Tipos de Procesamiento en Big Data

Daniel Marín López

Índice

1. Introducción

En la actualidad, los datos que se generan diariamente ha aumentado considerablemente, esto ha llevado a la necesidad de usar técnicas y herramientas para la gestión y análisis de los mismos. Esto hace que el campo del **Big Data** sea más importante que nunca si queremos analizar los volúmenes masivos de datos de manera rápida y eficaz.

Dentro del Big Data se encuentran distintos tipos de procesamiento de los datos que tiene sus ventajas y aplicaciones, teniendo en cuenta distintos factores como la velocidad con la que recibimos los datos, la veracidad de los mismos o su vulnerabilidad.

Dependiendo del tipo de procesamiento que usemos según las necesidades de la empresa en la que nos encontremos, hace esencial entender cuándo y cómo aplicarlos para aumentar su efectividad en futuros proyectos de Big Data.

2. Procesamiento por lotes

2.1. Definición y casos de uso

Se define el **procesamiento por lotes** como una técnica la cuál consiste en ejecutar un conjunto de tareas o procesos de manera automática y secuencial sin la ayuda de un usuario. Se suele usar para manejar grandes volúmenes de datos o para realizar tareas repetitivas que se configuran para ejecutarse en momentos concretos, como fuera del horario laboral o en períodos de baja actividad en el sistema. Los lotes son programados mediante **scripts** que especifican las acciones que deben realizar. Algunos ejemplos de programas que trabajan en este modo son: *Airflow, Delta Lake, Apache Drill, Apache Hadoop, Databricks LakehouseIQ, etc.*

Los casos de usos donde nos podemos encontrar el procesamiento por lotes son:

- **Análisis de datos**: Para hacer cálculos o generar informes detallados en grandes conjuntos de datos, como reportes financieros o estadísticas.
- **Procesos administrativos**: Se refiere a tareas como la facturación, nóminas o cálculos de intereses en bancos.
- **Gestión de inventarios**: Actualización de inventario y pedidos en sistemas de comercio.
- **Copias de seguridad y mantenimiento de sistemas**: Automatización de respaldos o actualizaciones de software.

- **Procesamiento científico**: Realizar simulaciones complejas o experimentos con grandes volúmenes de datos, como en modelado climático.

2.2. Aplicaciones

Las aplicaciones que puede tener el procesamiento por lotes son:

1. **Procesos Administrativos**
   - Las empresas de telecomunicaciones o servicios públicos manejan grandes volúmenes de datos que usan para generar facturas mensuales.
   - Recursos humanos necesitan calcular salarios, impuestos y deducciones usando este método. Lo que asegura pagos precisos y puntuales.
2. **Análisis de datos**
   - Se realizan análisis masivos de datos en las empresas para obtener reportes financieros, estudios de mercado o auditorías, como en un banco para calcular intereses acumulados en cuentas durante la noche.
3. **Gestión de inventarios**
   - En empresas de comercio minorista actualizan su inventario, procesan pedidos y realizan localizaciones de envíos mediante el procesamiento por lotes, normalmente en horas no operativas.
4. **Web scraping y extracción de datos**
   - Las empresas recopilan datos de distintos sitios web en grandes cantidades para un posterior análisis, como precios de productos o tendencias de consumo.

5. **Procesos industriales**
    - Las fábricas utilizan el procesamiento por lotes para programar tareas como los ensamblajes automatizados o los ajustes según las especificaciones de producción.

2.3. Beneficios y Desventajas

Algunos de los beneficios que se pueden obtener al usar el procesamiento por lotes son:

- **Eficiencia**: Se optimiza el uso de recursos al procesar grandes cantidades de datos a la vez.
- **Automatización**: Baja la intervención manual, disminuyen los errores humanos.
- **Flexibilidad temporal**: Se puede ejecutar en horarios no operativos, evitando interrupciones.
- **Coste reducido**: Minimiza gastos en el hardware y en el personal para las tareas repetitivas.

Por otro lado, las desventajas que supone son:

- **Complejidad inicial**: Diseñar y mantener los scripts puede requerir de personal cualificado.
- **Falta de interactividad**: No se permiten ajustes en tiempo real durante la ejecución del mismo.
- **Depuración difícil**: Localizar y arreglar los errores en los procesos por lotes puede ser complicado.

3. Procesamiento en Streaming

3.1. Definición y casos de uso

El **procesamiento en streaming** se caracteriza por procesar y analizar los datos de manera continua y casi en tiempo real a medida que se van generando. A diferencia del procesamiento por lotes, el streaming es bastante útil cuando los flujos de datos son constantes, como los sensores IoT, registros de servidores o redes sociales. Algunas herramientas que usan este tipo de procesamiento son Apache Kafka, Apache Flink, Amazon Kinesis, Azure Stream Analitycs, Apache Storm, etc.

