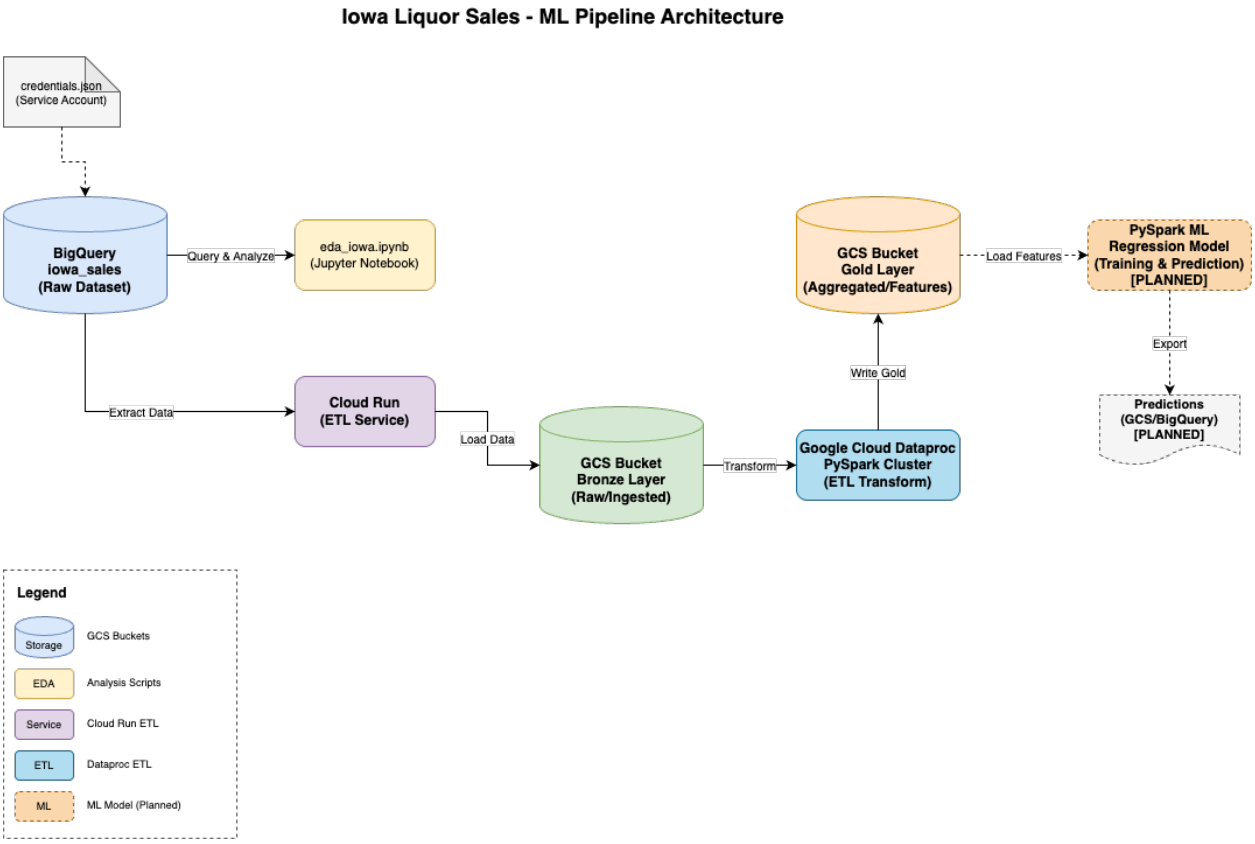

Figura 1: Arquitectura del pipeline distribuido: BigQuery → Cloud Run → GCS Bronze → Dataproc → GCS Gold → ML Model (***PLANNED***).

2.2. Flujo de datos

El pipeline implementa un flujo de datos end-to-end con las siguientes etapas:

1. **Fuente de datos (BigQuery):** El dataset público `iowa_liquor_sales` (32M+ registros) sirve como origen de datos. Los scripts de EDA (`eda_iowa.py` y `eda_iowa.ipynb`) realizan análisis exploratorio inicial directamente sobre BigQuery.

2. **Extracción (Cloud Run):** Un servicio ETL desplegado en Cloud Run ejecuta `bronze_extract.py`, que extrae datos desde BigQuery y los carga en formato Parquet particionado hacia la capa Bronze en Google Cloud Storage.

3. **Capa Bronze (GCS):** Almacenamiento de datos crudos en formato Parquet con particionamiento temporal, preservando la estructura original para trazabilidad y reproducibilidad.

4. **Transformación (Dataproc):** Un cluster de Dataproc ejecuta `gold_transform.py` con PySpark, aplicando limpieza, transformaciones y feature engineering sobre los datos Bronze. El procesamiento distribuido permite manejar el volumen masivo de forma eficiente.

