- Proyecto Final -
- Internship Teracloud -

Autor: Ezequiel Coggiola

Fecha: 03/12/2025

Plan de implementación

Objetivos:

- Proponer e implementar un proceso de datos completo en la nube para mi cliente (e-commerce).
- Automatizar la ingesta y procesamiento mensual de datos.
- Generar información analítica lista para negocio.
- Responder preguntas clave de negocio

Preguntas de negocio:

- **¿Cómo está funcionando el negocio este mes?**
- **¿Dónde se pierden más usuarios en el funnel?**
- **¿Cómo se comportan los usuarios según país, canal y horarios?**

Descripción de los datos

Volumen total datos historicos:

Más de 1 millón de filas entre todas las tablas, **~51 MB.**

Tablas principales

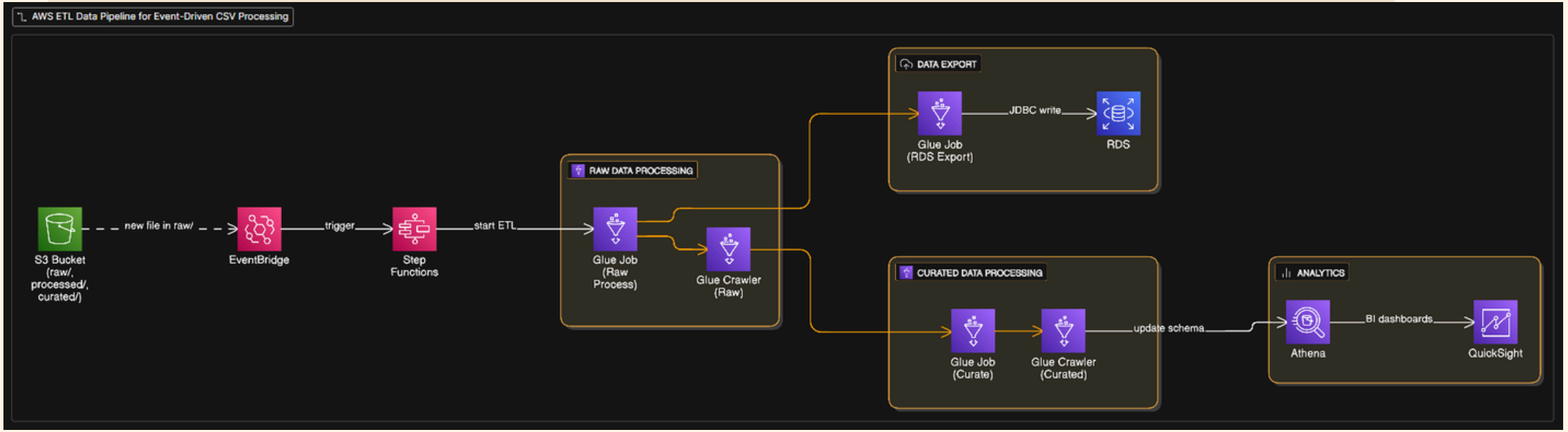
- **orders — información de compras**
- **order_items — detalle por producto**
- **products — catálogo y márgenes**
- **customers — datos de clientes**
- **reviews — reseñas**
- **sessions — sesiones de navegación**
- **events — eventos del funnel**

Adquisición → Navegación → Carrito → Checkout → Compra → Post-venta

Arquitectura de Datos

Modelo planteado:

- *Data Lake con 3 capas: raw → processed → curated.*
- *Procesamiento serverless con Glue.*
- *Orquestación automática con Step Functions.*
- Carga opcional a RDS para uso transaccional.
- Análisis con Athena + QuickSight.



Componentes del Proceso ETL

S3 Bucket (raw/processed/curated)

- Raw: llegan los CSV nuevos cada mes.
- Processed: limpieza, tipificación, Parquet.
- Curated: tablas analíticas finales.

EventBridge

- Detecta nuevos archivos en raw/.
- Dispara automáticamente la Step Function.

