

# PROCESAMIENTO E-COMMERCE

## 1. GENERACIÓN DE EVENTOS

Un script simula usuarios navegando, agregando productos al carrito y realizando compras. Cada acción se envía como un evento **JSON** hacia **Apache Kafka**.

## DATOS GENERADOS

```
source venv/bin/activate
cd my/e-commerce-project
python producers/producer_py.py

# This code is part of the exercise for the Apache Kafka course. It is deprecated and scheduled for removal in a future version. Use timezone-aware objects to represent dates/times in Python's datetime module instead.
# https://github.com/apache/beam/blob/master/sdks/python/apache_beam/examples/exercises/eCommerce_producer.py#L10

# An event is sent to Kafka with the following schema:
# {
#   "event_type": "page_view", "price": 100.0, "timestamp": "2023-12-15T09:30:00Z", "store_id": "ST005", "store_type": "physical_store", "city": "Quito", "province": "Azuay", "location": {"lat": -2.1708, "lon": -79.9224}
#   ...
# }

# Events are generated in three categories: page_views, cart_events, and orders.
# The page_view category includes events like 'add_to_cart' and 'purchase'.
# The cart_events category includes events like 'add_to_cart' and 'remove_from_cart'.
# The orders category includes events like 'purchase'.
```

producer\_py.py

## 2. INGESTA CON KAFKA

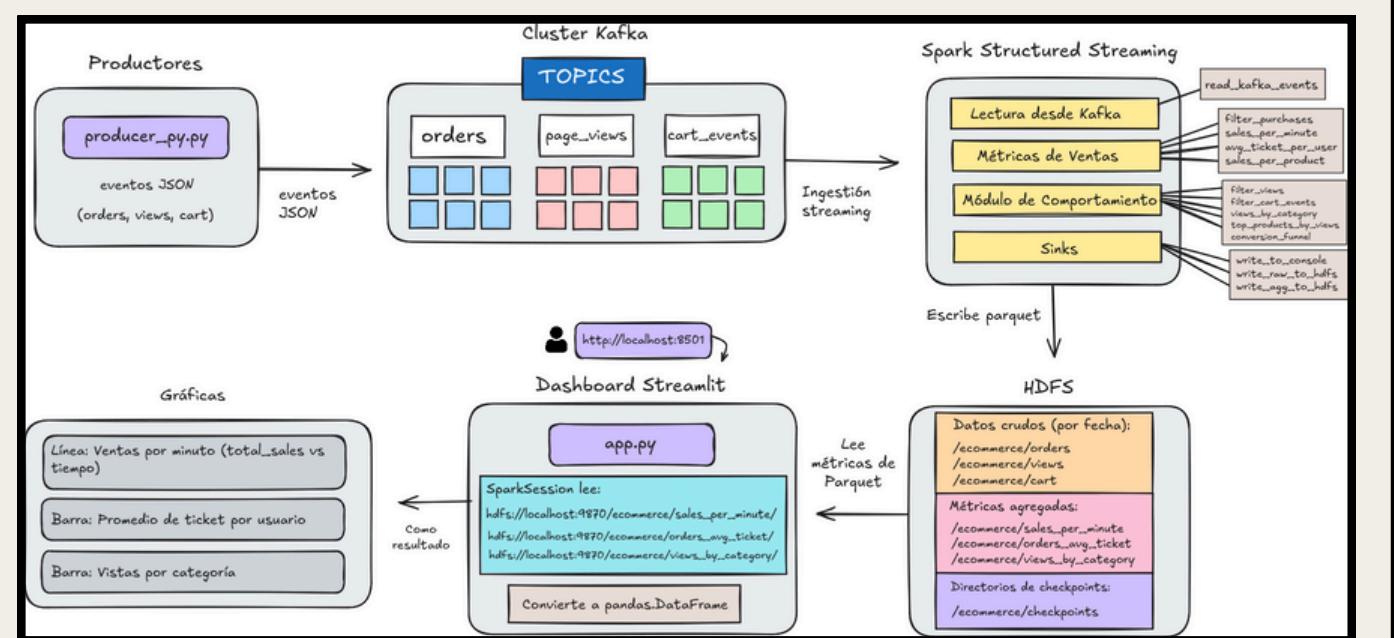
Los eventos se distribuyen en tres **topics**: **page\_views**, **cart\_events**, **orders** para mejor organización y flujo continuo de mensajes.

