Revista Difusiones, ISSN 2314-1662, Num. 21, 2(2) julio-diciembre 2021, pp.38-58
Fecha de recepción: 27-10-2021. Fecha de aceptación: 16-11-2021

Del BIG DATA al FAST DATA: enfoques modernos de streaming de datos para el procesamiento de datos masivos en tiempo real

From BIG DATA to FAST DATA: latest data streaming approaches for real-time mass data processing

María Laura Sanchez Piccardi[1]
lsanchezpiccardi@gmail.com
Universidad Católica de Santiago del Estero, Santiago del Estero, Argentina
Lilia Eugenia Palomo[2]
lilia.palomo@ucse.edu.ar
Universidad Católica de Santiago del Estero, Santiago del Estero, Argentina

1 Maestría en Informática, UNSTA 2017. Ing. en Computación, UCSE 2000, Técnico en Informática, UCSE., 1996. Profesor Adjunto FCID y FCSPyJ, UCSE. Investigación "Del Big Data al Fast Data: enfoques modernos de Streaming de datos para el procesamiento de datos masivos en tiempo real" (2019-2021). "Hacia una economía digital: integración de los negocios, las operaciones y la tecnología en nuevos modelos operativos para el futuro" (2019). "Aproximación teórica de las Estrategias de Delivery de Datos Unificados del ámbito organizacional" (2018). Producción: Big Data como tecnología disruptiva en los nuevos modelos operativos organizacionales (WICC-2019). El Big Data desde la perspectiva de sus implicaciones jurídicas en el Evidencia Digital. (WICC-2019). Técnicas de unificación de datos para la visualización de grandes volúmenes de datos (WICC-2018), Divulgación: Jornadas de Divulgación Científica de Informática y Tecnología, UCSE (2020). "Del Big Data al Fast Data: Enfoques modernos de streaming de datos para el procesamiento de datos masivos en tiempo real".

2 Esp. en Ing. Web; UCSE 2018. Esp. en Enseñanza de la Ed. Sup., UCC 2003. Ing. en Computación, UCSE 1999. Profesor

Resumen

Actualmente, impulsado por los objetivos claves de negocios de la nueva era digital, incluida la presión competitiva, la capacidad de actuar sobre nuevas oportunidades de negocio, la creciente necesidad de procesamiento de datos rápidos, y la necesidad de nuevas experiencias contextuales y relevantes para clientes más sofisticados, existe una demanda de técnicas de procesamiento de datos instantáneos, como Data Streaming, que sean capaces de entregar resultados en tiempo real.

El Data Streaming, también conocido como Fast Data, es más que la simple extracción de información que se importa en ese momento y más rápidamente. Se trata de aceptar el cambio general en la forma en que se construyen las aplicaciones centradas en datos.

Muchas empresas que tienen sus sistemas basados en Big Data han comenzado a cambiar sus sistemas de procesamiento por lotes para seguir el ritmo de la tercera dimensión de los datos, que es la velocidad. A pesar de su importancia, avanzar hacia arquitecturas de datos rápidas que proporcionen sistemas resilientes y escalables en tiempo real es una tarea desafiante.

Por ello, en este trabajo se presentará una exploración de los enfoques modernos de streaming de datos para el procesamiento masivo de datos en tiempo real, como una estrategia emergente que permita a las organizaciones obtener información de manera oportuna y reaccionar rápidamente en la toma de mejores decisiones.

Se emplearon instrumentos y técnicas para la recolección de datos y análisis documental, para aumentar el grado de familiaridad con este fenómeno relativamente desconocido.

Se brindarán aportes conceptuales, técnicos y prácticos, útiles para los distintos actores interesados en obtener un panorama general sobre las aplicaciones Fast Data; su arquitectura y aspectos más relevantes. Como resultado se proporcionará una guía de recomendaciones para impulsar iniciativas de datos rápidos con éxito.

Palabras clave

Datos rápidos, trasmisión de datos, procesamiento de datos sin límites, macrodatos.