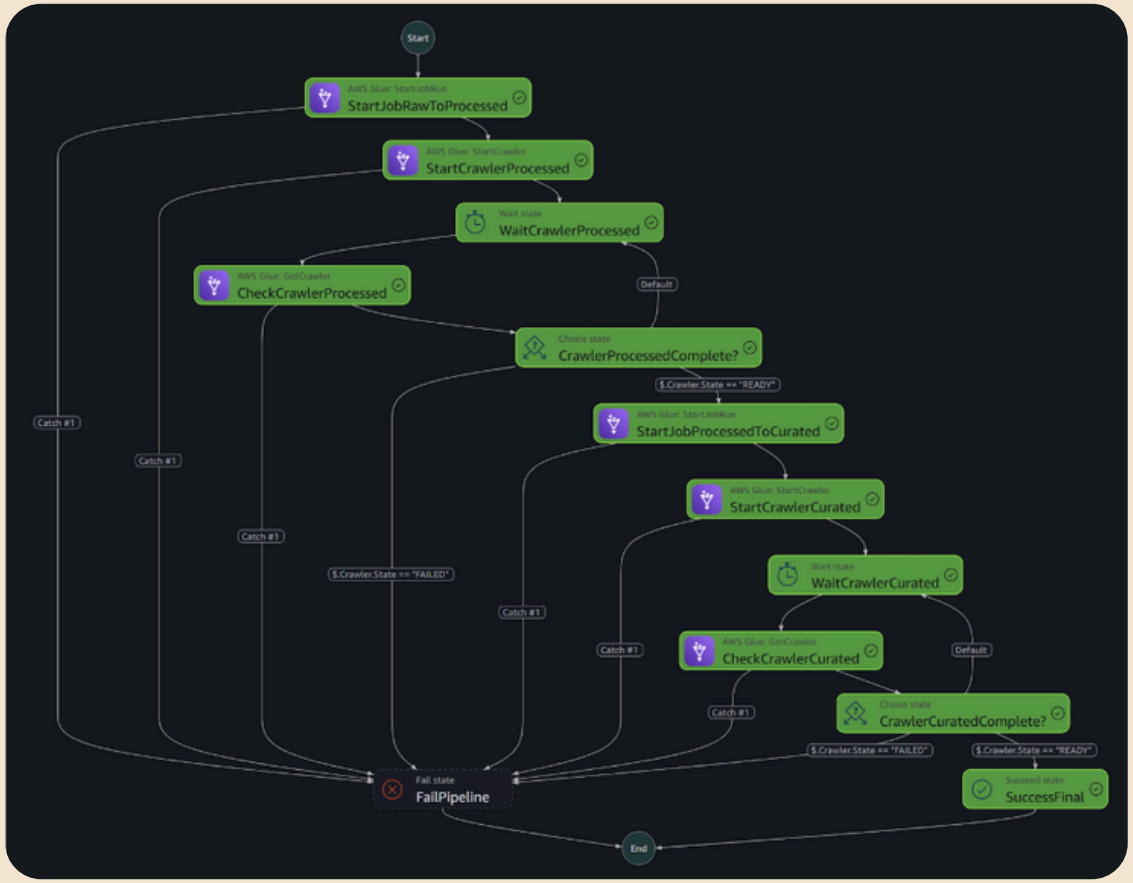
Step Functions

- Orquesta todo el pipeline.
- Corre jobs en orden y valida crawlers.

Glue Job 1 — raw → processed

- Limpieza, normalización, particiones, conversión a Parquet.

(Job Bookmarks para proceso incremental!)



Componentes del Proceso ETL

Glue Crawler (processed)

- Actualiza catálogo con schema limpio.

Glue Job 2 — processed → curated

- Enriquecimiento, joins, KPIs, deduplicación.

Crawler curated

- Publica tablas analíticas para Athena.

