

Comparación de las arquitecturas Kappa y Delta para el monitoreo remoto de pacientes basado en datos de sensores

Entregado como requisito para la obtencion del titulo
de Master en Big Data

Emiliano Conti - 289917

Tutor: Alejandro Bianchi

Universidad ORT

5 de abril de 2025

Disclaimer

Yo, Emiliano Conti declaro que el trabajo que se presenta en esta obra es de mi propia mano. Puedo asegurar que:

- La obra fue producida en su totalidad mientras realizaba el Proyecto Final del Master en Big Data;
- Cuando he consultado el trabajo publicado por otros, lo he atribuido con claridad;
- Cuando he citado obras de otros, he indicado las fuentes. Con excepción de estas citas, la obra es enteramente mía;
- En la obra, he acusado recibo de las ayudas recibidas;
- Cuando la obra se basa en trabajo realizado conjuntamente con otros, he explicado claramente qué fue contribuido por otros, y qué fue contribuido por mi;
- Ninguna parte de este trabajo ha sido publicada previamente a su entrega, excepto donde se han realizado las aclaraciones correspondientes.

Firma: _____

Fecha: _____

Abstract

Aquí va el resumen...

Índice general

1. Introducción	4
1.1. Descripción del Proyecto	4
1.2. Objetivos	5
1.3. Caso de Estudio: Big Data en Sistema de Salud	6
1.3.1. Contexto del Sistema	6
1.3.2. Descripción del Caso de Uso	6
1.3.3. Proceso	6
2. Marco Teórico	7
2.1. Introducción a Big Data y Streaming de Datos	7
2.1.1. Sistemas Distribuidos	7
2.1.2. Definición y características del Big Data	9
2.1.3. Streaming de datos	10
2.1.4. Teorema CAP	10
2.1.5. Desafíos en el manejo de datos de streaming	11
2.1.6. Conceptos clave en Streaming	13
2.1.7. Batch vs Streaming	14
2.1.8. Evolución de las arquitecturas de procesamiento de datos	15
2.2. Tecnologías para Streaming en Big Data	16
2.2.1. Mensajería Distribuida	16
2.2.2. Comparación	18
2.2.3. Motores de Procesamiento	21
2.2.4. Comparación	26
2.2.5. Almacenamiento de Datos	30
2.2.6. Comparación	41
2.2.7. Tecnologías Complementarias	42
2.2.8. Desafíos	43
2.3. Arquitecturas de Referencia	47
2.3.1. Instancias de Arquitectura	47
2.4. Arquitectura Lambda	48
2.4.1. Descripción General	48

2.4.2.	Componentes Principales	48
2.4.3.	Capacidades	50
2.4.4.	Desafíos	50
2.5.	Arquitectura Kappa	51
2.5.1.	Descripción General	51
2.5.2.	Componentes Principales	52
2.5.3.	Vista Lógica	53
2.5.4.	Capacidades	53
2.5.5.	Desafíos	54
2.6.	Arquitectura Delta	55
2.6.1.	Descripción General	55
2.6.2.	Componentes Principales	56
2.6.3.	Vista Lógica	57
2.6.4.	Capacidades	57
2.6.5.	Desafíos	57
2.7.	Monitoreo Remoto de Pacientes	58
2.7.1.	Monitoreo de Signos Vitales	58
2.7.2.	Identificación de Riesgo en Pacientes	60
2.7.3.	Desafíos	63
3.	Metodología	64
3.1.	Criterios de Evaluación	64
3.1.1.	Latencia y Rendimiento	65
3.1.2.	Manejo de Datos Históricos	65
3.1.3.	Costos Operativos	65
3.2.	Stack de Tecnologías a Utilizar	66
3.3.	Conjunto de Datos	68
3.3.1.	Conjunto de Datos	68
4.	Desarrollo	69
4.1.	Sistema de Monitoreo Remoto de Pacientes	69
4.1.1.	Introducción	69
4.1.2.	Fundamento Clínico: El Sistema de Puntuación NEWS2	70
4.1.3.	gdNEWS2	71
4.1.4.	Formato de Datos de Mediciones Brutas	71
4.1.5.	Algoritmo de Puntuación de Calidad	72
4.1.6.	Algoritmo de Puntuación de Frescura	73
4.1.7.	Algoritmo de Puntuación de Degradación	74
4.1.8.	Algoritmo de Puntuación de NEWS2	74
4.2.	Implementación	75
4.2.1.	Pipeline de Procesamiento	75

4.2.2.	Despliegue de Componentes	77
4.2.3.	Repositorio de Código	78
4.2.4.	Generación de Datos Sintéticos	79
4.2.5.	Limitaciones en la Implementación	81
4.3.	Arquitectura Kappa	82
4.3.1.	Principios de Diseño	82
4.3.2.	Stack Tecnológico	83
4.3.3.	Vista de Componentes	84
4.3.4.	Flujo de Procesamiento	85
4.4.	Arquitectura Delta	93
4.4.1.	Principios de Diseño	93
4.4.2.	Stack Tecnológico	94
4.4.3.	Vista de Componentes	96
4.4.4.	Flujo de Procesamiento	97
5.	Resultados	100
5.1.	Expectativas Iniciales	100
5.2.	Comparación Técnica	101
5.3.	Aspectos Operativos	103
5.3.1.	Throughput	103
5.3.2.	Latencia	104
5.3.3.	Uso de Recursos	106
5.3.4.	Resultados	108
5.4.	Costos	109
5.4.1.	Suposiciones	109
5.4.2.	Costos de la Arquitectura Kappa	110
5.4.3.	Costos de la Arquitectura Delta	111
5.4.4.	Resultados	112
6.	Conclusiones	113
A.	Primer Anexo	114

Capítulo 1

Introducción

1.1. Descripción del Proyecto

Este proyecto compara las arquitecturas Kappa y Delta en el contexto del monitoreo remoto de pacientes mediante sensores IoT. En la era de la salud digital, estos sistemas generan grandes volúmenes de datos en tiempo real que requieren procesamiento eficiente. Se analizarán ambas arquitecturas, se definirán métricas de comparación y se implementarán en un caso de uso de monitoreo de pacientes. El objetivo es determinar la arquitectura más adecuada, considerando factores como latencia, escalabilidad, complejidad de implementación y manejo de datos históricos y en tiempo real.

1.2. Objetivos

1. Realizar un análisis teórico exhaustivo de las arquitecturas Kappa y Delta, detallando sus componentes, flujos de datos y casos de uso típicos.
2. Definir un conjunto de métricas y criterios para la comparación objetiva de ambas arquitecturas en el contexto del monitoreo remoto de pacientes.
3. Implementar ambas arquitecturas utilizando un conjunto de datos simulado de sensores de monitoreo de pacientes.
4. Ejecutar pruebas de rendimiento y funcionalidad en ambas implementaciones.
5. Analizar los resultados obtenidos y determinar la arquitectura más adecuada para el caso de uso específico de monitoreo remoto de pacientes.
6. Proporcionar recomendaciones para la selección e implementación de arquitecturas de procesamiento de Big Data en el ámbito de la salud.

1.3. Caso de Estudio: Big Data en Sistema de Salud

1.3.1. Contexto del Sistema

Sistema de salud integral que incluye perfiles de pacientes, telemedicina e integración con dispositivos IoT. El objetivo es mejorar la atención médica, con foco en prevención, utilizando tecnología.

1.3.2. Descripción del Caso de Uso

Monitoreo continuo y en tiempo real de la salud del paciente mediante el uso de Big Data. Se espera además, tener la capacidad de identificar patrones y tendencias en los datos médicos. Así como también proporcionar recomendaciones personalizadas.

1.3.3. Proceso

1. Recopilación de Datos:

- Dispositivos IoT (datos en tiempo real)

2. Almacenamiento y Gestión:

- Almacenamiento de datos centralizada, segura y escalable

3. Análisis de Datos:

- Procesamiento en tiempo real
- Análisis históricos

4. Generación de Insights:

- Tableros
- Alertas en tiempo real

5. Intervención y Seguimiento:

- Monitoreo continuo
- Feedback y mejora continua del sistema

Capítulo 2

Marco Teórico

2.1. Introducción a Big Data y Streaming de Datos

2.1.1. Sistemas Distribuidos

Un sistema distribuido es una colección de elementos computacionales autónomos que para su usuario parecen un sistema único y coherente. (Tanenbaum & van Steen, 2020)

Los sistemas distribuidos tienen dos características que pueden regularse para escalar: Procesamiento y Almacenamiento.

En el último tiempo, ha habido una tendencia a preferir que la escala de ambas propiedades sea individual. Es decir, que se pueda escalar por un lado la potencia de procesamiento y por otro la capacidad de almacenamiento.

Consistencia

La Consistencia es la propiedad que tiene un sistema distribuido en la que todos los nodos ven los mismos datos al mismo tiempo. Esto significa que cualquier lectura en cualquier momento deberá devolver el valor más reciente escrito para ese dato. Si un sistema es consistente, una vez que se realiza una escritura, todas las lecturas subsiguientes deben reflejar esa escritura; sin importar desde que nodo se hagan. Esta propiedad garantiza que los clientes de los sistemas nunca vean datos desactualizados o inconsistentes.

Disponibilidad

La Disponibilidad es la propiedad que tiene un sistema distribuido para responder a todas las peticiones, ya sean de lectura o escritura, sin fallos. Un sistema disponible garantiza que cada solicitud reciba una respuesta sin importar el estado individual de cada nodo que lo compone. Esto significa que incluso si algunos nodos falla, el sistema en su conjunto debe poder seguir dando servicio a las peticiones que recibe.

Tolerancia a Particiones

La Tolerancia a Particiones es la propiedad que tiene un sistema distribuido en la que continua funcionando a pesar de la perdida de conectividad entre nodos. Una partición ocurre cuando hay una ruptura en la comunicación dentro de la red, lo que resulta en que dos o más segmentos de la red no puedan comunicarse entre sí. Un sistema tolerante a particiones puede seguir operando incluso cuando estas particiones ocurren, lo que significa que puede manejar retrasos o pérdidas de mensajes entre nodos sin fallar por completo.

2.1.2. Definición y características del Big Data

Big Data es un término paraguas que se usa para denominar a un conjunto de tecnologías que manejan grandes volúmenes de datos. La pregunta que se presenta entonces es: ¿qué tan grandes deberían ser estos volúmenes para ser considerados Big Data? O incluso, ¿existen otras características que definan lo que es Big Data? Una definición generalmente aceptada es la siguiente:

Las tecnologías de Big Data están orientadas a procesar datos de alto volumen, alta velocidad y alta variedad para extraer el valor de datos previsto y asegurar una alta veracidad de los datos originales y la información obtenida, lo que demanda formas de procesamiento de datos e información rentables e innovadoras para mejorar el conocimiento, la toma de decisiones y el control de procesos; todo esto exige nuevos modelos de datos que soporten todos los estados y etapas de los datos durante todo su ciclo de vida, y nuevos servicios y herramientas de infraestructura que permitan obtener y procesar datos de una variedad de fuentes y entregar datos en una variedad de formas a diferentes consumidores y dispositivos de datos e información. (Demchenko et al., 2014)

Por lo que podríamos considerar que es Big Data todo aquello que esté orientado a datos cuyo volumen, velocidad y variedad no puedan ser tratados por un modelo de procesamiento de datos tradicional (como podrían ser las bases de datos relacionales). Con el objetivo de generar valor, asegurando la veracidad de los datos originales y la información obtenida.

2.1.3. Streaming de datos

El streaming de datos, también conocido como procesamiento de flujo, es un paradigma de procesamiento de datos en el que los datos se tratan como un flujo continuo e ilimitado de eventos discretos. En el contexto de Big Data, el streaming permite procesar y analizar grandes volúmenes de datos en tiempo real o casi real, a medida que se generan o llegan al sistema. (Hueske & Kalavri, 2019)

2.1.4. Teorema CAP

El Teorema CAP es un concepto fundamental en el diseño de sistemas distribuidos. Este establece que es imposible garantizar al mismo tiempo, tanto la Consistencia (Consistency), Disponibilidad (Availability) y la Tolerancia a las Particiones (Partition Tolerance).

Según esto, un sistema distribuido sólo es capaz de garantizar dos de estas propiedades al mismo tiempo. En general, para los sistemas de Big Data de Streaming, la disponibilidad es una propiedad obligatoria, ya que cualquier inactividad puede resultar en la pérdida de datos valiosos o en la imposibilidad de realizar acciones.

Por otro lado, la Tolerancia a Particiones es también indispensable para estos sistemas, que por su naturaleza requieren que su capacidad de procesamiento este distribuida a través de múltiples nodos dispersos en una red no confiable; por lo que son susceptibles a que se genere una partición. Por lo tanto, si no tuviera esta propiedad el servicio podría dejar de ser disponible.

Entonces, como corolario, un sistema de Big Data de Streaming debe también ser tolerante a las particiones para poder ser disponible. Esto nos deja con una única opción: relajar el "grado de consistencia" hasta un punto razonable que permita que el sistema siga siendo eficaz. (Muñoz-Escóí et al., 2019)

Consistencia Eventual

La consistencia eventual es un modelo de consistencia en sistemas distribuidos que garantiza que, si no se realizan nuevas actualizaciones a un objeto, en algún momento (eventualmente) todos los accesos a ese objeto devolverán el último valor actualizado. La consistencia eventual se alinea con las compensaciones descritas por el teorema CAP, permitiendo que estos sistemas prioricen la disponibilidad y la tolerancia a particiones. Además, facilita la escalabilidad horizontal, crucial para manejar el crecimiento continuo de datos y clientes de los sistemas. Por último, es importante diseñar cuidadosamente el sistema para manejar las posibles inconsistencias temporales y asegurar que la aplicación pueda tolerar y resolver estas situaciones de manera apropiada (Muñoz-Escóí et al., 2019)

2.1.5. Desafíos en el manejo de datos de streaming

1. Procesamiento en tiempo real y baja latencia

El procesamiento de datos debe ocurrir con un retraso mínimo para proporcionar resultados en tiempo real.

Un desafío clave en el streaming es lograr equilibrar la latencia, el costo y la correctitud simultáneamente (Akidau et al., 2015).

2. Manejo de datos fuera de orden

Los datos pueden llegar en un orden diferente al que fueron generados, lo que complica el procesamiento.

El procesamiento de eventos fuera de orden es un desafío fundamental en los sistemas de streaming (Hueske & Kalavri, 2019, p. 87).

3. Escalabilidad

Los sistemas deben poder manejar volúmenes crecientes de datos sin degradación del rendimiento.

La escalabilidad en sistemas de streaming implica la capacidad de aumentar el rendimiento añadiendo recursos computacionales (Preuveneers et al., 2016).

4. Tolerancia a fallos y consistencia

El sistema debe poder recuperarse de fallos sin pérdida de datos y mantener la consistencia eventual de los resultados.

Garantizar la semántica de "exactly-once" en presencia de fallos es un desafío significativo en streaming (Carbone et al., 2015).

5. Procesamiento de ventanas temporales

Definir y procesar eficientemente ventanas de tiempo sobre streams de datos continuos.

El procesamiento de ventanas temporales es fundamental en aplicaciones de streaming y requiere consideraciones cuidadosas en cuanto a la semántica del tiempo y la completitud de los datos (Akidau et al., 2015).

6. Integración con sistemas batch

Combinar eficazmente el streaming con sistemas batch existentes.

La integración de paradigmas batch y streaming, a menudo referida como 'procesamiento híbrido', presenta desafíos únicos en términos de consistencia de datos y modelos de programación (Carbone et al., 2015).

2.1.6. Conceptos clave en Streaming

El streaming se refiere al análisis y manipulación de datos en tiempo real a medida que se generan o reciben. Según Carbone et al. (Carbone et al., 2015), los conceptos fundamentales incluyen:

- **Flujo de datos:** Una secuencia potencialmente infinita de registros que llegan continuamente (Akidau et al., 2015).
- **Latencia:** El tiempo entre la llegada de un dato y su procesamiento, crucial para aplicaciones en tiempo real (Akidau et al., 2015).
- **Ventanas:** Mecanismos para agrupar datos en intervalos finitos para su procesamiento (Akidau et al., 2015).
- **Estado:** Información que se mantiene entre eventos para cálculos incrementales (Carbone et al., 2015).
- **Watermarks:** Indicadores de progreso del tiempo en el flujo de datos (Akidau et al., 2015).

2.1.7. Batch vs Streaming

La elección entre batch y streaming depende de los requisitos específicos de la aplicación:

Característica	Batch	Streaming
Latencia	Alta (minutos a horas)	Baja (milisegundos a segundos)
Procesamiento	Alto	Moderado a alto
Complejidad	Menor	Mayor
Consistencia	Fuerte	Eventual
Uso típico	Análisis histórico, reportes	Monitoreo, alertas, decisiones inmediatas

Cuadro 2.1: Comparación de batch y streaming

Se puede decir que el streaming es esencial para aplicaciones que requieren decisiones inmediatas, mientras que el análisis batch es más adecuado para análisis profundos de grandes volúmenes de datos históricos. (Stonebraker et al., 2005)

2.1.8. Evolución de las arquitecturas de procesamiento de datos

La evolución de las arquitecturas de procesamiento de datos ha sido impulsada por la necesidad de manejar volúmenes cada vez mayores de datos en tiempo real:

1. **Arquitecturas por lotes:** Sistemas tradicionales como Hadoop MapReduce, diseñados para procesar grandes volúmenes de datos estáticos (Dean & Ghemawat, 2008).
2. **Arquitecturas de streaming puro:** Como Apache Storm, enfocadas en el procesamiento en tiempo real pero con limitaciones en la consistencia y exactitud (Toshniwal et al., 2014).
3. **Arquitectura Lambda:** Propuesta por Marz (Marz, 2011), combina procesamiento por lotes y en tiempo real para balancear latencia, throughput y tolerancia a fallos.
4. **Arquitectura Kappa:** Introducida por Kreps (Kreps, 2014), simplifica Lambda tratando todos los datos como streams.
5. **Arquitectura Delta:** Desarrollada por Databricks, también como respuesta a Lambda optimiza el procesamiento de datos tanto batch como streaming (Armbrust et al., 2020) (Leano, 2020).

2.2. Tecnologías para Streaming en Big Data

2.2.1. Mensajería Distribuida

Las tecnologías de mensajería distribuidas en tiempo real cumplen el crucial rol de actuar como intermediarios entre las fuentes de datos y los sistemas que efectivamente procesan estos datos. (Kleppmann, 2018)

Deben funcionar como un conducto de alta capacidad de almacenamiento y baja latencia, capturando y canalizando los flujos de información desde sus múltiples orígenes y hacia sus diversos destinos en tiempo real. (Marz & Warren, 2015)

Su papel es fundamentalmente el de un sistema nervioso central, coordinando y distribuyendo datos a través de complejas arquitecturas distribuidas. Actúan como amortiguadores, absorbiendo picos en el flujo de datos y garantizando un procesamiento constante y eficiente. Además, estas tecnologías sirven como una capa de abstracción, desacoplando los productores de datos de los consumidores, lo que permite una mayor flexibilidad y escalabilidad en el diseño del sistema.