Asociado FCID y FCSPyJ. Investigación “Del Big Data al Fast Data: enfoques modernos de Streaming de datos para el procesamiento de datos masivos en tiempo real” (2019-2021). “Hacia una economía digital: integración de los negocios, las operaciones y la tecnología en nuevos modelos operativos para el futuro” (2019). “Aproximación teórica de las Estrategias de Delivery de Datos Unificados del ámbito organizacional” (2018). Producción: WICC-2019 Big Data como tecnología disruptiva en los nuevos modelos operativos organizacionales. WICC-2018 Estudio y análisis de técnicas de modelado de grandes volúmenes de datos jurídicos. Divulgación: “Conversatorio sobre la enseñanza de la Ingeniería en tiempos de Covid-19 (2020). “El Big Data y sus Implicancias en el Contexto Jurídico y Empresarial” (2019). Gestión: Director Ing. en Informática, FCID, UCSE.

Abstract

Currently driven by the key business objectives of the new digital age, including competitive pressure, the ability to act on new business opportunities, the growing need for fast data processing, and the need for new contextual and relevant experiences for customers more sophisticated, there is a demand for instant data processing techniques, such as Data Streaming, that are capable of delivering results in real time.

Data Streaming, also known as Fast Data, is more than the simple extraction of information that is imported at that moment and more quickly. It's about embracing the general change in the way data-centric applications are built.

Many companies that have their systems based on Big Data have started to change their batch processing systems to keep up with the third dimension of data, which is speed. Despite its importance, moving towards fast data architectures that provide resilient and scalable systems in real time is a challenging task.

Therefore, this work presents an exploration of modern data streaming approaches for massive data processing in real time, as an emerging strategy that allows organizations to obtain information in a timely manner and react quickly in making better decisions.

Instruments and techniques for data collection and documentary analysis were used to increase the degree of familiarity with this relatively unknown phenomenon.

Conceptual, technical and practical contributions will be provided, useful for the different stakeholders interested in obtaining an overview of Fast Data applications; its architecture and most relevant aspects. As a result, a best practice guide will be provided to successfully drive fast data initiatives.

Key Words

Fast data, data streaming, unlimited data processing, big data.

Introducción

A medida que los volúmenes de datos crecen, las fuentes de datos se vuelven más variadas y los flujos de datos se aceleran, existe un creciente deseo de responder más rápidamente a las oportunidades de negocios emergentes. Ya sea que se trate de mantenimiento predictivo de máquinas, recomendaciones de productos para clientes, detección de fraudes o protección de ciberseguridad, un número creciente de procesos críticos para el negocio exigen decisiones en tiempo real (Amazon s.f.).

Asimismo, las necesidades de datos de la empresa cambian constantemente, pero a tasas incoherentes, y en los últimos años el cambio se ha producido a un ritmo cada vez mayor. Las herramientas que alguna vez se consideraron útiles para aplicaciones de Big Data ya no son suficientes. Cuando las operaciones por lotes predominaban, Hadoop[3] podía manejar la mayoría de las necesidades de una organización. El desarrollo en otras áreas de TI[4] (IoT[5], geolocalización, etc.) ha cambiado la forma en que se recopilan, almacenan, distribuyen, procesan y analizan los datos. Las decisiones en tiempo real complican este escenario y se necesitan nuevas herramientas y arquitecturas para manejar estos desafíos de manera eficiente (Estrada R., 2018).

Recordemos las 3 magnitudes claves (3 V) del Big Data: Volumen, Variedad y Velocidad. Durante un tiempo se enfatizó el volumen de datos; ahora las aplicaciones de datos rápidas significan que la velocidad y la variedad son la clave.

Dos tendencias han surgido de esta evolución: primero, la variedad y velocidad de los datos que las empresas necesitan para la toma de decisiones, que continúa creciendo. Estos datos incluyen no solo información transaccional, sino también datos comerciales, métricas IoT, información operacional y registros de aplicaciones. En segundo lugar, las empresas modernas necesitan tomar esas decisiones en tiempo real, en base a todos los datos recopilados.

En consecuencia, muchos segmentos de la industria están lidiando con datos rápidos, donde los datos llegan a gran volumen y velocidad. Las empresas de estos sectores necesitan procesar estos datos rápidos justo a tiempo para obtener ideas y actuar con celeridad.

Hoy en día, el procesamiento de datos en tiempo real es un gran problema en Big Data, y por buenas razones. Entre ellos se pueden mencionar (Tyler, 2015):

- Las empresas anhelan datos cada vez más oportunos, y cambiar al Streaming es una buena forma de lograr una latencia más baja.