Glue Job 3 — curated → RDS

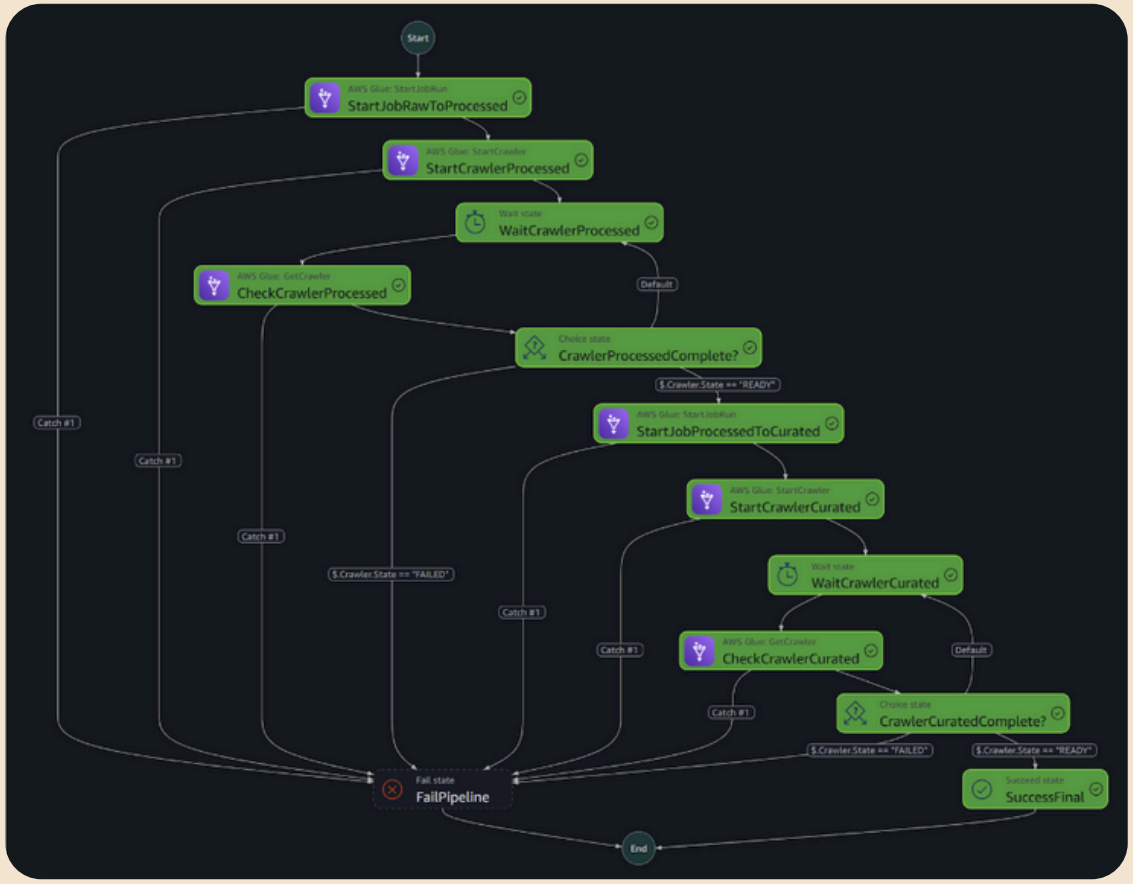
- Inserta datos limpios en base relacional para uso OLTP.

Athena

- Consultas SQL sobre el Data Lake.
- Fuente de datos para QuickSight.

QuickSight

- Dashboards finales para negocio.



Tablas Curated

- **fact_order_items_enriched**

  Rentabilidad, cantidades, precios, márgenes por producto/categoría.