Apache Kafka

Apache Kafka es el estándar de facto de este tipo de sistemas. Utiliza un modelo de publicación-subscripción (pub/sub) basado en logs, donde los datos se envían a "topics" y se almacenan en particiones distribuidas. Las particiones tienen una garantía de orden de los mensajes y permiten la retención de datos a largo plazo.

Además, proporciona conectores para integración con diversos sistemas y un amplio ecosistema. A nivel de seguridad ofrece encriptación en tránsito mediante TLS y permite ser configurado para soportar encriptación en reposo (aunque esto debe hacerse a nivel de sistema de archivos).

Por último, su arquitectura distribuida y replicada permite una alta disponibilidad y tolerancia a fallos.

Apache Pulsar

Apache Pulsar es también una plataforma de mensajería y streaming distribuida, al igual que Apache Kafka, pero que se distingue por tener una arquitectura basada en capas, separando la capa de almacenamiento de la capa de procesamiento; lo que permite escalar cada uno independientemente.

Apache Pulsar soporta modelos de entrega como colas, publicación y suscripción y puede ofrecer garantías de entregar un mensaje exactamente una vez ("exactly-once"). Ofrece encriptación a nivel de mensaje, lo que permite una granularidad fina en cuanto a que encriptar. También soporta TLS para la encriptación en tránsito y permite la configuración de encriptación en reposo.

Por último, también soporta almacenamiento de mensajes a largo plazo y soporte nativo para esquemas: Esto es, permite definir la estructura y el tipo de datos de los mensajes, lo que a su vez permite una validación automática de los datos y una serialización/deserialización más eficiente. Esto trae consigo además, la capacidad de evolucionar estos esquemas de mensajes, de forma que productores y consumidores evolucionen independientemente.

Amazon Kinesis

Amazon Kinesis es un servicio de streaming de datos administrado en la nube de AWS. Está diseñado para recopilar, procesar y analizar datos de streaming en tiempo real a gran escala. Kinesis, en realidad, se compone de varios servicios:

1. Kinesis Data Streams para ingestión de datos en tiempo real
2. Kinesis Data Firehose para cargar datos en los servicios de almacenamiento disponibles de AWS
3. Kinesis Data Analytics para procesar datos con SQL o Java
4. Kinesis Video Streams para streaming de video

Adicionalmente ofrece capacidades de auto-escalado, replicación entre zonas de disponibilidad para alta durabilidad, encriptación en reposo (y puede habilitarse la encriptación en tránsito) y permite la retención de datos hasta 365 días.

Como se puede ver, al ser tan completo permitiría implementar, al menos en principio, una gran parte de un sistema de Big Data en tiempo real.

Azure Event Hubs

Azure Event Hubs es un servicio de ingestión de datos en tiempo real administrado en la plataforma Microsoft Azure. Se supone que está diseñado para soportar millones de eventos por segundo con baja latencia. Al igual que Kinesis, ofrece una muy buena con otros servicios de Microsoft Azure, lo que permite por ejemplo capturar directamente los eventos en los servicios de almacenamiento disponibles en este proveedor de nube.

Event Hubs es compatible con el protocolo Kafka, lo que permite a las aplicaciones existentes de Kafka conectarse sin cambios de código. Ofrece una retención de mensajes por defecto de 1 día que puede aumentado hasta 7. En caso de necesitar una retención más a largo plazo se recomienda guardar los eventos en Azure Blob Storage o Azure Data Lake para su posterior procesamiento. Cuenta también con encriptación en tránsito con TLS y en reposo.

Proporciona además, características como el procesamiento batch para optimizar el rendimiento, control de acceso basado en roles, y encriptación en reposo y en tránsito.

2.2.2. Comparación

Escalabilidad

- **Apache Kafka:** Muy Alta escalabilidad horizontal, millones de mensajes/segundo.
- **Apache Pulsar:** Muy alta escalabilidad horizontal, millones de mensajes/segundo con separación de almacenamiento y cómputo.
- **Amazon Kinesis:** Buena escalabilidad, con 1.000 mensajes por segundo con la configuración por defecto aunque con configuración adicional puede llegar al millón por segundo hipotético.
- **Azure Event Hubs:** Buena escalabilidad, con 1.000 mensajes por segundo. También puede escalar con configuración adicional a los 20.000 mensajes. Existe la posibilidad de tener una instancia dedicada que permite escalar a millones de eventos por segundo de forma hipotética.

Retención de datos

- **Apache Kafka:** Configurable, potencialmente ilimitada.
- **Apache Pulsar:** Ilimitada por diseño.
- **Amazon Kinesis:** Hasta 365 días, configurable.
- **Azure Event Hubs:** Hasta 7 días, opción de Capture para largo plazo.

Garantías de entrega

- **Apache Kafka:** At-least-once por defecto, exactly-once configurable.
- **Apache Pulsar:** Exactly-once nativo.
- **Amazon Kinesis:** At-least-once.
- **Azure Event Hubs:** At-least-once.

Encriptación

- **En reposo:**
 - **Apache Kafka:** Configurable.
 - **Apache Pulsar:** Nativo.
 - **Amazon Kinesis:** Por defecto (AWS KMS).
 - **Azure Event Hubs:** Por defecto.
- **En tránsito:** Todos soportan TLS/SSL.

Observaciones clave

- Pulsar destaca en escalabilidad, retención, seguridad y consistencia nativos.
- Kafka ofrece el ecosistema más maduro y desarrollado que le otorga características muy completas.
- Servicios gestionados (Kinesis, Event Hubs) tienen encriptación en reposo por defecto, pero retención limitada.

Característica	Kafka	Pulsar	Kinesis	Event Hubs
Escalabilidad	Muy Alta	Muy alta	Alta	Alta
Retención	Configurable	Ilimitada	365 días máx.	7 días máx.
Garantía de entrega	Exactly-once	Exactly-once	At-least-once	At-least-once
Encriptación en reposo	Configurable	Sí	Sí (AWS KMS)	Sí
Encriptación en tránsito	Sí	Sí	Sí	Sí
Throughput	Muy alto	Muy alto	Alto	Alto

Cuadro 2.2: Comparación de Sistemas de Mensajería

2.2.3. Motores de Procesamiento

Si las tecnologías de mensajería distribuida son el sistema nervioso de un sistema de Big Data de Streaming, los motores de procesamiento podrían considerarse su cerebro.

Actúan como la capa que transforma, enriquece y analiza los datos cumpliendo varios roles:

- Aplicar la lógica de negocio sobre los datos mientras estos fluyen
- Detectar patrones y anomalías sobre el flujo de datos
- Mantener el contexto y estado necesario para las operaciones con históricos o agregaciones
- Garantizar la consistencia de las operaciones incluso ante fallos del sistema
- Distribuir la carga de trabajo entre diferentes nodos, paralelizando tareas

Estos componentes pueden enlazarse y programarse de diversas maneras. Permitiendo ensamblarlos de forma de cumplir con los requisitos de negocio.

Apache Spark

Apache Spark se destaca como uno de los motores de procesamiento más populares y versátiles en el ecosistema de Big Data. Ofrece dos APIs principales para el procesamiento en tiempo real: Spark Streaming y Structured Streaming, siendo esta última la más moderna y recomendada. Implementa el procesamiento en tiempo real tratando los datos streaming como micro-batches y su enfoque de procesamiento es en memoria, siendo capaz de mantener su estado distribuido a través de checkpoints, que son archivos que se guardan en un sistema de almacenamiento al que todos los nodos pueden acceder. Su modelo de procesamiento le permite ofrecer un cierto equilibrio entre latencia y throughput. La abstracción fundamental de Spark son los RDDs (Resilient Distributed Datasets), que son colecciones inmutables de datos distribuidos que pueden ser procesadas en paralelo, y los DataFrames, que proporcionan una abstracción de más alto nivel similar a una tabla de base de datos. Spark destaca por su amplio soporte de lenguajes de programación:

- Scala
- Python
- Java
- R
- SQL

Su API unificada y extenso catálogo de bibliotecas incluye MLlib para machine learning, GraphX para procesamiento de grafos, y Spark SQL para procesamiento estructurado.

Apache Flink

Apache Flink adopta un enfoque nativo de streaming, tratando al procesamiento batch como un caso especial de streaming con límites finitos. Su arquitectura está diseñada para mantener estado distribuido con garantías de consistencia muy fuertes y latencias extremadamente bajas. El motor gestiona automáticamente la distribución del estado y los checkpoints, asegurando semánticas de exactly-once y permitiendo recuperación exacta ante fallos sin duplicados. Su modelo de procesamiento se basa en dos conceptos fundamentales:

- Marcas de agua (Watermarks): Son metadatos que fluyen en el stream de datos indicando el progreso del tiempo del evento, permitiendo manejar datos desordenados.
- Ventanas de tiempo (Windows): Permiten agrupar y procesar datos en intervalos temporales definidos, soportando diversos tipos como tumbling, sliding y session windows.

Flink proporciona APIs de diferentes niveles:

- ProcessFunction: API de bajo nivel que ofrece máximo control sobre tiempo, estado y ventanas
- DataStream API: API de alto nivel para operaciones de streaming comunes
- Table API y SQL: APIs declarativas para operaciones relacionales

Los lenguajes soportados son:

- Java
- Scala
- Python
- SQL

Apache Beam

Apache Beam, por su lado, se distingue por proporcionar un modelo de programación unificado que abstrae el motor de ejecución subyacente. Su potencia radica en la capacidad de escribir la lógica de procesamiento una vez y ejecutarla en diferentes motores de procesamiento (runners) como Spark o Flink. Esta capacidad de abstracción es particularmente valiosa en escenarios donde la portabilidad y la flexibilidad de despliegue son requisitos clave, permitiendo cambiar de motor de procesamiento según evolucionen las necesidades y sin reescribir el código de procesamiento.

Apache Samza

Apache Samza se distingue por su estrecha integración con Apache Kafka y su arquitectura diseñada para mantener el estado de procesamiento de forma distribuida con un modelo de particionamiento que permite escalar horizontalmente. Samza proporciona un modelo de procesamiento simple pero potente, con fuerte énfasis en la gestión de estado local y la tolerancia a fallos. Se utiliza en LinkedIn y la arquitectura Kappa fué propuesta inicialmente pensando en la utilización de este motor de procesamiento.

Apache NiFi

Apache NiFi aborda el procesamiento de datos desde una perspectiva de orquestación y gobierno de datos; centrándose en la automatización del flujo de datos entre sistemas. Su arquitectura está orientada a la trazabilidad y auditabilidad de cada dato que fluye por el sistema, manteniendo un registro detallado de todas las transformaciones y movimientos. Los datos fluyen a través de un grafo de procesadores que pueden transformar, enrutar y mediar entre diferentes protocolos y formatos. NiFi se destaca por su capacidad para garantizar la entrega confiable de datos, proveer linaje de datos completo y permitir modificaciones de flujos en tiempo real sin necesidad de detener el sistema. Por último, NiFi proporciona una interfaz visual para diseñar, controlar y monitorizar flujos de datos.

Apache Kafka Streams

Apache Kafka Streams es una biblioteca de procesamiento de streaming que forma parte del ecosistema de Apache Kafka, diseñada para construir aplicaciones y microservicios de procesamiento en tiempo real. Opera con un modelo de procesamiento que permite operaciones con manejo de estado, incluyendo agregaciones por ventanas, manejo de múltiples streams de datos y transformaciones complejas mediante APIs de alto y bajo nivel. Tiene un manejo de estado distribuido que se gestionan con los mismos logs de Kafka. En cuanto a rendimiento, Kafka Streams alcanza una latencia típica de decenas de milisegundos a segundos, dependiendo de la complejidad del procesamiento y la configuración, mientras que su throughput puede escalar linealmente añadiendo más instancias. También ofrece garantías de procesamiento at-least-once con posibilidad de exactly-once mediante configuración.

Apache Pulsar Functions

Apache Pulsar Functions ofrece capacidad de cómputo integrándose directamente en la infraestructura de Apache Pulsar. Este framework permite implementar funciones livianas que procesan mensaje a mensaje. El manejo del estado es distribuido y se realiza mediante un almacenamiento basado en RocksDB, que permite mantener información entre invocaciones de funciones de manera consistente y tolerante a fallos. En términos de rendimiento, está optimizado para baja latencia, típicamente en el rango de milisegundos, gracias a su modelo de procesamiento directo. El throughput puede escalar horizontalmente añadiendo más instancias de funciones, y el sistema proporciona garantías de procesamiento exactly-once. La arquitectura de este sistema está diseñada para ser simple y eficiente, permitiendo casos de uso como enriquecimiento de datos, filtrado, y transformaciones en tiempo real.

2.2.4. Comparación

Modelo de Procesamiento

- **Apache Spark:** Micro-batches.
- **Apache Flink:** Streaming nativo con procesamiento registro a registro.
- **Apache Beam:** Mismo que el modelo de su motor asociado
- **Apache Samza:** Streaming nativo con procesamiento basado en tiempo
- **Apache NiFi:** Procesamiento basado en flujos de datos dirigidos
- **Kafka Streams:** Procesamiento de baja complejidad mensaje a mensaje integrado con Kafka
- **Pulsar Functions:** Procesamiento de baja complejidad mensaje a mensaje integrado con Pulsar

Manejo de Estado

- **Apache Spark:** Estado en memoria distribuido entre los nodos procesadores.
- **Apache Flink:** Estado distribuido utilizando snapshots.
- **Apache Beam:** Mismo que el manejo de su motor asociado
- **Apache Samza:** Estado local con respaldo en Kafka
- **Apache NiFi:** Se maneja localmente en memoria en el nodo que procesa el flujo
- **Kafka Streams:** Estado local con respaldo en Kafka
- **Pulsar Functions:** Almacenado y gestionado con la instancia de manejo de almacenamiento (BookKeeper) de Pulsar

Latencia

- **Apache Spark:** Dependiendo de la configuración del micro-batch puede ir desde milisegundos a segundos.
- **Apache Flink:** Microsegundos a milisegundos.
- **Apache Beam:** Mismo que el de su motor asociado
- **Apache Samza:** Milisegundos a segundos
- **Apache NiFi:** Segundos a Minutos
- **Kafka Streams:** Milisegundos a segundos
- **Pulsar Functions:** Milisegundos a segundos

Capacidad de Procesamiento

- **Apache Spark:** Cientos de miles de eventos por segundo por nodo
- **Apache Flink:** Millones de eventos por segundo por nodo
- **Apache Beam:** Mismo que el de su motor asociado
- **Apache Samza:** Cientos de miles de eventos por segundo por partición
- **Apache NiFi:** Miles de eventos por segundo por nodo
- **Kafka Streams:** Cientos de miles de eventos por segundo por partición
- **Pulsar Functions:** Decenas de miles de eventos por segundo por nodo

Garantías de Entrega

- **Apache Spark:** at-least-once por defecto aunque puede ser configurado para permitir exactly-once
- **Apache Flink:** exactly-once nativo
- **Apache Beam:** exactly-once, at-least-once y at-most-once dependiendo del motor utilizado
- **Apache Samza:** at-least-once por defecto aunque puede ser configurado para permitir exactly-once
- **Apache NiFi:** at-least-once con garantía de entrega a pesar de fallas del sistema
- **Kafka Streams:** at-least-once y exactly-once
- **Pulsar Functions:** mejor rendimiento con at-least-once pero soporta exactly-once y at-most-once

Observaciones clave

- Spark es el que tiene un ecosistema más extendido y es más accesible
- Flink tiene las mejores prestaciones a nivel de latencia y rendimiento
- NiFi no es una buena opción para streaming por su latencia
- Kafka Streams y Pulsar Functions pueden ser usados de forma ligera para ruteo de mensajes

Cuadro 2.3: Comparativa de Motores de Procesamiento Big Data

	Modelo	Estado	Latencia	Proc.	Garantías
Apache Spark	Micro-batches	En memoria	Baja	Alto	Exactly-once
Apache Flink	Streaming	Distribuido	Muy baja	Muy Alto	Exactly-once
Apache Beam	Variable	Variable	Variable	Variable	Variable
Apache Samza	Streaming	Por partición	Baja	Alto	Exactly-once
Apache NiFi	Flujo dirigido	Procesador	Media-Alta	Medio	At-least-once
Kafka Streams	Streaming	Distribuido	Baja	Alto	Exactly-once
Pulsar Functions	Mensaje	Distribuido	Baja	Alto	Exactly-once

2.2.5. Almacenamiento de Datos

Formatos de Almacenamiento

Los formatos de datos son un componente fundamental en cualquier arquitectura de sistemas de información moderna, ya que determinan no solo cómo se almacena la información, sino también cómo se procesa, transmite y analiza. La elección adecuada del formato de datos puede tener un impacto significativo en el rendimiento, la escalabilidad y la eficiencia del sistema en su conjunto. Para un sistema de Big Data, donde se manejan grandes volúmenes de información, la importancia de estos formatos se magnifica, ya que pueden significar la diferencia entre un sistema eficiente y uno que consume recursos excesivos. Además, los formatos de datos actúan como un lenguaje común entre diferentes componentes, facilitando la interoperabilidad y la integración de tecnologías.

Formatos Orientados a Filas

Los formatos orientados a filas representan la forma tradicional de almacenamiento de datos, donde cada registro se almacena de manera secuencial. Este enfoque ha sido la base de los sistemas de gestión de bases de datos durante décadas y sigue siendo crucial en muchos escenarios.

- Los registros completos se almacenan de manera contigua en disco
- Cada fila contiene todos los campos de un registro
- Optimizado para acceder a registros completos
- Los nuevos registros se añaden secuencialmente de forma eficiente
- Óptimo cuando las consultas necesitan todos los campos
- Fácil modificación de registros individuales
- Debe leer datos innecesarios cuando solo se necesitan algunas columnas
- Los datos heterogéneos juntos reducen la efectividad de los métodos de compresión
- Menos eficiente para análisis de columnas específicas

Formatos Orientados a Columnas

Los formatos de almacenamiento columnar representan un paradigma fundamental en el manejo de datos masivos, especialmente en entornos analíticos. A diferencia del almacenamiento tradicional orientado a filas, donde los registros se almacenan secuencialmente, el almacenamiento columnar organiza los datos por columnas, lo que ofrece ventajas significativas en ciertos escenarios.

- En lugar de almacenar registros completos de manera contigua, los datos se organizan por columnas
- Cada columna se almacena en bloques separados de memoria o disco
- Los valores similares se almacenan juntos, mejorando la compresión
- Los datos similares almacenados juntos permiten mayores tasas de compresión
- Solo se leen las columnas necesarias para una consulta
- Facilita operaciones como SUM, AVG, COUNT sobre columnas específicas
- Permite procesamiento eficiente de datos en hardware

Comparativa con Almacenamiento por Filas:

Consideremos una tabla simple de usuarios:

Formato por Filas:

[ID1, "Juan", 25] -> [ID2, "Ana", 30] -> [ID3, "Pedro", 28]

Formato Columnar:

IDs: [ID1 -> ID2 -> ID3]

Nombres: ["Juan" -> "Ana" -> "Pedro"]

Edades: [25 -> 30 -> 28]

Aspecto	Formato Columnar	Formato por Filas
Lectura parcial	Muy eficiente	Menos eficiente
Inserción de registros	Más lenta	Más rápida
Compresión	Alta	Moderada
Consultas analíticas	Excelente	Regular
Consultas transaccionales	Regular	Excelente

Formatos Específicos

JSON (JavaScript Object Notation) se ha convertido en el estándar de facto para el intercambio de datos en aplicaciones modernas, especialmente en entornos Web y APIs. Su popularidad se debe a su simplicidad, legibilidad humana y amplia compatibilidad con prácticamente todos los lenguajes de programación. A pesar de no ser el más eficiente en términos de espacio y rendimiento (ya que es un formato basado en filas), su flexibilidad para representar datos estructurados y semiestructurados lo hace invaluable en sistemas donde la interoperabilidad y la facilidad de desarrollo son prioritarias. Es particularmente útil en aplicaciones donde las transacciones individuales y la flexibilidad del esquema son más importantes que la eficiencia en el procesamiento de grandes volúmenes de datos.