Los casos de usos más comunes para el streaming son los siguientes:

- **Detección de fraudes**: Identificación de patrones sospechosos en transacciones financieras en tiempo real.
- **Gestión del tráfico**: Monitoreo y optimización del tráfico urbano o de las redes de datos en vivo.
- **Análisis de clientes**: Recomendaciones y promociones basadas en la actividad actual del usuario.
- **Monitoreo de sistemas**: La supervisión de servidores y aplicaciones para detectar y responder a problemas en tiempo real.
- **Aplicaciones en movilidad**: En servicios como Uber o Lyft hacen un análisis en tiempo real de la ubicación y el tráfico para poder emparejar a los conductores y los pasajeros.

3.2. Aplicaciones

Las aplicaciones que podemos relacionar con el procesamiento en streaming son:

1. **Análisis de redes sociales en tiempo real**: Las empresas pueden monitorear marcas, hashtags, nuevas tendencias, etc. para hacer ajustes a las estrategias de marketing rápidamente.
2. **Detección de fraudes financieros**: Los bancos y servicios financieros procesan en tiempo real las transacciones que se realizan para identificar patrones sospechosos y prevenir los fraudes.
3. **Monitoreo de sistemas críticos**: En campos como la sanidad se utiliza para monitorear los signos vitales de los pacientes en tiempo real para hacer alertas inmediatas ante anomalías, lo que mejora la capacidad de repuesta en situaciones de emergencia.
4. **Optimización de operaciones logísticas**: Las empresas que dan servicios a domicilio analizan los datos de localización y de tráfico en tiempo real para optimizar rutas y tiempos de entrega.
5. **Aplicaciones en IoT**: Los sensores en fábricas, vehículos autónomos y ciudades inteligentes recopilan los datos de manera continua que requieren de un análisis inmediato para ajustar los procesos, garantizar la seguridad y mejorar la eficiencia energética.

3.3. Beneficios y Desventajas

Algunos de los beneficios que podemos encontrar en el procesamiento en streaming son:

- **Desacoplamiento**: Los componentes no tienen que conocerse entre sí permitiendo incluso un intermediario que gestione las colas de mensajes.
- **Análisis en tiempo real**: El procesamiento en streaming permite que las herramientas analíticas procesen los datos en el momento de forma que los usuarios y clientes reaccionen a eventos de manera más ágil y continua en el tiempo.
- **Independencia**: Relacionado con el desacoplamiento, los equipos de desarrollo son más independientes. Por lo que no se necesita una gran coordinación para trabajar en cada uno de los extremos.

Aunque también hay limitaciones, estas son:

- **Disponibilidad**: Estas aplicaciones requieren de coherencia, baja latencia y disponibilidad. Los usuarios extraen datos nuevos de manera constante del flujo para ser procesados, si el productor de demora puede haber una ralentización en el sistema y generar errores.
- **Escalabilidad**: Las secuencias de datos no procesados pueden dispersarse rápida e inesperadamente.
- **Durabilidad**: Debido a la urgencia de los datos el sistema debe tolerar los errores, de lo contrario los datos se perderán para siempre debido a una interrupción o error.

4. Procesamiento en tiempo real

4.1. Definición y su relación con el procesamiento en streaming

El **procesamiento en tiempo real** es aquel con un análisis y respuesta a los datos en el mismo instante en que se generan o se reciben. Tiene el objetivo de garantizar que las decisiones y acciones se realicen casi de inmediato con una latencia mínima (milisegundos o segundos). Se usa en las mismas situaciones que en streaming. Este procesamiento requiere de infraestructuras capaces de manejar la velocidad y la continuidad de los datos generados.

Las diferencias clave entre este y el procesamiento en streaming son:

- **Latencia**: En tiempo real la latencia debe ser mínima y completamente garantizada. En streaming puede haber ligeras demoras dependiendo del caso.
- **Uso**: En tiempo real se enfoca en reacciones inmediatas, el streaming puede manejar datos en micro-lotes si prioriza la consistencia sobre la velocidad absoluta.
- **Aplicaciones**: En tiempo real se centra en decisiones críticas, como activar alarmas en sistemas médicos. En streaming se utiliza en flujos de trabajo como análisis en tiempo real de grandes volúmenes de datos.