- **fact_orders_enriched**

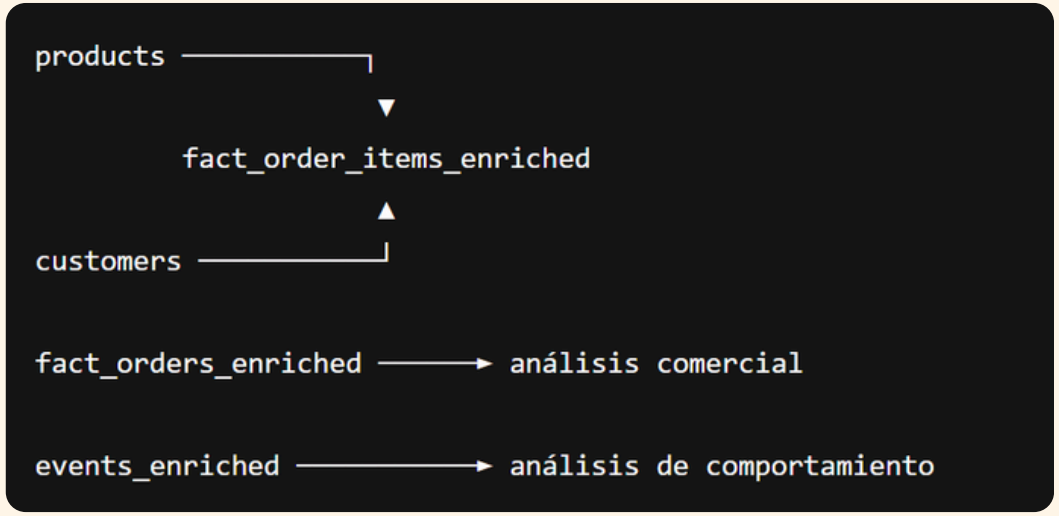
  Ingresos, país, canal, dispositivo, atributos del cliente.

- **events_enriched**

  Eventos del funnel: page views, add to cart, checkout, purchase.

- **Diseño listo para Athena y QuickSight**

  Tablas limpias, enriquecidas y orientadas al análisis.



IAM Roles y Security Groups

• **Roles separados por responsabilidad**

- GlueRole: acceso a S3 + permisos de Glue
- StepFunctionRole: ejecutar Glue + leer S3
- QuickSightRole: acceso limitado a Athena y bucket de resultados

• **Principio de privilegio mínimo**

Cada servicio recibe solo los permisos que necesita.