Apache AVRO destaca como un formato de serialización de datos binario que combina la eficiencia del almacenamiento binario con la flexibilidad de esquemas evolutivos. Su característica más distintiva es su capacidad para manejar cambios en el esquema de datos a lo largo del tiempo sin requerir cambios en el código o reescritura de datos existentes. Para esto, AVRO almacena el esquema junto con los datos, lo que permite una deserialización precisa y eficiente. Es especialmente valioso en sistemas de mensajería y streaming de datos, donde la evolución del esquema y la eficiencia en la transmisión son cruciales. Su formato binario compacto y su capacidad de compresión lo hacen ideal para sistemas distribuidos donde el ancho de banda y el almacenamiento son consideraciones importantes.

Apache Parquet se ha establecido como el formato columnar dominante en el ecosistema de Big Data, especialmente para cargas de trabajo analíticas. Su diseño columnar permite una compresión altamente eficiente y un muy buen rendimiento en consultas que involucran solo un subconjunto de columnas. Parquet destaca particularmente en escenarios de análisis de datos, donde su capacidad para manejar esquemas complejos anidados y su integración con casi todas las herramientas lo hacen indispensable. La adopción generalizada de Parquet en la industria, lo ha convertido en el estándar de facto para almacenamiento de datos analíticos.

Optimized Row Columnar (ORC) inicialmente fué desarrollado para optimizar Hive, y aunque ofrece excelentes capacidades de compresión y rendimiento en consultas, su relevancia ha disminuido significativamente en los últimos años frente a Parquet. Aunque ORC sigue siendo relevante en sistemas legacy y específicos de Hive, la tendencia de la industria se ha movido claramente hacia Parquet como el formato columnar preferido para análisis de datos a gran escala.

Almacenamiento de Objetos en la Nube

El almacenamiento de objetos en la nube constituye un paradigma de almacenamiento donde los datos se organizan y gestionan como objetos independientes dentro de una estructura plana. Cada objeto almacenado comprende tres elementos fundamentales: los datos en sí mismos, un conjunto extenso de metadatos que describen y categorizan la información, y un identificador único global que permite su localización y recuperación. Dicho paradigma, se caracteriza por su naturaleza distribuida y su capacidad para manejar tanto datos estructurados como no estructurados, permitiendo almacenar desde documentos, archivos multimedia y hasta flujos de datos en tiempo real.

Los sistemas de almacenamiento de objetos se destaca por su capacidad para satisfacer las demandas contemporáneas de procesamiento de datos a gran escala. Como por ejemplo, ofrece una escalabilidad prácticamente ilimitada, pudiendo crecer según las necesidades sin preocuparse por restricciones de capacidad. Por otro lado, la durabilidad y disponibilidad de los datos se garantiza a través de la replicación automática en múltiples ubicaciones de forma automática y transparente. Adicionalmente, proveen APIs normalmente basadas en el protocolo HTTP que permite acceder de forma interoperable y estandarizada a los recursos almacenados

El uso de estos sistemas también conlleva sus propios desafíos. El más importante puede ser la latencia; pero también deben los grados de consistencia que ofrecen.

Formato de Tabla Analítica

Los formatos de tabla analítica son tecnologías diseñadas para resolver los desafíos del manejo de datos a gran escala. Surgen como respuesta a las limitaciones de los formatos de archivo tradicionales como Parquet y ORC cuando se trabaja con ellos en la nube.

Las características más importantes que aporta un formato de tabla analítica son:

- Proveen abstracciones sobre la metadata de archivos
- Permiten el uso de tablas con semántica SQL y evolución de esquema en las mismas
- Transacciones ACID
- Actualizaciones y Borrados
- Optimización de datos para mejoras de rendimiento
- Compatibilidad con múltiples motores de procesamiento
- Control de versiones en los datos

Los formatos de tablas analíticas dan a sus sistemas subyacentes estas características a través de diferentes mecanismos y estrategias. La principal es la gestión de metadatos, ya que implementan estructuras de datos altamente optimizadas que permiten rastrear eficientemente los archivos y sus cambios; mientras mantienen un historial detallado de transacciones que garantiza la consistencia de los datos, complementando con estrategias efectivas de particionamiento y organización.

Para la optimización del rendimiento, estos formatos emplean diversas técnicas como la compactación automática de archivos, estrategias de caching de datos para acceso rápido, utilización de formatos de archivo columnares como Parquet u ORC, y la incorporación de capacidades avanzadas de indexación y filtrado optimizado. La consistencia de los datos se garantiza mediante la implementación de transacciones ACID completas, que proporcionan un sólido aislamiento entre operaciones de lectura y escritura, manejan conflictos de manera automática y aseguran la consistencia en escenarios de operaciones concurrentes. En cuanto a la evolución y mantenimiento, estos formatos facilitan cambios de esquema sin interrupciones en el servicio.

Todas estas mecanismos trabajan en conjunto para proporcionar una solución completa para el manejo de datos a escala masiva, que como subproducto permite tratar la escritura de los archivos como un canal de mensajes sobre el que se puede hacer streaming.

Los exponentes más importantes de estos formatos son: Delta Table, Apache Iceberg y Apache Hudi

Delta Lake es un sistema de almacenamiento de datos diseñado por Databricks que utiliza archivos Parquet como base, organizándolos en una estructura de directorios con dos componentes principales: los archivos de datos en formato Parquet y el directorio `_delta_log` para metadatos y registro de transacciones. Esta arquitectura mantiene las ventajas de Parquet mientras añade capacidades transaccionales y de control de versiones, implementando un sistema de checkpoints para optimizar el rendimiento y un manejo de concurrencia que combina control optimista con serialización de escrituras.

Las características fundamentales de Delta Lake incluyen transacciones ACID completas, capacidad de acceso a versiones anteriores, evolución de esquema controlada, operaciones de Merge sofisticadas y optimización automática de datos mediante compactación de archivos y mantenimiento de estadísticas. El sistema también proporciona soporte para procesamiento de datos en tiempo real con semántica exactly-once y una integración robusta principalmente con Spark.

Apache Hudi desarrollado inicialmente por Uber, es una plataforma enfocada en crear una plataforma analítica transaccional, disponibilizando dos tipos principales de formatos de tablas: Copy On Write (optimizado para lecturas) y Merge On Read (optimizado para escritura).

Utiliza una Timeline para gestionar metadatos y registrar cronológicamente todas las acciones, implementando un sistema de indexación que permite optimizar operaciones y facilitar búsquedas rápidas, además maneja control de concurrencia optimista que mantiene lecturas sin bloqueos mientras serializa escrituras. Las características principales de Hudi incluyen procesamiento incremental para manejar streams de datos, gestión sofisticada de registros individuales con versionado a nivel de registro, borrado lógico y evolución de esquemas. También proporciona garantías de consistencia con transacciones ACID y semántica exactly-once, ofreciendo buena integración con motores de procesamiento como Spark y Flink.

Hudi, al ser una plataforma, no ofrece solo su formato de tabla analítica, sino también "Table Services" que son componentes computacionales que permiten la optimización como compactación automática y limpieza de almacenamiento.

Apache Iceberg diseñado inicialmente por Netflix, implementa una arquitectura de almacenamiento que se caracteriza por un modelo de metadatos enfocado en la evolución del esquema.

Utiliza una estructura jerárquica que separa completamente los metadatos de los datos, implementando control de versiones basado en snapshots atómicos e inmutables, donde cada snapshot representa un punto en el tiempo de la tabla y contiene referencias a todos los archivos de datos válidos para esa versión, permitiendo operaciones concurrentes sin necesidad de bloqueos pesados.

Entre sus características principales destacan una gestión de esquemas flexible que permite evolucionar tanto el esquema como la estrategia de particionamiento sin reescribir datos, una optimización de consultas avanzada basada en estadísticas detalladas a nivel de columna y archivos, y un sistema de control de concurrencia optimista.

Además, disponibiliza herramientas de mantenimiento como expiración de snapshots, compactación de archivos y reescritura de datos para optimización física, junto con una robusta integración con diferentes motores de procesamiento como Spark y Flink.

Apache Paimon es un formato de tabla analítica que busca permitir la construcción de arquitecturas basadas en Lakehouse en tiempo real. Nació como un proyecto dentro del mismo Apache Flink por lo que tiene una integración muy fuerte con este motor de procesamiento. Paimon se destaca por su capacidad de manejar datos en tiempo real y batch, permitiendo la construcción de pipelines híbridos. Lo hace mediante el uso de su formato de datos propio en conjunto con un Log-structured Merge-Tree (LSM) para la gestión de metadatos; similar a Apache Hudi.

Permite actualizaciones en tiempo real a través del motor de procesamiento y deduplicar los datos de forma personalizada. También se encarga automáticamente de la limpieza y compactación de archivos, y ofrece soporte para la evolución de esquemas y control de versiones.

Por último, también permite almacenar metadata que otorga la capacidad de realizar transacciones ACID, evolución de esquema y control de versiones.

Catálogos de Metadatos

Los catálogos de metadatos actúan como un registro centralizado y organizado de toda la información sobre las tablas y conjuntos de datos en un sistema. Son una capa de abstracción que mantiene información crítica sobre la estructura, ubicación, esquema, particiones, historial de versiones y estadísticas de los datos, permitiendo una gestión eficiente y un acceso optimizado a los mismos.

Son necesarios para que sistemas externos conozcan la ubicación y la estructura de los datos almacenados.

Hive Metastore es la implementación más adoptada y el estándar de estos catálogos. Opera como un servicio centralizado que almacena la información sobre las estructuras de datos, típicamente usando una base de datos relacional como base (por ejemplo PostgreSQL o MySQL). Sin embargo, puede presentar problemas con operaciones concurrentes complejas. Sin embargo, se pueden implementar capas de abstracción que permitan manejar estos escenarios, y generalmente se prefiere utilizar este sistema ya que está ampliamente probado.

Project Nessie es otra implementación de catálogo de datos, aunque con un enfoque moderno, aplicando conceptos de control de versiones a los datos; similar a como lo hace git. No intenta mantener compatibilidad con Hive Metastore, sino que ofrece un nuevo paradigma de gestión de datos, por lo que puede implementar características más complejas. Sin embargo, esto también significa que requiere más esfuerzo en integración con herramientas existentes ya que su ecosistema aún no está tan desarrollado.

Apache Polaris es un proyecto relativamente nuevo que busca estandarizar y modernizar la gestión de metadatos. Es un esfuerzo para estandarizar y unificar los catálogos de metadatos. De esta manera, Apache Polaris busca ser una solución neutral y agnóstica a los distintos proveedores de catálogos.

Bases de Datos Analíticas

Las bases de datos analíticas, también conocidas como bases de datos OLAP (Online Analytical Processing) orientadas a tiempo real, son sistemas especializados diseñados para procesar y analizar grandes volúmenes de datos con énfasis particular en consultas complejas y agregaciones. Estos sistemas están arquitecturalmente optimizados para procesar rápidamente consultas que involucran múltiples dimensiones y métricas sobre conjuntos masivos de datos, permitiendo análisis en tiempo real o casi real.

Estas bases de datos se distinguen por su capacidad de manejar cargas de trabajo analíticas complejas mientras mantienen latencias bajas y consistentes; especializándose en resultados en segundos o milisegundos. Esta característica se debe a su arquitectura orientada específicamente al análisis, que contrasta con los sistemas diseñados primariamente para transacciones (OLTP) o almacenamiento general de datos.

Por diseño permiten un análisis multidimensional eficiente, soportan alta concurrencia de usuarios, y pueden integrarse efectivamente con fuentes de datos en streaming. Además, su diseño orientado a columnas permite una compresión más eficiente y mejor rendimiento en consultas analíticas que típicamente involucran solo un subconjunto de ellas. Proporcionando capacidades avanzadas de agregación y pueden manejar eficientemente tanto datos históricos como en tiempo real.

Ninguno de estos beneficios vienen sin sus propios desafíos, ya que se requiere una cuidadosa planificación, operación y mantenimiento para mantener su rendimiento. Además, más que nunca, es necesario organizar correctamente los índices y los esquemas de particionamiento son claves.

Ejemplos de estos sistemas son:

Apache Druid es una base de datos analítica distribuida diseñada principalmente para análisis en tiempo real de grandes volúmenes de datos de series temporales. Su arquitectura se distingue por su capacidad de ingesta en tiempo real combinada con consultas de baja latencia, utilizando un modelo de almacenamiento columnar híbrido que combina datos en memoria con almacenamiento en disco.

Apache Pinot fué desarrollado inicialmente por LinkedIn y se enfoca en proporcionar análisis en tiempo real con latencias extremadamente bajas, incluso en escenarios de alta concurrencia de usuarios. Su arquitectura está optimizada para consultas de lectura masivas y paralelas. Además, destaca por su modelo de consistencia eventual y su capacidad para manejar esquemas dinámicos.

Apache Doris inicialmente fue desarrollado por Baidu, integra capacidades de almacenamiento columnar MPP (Procesamiento Paralelo Masivo) con funcionalidades OLAP, ofreciendo una solución más cercana a una base de datos tradicional pero con capacidades analíticas avanzadas. Su arquitectura es más simple en comparación con Druid y Pinot, lo que facilita la operación y mantenimiento, manteniendo un buen rendimiento para consultas analíticas. Permite realizar consultas federadas sin necesidad de ingestar información duplicada.

2.2.6. Comparación

Observaciones clave

- Apache Parquet es el formato más utilizado para análisis de datos
- El almacenamiento de objetos especialmente en la nube es extremadamente barato y su uso permite extender los casos de uso de almacenamiento
- De entre los formatos de tabla analítica, Apache Iceberg es el que se encuentra en mejor posición de adopción debido a su naturaleza abierta
- Hive Metastore sigue siendo el estándar de facto de la industria debido a su amplia adopción y madurez
- Apache Doris permite reducir la duplicación de datos al tener capacidades tanto de almacenamiento en tiempo real como de consultas federadas

2.2.7. Tecnologías Complementarias

Docker

Docker es una plataforma de software que permite crear, desplegar y ejecutar aplicaciones en contenedores. Los contenedores son entornos ligeros y portátiles que encapsulan una aplicación y todas sus dependencias, lo que garantiza que funcionen de manera consistente en diferentes entornos. A nivel de desarrollo, también permite crear imágenes de contenedores que pueden ser compartidas y reutilizadas, facilitando la colaboración entre equipos y la integración continua. Es especialmente sencillo de usar y provee Docker Compose, una herramienta que permite definir y ejecutar aplicaciones multi-contenedor.

Prometheus

Prometheus es un sistema de monitorización y alerta open-source. Es especialmente eficaz para monitorizar entornos dinámicos como los que se encuentran en arquitecturas de microservicios y contenedores. Permite recolectar y analizar métricas en tiempo real de manera altamente eficiente y tiene un potente lenguaje de consultas. Destaca por su capacidad para manejar millones de métricas simultáneamente y su arquitectura pull-based, que le permite escalar horizontalmente sin problemas. Además, cuenta con un sistema de alertas flexible y puede integrarse fácilmente con Kubernetes y otras herramientas de orquestación, lo que lo convierte en una herramienta fundamental para monitorizar aplicaciones modernas a gran escala.

Grafana

Grafana es una plataforma de visualización y análisis de datos de open-source que destaca especialmente cuando se combina con Prometheus. Su principal valor radica en su capacidad para transformar datos complejos en visualizaciones claras e interactivas a través de tableros personalizables. Cuando se utiliza junto con Prometheus, Grafana actúa como la capa de visualización, permitiendo crear paneles intuitivos que muestran en tiempo real el estado y rendimiento de los sistemas. Esto facilita la detección temprana de problemas, el análisis de tendencias y la toma de decisiones basada en datos. Una de sus características más poderosas es la capacidad de combinar datos de múltiples fuentes en un único dashboard, proporcionando una vista unificada del sistema.

2.2.8. Desafíos

Las tecnologías que se han mencionado son parte de los bloques que componen un sistema de Big Data para Streaming. Sin embargo, solo con el hecho de usarlas no alcanza para construir estos sistemas; debido a que existen desafíos que si bien las tecnologías ayudan a mitigar, deben tenerse en cuenta para cumplir con éxito el propósito que tienen dichos sistemas.

A continuación se describirán brevemente los desafíos más importantes que deben tenerse en cuenta a la hora de diseñar estos sistemas.

Procesamiento fuera de Orden

Los sistemas distribuidos tienen la característica de que los eventos pueden no ser procesados en el orden en que fueron generados. Por ejemplo, podría pasar que dos eventos generados en lugares geográficamente distintos viajen por dos conexiones de red con velocidades dispares y lleguen al sistema desordenados. Más aún, los componentes de un sistema distribuido están asimismo distribuidos por lo que podría pasar que aunque lleguen en orden el procesamiento de dichos mensajes podría darse fuera de orden.

Esto provoca, fundamentalmente, una posible inconsistencia en los datos. Como se describió previamente, se puede decir que de todas formas eventualmente el sistema será consistente. El problema radica en que se necesita asegurar que el sistema es tan consistente como sea posible, tan seguido como sea posible. Para esto, lo que se suele hacer es tener una ventana de espera para los mensajes de forma de recibirlos si llegan un poco más tarde que otros mensajes o por otro lado complejizar los algoritmos de procesamiento utilizados para soportar procesamiento fuera de orden. En cualquier caso ambas soluciones llevan a un aumento de la latencia, por lo que se debe llegar a un compromiso entre una propiedad del sistema y la otra.

Manejo de Archivos Pequeños

Un desafío que puede parecer contraintuitivo es el Manejo de Archivos Pequeños. En este caso, los procesadores de datos reciben muchos archivos o eventos de poco tamaño, que no permiten aprovechar adecuadamente la capacidad de procesamiento y almacenamiento del sistema. Esto se debe a que leer y escribir muchos archivos pequeños es menos eficiente que leer y escribir pocos archivos grandes. Esto puede generar que se pase más tiempo gestionando estos archivos que procesando los datos en sí mismos. Esto por su parte genera un problema de latencia.

Gestión del Estado

La gestión del estado en representa un gran desafío técnico en el procesamiento de datos. Estos sistemas deben manejar un flujo constante de datos entrantes mientras mantienen y actualizan información histórica necesaria para realizar cálculos y análisis significativos. Esta tarea se vuelve particularmente desafiante debido al volumen de datos que debe procesarse, junto con la necesidad de mantener tiempos de respuesta bajos y garantizar la precisión de los resultados.