[3] Apache Hadoop es un entorno de trabajo para software, bajo licencia libre, para programar aplicaciones distribuidas que manejen grandes volúmenes de datos. Permite a las aplicaciones trabajar con miles de nodos en red y petabytes de datos.
[4] Tecnologías de la Información
[5] Internet of Things describe la red de objetos físicos (cosas) que incorporan sensores, software y otras tecnologías con el fin de conectar e intercambiar datos con otros dispositivos y sistemas a través de Internet.

● Los conjuntos de datos masivos e ilimitados que son cada vez más comunes en los negocios modernos, son más fáciles de dominar usando un sistema diseñado para tales volúmenes de datos interminables.
● El procesamiento de los datos a medida que llegan distribuye las cargas de trabajo de manera más uniforme a lo largo del tiempo, lo que genera un consumo de recursos más uniforme y predecible.
A pesar de este aumento de interés impulsado por el negocio del Streaming de datos, la mayoría de los sistemas existentes en la actualidad permanecen relativamente inmaduros en comparación con sus pares de lotes, lo que ha resultado en un gran desarrollo activo y emocionante en este espacio recientemente.
Entonces, en el presente trabajo se presenta el resultado de la exploración de los enfoques modernos de Streaming de datos para el procesamiento de datos masivos en tiempo real como una estrategia emergente que permita a las organizaciones ofrecer nuevos servicios, mejorar la experiencia de usuario, aumentar la eficiencia, lograr operaciones de mayor calidad y, en consecuencia, obtener una ventaja competitiva en el mercado.

Objetivos

La investigación se centró en el objetivo principal de estudiar y analizar los enfoques modernos de Streaming de datos para el procesamiento de datos masivos en tiempo real que proporcionen a las organizaciones conocimientos más rápidamente, para ser más eficientes y experimentar nuevas oportunidades de negocio.
Además, para el desarrollo del trabajo se formularon unos objetivos más específicos:
● Investigar y estudiar los aspectos claves del funcionamiento del Streaming de datos.
● Identificar las limitaciones existentes en los enfoques tradicionales o por lote para el procesamiento de datos rápidos.
● Relevar y analizar las diferentes opciones arquitectónicas emergentes y sus implicaciones.
● Presentar un breve mapa del panorama actual de las principales herramientas comerciales para el procesamiento de datos rápidos.
● Formular una serie de recomendaciones que guíen las acciones que se deben tomar para adoptar soluciones de datos rápidos con éxito.

Antecedentes

El surgimiento de Internet a mediados de la década de 1990 indujo la creación de conjuntos de datos de tamaño sin precedentes. Las herramientas existentes no eran lo suficientemente escalables ni rentables para estos conjuntos de datos, lo que forzó la creación de nuevas herramientas y técnicas. La naturaleza "siempre activada" de Internet

también elevó el estándar de disponibilidad y confiabilidad. El ecosistema de Big Data surgió en respuesta a estas presiones a fines de la década de 1990 y principios de la década de 2000, cuando las compañías de Internet más grandes se vieron obligadas a inventar nuevas formas de administrar inmensos volúmenes de datos en constante crecimiento.

Según Dean Wampler (2015): "La mayoría de las personas equiparan Big Data con las bases de datos Hadoop y NoSQL. Sin embargo, los componentes básicos originales de Hadoop, el sistema de archivos distribuidos Hadoop (HDFS[6]) para almacenamiento, el motor de cálculo MapReduce y el administrador de recursos (YARN[7]), están enraizados en modo por lotes (batch) o fuera de línea (sin conexión); arquitecturas de procesamiento que tienen dos décadas de antigüedad. Con el rápido aumento de las arquitecturas de Streaming como Apache Spark, las empresas quieren obtener una ventaja competitiva en el mercado con su infraestructura informática para reducir el intervalo de tiempo entre la entrada de datos y la extracción de información".

Las nuevas tendencias tecnológicas en informática están impulsando la transición a arquitecturas de datos rápidas (Fast Data) para admitir *aplicaciones reactivas*[8], incluida la proliferación de dispositivos inteligentes con el Internet de las Cosas (IoT), el cambio de cargas de trabajo informáticas a la nube y el aumento de BYOD[9] en el trabajo y móvil. Estas tendencias otorgan una nueva importancia a la velocidad y la flexibilidad para los canales de datos en la empresa para entregar aplicaciones que sean más confiables y puedan escalar.