5. **Capa Gold (GCS):** Datos limpios, transformados y enriquecidos con features derivadas, almacenados en formato Parquet particionado y optimizados para consumo analítico y modelado ML.

6. **Modelado ML (***PLANNED***):** Modelo de regresión lineal con PySpark MLlib entrenado sobre la capa Gold para predicción de ventas, con evaluación de métricas (R^2, RMSE, MAE) y comparación de performance entre configuraciones de cluster.

2.3. Componentes de la arquitectura

- **BigQuery:** Fuente de datos pública (`bigquery-public-data.iowa_liquor_sales.sales`)
- **Cloud Run:** Servicio ETL serverless para extracción batch hacia capa Bronze
- **GCS Bronze Layer:** Almacenamiento de datos crudos en formato Parquet particionado
- **Dataproc (PySpark):** Cluster de procesamiento distribuido para transformación y feature engineering
- **GCS Gold Layer:** Datos refinados listos para análisis y modelado
- **Terraform:** Infraestructura como código para provisionar clusters Dataproc con diferentes configuraciones
- **ML Model (***PLANNED***):** Modelo de regresión PySpark MLlib para predicción de ventas

3. Desarrollo de la ruta elegida: Procesamiento Distribuido con PySpark

3.1. Selección y exportación del dataset hacia GCS

3.1.1. Proceso de exportación

[Descripción del proceso de extracción desde BigQuery hacia GCS]

3.1.2. Verificación de estructura y consistencia

[Validaciones realizadas sobre los datos exportados]

Figura 2: Bucket de GCS con capas Bronze y Gold.

3.2. Procesamiento distribuido en Dataproc

3.2.1. Configuración del cluster

Cuadro 5: Configuración de clusters Dataproc.

Cluster	Tipo Nodo	Cantidad	vCPU	Memoria	Disco
Cluster 1	[tipo]	[n]	[vCPU]	[RAM]	[GB]
Cluster 2	[tipo]	[n]	[vCPU]	[RAM]	[GB]

3.2.2. Lectura del dataset desde GCS

[Código y descripción de lectura con PySpark]

3.2.3. Limpieza, filtrado y transformación

[Descripción de las transformaciones aplicadas]

- Limpieza de valores nulos
- Filtrado de registros inconsistentes
- Transformación de tipos de datos
- Feature engineering

Figura 3: Cluster de Dataproc ejecutando jobs de transformación.

3.3. Modelado predictivo en PySpark (***PLANNED***)

3.3.1. Modelo seleccionado

[Descripción del modelo de regresión lineal seleccionado]

3.3.2. Entrenamiento del modelo

[Proceso de entrenamiento sobre Gold layer]

3.3.3. Métricas de evaluación

[Tabla con métricas: R^2, RMSE, MAE]

Cuadro 6: Métricas de evaluación del modelo.

Métrica	Cluster 1	Cluster 2
R^2	[valor]	[valor]
RMSE	[valor]	[valor]
MAE	[valor]	[valor]

3.4. Evaluación comparativa entre configuraciones de cluster

3.4.1. Métricas de tiempo de ejecución

Cuadro 7: Comparativa de tiempos de ejecución.

Etapa	Cluster 1	Cluster 2	Diferencia
Lectura Bronze	[tiempo]	[tiempo]	[%]
Transformación	[tiempo]	[tiempo]	[%]
Escritura Gold	[tiempo]	[tiempo]	[%]
Total	[tiempo]	[tiempo]	[%]

Figura 4: Job UI de Dataproc mostrando métricas de tiempo y recursos.

3.4.2. Análisis de latencia, paralelismo y escalabilidad

[Interpretación de cómo el tamaño del cluster afecta la ejecución]

4. Métricas, gráficas y análisis de resultados

4.1. Interpretación de resultados

[Análisis de los resultados obtenidos]

4.2. Justificación del muestreo

[Explicación de las decisiones de muestreo si aplica]

4.3. Evaluación del desempeño del modelo

[Análisis crítico del desempeño]

5. Análisis crítico del enfoque

5.1. Ventajas del enfoque elegido

Ventaja 1

Ventaja 2

Ventaja 3

5.2. Limitaciones del enfoque elegido

Limitación 1

Limitación 2

Limitación 3

6. Conclusiones

[Conclusiones generales del proyecto]

7. Código utilizado

7.1. Script principal de PySpark

[Referencia al script principal]

7.2. Repositorio de código fuente

`https://github.com/[usuario]/ML-BigData`

8. Referencias

- Google LLC (s. f.). Google Cloud Console. `https://console.cloud.google.com/`
- Google Cloud. (2024). Crea un clúster de Dataproc con la consola de Google Cloud. `https://cloud.google.com/dataproc/docs/quickstarts/create-cluster-console?hl=es-419`
- BigQuery Public Data. Iowa Liquor Sales. `https://console.cloud.google.com/marketplace/product/iowa-department-of-commerce/iowa-liquor-sales`