La naturaleza distribuida de estos sistemas añade una capa adicional de complejidad ya que el estado debe distribuirse entre múltiples nodos de procesamiento, lo que introduce desafíos en términos de coordinación y consistencia. Cada nodo debe ser capaz de acceder y modificar su porción del estado de eficientemente, mientras mantiene una vista coherente del sistema en su conjunto.

Adicionalmente la tolerancia a fallos y la durabilidad del estado son aspectos críticos que no pueden ignorarse. Las limitaciones en los recursos representan otro desafío ya que la memoria disponible es finita, pero el estado puede crecer indefinidamente.

Todos estos desafíos deben abordarse pensando en mantener la disponibilidad, el rendimiento y la latencia del sistema en niveles aceptables.

Backpressure

El backpressure es un mecanismo que aborda el desafío del desequilibrio entre las tasas de producción y consumo de datos. Cuando los productores generan datos más rápidamente de lo que los consumidores pueden procesarlos, surge el riesgo de sobrecargar el sistema. El backpressure actúa como un sistema de control de flujo que permite evitar el desbordamiento de recursos.

La implementación efectiva de backpressure requiere coordinación entre los diferentes componentes del sistema de manera dinámica y adaptativo, ajustándose continuamente a las condiciones cambiantes del sistema. Sin embargo, implementar estos mecanismos de señalización de manera eficiente en sistemas distribuidos presenta desafíos significativos, especialmente cuando hay múltiples productores y consumidores involucrados.

La gestión del backpressure tiene implicaciones directas en la latencia y el rendimiento del sistema, ya que se puede generar un aumento en la latencia de procesamiento mientras el sistema se ajusta para manejar la carga. Esta compensación entre latencia y throughput debe manejarse cuidadosamente para mantener el sistema dentro de sus objetivos de nivel de servicio.

Reprocesamiento

El reprocesamiento surge cuando se necesita volver a procesar datos históricos debido a diversos factores como errores en el código, cambios en la lógica de negocio, o la necesidad de actualizar resultados anteriores. Este proceso, es complejo debido a la naturaleza continua y distribuida de los sistemas de streaming, donde el procesamiento de datos nuevos no puede detenerse mientras se realiza el reprocesamiento de datos históricos.

La gestión de recursos durante el reprocesamiento presenta desafíos particulares, ya que se debe balancear la carga entre el procesamiento de datos en tiempo real y el reprocesamiento de datos históricos. Esto es crítico para mantener el rendimiento del sistema y que el reprocesamiento termine en un tiempo razonable. La consistencia de los datos durante el reprocesamiento es otro aspecto crucial ya que se debe garantizar la coherencia con el estado actual del sistema y que no se degrade la calidad de los datos.

Conformidad Normativa

La conformidad normativa representa un desafío que va más allá de los aspectos puramente técnicos, abarcando requisitos legales, éticos y de privacidad que deben cumplirse rigurosamente. De las más importantes son HIPAA para datos de salud y GDPR para datos personales en Europa. Los sistemas deben implementar mecanismos robustos que garanticen la protección de datos sensibles mientras mantienen la funcionalidad y el rendimiento del procesamiento en tiempo real.

La gestión del consentimiento y los derechos de los usuarios presenta desafíos particulares en sistemas de tiempo real. Bajo GDPR, los usuarios tienen el derecho de acceder, modificar y eliminar sus datos personales (llamado "derecho al olvido"), lo que requiere que los sistemas puedan rastrear y gestionar estos datos a través de toda su estructura de procesamiento. En el caso de HIPAA, se requiere mantener registros detallados de quién accede a la información médica protegida y con qué propósito.

La seguridad y el cifrado de datos plantean otro conjunto de desafíos. Los datos sensibles deben estar cifrados tanto en tránsito como en reposo. El sistema debe ser capaz de cifrar y descifrar datos rápidamente sin introducir latencias significativas, mientras mantiene la integridad y la trazabilidad de los datos. Además, se deben implementar controles de acceso granulares y mecanismos de autenticación robustos que cumplan con los requisitos específicos de cada regulación, como las salvaguardas administrativas, físicas y técnicas requeridas por HIPAA.

Por su parte la retención y eliminación de datos sensibles presentan desafíos únicos. Las regulaciones pueden requerir que ciertos datos se conserven durante períodos específicos mientras otros deben eliminarse tan pronto como ya no sean necesarios. Se debe mantener un equilibrio entre cumplir con los requisitos de retención y mantener el rendimiento del sistema, especialmente cuando se trata de datos históricos que pueden ser necesarios para análisis o auditorías.

La auditoría y el monitoreo son aspectos cruciales para demostrar conformidad. Se debe mantener registros detallados de todas las operaciones de procesamiento de datos, especialmente aquellas que involucran datos sensibles o protegidos. Estos registros deben ser inmutables y seguros, permitiendo reconstruir quién accedió a qué datos, cuándo y con qué propósito. En el caso de HIPAA, esto incluye mantener registros de acceso a información médica protegida durante al menos seis años, mientras que GDPR requiere la capacidad de demostrar el cumplimiento en cualquier momento a través de registros detallados de procesamiento.

En sistemas distribuidos que procesan datos continuamente, implementar estas capacidades sin interrumpir el flujo de procesamiento o comprometer la integridad del sistema se vuelve extremadamente complejo.

2.3. Arquitecturas de Referencia

Una arquitectura de referencia representa una plantilla abstracta y probada que encapsula las decisiones arquitectónicas fundamentales de un sistema, mejores prácticas y experiencias acumuladas en un dominio específico. Esta proporciona un vocabulario común, además de componentes estandarizados y patrones de interacción que sirven como base para el desarrollo de los sistemas concretos. La arquitectura de referencia no solo define la estructura y comportamiento base del sistema, sino que también establece los principios de diseño, restricciones técnicas y mecanismos de extensibilidad que guiarán estas implementaciones.

2.3.1. Instancias de Arquitectura

Una plataforma que soporte el Monitoreo Remoto de Pacientes, necesita soportar grandes volúmenes de datos analizados como streaming. Además, para poder dar soporte al uso de la plataforma se requiere poder analizar el histórico de dichos datos. Por último, el tiempo en que el análisis de los datos de streaming es disponibilizado debe ser lo más cercano a tiempo real para poder tomar decisiones a tiempo para la salud del paciente.

Existen tres grandes familias de arquitecturas de referencia que cubren estos tres casos:

- Lambda
- Kappa
- Delta

De entre ellas, se instanciarán y se compararán Kappa y Delta; ya que son evoluciones propuestas sobre Lambda que siguen distintos caminos para alcanzar el mismo objetivo.

Realizar esta instanciación implica seleccionar las tecnologías que se usarán para cumplir las condiciones de la arquitectura de referencia; así como también componentes que si bien no son descritos por la arquitectura de referencia, son necesarios para cumplir con las características de calidad necesarias para el caso de uso propuesto.

Por último, las dos instancias de arquitectura implementadas utilizarán las mismas tecnologías, o tanto como sea posible, de modo que los resultados sean comparables.

2.4. Arquitectura Lambda

2.4.1. Descripción General

La Arquitectura Lambda es un paradigma de procesamiento de datos diseñado para manejar grandes cantidades de información en sistemas de Big Data. Propuesta por Nathan Marz en 2011, esta arquitectura busca abordar las limitaciones de los sistemas de procesamiento por lotes (batch) y en tiempo real, combinando ambos enfoques para proporcionar una vista completa y actualizada de los datos.

2.4.2. Componentes Principales

La Arquitectura Lambda se compone de tres capas fundamentales:

Batch Layer

- Almacena el conjunto completo de datos históricos.
- Procesa periódicamente volúmenes arbitrarios de datos.
- Genera vistas pre-computadas para consultas eficientes.

Serving Layer

- Almacena las vistas pre-computadas de la capa de lotes.
- Proporciona acceso de baja latencia a los resultados.

Speed Layer

- Procesa datos en tiempo real.
- Genera vistas de estos datos.
- Mantiene los datos guardados únicamente hasta que la Batch Layer haya hecho el reprocesamiento de los datos históricos.

Vista Lógica

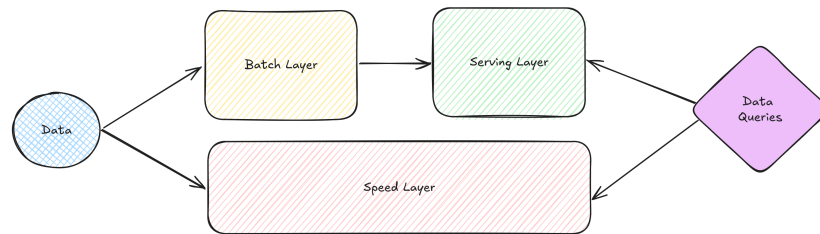


Figura 2.1: Diagrama de la Arquitectura Lambda

2.4.3. Capacidades

- **Procesamiento de datos a gran escala:** Maneja eficientemente volúmenes masivos de datos.
- **Baja latencia:** Proporciona resultados en tiempo real para consultas.
- **Tolerancia a fallos:** Mantiene la integridad de los datos incluso en caso de fallos del sistema.
- **Escalabilidad:** Se adapta fácilmente al crecimiento del volumen de datos.
- **Flexibilidad:** Permite el procesamiento tanto por lotes como en tiempo real.
- **Consistencia eventual:** Garantiza que los datos eventualmente reflejarán todos los cambios.
- **Reprocesamiento:** En caso de necesitar reprocesar los datos, este proceso es trivial, pues se tiene almacenado el histórico completo.

2.4.4. Desafíos

- **Complejidad:** La implementación y mantenimiento pueden ser complejos debido a la duplicación de lógica en las capas de lotes y velocidad.
- **Latencia:** El procesamiento batch genera latencia debido al tiempo de la actualización de vistas.
- **Costo:** Al utilizar recursos computacionales diferentes entre el procesamiento batch y en stream, esto puede requerir varios nodos computacionales, lo que incrementa los costos.

2.5. Arquitectura Kappa

2.5.1. Descripción General

La Arquitectura Kappa surge en 2014 como respuesta de parte de Jay Kreps a la Arquitectura Lambda. Si bien Lambda puede describirse de forma "sencilla" como una serie de transformaciones y además pone mucho énfasis en la posibilidad y facilidad de reprocesar los datos; tiene la desventaja de obligar a mantener código que debe producir el mismo resultado en dos sistemas distribuidos. Esto implica que cualquier cambio o mejora que reciba uno debe recibir un tratamiento de reingeniería para que el otro también lo tenga. Y, según argumenta Jay Kreps, la tendencia es a optimizar el código para uno de los motores (incluso si un mismo motor soporta dos modos de trabajo, la semántica que maneja será distinta por lo que terminará siendo una base de código distinta). (Kreps, 2014)

La Arquitectura Kappa busca responder la pregunta "Por qué un sistema de procesamiento de Streams no podría incrementar su paralelismo y reprocesar su historia muy rápido?"

La intuición detrás de esta arquitectura es la siguiente:

- Usar Kafka o algún otro sistema que permita tener la traza completa de los datos que se quieren reproducir para múltiples suscriptores
- Cuando se quiera hacer reprocesamiento, iniciar una segunda instancia de procesamiento que comience del principio de la historia
- Redirigir el procesamiento de la nueva historia a una tabla auxiliar
- Cuando el segundo procesamiento haya alcanzado al anterior, hacer que empiece a usarse la tabla auxiliar en lugar de la anterior
- Parar el procesamiento anterior y eliminar la tabla anterior

De todas maneras, el resultado del procesamiento o los estados intermedios pueden llegar a ser guardados a su vez por alguna otra herramienta para realizar procesamiento en batch.

2.5.2. Componentes Principales

Stream Store Layer

- Actúa como un registro inmutable de todos los eventos de datos entrantes.
- Permite la reproducción de datos históricos para reprocesamiento cuando se actualiza la lógica de procesamiento.

Stream Processing Layer

- Ingiere datos en tiempo real desde diversas fuentes.
- Procesa estos datos utilizando un sistema de procesamiento de streams.
- Aplica la lógica de negocio y las transformaciones necesarias a los datos entrantes.

Serving Layer

- Almacena los resultados procesados del stream.
- Proporciona acceso de baja latencia a los resultados.

2.5.3. Vista Lógica

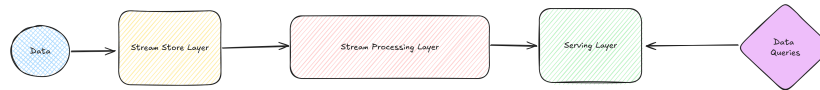


Figura 2.2: Diagrama de la Arquitectura Kappa

2.5.4. Capacidades

La Arquitectura Kappa ofrece varias capacidades clave:

- Reduce la complejidad del sistema al unificar el procesamiento batch y streaming
- Mejora la mantenibilidad al no tener que duplicar lógica para los distintos esquemas de procesamiento
- Garantiza la coherencia entre los resultados del streaming y el reprocesamiento.

2.5.5. Desafíos

A pesar de sus ventajas, la Arquitectura Kappa presenta algunos desafíos:

- Requiere un incremento en el uso de recursos muy grande cuando es necesario reprocesar los datos históricos.
- El costo de retención de eventos a largo plazo son enormes y vuelven prohibitiva esta arquitectura sin cambios.
- Es necesario reprocesar toda la historia en caso de que exista la necesidad de borrado de información

2.6. Arquitectura Delta

La Arquitectura Delta surge desde Databricks, al igual que la Arquitectura Kappa, como una respuesta a los desafíos que presenta la arquitectura Lambda. A diferencia de la suposición que hace Kappa de que todo puede ser tratado como un Stream, Delta por su parte, utiliza el almacenamiento de objetos en la nube como base y agrega por encima tecnología de metadata que otorga la posibilidad de utilizar este sistema de almacenamiento como un canal de mensajes de modo que se puede hacer Streaming sobre él; además de agregar otras capacidades. Esto permite utilizar un único flujo de datos tanto para análisis en tiempo real como análisis histórico.

2.6.1. Descripción General

Delta surge de la necesidad de procesar datos masivos a bajo costo; intentando aprovechar la estructura existente en las organizaciones que utilizan almacenamiento en la nube. Los formatos de tabla analítica permiten abstraer el almacenamiento y tratarlo como si fuera una tabla, por lo que se ingestan y luego, mediante el uso de motores de procesamiento se realiza el análisis de dichos datos, que se vuelcan en otras tablas dentro de la misma infraestructura. Esto permite no tener que distinguir entre batch y streaming, ya que los formatos de tabla analítica proveen mecanismos para detectar los cambios en las tablas y transmitirlos, generando un stream interno que puede ser aprovechado para realizar un análisis incremental.

Por lo general se utiliza un patrón de diseño de datos llamado Medallion, que define tres niveles de calidad de datos:

- Bronze: Donde se almacenan los datos que llegan en crudo
- Silver: Donde se filtran, limpian y enriquecen los datos de Bronze
- Gold: Donde se analizan los datos para generar información valiosa para el negocio

2.6.2. Componentes Principales

Stream Store Layer

- Similar a su función en la Arquitectura Kappa
- Recibe eventos y los envía al Data Lakehouse Layer

Data Lakehouse Layer

- Es una capa montada sobre almacenamiento barato como los servicios de almacenamiento de objetos en la nube
- Los formatos de tabla analítica se montan sobre este almacenamiento
- Recursos de cómputo llamados Table Services pertenecen a esta capa y dan mantenimiento al almacenamiento
- Los motores de procesamiento analizan los datos en varias etapas y las vuelvan nuevamente sobre el almacenamiento
- Generalmente se opta por una sub-arquitectura en niveles, cuyo último nivel son los datos procesados disponibles para el negocio

Catalog Layer

- Provee una capa de gobernanza permitiendo acceso granular a los datos, auditoría y políticas de retención
- Permite interoperar con terceros, permitiéndoles descubrir tablas en base a metadatos
- Ofrece estadísticas de las tablas y herramientas para optimizar las consultas sobre ellas
- Es necesario para acceder a los datos históricos

Serving Layer

- Almacena los resultados procesados del stream por un periodo de tiempo.
- Proporciona acceso de baja latencia a los resultados del procesamiento y al histórico de datos disponibles.

2.6.3. Vista Lógica

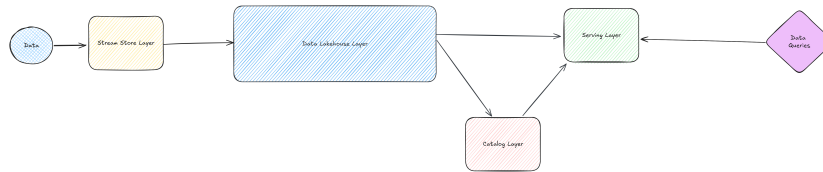


Figura 2.3: Diagrama de la Arquitectura Delta

2.6.4. Capacidades

- Garantiza características ACID para un sistema distribuido
- Reduce los costos de almacenamiento y procesamiento de la información
- Define una fuente de verdad única que puede ser usada por todos los procesos de análisis
- Reduce la cantidad de código que se debe mantener
- Permite agregar nuevas fuentes de datos sin necesidad de cambios en los procesos de análisis

2.6.5. Desafíos

- La latencia es un problema si se necesitan capacidades de análisis en tiempo real
- Se requieren compromisos de latencia y rendimiento por el problema de "Manejo de archivos pequeños"
- Si el Catalog Layer no se construye correctamente no es posible que la arquitectura escale

2.7. Monitoreo Remoto de Pacientes

Los Sistemas de Monitorización Remota de Pacientes (RPM, por sus siglas en inglés) constituyen un paradigma tecnológico de salud en el área de la Telemedicina que permite la adquisición, transmisión y análisis, idealmente en tiempo real, de datos fisiológicos del paciente fuera de los entornos clínicos tradicionales, mediante una red de dispositivos médicos y sensores de dispositivos inteligentes.

Este enfoque contribuye a mejores resultados para los pacientes, disminuye costos para las instituciones de salud y permite dar un acceso más generalizado a los servicios médicos. Es especialmente para el seguimiento de condiciones pre-existente, población anciana y monitoreo luego de intervenciones quirúrgicas.(V et al., 2024)

2.7.1. Monitoreo de Signos Vitales

Frecuencia Respiratoria

- Número de ciclos respiratorios (inspiración/expiración) por minuto
- Valores normales adulto: 12-20 respiraciones/min

Saturación de Oxígeno

- Porcentaje de hemoglobina unida a oxígeno en sangre arterial
- Valores normales: 95-100 %

Presión Sistólica

- Presión máxima ejercida por la sangre sobre las paredes arteriales durante la sístole
- Valores normales: 90-120 mmHg
-

Frecuencia Cardíaca

- Número de contracciones cardíacas por minuto
- Valores normales adulto: 60-100 latidos/min

Temperatura

- Medida del calor corporal
- Valores normales: 36.5-37.5°C

Escala Glasgow

- Escala neurológica que evalúa nivel de consciencia
- Evalúa la apertura ocular, la respuesta verbal y la respuesta motora en distintos rangos
- Valores normales: 15
- Es difícil de automatizar

Nivel de Conciencia

- Sistema simplificado de evaluación del estado de consciencia utilizado en valoración inicial y monitoreo
- Utiliza el sistema APVU: Alerta, Respuesta a estímulos verbales, Respuesta a estímulos dolorosos, Sin respuesta
- Valores normales: 0
- Al igual que la escala Glasgow es difícil de automatizar

2.7.2. Identificación de Riesgo en Pacientes

En un entorno hospitalario, la monitorización de los signos vitales constituye un pilar fundamental en la evaluación del estado clínico de un paciente. Estos parámetros fisiológicos esenciales incluyen la presión arterial (PA), la saturación de oxígeno en sangre (SpO₂), la temperatura corporal (T), la frecuencia cardíaca (FC) y la frecuencia respiratoria (FR), los cuales son registrados sistemáticamente en intervalos de 4 a 6 horas como parte del protocolo estándar de vigilancia para detectar posibles deterioros en la condición del paciente.