La importancia del Streaming de datos ha crecido en los últimos años, incluso para datos que no lo requieren estrictamente, ya que brinda a las empresas una ventaja competitiva al reducir la brecha de tiempo entre la entrada de datos y la extracción de información. (Lopez, 2020). Por ejemplo, si escucha una noticia de última hora y busca información en Google y Bing, desea que los resultados de búsqueda muestren las últimas actualizaciones en sitios web de noticias. Por lo tanto, las actualizaciones en modo batch para los motores de búsqueda ya no son aceptables, aunque una demora de unos segundos a minutos está bien.

En consecuencia, las arquitecturas clásicas de Big Data están evolucionando para admitir los nuevos escenarios de procesamiento streaming. El término Fast Data (datos rápidos) se ha acuñado para hacer referencia a estas nuevas arquitecturas e involucra una amplia gama de nuevos sistemas y enfoques, que equilibran diversos intercambios para ofrecer un procesamiento de datos oportuno y rentable, así como una mayor productividad del desarrollador (Perera, 2018).

Los datos rápidos (Fast Data) son la evolución natural derivada de los grandes datos (Big Data) para que se oriente a los flujos de datos y se procesen rápidamente, a la vez que se

6 Hadoop Distributed File System
7 Yet Another Resource Negotiator
8 Sistemas flexibles, débilmente acoplados y escalables¸ que los hace más fáciles de desarrollar y susceptibles de cambiar.
9 BringYourOwnDevice: Nueva tendencia tecnológica que permite a los trabajadores llevar sus dispositivos portátiles personales para realizar tareas del trabajo y conectarse a la red y recursos corporativos.

habilitan los análisis clásicos en modo por lotes, el almacenamiento de datos y las consultas interactivas (Wampler, 2016).

El streaming es una tecnología que permite a los usuarios consultar flujos continuos de datos y detectar condiciones rápidamente en un período de tiempo pequeño desde el momento en que se reciben los datos. El período de tiempo de detección varía de pocos milisegundos a minutos. Por ejemplo, con el procesamiento streaming, se puede consultar flujos de datos provenientes del sensor de temperatura y recibir una alerta cuando la temperatura haya alcanzado el punto de congelación.

El streaming de datos también es conocido bajo nombres como análisis en tiempo real, análisis de transmisión, procesamiento de eventos complejos, análisis de transmisión en tiempo real y procesamiento de eventos.

En general, el procesamiento streaming es útil en sistemas que manejan grandes volúmenes de datos y donde importan los resultados en tiempo real; es decir, cuando el valor de la información contenida en el flujo de datos disminuye rápidamente a medida que envejece. El procesamiento streaming se aplica a escenarios difíciles, principalmente a (Schreiner, 2018):

- Análisis en tiempo real (para obtener información empresarial rápida y toma de decisiones).
- Detección de anomalías, fraude, problemas de rendimiento, et.
- Procesamiento de eventos complejos.
- Estadísticas en tiempo real (monitoreo, alimentación de los paneles de control en tiempo real).
- Sistemas ETL (extraer, transformar, cargar) en tiempo real.
- Implementación de arquitecturas basadas en eventos.

La fortaleza clave del streaming de datos es que puede proporcionar conocimientos más rápidamente, a menudo en cuestión de milisegundos a segundos.

Por ejemplo, para la famosa casa de apuestas deportivas y casino online William Hill[10], brindando servicio a millones de clientes, *el tiempo es realmente la esencia*. Conocer los últimos detalles de los eventos más populares y estar preparado para manejar picos masivos en el tráfico durante los eventos es un diferenciador clave en cualquier parte del mundo (Wampler, 2015).

Importancia y tendencia actual del Fast Data

¿Por qué aprender sobre otra tendencia en Big Data, e invertir el tiempo y los recursos para adoptarla e implementarla? La respuesta es simple: El procesamiento y análisis en tiempo real conllevan la promesa de hacer que las organizaciones sean más eficientes y abran

[10] http://www.williamhill.es/

nuevas oportunidades (Anadiotis, 2017).

