Sistema Big Data para Interacciones con Películas

Descripción de proyecto, arquitectura y componentes

Equipo de Proyecto

17 de octubre de 2025

---

1. DESCRIPCIÓN DEL PROYECTO

Objetivo Central

Analizar y visualizar en tiempo real las interacciones de usuarios con películas mediante un *pipeline* completo de Big Data que combina procesamiento en **streaming** para métricas inmediatas y procesamiento **batch** para análisis históricos, permitiendo la detección de tendencias y patrones de comportamiento.

Dataset Seleccionado

- **Nombre:** *Movies Dataset*
- **Formato:** JSON

Volumen

- 100 películas con información completa.
- **Streaming continuo:** generación de 1 interacción cada 2 segundos (1,800 interacciones/hora).
- **Almacenamiento histórico:** todos los datos se persisten en HDFS para análisis batch.
- **Simulación de Big Data:** arquitectura escalable que puede manejar millones de registros.

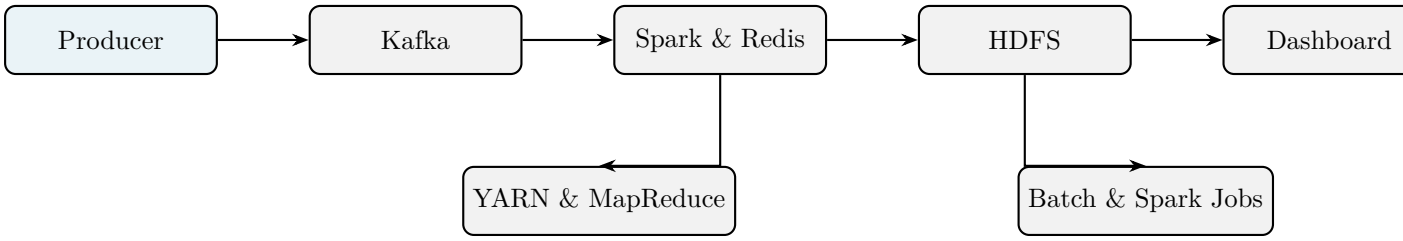
Características

- **Atributos estructurados:** ID, nombre, género, *rating*, popularidad, descripción.
- **Temporalidad:** marcas de tiempo (*timestamps*) precisas para cada interacción.
- **Etiquetas:** tipos de interacción (`click`, `view`, `rating`, `purchase`).
- **Datos limpios:** estructura consistente y validada.
- **Metadatos ricos:** información completa de películas para análisis multidimensional.

Pertinencia

- Relación directa con el objetivo de analizar comportamiento de usuarios.
- Datos realistas que simulan plataformas de *streaming* reales.
- Estructura flexible que permite múltiples tipos de análisis.
- Escalabilidad demostrada para crecer en volumen y complejidad.

Diagrama del Pipeline



Enfoque: Combinación Batch + Streaming

Arquitectura **Lambda** que combina procesamiento en tiempo real para métricas inmediatas y procesamiento **batch** para análisis profundos.

2. Componentes y Funciones

Capa de Ingesta (Streaming)

- **Producer:** genera datos simulados de interacciones de usuarios cada 2 segundos.
- **Kafka:** sistema de mensajería distribuido que *bufferiza* los datos en tiempo real.

Capa de Procesamiento (Streaming + Batch)

- **Spark Consumer:** procesa datos en tiempo real, calcula métricas y almacena en Redis.
- **Redis:** base de datos en memoria para métricas en tiempo real del *dashboard*.
- **HDFS:** almacenamiento distribuido para todos los datos históricos.

Capa de Análisis (Batch)

- **Batch Processor:** ejecuta análisis periódicos sobre datos históricos en HDFS.
- **MapReduce con YARN:** procesamiento distribuido de grandes volúmenes de datos.
- **Spark Jobs:** análisis avanzados y agregaciones complejas.

Capa de Visualización

- **Dashboard:** interfaz web en tiempo real que muestra métricas actualizadas cada 3 segundos.

Resumen

Este documento describe un *pipeline* híbrido (Lambda) que integra **streaming** (métricas inmediatas con Kafka, Spark y Redis) y **batch** (análisis históricos con HDFS, YARN/MapReduce y *Spark jobs*) para modelar y visualizar el comportamiento de usuarios en una plataforma de películas.