En diversos establecimientos sanitarios a nivel global, el personal médico y de enfermería implementa metodologías estandarizadas de evaluación, conocidas como sistemas de alerta temprana (SAT). Estos sistemas utilizan algoritmos validados que asignan puntuaciones específicas a las desviaciones de los rangos normales de los signos vitales, permitiendo la activación de alertas cuando se detectan patrones que indican un deterioro clínico.

Esta práctica sistemática facilita la identificación a tiempo de pacientes en riesgo y permite la intervención terapéutica a tiempo, contribuyendo significativamente a la reducción de eventos adversos y a la optimización de los resultados clínicos.

Existen diferentes estándares para la detección, muchos dependientes del contexto de la unidad donde se atiende al paciente. (Arora et al., 2024)

MEWS

MEWS (Modified Early Warning Score) es un sistema de puntuación fisiológica validado para la detección temprana del deterioro clínico en pacientes hospitalizados, que evalúa cinco parámetros vitales fundamentales: frecuencia respiratoria, frecuencia cardíaca, presión arterial sistólica, temperatura y nivel de consciencia.

Cada parámetro recibe una puntuación de 0 a 3 según la gravedad de su alteración, siendo 0 el valor normal y 3 el más patológico; La suma total de estos valores genera una puntuación que oscila entre 0 y 14, categorizando el riesgo del paciente en bajo (0–1), medio (2–3), alto (4–5) o crítico (≥ 6), lo que determina la frecuencia de monitorización necesaria y las intervenciones requeridas, desde una vigilancia rutinaria cada 8-12 horas en puntuaciones bajas hasta la activación inmediata del equipo de respuesta rápida y posible traslado a UCI en puntuaciones críticas.

NEWS2

El NEWS2 (National Early Warning Score 2) es una versión mejorada y actualizada del sistema de alerta temprana, adoptado como estándar por el Servicio Nacional de Salud del Reino Unido, que evalúa siete parámetros fisiológicos: frecuencia respiratoria (3-0 puntos), saturación de oxígeno (con dos escalas distintas según el riesgo de insuficiencia respiratoria hipercápnica, 3-0 puntos), uso de oxígeno suplementario (2 puntos si requiere), temperatura (3-0 puntos), presión arterial sistólica (3-0 puntos), frecuencia cardíaca (3-0 puntos) y nivel de consciencia utilizando la escala ACVPU; la puntuación total varía de 0 a 20, estratificando el riesgo en bajo (0-4), medio (5-6), alto (7 o más, o cualquier parámetro individual con puntuación de 3) y determinando la respuesta clínica necesaria, desde monitorización estándar hasta evaluación urgente por equipo de cuidados críticos, representando una mejora significativa respecto al MEWS al incluir la saturación de oxígeno y la confusión como nuevo nivel de consciencia.

SOFA

El SOFA (Sequential Organ Failure Assessment Score) es un sistema de puntuación diseñado para evaluación diaria (cada 24 horas) de la disfunción/fallo multiorgánico en unidades de cuidados intensivos, evaluando seis sistemas orgánicos: respiratorio (mediante la relación $\text{PaO}_2/\text{FiO}_2$ evaluada con cada gasometría, 0-4 puntos), cardiovascular (mediante presión arterial media y requerimiento de vasopresores monitorizados continuamente, 0-4 puntos), hepático (mediante bilirrubina sérica medida diariamente, 0-4 puntos), coagulación (mediante recuento plaquetario diario, 0-4 puntos), renal (mediante creatinina sérica diaria o gasto urinario horario, 0-4 puntos) y neurológico (mediante la escala de Glasgow evaluada cada 4 horas o con cambios clínicos, 0-4 puntos); cada sistema recibe una puntuación de 0 (normal) a 4 (máxima disfunción), con una puntuación total que varía de 0 a 24 puntos, calculándose cada 24 horas o antes si hay deterioro clínico significativo, siendo especialmente relevante el cambio en la puntuación a lo largo del tiempo.

qSOFA

El qSOFA (quick Sequential Organ Failure Assessment) es una versión simplificada del SOFA, diseñada para la identificación rápida de pacientes con sospecha de sepsis y alto riesgo de mortalidad fuera de la UCI, evaluando únicamente tres parámetros clínicos que se pueden medir de manera inmediata a pie de cama, sin necesidad de pruebas de laboratorio: alteración del estado mental (escala de Glasgow ≤ 13 puntos, 1 punto), frecuencia

respiratoria elevada (≥ 22 respiraciones/minuto, 1 punto) y presión arterial sistólica baja (≤ 100 mmHg, 1 punto); la puntuación total varía de 0 a 3 puntos, donde una puntuación ≥ 2 indica alto riesgo de mortalidad y la necesidad de evaluación más exhaustiva, monitorización estrecha y consideración de traslado a un nivel superior de cuidados; el qSOFA debe reevaluarse con cada valoración del paciente o ante cualquier cambio en su estado clínico, típicamente cada 1-2 horas en pacientes inestables o con sospecha de sepsis, siendo una herramienta especialmente útil en servicios de urgencias, plantas de hospitalización y entornos extrahospitalarios.

2.7.3. Desafíos

Los sensores empleados en la monitorización remota permiten un monitoreo continuo de los parámetros fisiológicos correspondientes, proporcionando una frecuencia de recolección de datos significativamente superior a los intervalos tradicionales establecidos en entornos convencional. Sin embargo, la implementación de estos sistemas de monitorización remota presenta diversos desafíos técnicos y operativos que requieren consideración.

Entre las principales se encuentran:

- Movilidad del usuario que puede afectar la calidad de la señal
- Uso correcto del dispositivo
- Variabilidad en las condiciones ambientales
- Fallos intermitentes de los sensores
- Agotamiento de la batería de los dispositivos
- Pérdida de conectividad en la transmisión de datos
- Falta de datos por dificultad de medición como en el caso de la escala Glasgow o niveles de conciencia

Esto genera la necesidad de procesar los datos adecuadamente y desarrollar técnicas para la limpieza y validación de datos. A su vez, se deben implementar estrategias para el manejo de datos faltantes e incompletos. (Arora et al., 2024)

Capítulo 3

Metodología

3.1. Criterios de Evaluación

Para evaluar y comparar las arquitecturas Kappa y Delta en el contexto del monitoreo remoto de pacientes, se considerarán creiterios unificados y medibles. Estos servirán como base para una evaluación objetiva de las arquitecturas Kappa y Delta en el contexto del monitoreo remoto de pacientes, permitiendo tomar criterios fundamentados para la elección de una u otra según el escenario.

3.1.1. Latencia y Rendimiento

Se implementarán mediciones a través de puntos de instrumentación estratégicos a lo largo de los componentes de la arquitectura. Estos puntos de medición incluirían timestamps en los mensajes, métricas de procesamiento en los componentes intermedios, y el tiempo de escritura/lectura en la capa final.

Las métricas serán tomadas utilizando Prometheus y mostradas en un tablero de Grafana.

Las métricas a considerar serán:

- Tiempo de ingesta de histórico
- Latencia de procesamiento

3.1.2. Manejo de Datos Históricos

Se medirán específicamente la efectividad con la que el sistema permite consultar datos históricos y reprocesar toda la historia. Las métricas serán tomadas utilizando Prometheus y mostradas en un tablero de Grafana.

- Tiempo de reprocesamiento de la historia completa
- Uso de recursos en reprocesamiento de la historia completa

Se omitirá la implementación de un cambio en el modelo de procesamiento que implique reprocesar la historia completa, pero se dará una propuesta de como hacerlo y se evaluará las implicancias de su puesta en producción manteniendo los dos sistemas funcionando e integrandolos eventualmente.

3.1.3. Costos Operativos

La implementación del sistema en contenedores, permitirá un monitoreo del uso de los recursos del sistema. Además, se plantearán cálculos, utilizando las calculadoras de costo que proveen los sistemas de nube Azure y AWS, que permitan dimensionar el costo operativo de estos sistemas. Las métricas serán tomadas utilizando Prometheus y mostradas en un tablero de Grafana.

Las métricas a considerar serán:

- Uso de memoria RAM
- Uso de CPU
- Uso de disco
- Uso de red

3.2. Stack de Tecnologías a Utilizar

Para que la implementación de las instancias de arquitecturas de referencia sean comparables es necesario mantener criterios similares para no favorecer a una de ellas. Estas instancias serán definidas mediante contenedores y desplegadas en Kubernetes de forma local utilizando la herramienta Minikube; lo que permitirá definir un ambiente Cloud-Native para la comparación de las mismas.

Se realizó un análisis exhaustivo de las tecnologías, de modo de utilizar aquellas que puedan ser compartidas entre ambas implementaciones y a su vez permitan optimizar su uso. Las seleccionadas son:

- Apache Kafka como punto de ingestión de datos y ruteo de mensajes
 - Estandar de la industria
 - Ecosistema maduro con fuertes integraciones
- Apache Flink como motor de procesamiento distribuido
 - Su modelo de procesamiento ofrece la menor latencia posible
 - Tiene amplia compatibilidad con los diferentes ecosistemas de tecnologías
- MinIO como almacenamiento de objetos
 - Compatible con API S3
 - Alta disponibilidad
 - Gestión de datos distribuida
- Apache Paimon como formato de tabla analítica
 - Formato de tabla analítica especialmente pensado para streaming de datos
 - Fuerte integración con Apache Flink
- Apache Doris como base de datos de análisis
 - Consultas con baja latencia
 - Tablas en tiempo real
 - Integración con Apache Paimon mediante un catálogo
 - Tablas híbridas

- Grafana como tablero que consume de los sistemas implementados
 - Visualizaciones interactivas
 - Múltiples fuentes de datos
- Prometheus para monitoreo de métricas de los sistemas
 - Estandar de la industria
 - Muy buena integración con diferentes ecosistemas
- Grafana para creación de tableros de métricas comparativas
 - Estandar de la industria
 - Tableros personalizables
- Docker Compose para el despliegue de los componentes del sistema
 - Orquestación de contenedores

3.3. Conjunto de Datos

3.3.1. Conjunto de Datos

El conjunto de datos utilizado para evaluar el sistema será de datos sintéticos. Se desarrollará un script en Python que genere datos sintéticos para los sensores de signos vitales. Estos datos serán generados de acuerdo a la naturaleza de la medición de cada sensor y se interconectarán para simular un flujo de datos continuo. El script generará datos para tres tipos de pacientes: Sanos, Sanos pero que se deterioran con el tiempo, y enfermos pero estables. De esta manera, se podrá evaluar el sistema en diferentes escenarios.

Existen dos razones principales para el uso de datos sintéticos:

- Los conjuntos de datos de sensores disponibles son heterogéneos y muy pequeños; lo que de todos modos implicaría generar datos sintéticos.
- No es necesario contar con datos que muestren tendencias reales, sino que se comporten como lo harían en realidad para demostrar los atributos de calidad del sistema.

Para esto, se definirá un flujo de datos para cada sensor que será interconectado pero definiendo comportamientos específicos que tengan que ver con la naturaleza de su medición. Los signos vitales tomados para esto son:

- Frecuencia respiratoria
- Presión sistólica
- Frecuencia cardíaca
- Temperatura
- Saturación de oxígeno
- Nivel de conciencia en escala APVU

Capítulo 4

Desarrollo

4.1. Sistema de Monitoreo Remoto de Pacientes

4.1.1. Introducción

Antecedentes

El Monitoreo Remoto de Pacientes ha emergido como una tecnología para mejorar la atención médica moderna, permitiendo la vigilancia continua del paciente fuera de los entornos clínicos tradicionales. Sin embargo, estos sistemas enfrentan desafíos en el mantenimiento de la calidad consistente de datos y la integración de mediciones de múltiples dispositivos con diferentes niveles de confiabilidad y tasas de muestreo.

Planteamiento del Problema

Los sistemas tradicionales de monitoreo de signos vitales frecuentemente enfrentan dificultades con:

- Temporización inconsistente de mediciones entre diferentes signos vitales
- Variación en la confiabilidad y precisión de los dispositivos
- Integración de múltiples fuentes de datos para el mismo signo vital
- Mantenimiento de la validez clínica con datos incompletos

Solución Propuesta

Se propone abordar estos desafíos mediante:

- Fusión de datos multi-dispositivo
- Puntaje de calidad del dato
- Puntaje de frescura del dato
- Capacidad de degradación gradual

4.1.2. Fundamento Clínico: El Sistema de Puntuación NEWS2

Descripción General de NEWS2

El National Early Warning Score 2 (NEWS2) es una herramienta de evaluación estandarizada utilizada para detectar el deterioro clínico. Evalúa seis parámetros fisiológicos:

- Frecuencia respiratoria
- Saturación de oxígeno
- Presión arterial sistólica
- Frecuencia cardíaca
- Nivel de consciencia
- Temperatura

Desafíos de Implementación Tradicional

NEWS2 fue originalmente diseñado para mediciones manuales periódicas en entornos clínicos. Su adaptación para monitoreo remoto continuo presenta varios desafíos:

- Diferentes frecuencias de medición para diferentes parámetros
- Calidad y confiabilidad variable de las mediciones
- Necesidad de actualizaciones de puntuación en tiempo real
- Manejo de datos faltantes o degradados

4.1.3. gdNEWS2

Concepto y Fundamentos

Se propone aumentar el sistema NEWS2 con el concepto de degradación gradual (graceful degradation) de modo que la puntuación aún sea útil incluso cuando la medición de alguno de los datos de signos vitales no se encuentre presente o no se confíe del todo en la calidad del mismo. Cada parámetro de NEWS2 tiene un puntaje de calidad y de frescura asociado, que se utilizarán para definir el nivel de confianza que se le tiene a dicho parámetro. De esta forma, se puede calcular una puntuación de alerta temprana incluso cuando no se tienen todos los datos disponibles y brindarle transparencia al profesional de la salud para que tome las decisiones correspondientes en cuanto al tratamiento del paciente.

4.1.4. Formato de Datos de Mediciones Brutas

Los datos serán recibidos en formato JSON, con un esquema común para todos los tipos de mediciones. El esquema general es el siguiente:

```
1 {  
2   "measurement_type": "RESPIRATORY_RATE" | "HEART_RATE" |  
   "OXYGEN_SATURATION" | "BLOOD_PRESSURE_SYSTOLIC" | "  
   TEMPERATURE" | "CONSCIOUSNESS",  
3   "measurement_timestamp": "datetime",  
4   "device_id": "string",  
5   "raw_value": "number",  
6   "battery": "number",  
7   "signal_strength": "number"  
8 }
```

Listing 4.1: JSON example

4.1.5. Algoritmo de Puntuación de Calidad

Un nuevo algoritmo de puntuación de calidad debería basarse en la experiencia clínica y la evidencia científica. Para el caso de estudio presentado, se propone un algoritmo sencillo a fin de ilustrar su potencial y mantener limitado el enfoque de este trabajo.

Se tomarán los identificadores de los dispositivos, que luego serán clasificados en 3 grupos cada uno con un peso asociado:

- Dispositivos de calidad médica: 1.0
- Dispositivos de calidad premium: 0.7
- Dispositivos de calidad de consumo: 0.4

Además, se tomará en cuenta la señal de batería y la intensidad de la señal, que serán clasificados en 3 grupos cada uno con un peso asociado:

- Batería a mas del 80 %: 1.0
- Batería entre 80 % y 50 %: 0.7
- Batería entre 50 % y 20 %: 0.6
- Batería a menos de 20 %: 0.4

Por último, se tomará en cuenta la intensidad de la señal:

- Valor de señal de mas de 0.8: 1.0
- Valor de señal entre 0.8 y 0.6: 0.8
- Valor de señal entre 0.5 y 0.6: 0.6
- Valor de señal menor a 0.5: 0.4

De esta manera, el cálculo de la puntuación de calidad se realizará de la siguiente manera:

$$\text{Quality Score} = 0.7 \times \text{Device Quality} + 0.2 \times \text{Battery Quality} + 0.1 \times \text{Signal Quality} \quad (4.1)$$

En este momento, los valores tanto de los pesos como de los parametros son arbitrarios y se espera que en caso de encontrar útil este acercamiento, futuras iteraciones ajusten estos parámetros o utilicen un criterio diferente para su cálculo.

4.1.6. Algoritmo de Puntuación de Frescura

Para medir la frescura de los datos, se propone un enfoque simple que considera el tiempo transcurrido desde la última medición y el tiempo que le tomó a la medición actual llegar a ser procesada.

Tiempo desde medición hasta procesamiento:

- Menos de una hora: 1.0
- Entre una y seis horas: 0.9
- Entre seis y doce horas: 0.7
- Entre doce y veinticuatro horas: 0.5
- Entre veinticuatro y cuarenta y ocho horas: 0.3
- En cualquier otro caso: 0.2

Tiempos entre mediciones:

- Menos de cuatro horas: 1.0
- Entre cuatro y ocho horas: 0.8
- Entre ocho y doce horas: 0.6
- Entre doce y veinticuatro horas: 0.4
- Más de veinticuatro horas: 0.2

De esta manera, se propone el siguiente algoritmo de puntuación de frescura:

$$\text{Freshness Score} = 0.5 \times \text{Time Since Last Measurement} + 0.5 \times \text{Time Since Measurement} \quad (4.2)$$

4.1.7. Algoritmo de Puntuación de Degradación

Este algoritmo se encargará de calcular la puntuación de degradación de la puntuación NEWS2 en caso de que no se tengan todos los datos disponibles. Simboliza la confianza que se le tiene a la puntuación NEWS2 en base a la calidad y frescura de los datos.

Se propone utilizar la siguiente fórmula para calcular la puntuación de degradación:

$$\text{Degradation Score} = 0.7 \times \text{Quality Score} + 0.3 \times \text{Freshness Score} \quad (4.3)$$

Se calcula este puntaje para cada uno de los parámetros de NEWS2 y se promedia para obtener la puntuación de degradación final.

4.1.8. Algoritmo de Puntuación de NEWS2

El algoritmo de puntuación NEWS2 se basa en la suma de los puntajes de cada uno de los parámetros,

$$\text{NEWS2 Score} = \sum_{i=1}^n \text{Parameter Score}_i \quad (4.4)$$

donde n es la cantidad de parámetros que se tienen disponibles. En caso de que no se tenga un parámetro disponible, se asumirá una puntuación de cero para ese parámetro en su lugar. De esta manera, se puede calcular la puntuación NEWS2 incluso cuando no se tienen todos los datos disponibles.

Por otro lado, se propone utilizar la puntuación de degradación para dar contexto sobre la puntuación NEWS2 final. Así, se puede explicar que tan confiable es la puntuación NEWS2 calculada en base a los datos disponibles.