Los datos son el mayor activo de una empresa moderna, si se utilizan de manera efectiva.

Después de todo, en la economía actual siempre conectada, las aplicaciones de negocios más valiosas están basadas en datos. Los clientes esperan interacciones en tiempo real impulsadas por millones de puntos finales y cantidades masivas de datos. Y es más que solo compañías como Facebook, Amazon y Uber alimentando estas experiencias: las compañías de tarjetas de crédito alertan a los clientes sobre posibles fraudes en tiempo real, los autos conectados brindan información actualizada sobre el tráfico, los médicos brindan recomendaciones de atención perspicaces basadas en los modelos predictivos y las plantas de fabricación detectan problemas de calidad del producto incluso antes de que ocurran.

Como resultado, las aplicaciones empresariales están cambiando de arquitecturas monolíticas a sistemas distribuidos compuestos por microservicios[11] implementados en contenedores y servicios de plataforma como colas de mensajes, bases de datos distribuidas y motores de análisis. Diseñar, construir y escalar estas aplicaciones distribuidas modernas es un proceso complejo, y muchas empresas están luchando para mantenerse al día con un panorama evolutivo de tecnologías (Tripathi, 2018).

Pero al igual que cualquier otra técnica, hay algunos desafíos que los analistas y especialistas de Big Data encuentran en el Streaming de datos también.

Uno de los desafíos más importantes que definen todo el proceso es la velocidad seguida de su construcción. Además de estos, los desafíos también son evidentes en la planificación de la escalabilidad, tolerancia a fallas y durabilidad de los datos (Kleppmann, 2017).

Sin embargo, muchas organizaciones cada vez más conscientes del valor de los datos y el impacto significativo que puede traer a su negocio, ya la han adoptado, otros lo tienen en su radar, y casi todas las predicciones para 2019 lo mencionan de una forma u otra. Entre las recomendaciones de los analistas, Forrester[12] ve en el Streaming un paso clave en el camino hacia una plataforma de datos de autoservicio ágil y en tiempo real. Según un reporte de esta consultora, tres cuartas partes de los encuestados afirman que los datos de fuentes móviles y los datos de Internet de las cosas (IoT) son una prioridad alta o crítica para las estrategias de datos de sus empresas. Casi las tres cuartas partes están de acuerdo en que las fuentes de datos internas de aplicaciones empresariales u operativas son una prioridad alta, y más de dos tercios están de acuerdo en que los datos de fuentes externas, como proveedores de datos de terceros o clientes, son sumamente importantes o de importancia crítica (Forrester, 2018)

Al mismo tiempo, administrar todas estas fuentes de datos se ha vuelto más complicado en los últimos años, ya que la variedad de fuentes de datos, el volumen de datos que ingresan y

11 Pequeños servicios, los cuales se ejecutan en su propio proceso y se comunican con mecanismos ligeros (normalmente una API de recursos HTTP).

12 https://go.forrester.com/

la velocidad han aumentado significativamente para la mayoría de las organizaciones. Se prevé que la cantidad total de datos creados, capturados, copiados y consumidos a nivel mundial aumente rápidamente durante los próximos años hasta 2025, estimando que la creación de datos globales crezca a más de 180 zettabytes. En 2020, la cantidad de datos creados y replicados alcanzó un nuevo récord. El crecimiento fue mayor de lo esperado anteriormente debido al aumento de la demanda debido a la pandemia de COVID-19, ya que más personas trabajaron y aprendieron en el hogar y utilizaron las opciones de entretenimiento en el hogar con más frecuencia (Holst,2021).

Diseño de arquitecturas de datos rápidos

En las arquitecturas de Fast Data, los eventos individuales se procesan a medida que llegan, en tiempos de procesamiento de milisegundos, incluso microsegundos.

Crear este tipo de arquitecturas que puedan hacer este tipo de procesamiento de milisegundos significa usar sistemas y enfoques que ofrecen un procesamiento de datos oportuno y rentable centrado en la productividad del desarrollador. Cualquier arquitectura de datos rápida y exitosa debe satisfacer estos requisitos de alto nivel: adquisición o ingesta de datos de alto rendimiento y confiable, almacenamiento y consultas flexibles y sofisticadas herramientas de análisis (Estrada, 2018).