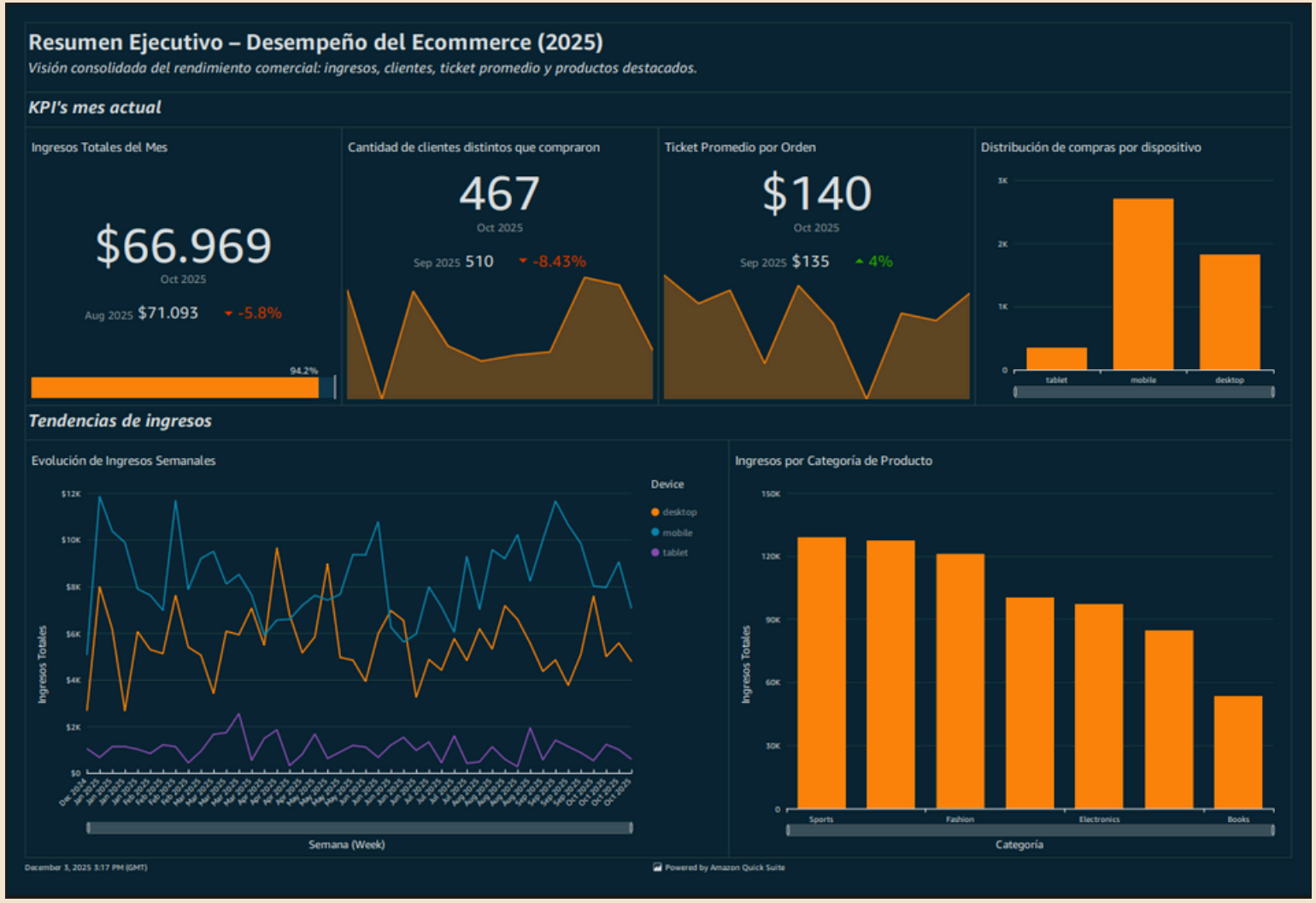
Seguridad en red con Security Groups

- SG de RDS: acceso solo desde Glue y mi IP
- SG de Glue: comunicación interna habilitada

Análisis de Negocio

Análisis Comercial – KPIs del Mes

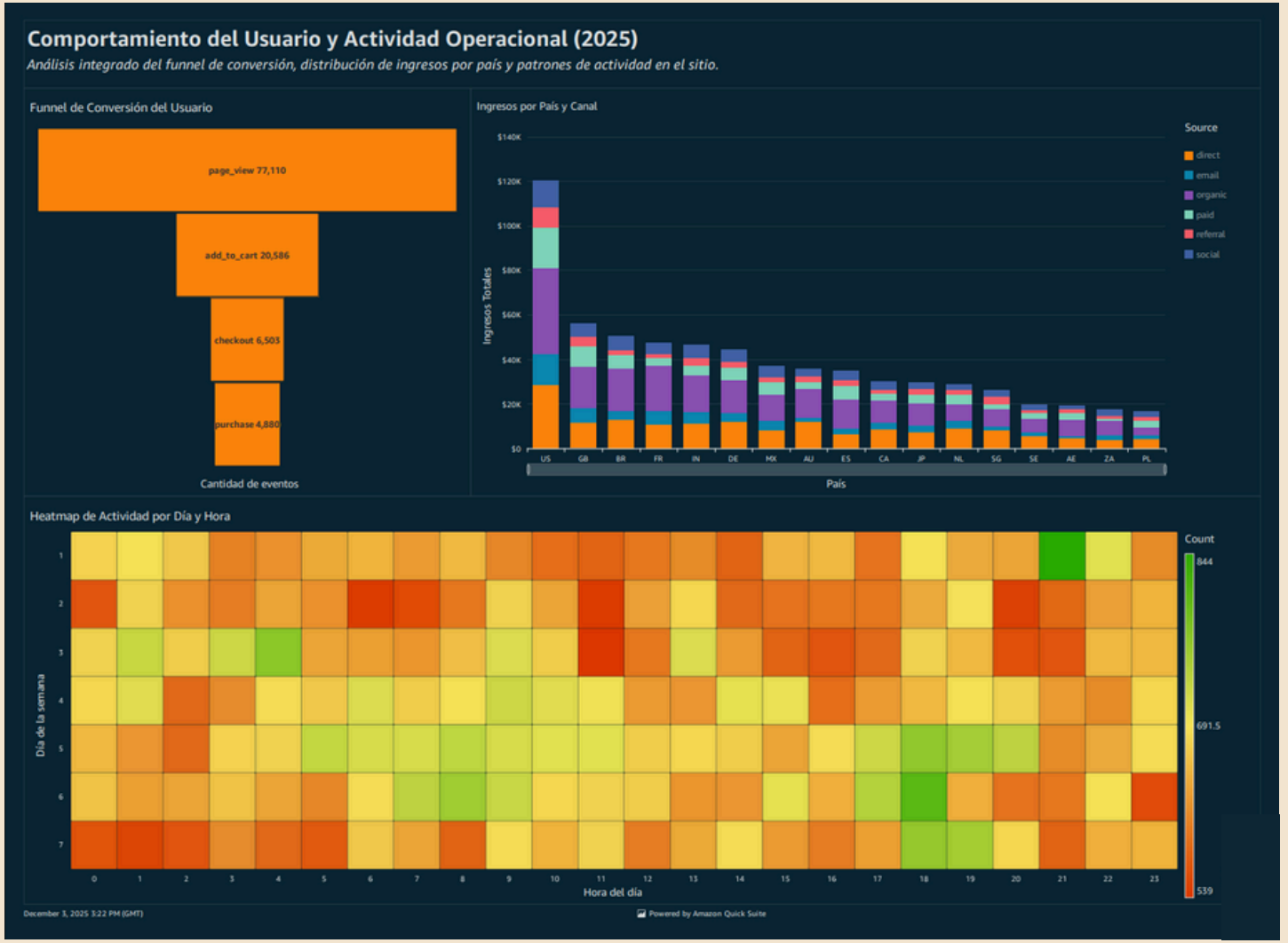
- Ingresos totales del mes
- Clientes únicos que compraron
- Ticket promedio por orden
- Tendencias semanales de ingresos
- Categorías con mayor impacto en ventas