4.2. Implementación

Se implementará la especificación definida en el capítulo anterior de modo tal que para ambas arquitecturas los detalles de implementación sean lo más similares posible. Para esto, se utilizará como lenguaje de procesamiento Flink SQL, que permite desarrollar los trabajos de procesamiento utilizando un lenguaje agnóstico a las plataformas subyacentes.

4.2.1. Pipeline de Procesamiento

El pipeline de procesamiento se encargará de recibir los datos en formato JSON, realizar el procesamiento de los mismos y devolver la puntuación NEWS2 calculada.

Esto se realizará de la siguiente manera:

- Recepción de datos en formato JSON mediante un topico de Kafka
- Enriquecimiento de los datos con los puntajes de calidad y frescura
- Enrutamiento de los datos para su procesamiento particular según el signo vital
- Cálculo de la puntuación NEWS2 para cada una de las Componentes
- Unión y agrupación según una ventana de tiempo
- Cálculo de valores de agregación de los puntajes de NEWS2 y de degradación

Debido a una limitante en el hardware de procesamiento, se simplificó el cálculo de frescura de los datos para no tener en cuenta los anteriores. Esto provocaba que el sistema se quedara sin memoria y no pudiera procesar los datos para ambas arquitecturas. Por lo que se optó por no tener en cuenta los datos anteriores y solo calcular la frescura de los datos con las medidas de tiempo propias de cada registro.

A continuación se presenta un diagrama de flujo del pipeline de procesamiento:

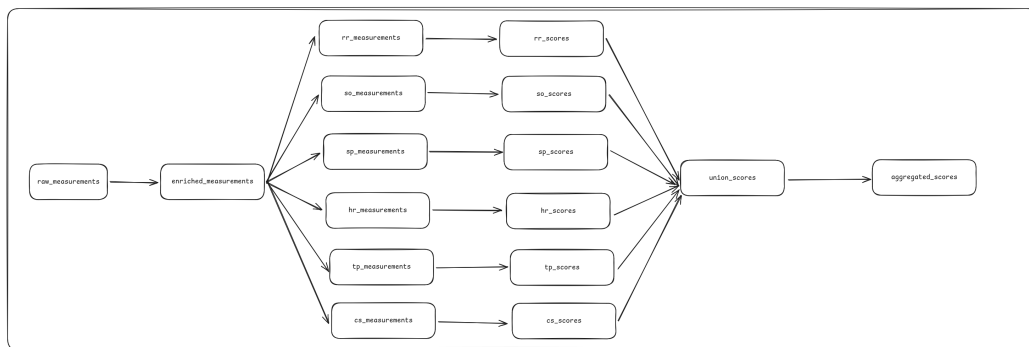


Figura 4.1: Diagrama de flujo del pipeline de procesamiento

4.2.2. Despliegue de Componentes

El despliegue de los componentes se realizó mediante el uso de **Docker Compose**, y se midió según las métricas expuestas por esta herramienta. Por otro lado, para el calculo de costos, se asume un despliegue de alta disponibilidad en la nube de AWS basado en el siguiente diagrama:

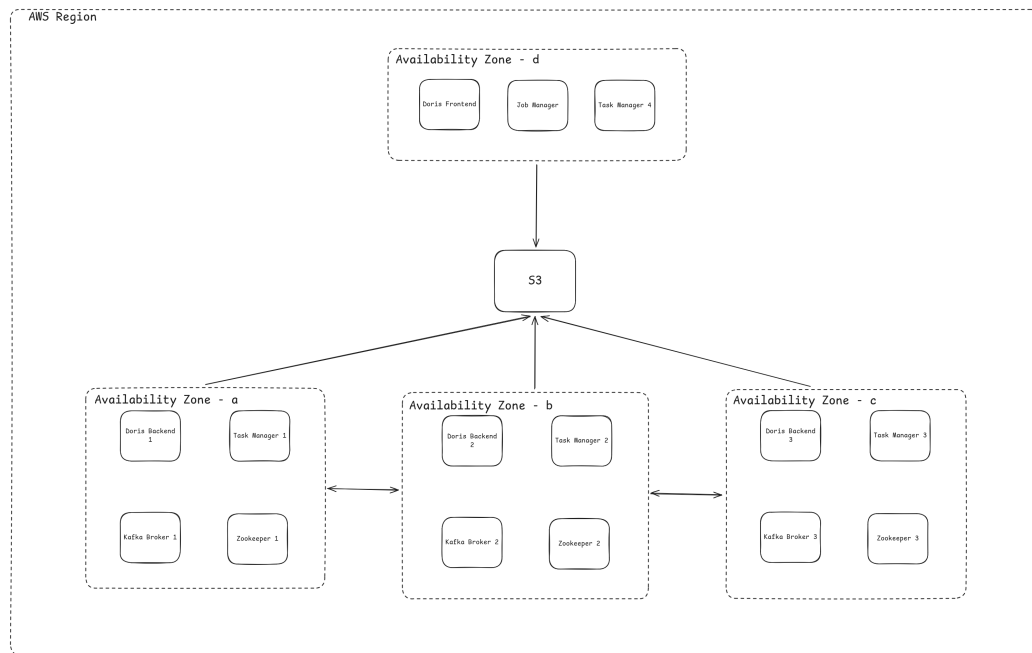


Figura 4.2: Diagrama de despliegue de componentes

La razón de este despliegue es la alta disponibilidad y la tolerancia a fallos, por lo que se despliega en una única región y se dividen los servicios en zonas de disponibilidad para asegurar que si una de ellas falla, el sistema siga funcionando lo mejor posible.

Tres de las zonas de disponibilidad (a, b y c) son idempotentes en cuanto a su funcionamiento, cada una cuenta con un nodo de Kafka, un nodo Zookeeper, un nodo de procesamiento de Flink y un nodo de backend de Doris. La cuarta zona de disponibilidad (d) tiene el nodo de frontend de Doris y el nodo de gestión de Flink; así como también un nodo extra de procesamiento de Flink. No se incluye MinIO en este despliegue porque se utiliza S3 nativo, que se define en una región y esta igualmente comunicado con todas las zonas de disponibilidad.

A su vez, estarían idealmente desplegados mediante un orquestador de contenedores como Kubernetes, utilizando algún servicio como Elastic Kubernetes Service (EKS).

4.2.3. Repositorio de Código

Se definieron tres repositorios de código para el desarrollo de la arquitectura Kappa y Delta. El primero de ellos es el repositorio de la arquitectura Kappa, que contiene el código de procesamiento de datos y la configuración de los componentes. El segundo es el repositorio de la arquitectura Delta, que contiene el código de procesamiento de datos y la configuración de los componentes. El tercero es el repositorio del generador de datos sintéticos que se utilizó para realizar las pruebas de carga y estrés.

El código de cada uno de los repositorios se encuentra disponible en el siguiente enlace:

- <https://github.com/Rekeyea/Tesis-Kappa>
- <https://github.com/Rekeyea/Tesis-Delta>
- <https://github.com/Rekeyea/Tesis-SynthDS>

4.2.4. Generación de Datos Sintéticos

Se utilizaron datos sintéticos para realizar las pruebas de carga y estrés de ambas arquitecturas. Para la generación de los datos sintéticos se utilizó un script en Python que permite definir perfiles de pacientes con diferentes características, así como también el rango de tiempo para el que se genera la información. Se generaron datos sintéticos para 32 pacientes a lo largo de un año, con un total de 110.122.654 registros; que implica un archivo CSV de aproximadamente 7GB.

El proceso de generación de datos es el siguiente:

- Se genera un archivo CSV con los datos sintéticos ordenados por signo vital.
- Se separa este archivo CSV en un archivo por paciente y se ordena cada uno.
- Se vuelve a juntar la información de cada paciente de forma ordenada en un único archivo CSV.
- Se envía cada línea del CSV a los nodos Kafka de la cada una de las arquitecturas para su ingestión

Se debieron definir estos pasos debido al tamaño del archivo CSV, ya que al ser tan grande no se puede cargar en memoria. Por otro lado, para simular demoras, se agregó una columna de delay al archivo inicial que luego fué tomada en cuenta al momento de ordenar las filas. El archivo CSV final resultante tiene el siguiente formato:

```
1 device_id,measurement_type,timestamp,raw_value,battery,signal
2 DEVICE_002_P0001,RESPIRATORY_RATE,-19.48,16.27,99.59,0.48
3 DEVICE_002_P0001,TEMPERATURE,-3.46,36.43,99.67,0.72
4 DEVICE_002_P0001,BLOOD_PRESSURE_SYSTOLIC,1.22,119.03,99.71,0.75
5 DEVICE_002_P0001,HEART_RATE,3.0,70.31,99.8,0.71
6 DEVICE_002_P0001,HEART_RATE,173.27,73.08,100.0,0.48
7 DEVICE_002_P0001,RESPIRATORY_RATE,247.35,16.36,99.32,0.71
```

El archivo de configuración usado tiene la siguiente forma:

```
1  {
2    "time_range": "1_year",
3    "patients": [
4      {"patient_id": "P0001", "category": "HEALTHY"},
5      ...
6      {"patient_id": "P0005", "category": "ILL_STABLE"},
7      ...
8      {"patient_id": "P0009", "category": "HEALTHY_DETERIORATING"},
9      ...
10   ],
11   "devices": [
12     {
13       "device_id": "MEDICAL_RR",
14       "patient_ids": [
15         ...
16       ],
17       "vitals": {
18         "RESPIRATORY_RATE": {"measurement_rate": 60}
19       },
20       "battery": {"initial": 100, "drain_rate": 0.05},
21       "signal_strength": {"base": 0.9, "variation": 0.1}
22     },
23     ...
24     {
25       "device_id": "CONSUMER_SMARTWATCH_002",
26       "patient_ids": [
27         ...
28       ],
29       "vitals": {
30         "RESPIRATORY_RATE": {"measurement_rate": 240},
31         "BLOOD_PRESSURE_SYSTOLIC": {"measurement_rate": 300},
32         "HEART_RATE": {"measurement_rate": 180},
33         "TEMPERATURE": {"measurement_rate": 600}
34       },
35       "battery": {"initial": 100, "drain_rate": 0.02},
36       "signal_strength": {"base": 0.6, "variation": 0.2}
37     }
38   ]
39 }
```

4.2.5. Limitaciones en la Implementación

Se definieron algunas limitaciones en la implementación de la arquitectura Kappa y Delta, que se detallan a continuación:

- No se implementará una solución de Governanza de datos
- No se implementará una solución de Calidad de Datos
- No se implementará una solución de Linaje de Datos
- No se implementará una solución de Gestión de Plataforma

Estas limitaciones se deben a que el objetivo de este trabajo es evaluar las arquitecturas Kappa y Delta, e incluir esas soluciones haría que el trabajo se extienda mucho más allá de lo esperado.

4.3. Arquitectura Kappa

4.3.1. Principios de Diseño

La principal característica de esta arquitectura es su fuerte uso de un registro de eventos inmutable y ordenado cronológicamente que actúa como única fuente de verdad sobre los datos ingresados al sistema.

De esta manera, se logra unificar el procesamiento de datos en batch y streaming tratándolos como un flujo continuo de eventos, eliminando la dualidad de código y reduciendo la complejidad operativa.

El procesamiento de estos datos se realiza mediante motores de procesamiento de eventos que leen este registro, aplican transformaciones determinísticas y generan resultados derivados que pueden recomputarse en cualquier momento desde el inicio del log.

Este principio de reproducibilidad permite regenerar el estado completo del sistema cuando cambian los requisitos o algoritmos de procesamiento, sin necesidad de mantener rutas de código separadas.

Las vistas materializadas son otro principio fundamental, donde los resultados procesados se almacenan en sistemas optimizados para consultas, proporcionando acceso eficiente al estado actual sin necesidad de reprocesar todo el historial de eventos.

4.3.2. Stack Tecnológico

Para la capa de ingesta y transporte de datos, la Arquitectura Kappa implementa **Apache Kafka** como componente central, funcionando no solo como sistema de mensajería sino como la fuente única de verdad y almacén principal de eventos. En Kappa se configura Kafka con períodos de retención extendidos, aprovechando la capacidad de compactación de logs para mantener el historial completo de eventos mientras se optimiza el espacio de almacenamiento. Esto se logra agregando la capacidad de almacenamiento en capas, mediante la cual se pueden mantener los eventos en Object Storage (utilizando **MinIO**), cuando pasa un tiempo definido de mantención en almacenamiento local.

El procesamiento de datos se realiza mediante **Apache Flink**, se despliega en un cluster con un nodo Job Manager y cuatro nodos Task Manager; de forma de distribuir la carga de trabajo lo mejor posible. En este caso, se define como punto de entrada un tópico de Kafka, para procesar los datos en tiempo real y enviarlos a un nuevo topico y continuar con el procesamiento más adelante en la arquitectura.

En el último paso, se guarda el resultado del procesamiento en **Apache Doris**, un motor de análisis de datos distribuido que permite realizar consultas SQL en tiempo real sobre grandes volúmenes de datos con una interfaz basada en MySQL. Este componente permite escalar de forma diferente el acceso a los datos del procesamiento, siendo desplegado como un nodo frontend y tres nodos backend. Estos comparten el trabajo de procesamiento de consultas y almacenamiento de datos, mientras que el frontend se encarga de la distribución de las mismas.

4.3.3. Vista de Componentes

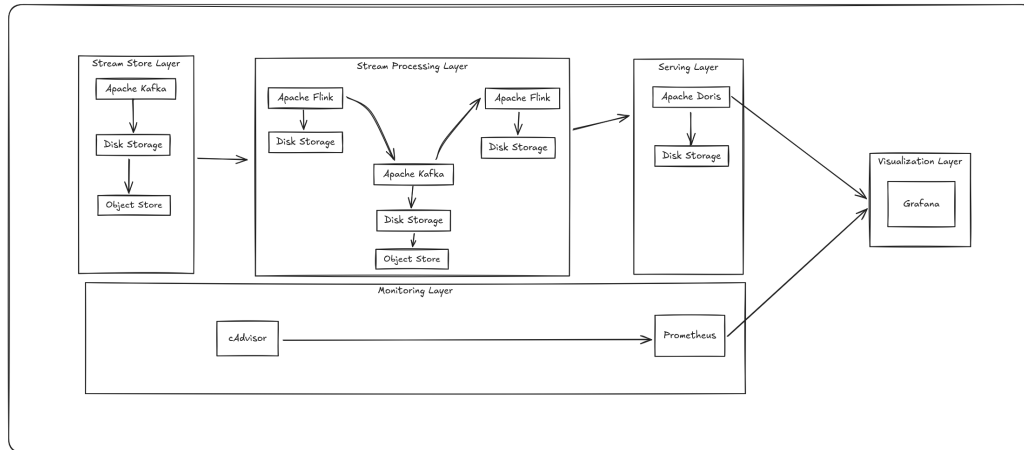


Figura 4.3: Diagrama de la Arquitectura Kappa

4.3.4. Flujo de Procesamiento

El siguiente es un ejemplo de uno de los trabajos de procesamiento de datos desarrollados:

```
1 SET 'execution.runtime-mode' = 'streaming';
2 SET 'execution.checkpointing.mode' = 'EXACTLY_ONCE';
3 SET 'table.local-time-zone' = 'UTC';
4 SET 'execution.checkpointing.interval' = '60000';
5 SET 'execution.checkpointing.timeout' = '30000';
6 SET 'state.backend' = 'hashmap';
7 SET 'table.exec.state.ttl' = '300000';
8 SET 'parallelism.default' = '4';

-- Raw measurements table with original timestamps and device metrics
9 CREATE TABLE raw_measurements (
10     measurement_timestamp TIMESTAMP(3),
11     measurement_type STRING,
12     raw_value STRING,
13     device_id STRING,
14     battery DOUBLE,
15     signal_strength DOUBLE,
16     ingestion_timestamp TIMESTAMP(3) METADATA FROM 'timestamp' VIRTUAL
17     ,
18     WATERMARK FOR measurement_timestamp AS measurement_timestamp -
        INTERVAL '10' SECONDS
19 ) WITH (
20     'topic' = 'raw.measurements',
21     'connector' = 'kafka',
22     'properties.bootstrap.servers' = 'kafka-1:19091,kafka-2:19092,
        kafka-3:19093',
23     'format' = 'json',
24     'json.timestamp-format.standard' = 'ISO-8601',
25     'scan.startup.mode' = 'latest-offset'
26 );
```



```

1 CREATE TABLE enriched_measurements (
2     measurement_type STRING,
3     'value' DOUBLE,
4     device_id STRING,
5     patient_id STRING,
6
7     -- Weights
8     quality_weight DOUBLE,
9     freshness_weight DOUBLE,
10
11     -- Timestamps
12     measurement_timestamp TIMESTAMP(3),
13     ingestion_timestamp TIMESTAMP(3),
14     enrichment_timestamp TIMESTAMP(3) METADATA FROM 'timestamp'
15     VIRTUAL,
16     WATERMARK FOR measurement_timestamp AS measurement_timestamp -
17     INTERVAL '10' SECONDS
18 ) WITH (
19     'topic' = 'enriched.measurements',
20     'connector' = 'kafka',
21     'properties.bootstrap.servers' = 'kafka-1:19091,kafka-2:19092,
22     kafka-3:19093',
23     'format' = 'json',
24     'json.timestamp-format.standard' = 'ISO-8601',
25     'scan.startup.mode' = 'latest-offset'
26 );

```

```

1  -- Insert with quality and freshness calculations
2  INSERT INTO enriched_measurements
3  SELECT
4      measurement_type,
5      CAST(raw_value AS DOUBLE) AS 'value',
6      device_id,
7      REGEXP_EXTRACT(device_id, '.*_(P\d+)$', 1) AS patient_id,
8
9      -- Quality components
10     CAST((
11         CASE
12             WHEN device_id LIKE 'MEDICAL%' THEN 1.0
13             WHEN device_id LIKE 'PREMIUM%' THEN 0.7
14             ELSE 0.4
15         END * 0.7 +
16         CASE
17             WHEN battery >= 80 THEN 1.0
18             WHEN battery >= 50 THEN 0.8
19             WHEN battery >= 20 THEN 0.6
20             ELSE 0.4
21         END * 0.2 +
22         CASE
23             WHEN signal_strength >= 0.8 THEN 1.0
24             WHEN signal_strength >= 0.6 THEN 0.8
25             WHEN signal_strength >= 0.4 THEN 0.6
26             ELSE 0.4
27         END * 0.1
28     ) AS DECIMAL(7,2)) AS quality_weight,
29
30     -- Combined freshness calculation
31     CASE
32         WHEN TIMESTAMPDIFF(HOUR, measurement_timestamp,
33             ingestion_timestamp) <= 1 THEN 1.0
34         WHEN TIMESTAMPDIFF(HOUR, measurement_timestamp,
35             ingestion_timestamp) <= 6 THEN 0.9
36         WHEN TIMESTAMPDIFF(HOUR, measurement_timestamp,
37             ingestion_timestamp) <= 12 THEN 0.7
38         WHEN TIMESTAMPDIFF(HOUR, measurement_timestamp,
39             ingestion_timestamp) <= 24 THEN 0.5
40         WHEN TIMESTAMPDIFF(HOUR, measurement_timestamp,
41             ingestion_timestamp) <= 48 THEN 0.3
42         ELSE 0.2
43     END AS freshness_weight,
44
45     -- Timestamps
46     measurement_timestamp,
47     ingestion_timestamp
48 FROM raw_measurements;

```

Como se puede ver, FLink SQL permite tratar a los tópicos de Kafka como tablas, pudiendose así leer y escribir sobre ellos. Esto permite realizar un procesamiento de datos en tiempo real, enriquecerlos y enviarlos a otro tópico de Kafka para su posterior procesamiento.