También es importante mencionar que los componentes de la arquitectura deben cumplir con las R: reactivos (escalado hacia arriba y hacia abajo en función de la demanda), resistentes (frente a errores en todos los sistemas distribuidos) y con capacidad de respuesta (incluso si los errores limitan la capacidad de prestar servicios).

Tradicionalmente, los sistemas interactivos y de modo por lotes han tenido requisitos menos estrictos para estas cualidades. Las arquitecturas de datos rápidas son como otros sistemas en línea donde el tiempo de inactividad y la pérdida de datos son problemas serios y costosos. Al implementar estas arquitecturas, los desarrolladores que se han centrado en las herramientas de análisis que se ejecutan en el back office se ven obligados continuamente a aprender nuevas habilidades para la programación y las operaciones de sistemas distribuidos.

Una estrategia de datos rápida debe incluir la arquitectura de una solución que permita la ingestión de datos orientados a eventos de alta velocidad; la capacidad de realizar análisis en tiempo real en la alimentación de datos en vivo; la capacidad de actuar sobre esos datos; y exportación rápida a sumideros de análisis a largo plazo.

Esta arquitectura supone una ruptura con la aplicación de datos en silos tradicional, donde los datos se desconectan de la analítica y otras aplicaciones y datos; admite datos rápidos creados en una multitud de nuevos puntos finales, operacionaliza el uso de esos datos en aplicaciones y mueve los datos a un sumidero de datos donde los servicios están disponibles para las necesidades de análisis y almacenamiento a largo plazo de la empresa.

Esta arquitectura de datos se puede representar como una canalización de datos que unifica aplicaciones, análisis e interacción de aplicaciones en múltiples funciones, productos y disciplinas y los pone a disposición para ser utilizados estratégicamente (Figura1).

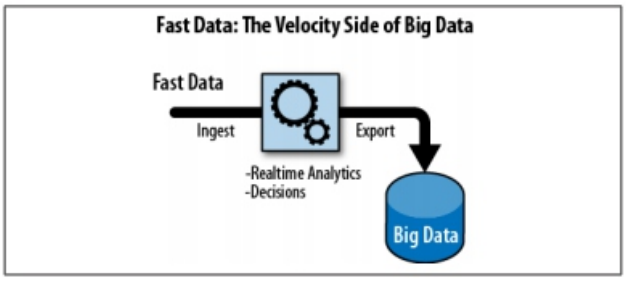

Figura 1. Los datos rápidos representan el aspecto de la velocidad del Big Data (Jarr, 2015)

Componentes de una arquitectura Fast Data

Las arquitecturas de datos rápidos definen el conjunto de componentes integrados que proporcionan los mecanismos básicos para crear, implementar y operar aplicaciones escalables, eficaces y resistentes las 24 horas del día.

Desde una perspectiva de alto nivel, los sistemas de datos rápidos se componen de tres bloques funcionales o etapas de transformación que proporcionan las siguientes capacidades de datos (Maas&Kontopoulos, 2018):

Figura 2. Etapas de transformación de los datos (desarrollo propio)

A continuación, se hará referencia a cada una de estas áreas funcionales y el abanico de tecnologías líderes disponibles que implementan estas funciones según los diferentes escenarios o plataformas.

1. Adquisición

La adquisición o ingestión es la primera etapa en la arquitectura Fast Data, en donde se proporcionan las interfaces para interactuar con diversas fuentes externas de datos y aceptar y transformar o normalizar los datos entrantes. También, en arquitecturas de datos rápidos define cómo y dónde se transmiten los datos.

Es el primer punto en el que se pueden realizar transacciones de datos, aplicando funciones y procesos clave para extraer valor de los datos; valor que incluye conocimiento, inteligencia y acción.

El enfoque clave de esta etapa es el rendimiento[13] ya que este paso afecta la cantidad de datos que puede recibir todo el sistema en un momento dado.

La ingestión de datos representa la fuente de todos los mensajes que ingresan al sistema. Algunos ejemplos de fuentes de ingestión incluyen una API RESTful[14] orientada al usuario que se encuentra en la periferia del sistema, respondiendo a las solicitudes HTTP que se originan en los usuarios finales; un registro de captura de datos modificados de una base de datos que registra operaciones de actualización o un directorio del sistema de archivos desde el que se leen los archivos.