## 3. STRUCTURED STREAMING

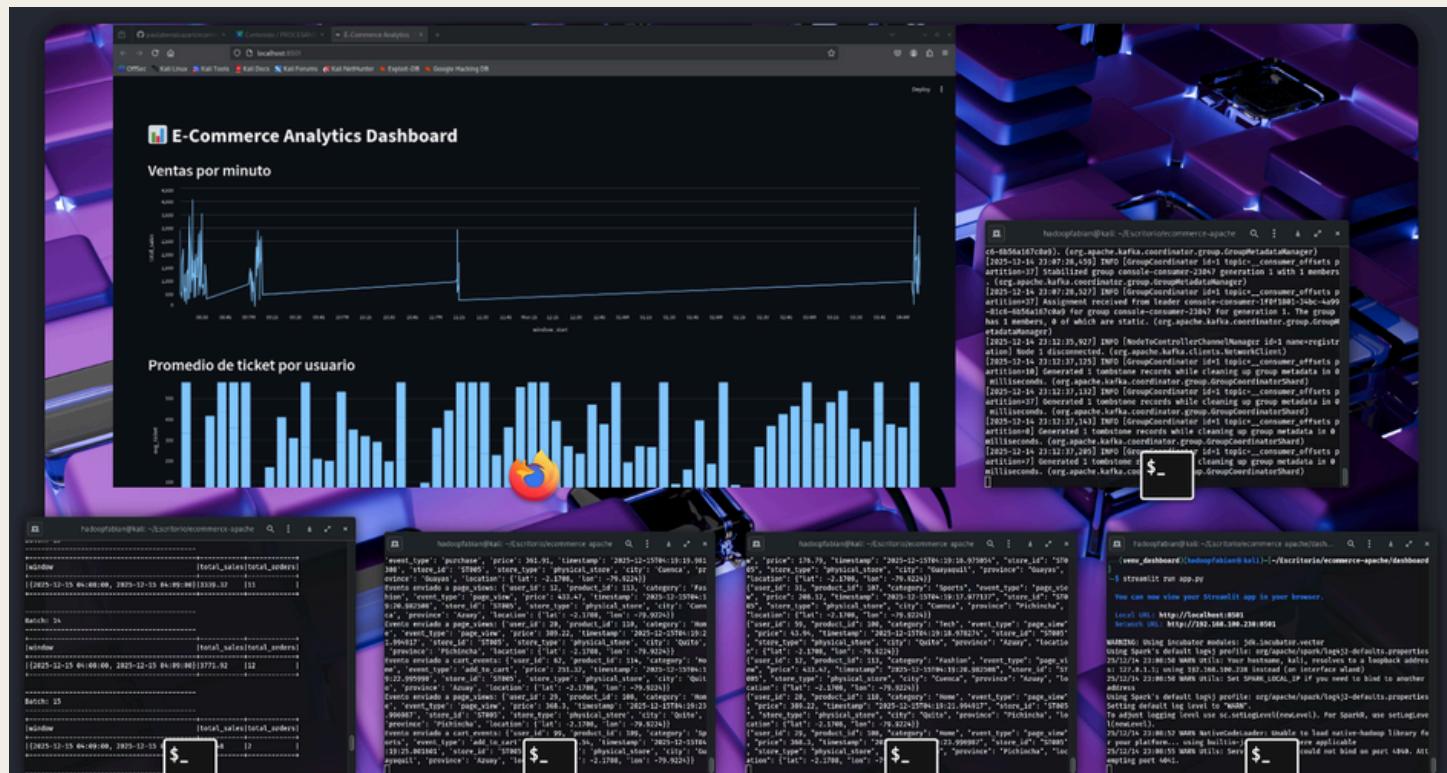
**Spark Structured Streaming** consume los eventos desde Kafka y ejecuta:

- Filtrado por tipo de evento
- Agregaciones por ventanas de tiempo
- Cálculo de métricas clave (ventas por minuto, vistas por categoría, ticket promedio)
- Join entre streams para analizar la conversión vista → carrito → compra

## ARQUITECTURA



Fabián Rodas y Paula Benalcázar



## 4. ALMACENAMIENTO HDFS

Tanto los datos crudos como las métricas agregadas se guardan en HDFS en formato **Parquet**, organizados por **fecha**.

/orders, /views, /cart, /sales\_per\_minute

## 5. VISUALIZACIÓN

Un **dashboard** en **Streamlit** lee los datos procesados desde HDFS y muestra métricas como:

- Ventas por minuto
- Actividad por categoría
- Productos más vistos.

## 6. CONCLUSIONES

- Kafka y Spark procesan eventos en tiempo real eficazmente.
- HDFS posee almacenamiento histórico organizado y accesible.
- La arquitectura demuestra alta escalabilidad y tolerancia a fallos.
- Las métricas permiten comprender mejor el comportamiento del usuario.