Para esta arquitectura se utilizaron dos conectores diferentes de Kafka. El primero, visto en los ejemplos, permite leer y escribir pero no modificar. Por otro lado, para las agregaciones, se utilizó **upsert-kafka** que agrega la semántica de actualización y borrado de mensajes, que es muy útil para cuando se necesita un procesamiento incremental de la información, como es el caso de las agregaciones. Aunque cabe destacar que la potencia de Flink permite que se pueda hacer esto incluso para otros destinos de datos como se verá más adelante para Paimon. Todo esto sin cambiar el código del trabajo de procesamiento.

```

1 CREATE TABLE scores (
2     patient_id STRING,
3     window_start TIMESTAMP(3),
4     window_end TIMESTAMP(3),
5
6     respiratory_rate_value DOUBLE,
7     oxygen_saturation_value DOUBLE,
8     blood_pressure_value DOUBLE,
9     heart_rate_value DOUBLE,
10    temperature_value DOUBLE,
11    consciousness_value DOUBLE,
12
13    respiratory_rate_score DOUBLE,
14    oxygen_saturation_score DOUBLE,
15    blood_pressure_score DOUBLE,
16    heart_rate_score DOUBLE,
17    temperature_score DOUBLE,
18    consciousness_score DOUBLE,
19
20    respiratory_rate_trust_score DOUBLE,
21    oxygen_saturation_trust_score DOUBLE,
22    blood_pressure_trust_score DOUBLE,
23    heart_rate_trust_score DOUBLE,
24    temperature_trust_score DOUBLE,
25    consciousness_trust_score DOUBLE,
26
27    measurement_timestamp TIMESTAMP(3),
28    ingestion_timestamp TIMESTAMP(3),
29    enrichment_timestamp TIMESTAMP(3),
30    routing_timestamp TIMESTAMP(3),
31    scoring_timestamp TIMESTAMP(3),
32    union_timestamp TIMESTAMP(3),
33    WATERMARK FOR union_timestamp AS union_timestamp - INTERVAL '10'
        SECONDS,
34    PRIMARY KEY (patient_id, window_start) NOT ENFORCED
35 ) WITH (
36     'connector' = 'upsert-kafka',
37     'topic' = 'scores',
38     'properties.bootstrap.servers' = 'kafka-1:19091,kafka-2:19092,
        kafka-3:19093',
39     'key.format' = 'json',
40     'value.format' = 'json'
41 );

```

```

1  INSERT INTO scores
2  SELECT * FROM (
3      WITH unions as (
4          ...
5      )
6      SELECT
7          patient_id,
8          window_start,
9          MAX(window_end) as window_end,
10
11         MAX(CASE WHEN measurement_type = 'RESPIRATORY_RATE' THEN '
            value' END) as respiratory_rate_value,
12         MAX(CASE WHEN measurement_type = 'OXYGEN_SATURATION' THEN '
            value' END) as oxygen_saturation_value,
13         MAX(CASE WHEN measurement_type = 'BLOOD_PRESSURE_SYSTOLIC'
            THEN 'value' END) as blood_pressure_value,
14         MAX(CASE WHEN measurement_type = 'HEART_RATE' THEN 'value' END
            ) as heart_rate_value,
15         MAX(CASE WHEN measurement_type = 'TEMPERATURE' THEN 'value'
            END) as temperature_value,
16         MAX(CASE WHEN measurement_type = 'CONSCIOUSNESS' THEN 'value'
            END) as consciousness_value,
17
18         ...
19
20         MIN(measurement_timestamp) AS measurement_timestamp,
21         MIN(ingestion_timestamp) AS ingestion_timestamp,
22         MIN(enrichment_timestamp) AS enrichment_timestamp,
23         MIN(routing_timestamp) AS routing_timestamp,
24         MIN(scoring_timestamp) AS scoring_timestamp,
25         CURRENT_TIMESTAMP as union_timestamp
26     FROM TABLE(
27         TUMBLE(
28             TABLE unions,
29             DESCRIPTOR(measurement_timestamp),
30             INTERVAL '1' MINUTES
31         )
32     ) AS unions
33     GROUP BY patient_id, window_start
34 ) as t;

```

Por último, se guarda el resultado del procesamiento en **Apache Doris** directamente desde Flink. Para esto, es necesario que la tabla en Doris haya sido creada previamente y además definir un nombre con el que llamarla en el trabajo de procesamiento. Luego, se puede insertar los datos y Flink y Doris acordarán la forma de hacerlo. Según las pruebas realizadas, esto se hace en batches. El tiempo, entre que se terminó de procesar y fue insertado en Doris no fué posible de medir ya que no se encontró una forma de definir la fecha de inserción real.

```

1 CREATE TABLE doris_gdnews2_scores (
2     patient_id STRING,
3     window_start TIMESTAMP(3),
4     window_end TIMESTAMP(3),
5
6     -- AVG Raw measurements
7     ...
8
9     -- Raw NEWS2 scores
10    ...
11    news2_score DOUBLE,
12
13    -- Trust gdNEWS2 scores
14    ...
15
16    news2_trust_score DOUBLE,
17
18    -- Timestamps
19    measurement_timestamp TIMESTAMP(3),
20    ingestion_timestamp TIMESTAMP(3),
21    enrichment_timestamp TIMESTAMP(3),
22    routing_timestamp TIMESTAMP(3),
23    scoring_timestamp TIMESTAMP(3),
24
25    flink_timestamp TIMESTAMP(3),
26    aggregation_timestamp TIMESTAMP(3),
27    PRIMARY KEY (patient_id, window_start) NOT ENFORCED
28 ) WITH (
29     'connector' = 'doris',
30     'fenodes' = '172.20.4.2:8030',
31     'table.identifier' = 'kappa.gdnews2_scores',
32     'username' = 'kappa',
33     'password' = 'kappa',
34     'sink.label-prefix' = 'doris_sink_gdnews2',
35     'sink.properties.format' = 'json',
36     'sink.properties.timezone' = 'UTC'
37 );

```

```

1 INSERT INTO doris_gdnews2_scores
2 SELECT *
3 FROM gdnews2_scores;

```

4.4. Arquitectura Delta

4.4.1. Principios de Diseño

Todo el almacenamiento de datos a largo plazo se realiza en un formato de tabla abierta, que combina archivos en formato Parquet con un registro de transacciones. Esto garantiza propiedades ACID y permite operaciones confiables en entornos distribuidos.

Al utilizar object storage, se logra una separación clara entre los recursos de procesamiento y los de almacenamiento, lo que permite escalarlos de manera independiente.

Además, múltiples clientes pueden acceder a los mismos datos de forma simultánea sin interferencias, incluso utilizando herramientas diferentes, siempre que sean compatibles con el formato subyacente.

El formato Parquet, al ser columnar, permite ejecutar consultas SQL de manera eficiente y es compatible con la mayoría de los motores de análisis modernos.

El procesamiento de datos se gestiona como un flujo continuo de eventos, donde el motor de procesamiento utiliza el log de transacciones como fuente de verdad para mantener el estado de los datos. Esto permite unificar el procesamiento batch y streaming en una misma arquitectura. Como efecto secundario de esto, todos los datos son guardados para cada etapa del flujo de procesamiento. Esto significa que están disponibles para su fácil consumo en caso de que se quieran analizar; pero como contraparte, consumen más espacio de almacenamiento ya que se almacena potencialmente varias veces el mismo dato (aunque enriquecido).

Además, el sistema se encarga automáticamente de optimizaciones como la compactación de archivos pequeños y la gestión eficiente de metadatos, asegurando un rendimiento óptimo sin intervención manual.

Por último, al estar basado en estándares abiertos, el sistema evita el vendor lock-in y permite integración con diversas herramientas de BI, machine learning y ETL.

4.4.2. Stack Tecnológico

Para la capa de ingesta y transporte de datos se implementó **Apache Kafka**, un sistema de mensajería distribuido que proporciona alta durabilidad, replicación y garantía en el orden de los eventos. Kafka actúa como el punto de entrada de la arquitectura, permitiendo desacoplar la ingesta de datos del procesamiento y asegurando una capa de buffer que absorbe picos de tráfico mientras mantiene los datos disponibles para su consumo. En este caso, Kafka se despliega en un cluster de tres nodos, con un factor de replicación de tres y un factor de partición de tres. Por otro lado, se definió un tiempo de retención de mensajes acotado, en este caso de 7 días, para evitar la acumulación de datos pero a su vez asegurar la disponibilidad de los mismos para su procesamiento.

El procesamiento de datos se realiza mediante **Apache Flink**, se despliega en un cluster con un nodo Job Manager y cinco nodos Task Manager; de forma de distribuir la carga de trabajo lo mejor posible. En este caso, se define como punto de entrada un tópico de Kafka, para luego continuar procesando los datos, no utilizando Kafka sino aprovechando las capacidades de **Apache Paimon**. Este adopta un enfoque log-structured para las escrituras, lo que lo hace especialmente eficiente para cargas de trabajo de streaming con alta frecuencia de actualizaciones.

Este permite tratar una tabla de datos como un flujo continuo de eventos, funcionando de forma efectiva como una cola de mensajes, pero con la ventaja de tener los datos materializados en un almacenamiento persistente y barato.

Por último, este formato de almacenamiento permite ser leído por **Apache Doris**, un motor de análisis de datos distribuido que permite realizar consultas SQL en tiempo real sobre grandes volúmenes de datos con una interfaz basada en MySQL.

Para esto, tanto como para el uso de Flink, se necesita definir un catálogo de tablas. Normalmente, esto podría hacerse utilizando Apache Hive, pero se optó por simplificar el sistema tanto como sea posible, y dado que no se necesitaba interactuar con un sistema existente; por lo que el catálogo se almacenó en Object Storage.

Esto último se logró mediante el uso de **MinIO**, un servidor de almacenamiento de objetos de código abierto que permite almacenar datos de forma segura y eficiente, y que además es compatible con el protocolo S3 de Amazon Web Services.

Esta combinación tecnológica permite implementar efectivamente los principios de la Arquitectura Delta, donde los datos fluyen desde las fuentes a través de Kafka, son procesados por Flink, almacenados en diferentes capas mediante Paimon sobre MinIO, y finalmente consultados a través de Doris, manteniendo en todo momento las propiedades ACID y permitiendo el procesamiento continuo así como análisis retrospectivos sobre datos históricos.

4.4.3. Vista de Componentes

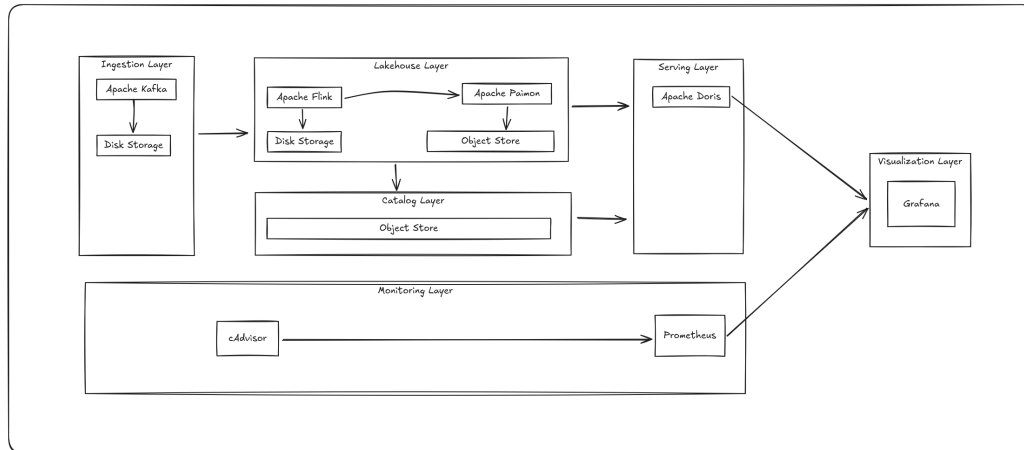


Figura 4.4: Diagrama de la Arquitectura Delta

4.4.4. Flujo de Procesamiento

El siguiente es un ejemplo de uno de los trabajos de procesamiento de datos desarrollados:

```
1  SET 'execution.runtime-mode' = 'streaming';
2  SET 'execution.checkpointing.mode' = 'EXACTLY_ONCE';
3  SET 'table.local-time-zone' = 'UTC';
4  SET 'execution.checkpointing.interval' = '60000';
5  SET 'execution.checkpointing.timeout' = '30000';
6  SET 'state.backend' = 'hashmap';
7  SET 'table.exec.state.ttl' = '300000';
8  SET 'parallelism.default' = '2';
9
10 CREATE CATALOG paimon WITH (
11     'type' = 'paimon',
12     'warehouse' = 's3://datalake/paimon',
13     's3.endpoint' = 'http://minio:9000',
14     's3.access-key' = 'minioadmin',
15     's3.secret-key' = 'minioadmin',
16     's3.path.style.access' = 'true',
17     'location-in-properties' = 'true'
18 );
19
20 CREATE TABLE paimon.delta.raw_measurements (
21     measurement_timestamp TIMESTAMP(3),
22     measurement_type STRING,
23     raw_value STRING,
24     device_id STRING,
25     battery DOUBLE,
26     signal_strength DOUBLE,
27     ingestion_timestamp TIMESTAMP(3),
28     WATERMARK FOR measurement_timestamp AS measurement_timestamp -
29         INTERVAL '10' SECONDS
30 );
```

```

1 CREATE TABLE default_catalog.default_database.raw_measurements (
2     measurement_timestamp TIMESTAMP(3),
3     measurement_type STRING,
4     raw_value STRING,
5     device_id STRING,
6     battery DOUBLE,
7     signal_strength DOUBLE,
8     ingestion_timestamp TIMESTAMP(3) METADATA FROM 'timestamp' VIRTUAL
9     ,
10    WATERMARK FOR measurement_timestamp AS measurement_timestamp -
11        INTERVAL '10' SECONDS
12 ) WITH (
13     'topic' = 'raw.measurements',
14     'connector' = 'kafka',
15     'properties.bootstrap.servers' = 'kafka-1:19091,kafka-2:19092,
16         kafka-3:19093',
17     'format' = 'json',
18     'json.timestamp-format.standard' = 'ISO-8601',
19     'scan.startup.mode' = 'latest-offset'
20 );

```

```

1 INSERT INTO paimon.delta.raw_measurements
2 SELECT * FROM default_catalog.default_database.raw_measurements;

```

Este código muestra algunas de las características principales del uso de Flink SQL en conjunto con Paimon. Las primeras tres reglas son las estándares para un procesamiento de streaming con manejo de mensajes en tiempo real y que además utilice una hora estándar para tener sincronizadas las marcas de tiempo de todos los sistemas que intervienen.

Luego se define el 'execution.checkpointing.interval' que es de los parámetros más importantes para el procesamiento de streaming, ya que define cada cuanto tiempo se guardan los estados intermedios de los datos procesados. Para el caso de Paimon particularmente, marca cada cuanto se impactan los datos procesados en el almacenamiento, por lo que define que tan pronto estarán disponibles estos para su consumo.

Esto afecta directamente a la latencia, por lo que es algo que se tiene que balancear fuertemente con los recursos del sistema para evitar retrasos en el procesamiento.

Algo importante a destacar es que si bien en este caso se define el catalogo en cada uno de los scripts, esto no debería ser necesario pues Flink permite definir mediante configuración sus catálogos disponibles. Sin embargo, no fue posible configurarlo de esta manera, por lo que se optó por definirlo en cada uno de los scripts.

Un último detalle a destacar, es que se define el paralelismo en 2 de modo tal que se pueda aprovechar al máximo las particiones del tópico de Kafka. Lo ideal sería definirlo en 3, pero esto no es posible por una limitante de hardware en cuanto a memoria en el nodo que ejecuta el trabajo de procesamiento.

Esta primer inserción de datos que se encarga de recibir los datos en formato JSON desde el tópico de Kafka y almacenarlos en el formato de Paimon es una de las dos grandes diferencias de esta arquitectura respecto a la anterior. La segunda diferencia es que en este caso no es necesaria la inserción en Doris, ya que esta es sólo una herramienta que se usa para acceder a los datos pero no los almacena. Esto hace que el flujo de procesamiento sea mas simple y liviano al no haber un tercer componente que tenga que procesar información.

Capítulo 5

Resultados

5.1. Expectativas Iniciales

Inicialmente, se esperaba que la arquitectura Delta proporcionara una mayor flexibilidad y escalabilidad en comparación con la arquitectura Kappa. Sobre todo porque a nivel publicitario, en los últimos tiempos la industria de ingeniería de datos se ha encargado de promover la arquitectura Delta, en particular los Data Lakehouse, como la solución definitiva para el procesamiento de datos en tiempo real y batch.

Por lo que a priori, se esperaba que Delta fuera más eficiente en términos de rendimiento y costos, y en simplicidad de implementación y mantenimiento.

Por otro lado, dado que Delta utiliza de forma extensiva el Object Storage, se esperaba una pérdida de rendimiento, y en particular más latencia, que su contraparte. Sin embargo, no se esperaba una diferencia demasiado grande ya que muchas veces se propone utilizar un Data Lakehouse como un sistema de almacenamiento de datos en tiempo real.

Por último, dado el apogeo de los Data Lakehouse se suponía una facilidad en el ensamblado y funcionamiento conjunto de los distintos componentes de la arquitectura, asumiendo que hubieran soluciones estandarizadas y abiertas que permitieran la integración de los distintos componentes de forma sencilla y rápida.

5.2. Comparación Técnica

Desde el punto de vista técnico se eligió realizar una comparación basándose en el común denominador tecnológico para ambas arquitecturas. Esto implica que se realizó una optimización profunda para aprovechar al máximo las capacidades de ambas. Sin embargo, se pudieron ver las características más notables en los ecosistemas tecnológicos en los que se basan sus implementaciones.

Un caso interesante es Kafka. Kappa surge del mismísimo creador de Kafka y por lo tanto, aprovecha al máximo las capacidades de baja latencia de esta tecnología. Aún años después de introducida, es la tecnología de mensajería y streaming más popular; habiendo cosechado una enorme comunidad. Esto permitió que para ambas arquitecturas se pudiera implementar un cluster altamente distribuido, con las mejores prácticas posibles. Es de destacar que inicialmente se intentó utilizar Apache Pulsar ya que es otra plataforma que cumple con los mismos casos de uso que Kafka y promete mayor escalabilidad. Sin embargo, no existe suficiente documentación para llegar rápidamente a la misma calidad con la que se llegó al usar Kafka. Otra desventaja es en su uso en conjunto con Flink ya que los conectores existentes no funcionaban correctamente. Es importante mencionar también, que la comunidad de Kafka es tremendamente activa y en el tiempo que se desarrolló esta tesis mejoraron los protocolos de comunicación entre nodos por lo que Apache Zookeeper dejó de ser necesario en las últimas versiones.

