Trabajo Final: Procesamiento de Grandes Volúmenes de Datos

Maestría de Ciencia de Datos - UNAJ

Raúl Burgos Mauro Cejas Marcovecchio

2026-02-01

Introducción

En el contexto del análisis de grandes volúmenes de datos, las arquitecturas distribuidas juegan un rol fundamental para el procesamiento eficiente y escalable de información. Tecnologías como Apache Spark y Apache Kafka se han convertido en estándares de facto para el procesamiento batch, streaming y la ingestión de datos en tiempo real.

El objetivo de este trabajo es diseñar e implementar un clúster virtualizado que permita simular una infraestructura Big Data orientada al análisis de datos del mercado financiero. Para ello, se utilizó Docker como tecnología de virtualización liviana, desplegando un clúster compuesto por tres nodos que integran Apache Spark, Apache Kafka y Apache Zookeeper.

El caso de uso se inspira en un concurso académico realizado en 2022, donde se propuso el análisis en tiempo real de datos de mercado. Dado que la API original ya no se encuentra disponible, se optó por simular la ingestión de datos utilizando Kafka como sistema de mensajería distribuido.

Desarrollo

El sistema fue diseñado bajo un esquema de microservicios que garantiza que cada componente cumpla una función específica sin cuellos de botella. La arquitectura implementada consta de tres nodos virtuales, cada uno representado por un contenedor Docker: - Nodo 1 – Spark Master: responsable de la gestión y coordinación de aplicaciones Spark. - Nodo 2 – Kafka Broker: encargado de la ingestión y distribución de eventos que simulan datos del mercado financiero. - Nodo 3 – Zookeeper: utilizado por Kafka para la coordinación, gestión de metadatos y elección de líderes.

La virtualización mediante Docker permitió cumplir con el requerimiento de tres nodos sin necesidad de máquinas virtuales completas, reduciendo el consumo de recursos y simplificando el despliegue (ver imagen 1 en el anexo A).

Configuración del clúster

El despliegue se realizó mediante un archivo docker-compose.yml (ver script 1 en el anexo B), donde se definieron: - Imágenes oficiales de Spark, Kafka y Zookeeper - Variables de entorno necesarias para la correcta inicialización - Puertos expuestos para acceso a servicios (Spark UI y Kafka) - Configuración específica de Kafka para operar con un único broker, ajustando los factores de replicación internos

Un aspecto clave del desarrollo fue la correcta configuración de "advertised.listeners" y de los factores de replicación internos de Kafka, necesarios para evitar errores de liderazgo en entornos de un solo nodo.

Flujo de información

En primer lugar, la generación de datos está a cargo de un productor desarrollado en Python que actúa como nuestro "simulador de mercado", emitiendo eventos constantes (Ticker, Timestamp, Precio). Cada evento contiene el identificador del activo, un timestamp y un valor de precio, permitiendo evaluar el comportamiento del sistema ante un flujo continuo de datos en tiempo real.

En segundo lugar, el transporte se realiza mediante Apache Kafka que recibe estos eventos en el tópico "market-data" (ver script 2 en el anexo B), asegurando que ningún dato se pierda en el camino.

Por último, el procesamiento es realizado con Spark Structured Streaming que se encarga de la parte "pesada": lee el flujo, limpia el formato CSV y realiza cálculos matemáticos (ver script 3 en el anexo B).

Cabe destacar que todo este ecosistema (Zookeeper, Kafka y Spark) conviva en una red aislada y lista para desplegarse en cualquier entorno se debe a la orquestación Mediante Docker Compose.

Estrategia de análisis

En este punto es importante resaltar que no nos limitamos solo a leer datos, sino que también los organizamos. Para ello implementamos ventanas temporales de 30 segundos, lo que nos permite observar el comportamiento de activos en bloques de tiempo manejables.

Además, añadimos un Watermark de 1 minuto (vital en sistemas reales) que nos permite ser tolerantes y esperar por datos que puedan llegar con un ligero retraso debido a la red, sin detener el procesamiento global.

Pruebas

En primer lugar se verificó el correcto funcionamiento de cada componente: - Spark Master: Acceso exitoso a la interfaz web de Spark (http://localhost:8080), confirmando que el servicio se encontraba operativo. - Kafka Broker: Ejecución de comandos administrativos para listar y crear tópicos. - Zookeeper: Verificación indirecta a través del correcto funcionamiento de Kafka.

Como prueba funcional del sistema se implementó un caso de uso simple de ingestión de datos. Se enviaron mensajes simulando datos de mercado financiero y, mediante un consumidor Kafka, se verificó la recepción exitosa de los mensajes producidos, confirmando el correcto funcionamiento del flujo productor–broker–consumidor. Este procedimiento valida que el clúster es capaz de manejar eventos en tiempo real, cumpliendo con el objetivo de simular una arquitectura de streaming de datos.

Durante las pruebas el sistema demostró una sincronía impecable. Validamos que el flujo end-to-end funcionara correctamente, desde la creación del mensaje en Python hasta la actualización del promedio de precios en la consola de Spark. Asimismo, al agrupar los datos, logramos una visión clara del mercado en ventanas específicas, obteniendo promedios de precios precisos para cada Ticker de forma casi instantánea.

Conclusiones

En este trabajo se logró implementar exitosamente un clúster Big Data compuesto por tres nodos virtualizados utilizando Docker. La infraestructura desplegada permite simular un entorno realista de ingestión y procesamiento de datos, integrando Apache Spark y Apache Kafka (ver imagen 2 en el anexo A). Podríamos destacar como punto fuerte del proyecto:

- El despliegue de un clúster distribuido con tecnologías ampliamente utilizadas en la industria.
- La implementación de un caso de uso funcional de ingestión de datos en tiempo real.
- La resolución de problemas reales asociados a la configuración de Kafka en entornos Docker, demostrando comprensión del funcionamiento de la plataforma.

La solución desarrollada cumple con los requerimientos planteados y sienta las bases para la extensión hacia escenarios de procesamiento más complejos.

En síntesis, hemos logrado construir una base sólida y reproducible. El uso de Docker ha sido un acierto para la portabilidad, aunque aprendimos que la gestión de scripts mediante volúmenes es un punto clave para la persistencia.

Trabajo a futuro

Este pipeline es solo el comienzo. Para llevar esta solución al siguiente nivel, nuestras metas a corto plazo son:

- Ampliación: Crear otras métricas financieras que permitan un análisis más amplio para la toma de decisiones.
- Visualización: Conectar los resultados a un dashboard para ver las curvas de precios en tiempo real.
- Almacenamiento: Persistir las métricas en una base de datos No SQL (como Cassandra) para análisis históricos.

Bibliografía

- Apache Kafka Documentation: https://kafka.apache.org/documentation/
- Apache Spark Structured Streaming Guide: https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html
- Docker Compose Documentation: https://docs.docker.com/compose/
- Confluent Kafka Configuration Guide: https://docs.confluent.io/
- Calavaro, Russo, Cardellini (2022). Real-time analysis of market data leveraging Apache Flink.