Desde la perspectiva de una plataforma Fast Data, la transmisión de datos[15] llega a través de la red, generalmente terminada por un adaptador escalable que puede conservar los datos dentro de la infraestructura interna. Este proceso de captura debe escalar a las mismas características de rendimiento de la fuente de transmisión o proporcionar algún medio de retroalimentación a la parte de origen para permitirle adaptar su producción de datos a la capacidad del receptor. En muchos escenarios distribuidos, la adaptación por parte de la parte de origen no siempre es posible, ya que los dispositivos de borde a menudo consideran que el backend de procesamiento está siempre disponible.

Una vez que los mensajes de eventos están dentro de la infraestructura de backend, el control de flujo de transmisión puede proporcionar señalización bidireccional para mantener una serie de aplicaciones de streaming funcionando a su carga óptima.

La cantidad de datos que se pueden recibir generalmente está limitada por la cantidad de datos que se pueden procesar y qué tan rápido debe ser ese proceso para mantener un sistema estable. Esto conduce a la siguiente área de la arquitectura Fast Data: los motores de procesamiento.

Para esta etapa, se debe considerar APIs de streamingy soluciones de mensajería como: Apache Kafka[16], AkkaStreams[17],Amazon Kinesis[18], ActiveMQ[19], RabbitMQ[20], Red Hat AMQ[21], Oracle Tuxedo[22].

2. Análisis y procesamiento.

A medida que se crean los datos, llegan a la organización velozmente.

13 Cantidad de mensajes recibidos durante un período de tiempo.

14 Una API de REST, o API de RESTful, es una interfaz de programación de aplicaciones que se ajusta a los límites de la arquitectura REST y permite la interacción con los servicios web de RESTful.

15 La transmisión(streaming) de datos es un flujo infinito de datos que se genera mediante una o más fuentes de datos y se recopila para su entrega a un consumidor a través de un transporte. Este flujo de datos podría provenir de fuentes distribuidas, como el flujo proveniente de una cámara de seguridad, o estar centralizado en un lugar, como un servidor web en un centro de datos.

16 https://kafka.apache.org/

17 https://akka.io/

18 https://aws.amazon.com/es/kinesis/

19 http://activemq.apache.org/

20 https://www.rabbitmq.com/

21 https://developers.redhat.com/products/amq/overview/

22 https://www.oracle.com/middleware/technologies/tuxedo.html

Los datos de una secuencia pueden llegar en muchos tipos y formatos. La mayoría de las veces, los datos proporcionan información sobre el proceso que los generó; esta información puede denominarse mensajes o eventos. Esto incluye datos de nuevas fuentes, como datos de sensores, así como flujos de clics de servidores web, datos de máquinas y datos de dispositivos, transacciones e interacciones con los clientes.

El aumento de los datos rápidos presenta la oportunidad de realizar análisis de los datos a medida que se transmiten, en lugar de a posteriori, después de que se hayan enviado a un almacén de datos para un análisis a largo plazo. La capacidad de analizar flujos de datos y tomar decisiones durante la transacción sobre estos datos nuevos es la visión más convincente para los diseñadores de aplicaciones basadas en datos.

Los motores de procesamiento son un componente muy importante de una arquitectura de Fast Data ya que el valor real de los datos se captura y se extrae en activos valiosos. Es el lugar donde la lógica empresarial se aplica e implementa de acuerdo con los requisitos del sistema que satisfacen algunos objetivos comerciales.

Cuando se caracterizan por métodos utilizados para manejar mensajes, los motores de procesamiento de transmisión se pueden clasificar en dos tipos principales (Maas & Kontopoulos, 2018):

a) Uno a la vez: Estos motores de transmisión procesan los flujos de datos individualmente a medida que llegan, lo que se optimiza para la latencia[23] a expensas de un mayor consumo de recursos del sistema o de un menor rendimiento en comparación con los micro lotes.

b) Micro Batch: Los datos se procesan utilizando motores de micro lotes que internamente acumulan registros siguiendo ciertos criterios. Cuando se cumplen los criterios, se cierra el lote y se envía a para su ejecución y todos los registros recibidos del lote se someten a la misma serie de procesamiento.