También Flink fué un caso problemático ya que si bien promete unas muy buenas prestaciones para el caso de uso de streaming, su comunidad es muchísimo más pequeña que la de la alternativa más clara (Apache Spark) lo que implicó mucho tiempo de investigación inicial y poco de profundización en la misma. Teniendo en cuenta esto, a menos que el caso de uso específico requiera un procesamiento de streaming puro y tiempos de latencia del nivel de milisegundos, se recomienda utilizar Spark ya que es más fácil de usar y tiene una comunidad mucho más activa.

Doris por su parte, es una tecnología relativamente nueva. Sumado a esto, la mayoría de la documentación está en chino, lo que dificultó mucho su implementación. Sin embargo, las capacidades que ofrece son muy interesantes y prometedoras. Parece tener una comunidad bastante activa y en crecimiento, que apuesta por la adopción y evolución de esta tecnología. Sin embargo, no se pudo encontrar una comunidad de usuarios en inglés que permita una rápida adopción y resolución de problemas. Por lo que tampoco se pudo utilizar las últimas versiones de la misma y por ende se perdió la posibilidad de poner a prueba las últimas mejoras que se han ido introduciendo.

A nivel tecnológico Apache Paimon, fué un hallazgo inesperado pero afortunado ya que es una herramienta que funciona nativamente con Flink (y de hecho nace como un subproyecto de esa comunidad). Esto permitió que si bien la comunidad es pequeña y también en su mayoría en chino, fuera fácilmente integrable para la arquitectura Delta. También su caso de uso más importante era exactamente el que se buscaba en este caso en un formato de tabla analítica. De no haberlo encontrado, se hubiera tenido que utilizar Apache Iceberg, que es la alternativa más popular y que está tomando más relevancia en el mundo de la ingeniería de datos. De todas maneras, en base a las pruebas realizadas inicialmente, Apache Hudi hubiera sido la mejor opción por su enfoque en streaming de datos. Su mayor problema es que tiene una comunidad mucho más pequeña que Iceberg (aunque muchísimo más grande que Paimon), que no es tan fácil de integrar con Flink y que por su parte también presenta complejidades en su despliegue ya que no es solamente un formato de tabla analítica sino que se proveen de forma separada servicios de Lakehouse que se deben integrar.

El catálogo de metadatos fué otro punto importante a tener en cuenta. En este caso, al no necesitar integración con otras herramientas, se optó por un catálogo en sistema de archivos. Sin embargo, normalmente sería preferible por cuestiones de integración exponer un servicio de metadatos como Apache Hive o AWS Glue. Este es otro aspecto muy emergente, sobre todo en el mundo de los Data Lakehouse, donde se están desarrollando estándares abiertos para la interoperabilidad entre distintas herramientas. Sin embargo, todas las iniciativas existentes son extremadamente nuevas y no están maduras; o no son soportadas por algún otro componente de la arquitectura. Tal fué el caso de Project Nessie que no tenía una forma de ser integrado a Apache Doris para su uso; y que de todas maneras será absorbido por Apache Polaris en un futuro.

Desde el punto de vista de su componentes técnicos, las arquitecturas extremadamente similares y no presentan ventajas significativas una sobre otra (más allá de que para utilizar Paimon en Flink es necesario incluir algunos JAR adicionales). De hecho, una vez resuelta la implementación de la arquitectura Kappa (que fué la primera en implementarse), fué trivial implementar la arquitectura Delta.

5.3. Aspectos Operativos

Se realizó una prueba de carga de ambas arquitecturas con datos sintéticos de 32 pacientes a lo largo de un año, con un total de 110.122.654 registros; que implica un archivo CSV de aproximadamente 7GB. El hardware donde se realizó la prueba cuenta con 64GB de RAM y 24 núcleos de CPU, y se utilizó Docker Compose para la ejecución de las pruebas.

Variaron no solo los tiempos de carga del total del conjunto de datos, sino también el uso de recursos.

5.3.1. Throughput

El throughput se refiere a la cantidad de datos procesados por unidad de tiempo. En este caso, se midió en base a los mismos límites en el uso de recursos, la cantidad de datos que podían ser ingestados por el sistema hasta completar la carga total.

Para **Delta** se pudieron ingestar en un promedio de 1300 registros por segundo; y la carga completa llevó 88438 segundos. Para **Kappa** se pudieron ingestar en un promedio de 792 registros por segundo; y la carga del 61 % de los datos llevó 84879 segundos.

La carga de datos en Kappa no pudo completarse completamente ya que en todas las instancias en que se realizó, el hardware usado para el despliegue de la arquitectura consumió toda la memoria de su disco duro.

Se entiende entonces que a nivel de gestión (al menos de disco), Delta es más eficiente que Kappa. A su vez, a nivel de cantidad de mensajes que se pueden consumir por segundo, Delta fué dos veces más eficiente que Kappa.

5.3.2. Latencia

La latencia se refiere al tiempo que tarda un dato en ser procesado por el sistema. En este caso, se midió en base a los mismos límites en el uso de recursos, la cantidad de tiempo que tardó un dato en pasar por todo el flujo de datos.

La latencia se midió en segundos y los resultados fueron los siguientes:

Arquitectura	Mínima	Máxima	Promedio
Kappa	-1	322	0.18
Delta	91	610	180

Rango (segundos)	Registros	Porcentaje (%)
90-120	12247607	64,1047
120-150	13210708	24,1907
150-180	7582878	47,3081
180-210	7892041	49,2370
210-240	299165	1,8664
240-270	16,24	8,1E-06
270-300	116,86	6,9E-05
330-360	95,61	9,3E-05
390-420	31,87	6,4E-05
420-450	16,24	8,1E-06
600-630	16,24	8,1E-06

Cuadro 5.2: Distribución de Latencia en Delta

Rango (segundos)	Registros	Porcentaje (%)
<1	9962080	81,48
1-10	2264616	18,52
10-60	60	0,0005
60-180	17	0,00014
180-300	17	0,00014
>300	2	1,63575E-05

Cuadro 5.3: Distribución de Latencia en Kappa

Es interesante ver los resultados. Ya que la latencia mínima de Kappa es negativa. Esto se debe a que el conector de kafka que se usa, no actualiza el campo **aggregation_timestamp** cuando se actualiza el registro porque llegan nuevos datos. Esto muestra que también muestra un problema de diseño del trabajo de inserción al calcular los datos pero no se encontró una solución al mismo. Por otro lado, se registró el tiempo de latencia de los registros que se insertaron en el sistema; sin embargo, debido al conector de Flink con Doris, este proceso se realiza en batch, por lo que nuevamente es posible que el registro de tiempo de inserción sea anterior al de procesamiento.

Por otro lado, Delta muestra una latencia de varias magnitudes superior a Kafka. Esto es esperable debido al uso extensivo de Object Storage; ya que las lecturas en disco en conjunto con la transferencia por red suelen ser mucho más lentas que las lecturas en memoria. Sin embargo, es importante destacar que la latencia máxima de Delta es de 610 segundos, lo que implica que en el peor de los casos, el sistema tardó 10 minutos en procesar un dato. Esto es un tiempo aceptable para la mayoría de los casos de uso en los que se implementan flujos de datos; aunque no se puede decir que lo sea para un sistema de procesamiento de datos en tiempo real.

5.3.3. Uso de Recursos

El uso de recursos se refiere a la cantidad de recursos que utiliza el sistema para procesar los datos. En este caso, se midió en base a los mismos límites en el uso de recursos, la cantidad máxima de recursos que utilizó el sistema para procesar los datos.

Para **Kappa** se presentan los siguientes resultados:

Componente	Uso de CPU	Memoria	Almacenamiento	Transferencia
Job Manager	134 %	1.8 GB	-	54 GB
Task Manager 1	260 %	6.2 GB	-	230 GB
Task Manager 2	244 %	6.3 GB	-	252 GB
Task Manager 3	272 %	6.2 GB	-	243 GB
Task Manager 4	256 %	6.4 GB	-	240 GB
Kafka 1	250 %	8.2 GB	234 GB	255 GB
Kafka 2	282 %	8.4 GB	234 GB	255 GB
Kafka 3	268 %	8.5 GB	234 GB	255 GB
Doris Frontend	370 %	2.4 GB	-	0.55 GB
Doris Backend 1	374 %	4.2 GB	0.234 GB	0.1 GB
Doris Backend 2	345 %	4.2 GB	0.234 GB	0.1 GB
Doris Backend 3	356 %	4.3 GB	0.234 GB	0.1 GB
MinIO	127 %	4.2 GB	384 GB	386 GB

Para **Delta** se presentan los siguientes resultados:

Componente	Uso de CPU	Memoria	Almacenamiento	Transferencia
Job Manager	250 %	3.5 GB	-	5.9 GB
Task Manager 1	208 %	9.7 GB	-	86.1 GB
Task Manager 2	166 %	7.6 GB	-	45.2 GB
Task Manager 3	194 %	8.5 GB	-	40.2 GB
Task Manager 4	206 %	7.5 GB	-	10.9 GB
Kafka 1	234 %	2.7 GB	38 GB	101.8 GB
Kafka 2	274 %	2.7 GB	38 GB	101.9 GB
Kafka 3	176 %	2.7 GB	38 GB	102.1 GB
Doris Frontend	365 %	3.6 GB	-	1.1 GB
Doris Backend 1	331 %	4.1 GB	0 GB	0.2 GB
Doris Backend 2	351 %	4.1 GB	0 GB	0.2 GB
Doris Backend 3	340 %	4.1 GB	0 GB	0.2 GB
MinIO	127 %	3.8 GB	8.4 GB	374 GB

Acorde con lo esperado, Kappa requiere un mayor uso de recursos para los nodos de Kafka que Delta. Por su lado, Delta requiere un mayor uso de recursos para los nodos de procesamiento. También para el nodo de frontend de Doris. Aunque es notablemente menor el uso de memoria que requiere de los nodos Kafka.

El uso de CPU es similar en las dos arquitecturas, aunque Delta tiene más diferencias entre nodos del mismo tipo. Probablemente sea debido a un desbalanceo en el paralelismo de los trabajos de procesamiento.

La sorpresa más grande es el uso de almacenamiento y de transferencia de red. Kappa requiere un uso de almacenamiento mucho mayor que Delta, y también un uso de transferencia de red mucho mayor. Esto se debe a que Delta guarda sus datos en Parquet, lo que lo hace muchísimo más eficiente que el formato de Kafka. Incluso, en el caso de Delta, todas las transformaciones de los datos están guardadas en Object Storage y aún así el uso de almacenamiento es notablemente menor que en Kappa.

5.3.4. Resultados

A nivel de throughput, Delta es más eficiente que Kappa. No solo eso sino que es más fácil de operar y consume menos recursos. A nivel de latencia, Kappa es más eficiente que Delta. Sin embargo, la latencia máxima de Delta es aceptable para la mayoría de los casos de uso. En cuanto a uso de memoria y procesamiento, ambas son similares. Aunque Delta es muchísimo mas eficiente en el uso de almacenamiento y transferencia de red.

5.4. Costos

5.4.1. Suposiciones

Se calcularán los costos basados en el despliegue de cada arquitectura en AWS.

La arquitectura desplegada en AWS consiste en:

- Cluster de EKS (Elastic Kubernetes Service) distribuido en 4 Availability Zones (a, b, c, d)
- AZ-a: Doris Backend 1, Task Manager 1, Kafka Broker 1, Zookeeper 1
- AZ-b: Doris Backend 2, Task Manager 2, Kafka Broker 2, Zookeeper 2
- AZ-c: Doris Backend 3, Task Manager 3, Kafka Broker 3, Zookeeper 3
- AZ-d: Doris Frontend, Job Manager, Task Manager 4
- Almacenamiento S3 nativo (sustituyendo MinIO)
- Interconexión entre zonas de disponibilidad

El costo de EKS en AWS es de \$0.10 por hora por cada nodo del cluster. Por lo que el costo mensual base es de U\$S 72.00.

Se asume además que todos los nodos experimentan una carga constante igual a la mayor carga registrada en las pruebas. Esto asegura que se mantenga la disponibilidad del sistema.

Por otro lado, Zookeeper no fué incluido en la comparación de costos operativos ya que no hace una diferencia. Sin embargo, será incluido en el costo total de ambas arquitecturas.

Se asume que el tráfico de red entre los nodos de la misma zona de disponibilidad es gratuito. Sin embargo, el tráfico entre zonas de disponibilidad tiene un costo de \$0.01 por GB. Y habiendo 4 zonas de disponibilidad que practicamente simétricas, se asume que se deistribuye el tráfico de forma equitativa entre todas ellas.

Se le dará a cada nodo una instancia de AWS que permita de forma óptima el procesamiento de los datos. Además, se asignará un volumen EBS de mínimo 50 GB a cada nodo.

5.4.2. Costos de la Arquitectura Kappa

Componente	Instancia	Base	Almacenamiento	Red	Total
Job Manager	c5.large	\$47	\$5	\$0.4	\$52.4
Task Manager	r5.xlarge	\$137	\$5	\$1.7	\$143.7
Kafka Broker	c5.x2large	\$188	\$23.4	\$5.74	\$217.14
Zookeeper	m5.large	\$69.12	\$5	\$1.9	\$76.02
Doris Frontend	c5.xlarge	\$94	\$5	\$0.01	\$99.01
Doris Backend	c5.x2large	\$188	\$5	\$0.05	\$193.05
Almacenamiento S3	-	\$0	\$38.4	\$2.9	\$41.3

El costo total de la arquitectura Kappa es de:

$$\$52.4 + 4 * \$143.7 + 3 * \$217.14 + 3 * \$76.02 + \$99.01 + 3 * \$193.05 + \$41.3 + \$72.00 = 2298.11$$

5.4.3. Costos de la Arquitectura Delta

Componente	Instancia	Base	Almacenamiento	Red	Total
Job Manager	c5.large	\$47	\$5	\$0.4	\$52.4
Task Manager	r5.xlarge	\$137	\$5	\$0.65	\$142.65
Kafka Broker	c5.xlarge	\$94	\$3.8	\$2.30	\$100,1
Zookeeper	m5.large	\$69.12	\$5	\$1.9	\$76.02
Doris Frontend	c5.xlarge	\$94	\$5	\$0.01	\$99.01
Doris Backend	c5.x2large	\$188	\$5	\$0.05	\$193.05
Almacenamiento S3	-	\$0	\$0.84	\$2.8	\$3.64

El costo total de la arquitectura Delta es de:

$$\$52.4 + 4 * \$142.65 + 3 * \$100.1 + 3 * \$76.02 + \$99.01 + 3 * \$193.05 + \$3.64 + \$72.00 = 1905.16$$

5.4.4. Resultados

Se puede ver que la arquitectura Delta es más económica que la arquitectura Kappa. Además, a medida que crecen los datos este diferencial se hace más grande. La diferencia más grande se da en el uso extensivo de Kafka por parte de Kappa para el ruteo de mensajes. Esto se traduce no sólo en un mayor costo de almacenamiento (aún usando almacenamiento en capas) sino también en un mayor costo de procesamiento.

Capítulo 6

Conclusiones

Contenido del capítulo...

Apéndice A

Primer Anexo

Contenido del anexo...

Bibliografía

- Stonebraker, M., Çetintemel, U., & Zdonik, S. (2005). The 8 requirements of real-time stream processing. *ACM SIGMOD Record*, 34(4), 42-47.
- Dean, J., & Ghemawat, S. (2008). MAPREDUCE: SIMPLIFIED DATA PROCESSING ON LARGE CLUSTERS. *Communications of the ACM*, 51(1), 107-113. <https://research.ebsco.com/linkprocessor/plink?id=1a5fefb3-a714-300d-bc4c-94603fe83a6f>
- Marz, N. (2011). *How to beat the CAP theorem* [Accessed on 2024-10-08]. Nathan Marz's Blog. Consultado el 8 de octubre de 2024, desde <http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html>
- Demchenko, Y., Laat, C. D., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. *2014 International Conference on Collaboration Technologies and Systems (CTS)*, 1(2), 104-112.
- Kreps, J. (2014). *Questioning the Lambda Architecture*. Consultado el 7 de octubre de 2024, desde <https://www.oreilly.com/radar/questioning-the-lambda-architecture/>
- Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J. M., Kulkarni, S., Jackson, J., Gade, K., Fu, M., Donham, J., et al. (2014). Storm@twitter. *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 147-156.
- Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Moctezuma, R. J., Lax, R., McVeety, S., Mills, D., Perry, F., Schmidt, E., et al. (2015). The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8(12), 1792-1803.
- Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache flink: Stream and batch processing in a single engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 36(4).
- Marz, N., & Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems* (1st). Manning Publications Co.

- Preuveneers, D., Berbers, Y., & Joosen, W. (2016). SAMURAI: A batch and streaming context architecture for large-scale intelligent applications and environments. *Journal of Ambient Intelligence, Smart Environments*, 8(1), 63-78. <https://research.ebsco.com/linkprocessor/plink?id=afac9a51-b38e-3302-b299-be23cea08349>
- Kleppmann, M. (2018). *Designing Data-Intensive Applications* (1.^a ed., Vol. 1). O'Reilly Media Inc.
- Hueske, F., & Kalavri, V. (2019). *Stream Processing with Apache Flink : Fundamentals, Implementation, and Operation of Streaming Applications* (First edition). O'Reilly Media.
- Muñoz-Escóí, F. D., Juan-Marín, R. d., García-Escrivá, J.-R., Mendívil, J. R. G. d., & Bernabéu-Aubán, J. M. (2019). CAP Theorem: Revision of Its Related Consistency Models. *Computer Journal*, 62(6), 943-960. <https://research.ebsco.com/linkprocessor/plink?id=3508200f-f883-34a0-a32c-3dea6a6209e0>
- Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A., undefinedwitakowski, M., Szafranski, M., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., ... Zaharia, M. (2020). Delta lake: high-performance ACID table storage over cloud object stores. *Proc. VLDB Endow.*, 13(12), 3411-3424. <https://doi.org/10.14778/3415478.3415560>
- Leano, H. (2020, 20 de noviembre). *Delta vs. Lambda: Why Simplicity Trumps Complexity for Data Pipelines*. Databricks. <https://www.databricks.com/blog/2020/11/20/delta-vs-lambda-why-simplicity-trumps-complexity-for-data-pipelines.html>
- Tanenbaum, A. S., & van Steen, M. (2020). *Distributed Systems: Principles and Paradigms* (Third edition). Pearson Education, Inc.
- Arora, T., Balasubramanian, V., Stranieri, A., & Menon, V. G. (2024). Modified Early Warning Score (MEWS) Visualization and Pattern Matching Imputation in Remote Patient Monitoring. *IEEE Access*, 12, 74784-74794. <https://doi.org/10.1109/ACCESS.2024.3396274>
- V, D. G., Joshi, D., C, S. k., G, M. R., D, S., & Habeeb, M. (2024). Impact of IoT on Remote Patient Monitoring and Advancements in Telemedicine. *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, 319-326. <https://doi.org/10.1109/ICoICI62503.2024.10696558>