4.2. Aplicaciones

Algunas de las aplicaciones que usan procesamiento real son:

1. **Conducción autónoma y sistemas de transportes inteligentes**: Estos sistemas dependen de las decisiones inmediatas basadas en datos de sensores para evitar accidentes y optimizar rutas. Una latencia mínima garantiza evitar situaciones peligrosas.
2. **Sanidad**: En cirugía asistida por bots o monitorización remota, cualquier retraso en la transmisión de los datos podría poner en peligro la vida del paciente. Una latencia baja garantiza una respuesta oportuna en situaciones críticas.
3. **Finanzas y comercio electrónico**: Aplicaciones como el comercio algorítmico[1] requieren latencia baja para realizar transacciones en mercados financieros donde una variación de un milisegundo puede implicar ganancias o pérdidas significativas.
4. **Videojuegos en línea y realidad aumentada**: En estos entornos donde las experiencias deben der fluidas y evitar los problemas de red necesitan una latencia baja para que la jugabilidad y la inmersión no deben quedar afectadas.
5. **Internet de las cosas (IoT)**: Dispositivos industriales y domésticos como los sistemas de alarmas inteligentes requieren respuestas inmediatas para garantizar una eficiencia operativa y seguridad.

1 El **trading algorítmico**, o trading basado en reglas y procesos, es una modalidad de operación en mercados financieros (trading) que se caracteriza por el uso de algoritmos, reglas y procedimientos automatizados en diferentes grados, para ejecutar operaciones de compra o venta de instrumentos financieros.

4.3. Diferencias con el procesamiento en streaming

Algunas de las diferencias que hay entre el procesamiento en tiempo real y en streaming son:

Característica	Streaming	Tiempo real
Latencia	Baja pero tolera pequeñas demoras	Mínima e inmediata
Prioridad	Continuidad del flujo de datos	Respuesta inmediata
Aplicaciones	Análisis de tendencias, monitoreo de IoT	Detección de fraudes, conducción autónoma
Herramientas comunes	Apache Kafka, Flink, Spark Streaming	Redis, tecnologías edge computing[2]

2 El **Edge Computing** es un tipo de arquitectura de Tecnología de la Información (TI) que le permite a las empresas y organizaciones obtener servicios confiables y seguros de sus aplicaciones y soluciones de Cloud Computing, o Computación en la Nube, y distribuirlas en una gran cantidad de ubicaciones.

5. Comparativa entre los tres procesamiento

A continuación se muestra una tabla para ver la comparativa entre los tres tipos de procesamiento:

Aspecto	Procesamiento por lotes	Procesamiento en streaming	Procesamiento en tiempo real
Latencia	Alta, ya que los datos se procesan en bloques en intervalos específicos.	Baja, los datos se procesan continuamente a medida que llegan.	Mínima, ideal para decisiones inmediatas.
Frecuencia de actualización	Alta, pero periódica (diaria, semanal, etc.)	Continua, según los eventos.	Inmediata, datos procesados al instante.
Aplicaciones	Análisis históricos, informes financieros periódicos, procesamiento masivo de datos.	Detección de fraudes, análisis de redes sociales, monitoreo de sensores.	Prevención de fraudes, monitoreo en tiempo real, análisis de comportamiento en vivo.
Beneficios	Eficaz para grandes volúmenes de datos, bajo costo en recursos computacionales.	Escalable, procesado continuo, ideal para datos en vivo.	Capacidad de tomar decisiones basadas en tiempo real.
Limitaciones	No adecuado para decisiones inmediatas, puede haber retrasos en la entrega de información crítica.	Requiere infraestructuras robustas y pueden ser complejos de implementar.	Requiere infraestructuras potentes y con manejo de alta carga de datos.

6. Conclusión

La conclusión a la que se puede llegar es que los tres procesamientos son útiles dentro de sus entornos. Solo hay que saber el entorno en el que nos vamos a encontrar, que empresa lo va a utilizar y en que tipo de situaciones se manejarán los datos.

En este mundo en constante cambio donde el flujo de datos cambia en su propio ecosistema, nosotros debemos ser capaces de adaptarnos a tales cambios y poder manejar la capacidad de controlar los grandes volúmenes de datos con las herramientas que nos proporciona el campo del Big Data.

7. Bibliografía

7.1. Procesamiento por lotes

- Wikipedia
- Ejemplos de programas que trabajan por lotes
- Profesional Review
- PhoenixNAP

7.2. Procesamiento en streaming

- Confluent.io
- Cloudera
- TechFormacion
- Ejemplos de programas que trabajan en streaming
- Nativos Digitales
- ITDO
- Aprender Big Data
- AWS

7.3. Procesamiento en tiempo real

- Cloud Levante
- Data Camp
- Seidor
- TechFormacion
- Appmaster
- Axial ERP