UNIVERSIDAD EAFIT
ST0263: Tópicos Especiales en Telemática, 2025-2
Trabajo 3 – Procesamiento distribuido con MapReduce
(Arquitectura Batch con Hadoop)
Fecha de entrega: 23 de Noviembre de 2025

Descripción

Durante el curso hemos abordado diferentes temas relacionados con los sistemas distribuidos y en la parte final el procesamiento de grandes volúmenes de datos.

En este trabajo final, se propone construir una solución de arquitectura batch basada en Hadoop para simular un flujo real de procesamiento distribuido utilizando únicamente el modelo MapReduce.

Este proyecto permite al estudiante experimentar con todas las etapas esenciales de un flujo batch: obtención de datos, almacenamiento distribuido, procesamiento paralelo y entrega de resultados.

Objetivo general

Implementar un flujo completo de procesamiento distribuido usando HDFS y MapReduce. El propósito es comprender cómo funcionan los sistemas distribuidos de almacenamiento y procesamiento batch desde sus fundamentos.

Etapas del proyecto

El proyecto consta de las siguientes etapas:

- **Obtención de datos (manual):** El estudiante selecciona una fuente de datos abierta y descarga localmente archivos CSV, JSON o de texto plano. No se hace necesario la automatización de esta etapa.
- **Carga a HDFS:** Se cargan los archivos al sistema distribuido de archivos (HDFS), en la nube (como Amazon EMR). Esta carga puede realizarse manualmente o mediante un script reproducible.
- **Procesamiento con MapReduce:** El análisis de los datos se realiza mediante uno o varios programas MapReduce. Los programas pueden implementarse en Java (nativo de Hadoop) o Python usando (MRJOB.). Debe incluirse al menos un job MapReduce que produzca resultados significativos (agregación, filtrado, conteo, análisis estadístico, etc.).

- **Salida y consulta de resultados:** Los resultados deben almacenarse nuevamente en HDFS y ser legibles como salida final. Se deben exportar los resultados a un archivo CSV y servirlos mediante una pequeña API (Flask/FastAPI)

Alcance

- Implementar y ejecutar programas MapReduce en un entorno Hadoop.
- Trabajar con archivos reales, en un formato estructurado o semi-estructurado.
- Utilizar HDFS como sistema de almacenamiento principal.
- Mostrar todo el flujo funcionando: carga → procesamiento → salida.
- Visualización de los datos vía una API.

Entrega

- Repositorio en GitHub, que debe contener:
  - Código de MapReduce (.java o .py)
  - Script(s) de carga a HDFS si se usaron.
  - Ejemplo de archivos de entrada y salida.
  - Código de la api para visualización de los resultados.
  - Instrucciones claras en un README.md para ejecutar todo.
- Video de sustentación (máx. 10 minutos), donde se explique:
  - Qué datos se usaron y por qué.
  - Cómo se cargaron al sistema.
  - Explicación detallada de cómo funciona el programa MapReduce.
  - Qué resultados se obtuvieron.
- Sustentación (en caso de ser requerida): Lunes 24 de noviembre de 2025, 8:00 a.m. a 12:00 m.

Sugerencias de temas o fuentes de datos

Fuentes de datos:

Cada equipo de trabajo debe explorar y seleccionar los datos insumo para el proyecto 3 de las siguientes fuentes en línea y gratuitas, que le permita definir una problemática concreta que quiere analizar con alguno de las siguientes fuentes de datos:

1. Datos del tiempo en línea y datos históricos de cualquier parte del mundo

Open-Meteo

Url: https://open-meteo.com
Acceso: API gratuita (sin autenticación)
Variables: clima actual, pronóstico, históricos

Ideal para carga masiva en S3
Ej: https://archive-api.open-meteo.com/v1/archive?latitude=6.25&longitude=-75.56&start_date=2022-01-01&end_date=2022-12-31&daily=temperature_2m_max,precipitation_sum&timezone=America/Bogota

WeatherAPI

Url: https://www.weatherapi.com
API Key requerida (gratis limitado)
Variables: clima actual, pronóstico, históricos hasta 2010
formatos: CSV/JSON/XML

Meteostat

Url: https://dev.meteostat.net/
Acceso: API gratuita (requiere token)
Datos históricos desde 1973
Variables: temperatura, viento, nubosidad, precipitaciones, presión, históricos desde 1973
Formato: JSON

2. Transporte y Movilidad

Datos Abiertos de TransMilenio

URL: https://datosabiertos.transmilenio.gov.co/
Datos de estaciones, rutas, horarios, validaciones
Formato: CSV y APIs JSON
Uso: cruzar con base de datos simulada de usuarios o incidentes

OpenTraffic (Uber Movement)

URL: https://movement.uber.com/
Datos de velocidad y tráfico por ciudad (requiere inscripción)
Uso: análisis de movilidad urbana y predicción de congestión

3. Datos financieros:

Datos de la Superintendencia Financiera de Colombia

Url: https://www.superfinanciera.gov.co
Formato: CSV, Excel, JSON (con algunas APIs disponibles)
Ejemplo: precios históricos de acciones, tasas de interés, datos de bancos

Uso: integración con bases de datos relacional simulada de clientes e inversiones

Banco Mundial - Open Data

URL: https://data.worldbank.org/
Acceso por archivo y API
Ejemplo: PIB, desempleo, acceso a internet por país
Uso: análisis comparativo internacional o regional

4. Salud publica

Datos abiertos del Ministerio de Salud Colombia

URL: https://www.datos.gov.co/Salud-y-Protecci-n-Social
Ejemplo: vacunación COVID, morbilidad, atención EPS
Formato: CSV, JSON
Uso: análisis epidemiológico por región, cruzado con base de pacientes

5. Datos sobre ecommerce

MercadoLibre Public APIs

Url: https://developers.mercadolibre.com.co
Acceso: API pública con autenticación (algunos endpoints sin token)
Datos disponibles:

- Productos por categoría o palabra clave
- Detalles de publicaciones (precio, título, vendedor)
- Comentarios y valoraciones

Uso educativo:

- Recolectar precios de productos por región o categoría
- Análisis de oferta y demanda

Fake Store API (simulación estilo Amazon)

Url: https://fakestoreapi.com
Acceso libre, sin autenticación
Datos disponibles:
Productos (título, categoría, precio, rating)
Carritos de compra, usuarios simulados
Útil para: prácticas de análisis, dashboards, simulación de compras
Ejemplo: https://fakestoreapi.com/products

Best Buy Developer API

Sitio web: https://developer.bestbuy.com/
API Key gratuita
Datos en tiempo real: catálogo de productos, precios, disponibilidad
Útil para: construir dashboards o análisis de precio por categoría