Análisis de Negocio

Funnel y Comportamiento del Usuario

- Funnel completo: del page view a la compra
- Paso con mayor pérdida: Add to Cart → Checkout
- Ingresos por país y canal
- Actividad por día y hora (heatmap)



Optimización de Costos

Glue Jobs

- USD 0.44 por DPU-hora

Primeras ejecuciones:

- Job 1: 30 min, 2 DPUs
- Job 2: 19 min, 2 DPUs
- Job 3: 25 min, 2 DPUs

Estimación: USD 1.09

Glue Crawlers

Crawler cobra por DPU-hora (1 DPU)

- 2 crawlers × 1.5 min

Estimación: USD 0.02

S3 Storage

- USD 0.023 por GB-mes

Datos:

- 133 MB ≈ 0.133 GB

Estimación: USD 0.003 / mes

RDS MySQL (db.t3.micro)

- ~USD 0.018 por hora aprox.

Tiempo encendido optimizado:

- 8 h/día × 5 días × 4 semanas = 160 horas/mes

Estimación: USD 2.88 / mes

• **Formato Parquet → hasta 80% menos costo en Athena**

Reduce volumen escaneado y acelera consultas.

• **Particionamiento por fecha**

Evita leer datos innecesarios y mejora performance.

• **Glue Jobs: 2 DPUs + timeouts configurados**

Procesamiento eficiente y sin ejecuciones colgadas.

• **RDS solo 8 h/día (EventBridge + Lambda)**

Instancia encendida únicamente en horario laboral.

Uso del Sistema

1. Terraform despliega toda la infraestructura

Pipeline completo creado con un solo comando.

2. El cliente solo ejecuta upload_files.py

Carga los nuevos archivos del mes a raw/.

3. EventBridge detecta la carga y dispara el pipeline

Orquestación automática con Step Functions.

4. Todo el procesamiento corre sin intervención manual

Glue Jobs + Crawlers completan el ETL end-to-end.

5. QuickSight muestra los dashboards actualizados

El cliente analiza métricas sin entrar a AWS.

Desafíos y Aprendizajes

Teoria =/= Practica

- **Planeamiento Inicial**

Diseñar bien las capas, rutas y dependencias evita retrabajo.

- **Debugging Real = Logs + Logs + Logs**

Interpretar logs de Glue, Step Functions y RDS es parte esencial del trabajo.

- **Optimizar Costos desde el Testeo**

Configurar DPUs, timeouts y límites evita gastos innecesarios.

¡Muchas Gracias!