Los motores de procesamiento ofrecen una API y un modelo de programación mediante el cual los requisitos se pueden traducir a código ejecutable. También brindan garantías con respecto a la integridad de los datos, como la ausencia de pérdida de datos o la recuperación de fallas sin problemas.

En todos los casos de uso, el motor de procesamiento para datos rápidos debe tener las siguientes propiedades (Piekos, 2015):

· Alta tasa de ingestión: se debe ingerir datos a tasas de transacción históricamente desafiantes.

· Análisis en tiempo real.

· Decisiones en tiempo real.

Además, se requiere un sistema diseñado para ofrecer:

23 Métrica que suele medir el tiempo entre dos puntos en un flujo de datos, el borde y el servidor. La latencia describe qué tan rápido llegan los datos del productor real al punto de recolección

· Alta disponibilidad.
· Escalable elástica y horizontalmente.

Para esta etapa, se debe considerar soluciones de procesamiento de datos como: Apache Spark[24](micro lotes), Apache Flink[25] (transmisión), Apache Storm[26], Apache Beam[27].

3. Almacenamiento

Una vez que se han capturado y procesado los datos, hay que crear su valor real, siendo necesario almacenarlos en un subsistema de almacenamiento, como bases de datos, sistemas de archivos o cachés distribuidos (Maas & Kontopoulos, 2018).

En el caso particular de las arquitecturas Fast Data, el almacenamiento generalmente marca los límites de transición entre el núcleo Fast Data y las aplicaciones tradicionales que pueden consumir los datos producidos.

La elección de la tecnología de almacenamiento se basa en los requisitos particulares de esta transición entre datos en movimiento y en reposo.

Si necesitamos almacenar el flujo de datos completo a medida que ingresan, y precisamos acceder a cada registro individual o porciones secuenciales de ellos, se requiere un backend altamente escalable con escrituras de baja latencia y capacidades de consulta basadas en claves.

Los datos también podrían cargarse en un almacén de datos tradicional o podrían usarse para construir un lago de datos que pueda admitir diferentes capacidades, incluido el aprendizaje automático, la generación de informes o el análisis ad hoc.

En el otro lado del espectro, se tendrán agregados predigeridos que son solicitados por un sistema de visualización frontend. En este caso, probablemente se requiera una consulta SQL completa y el soporte de indexación para localizar rápidamente esos registros para su visualización. Un PostgreSQL[28], un MySQL[29] o las contrapartes comerciales de un sistema de gestión de bases de datos relacionales (RDBMS) serían una opción razonable.

Entre estos dos casos hay una amplia gama de opciones, que van desde bases de datos especializadas (como InfluxDB[30] para series de tiempo o Redis[31] para búsquedas rápidas en memoria), hasta almacenamiento en bruto (como HDFS[32] local) o las ofertas de almacenamiento en la nube como Amazon S3[33], Almacenamiento de Azure[34], Google Cloud Storage[35] entre otras.

24 https://spark.apache.org/
25 https://flink.apache.org/
26 http://storm.apache.org/
27 https://beam.apache.org/
28 https://www.postgresql.org/
29 https://www.mysql.com/
30 https://www.influxdata.com/
31 https://redis.io/
32 https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
33 https://aws.amazon.com/es/s3/
34 https://azure.microsoft.com/es-es/
35 https://cloud.google.com/

Para esta etapa, se puede considerar también soluciones de almacenamiento de bases de datos distribuidas como:Apache Cassandra[36], Couchbase[37], Amazon DynamoDB[38], Apache Hive[39], Riak[40], Neo4J[41], MongoDB[42], MemSQL[43].

Modelo de una arquitectura Fast Data exitosa: la Pila SMACK

En la actualidad ha surgido la pila SMACK como una tecnología de código abierto crucial para manejar un gran volumen y velocidad en la era de los datos rápidos. Es la plataforma ideal para construir aplicaciones Fast Data.

SMACK es el acrónimo que representa las partes individuales de la colección: Spark, Mesos, Akka, Cassandra y Kafka. Todas estas tecnologías son de código abierto y con la expectativa de Akka, todos son proyectos de software Apache.

En el mundo de hoy, conectado y basado en datos, el stack SMACK se utiliza para crear aplicaciones empresariales modernas que cumplen con un nuevo conjunto de requisitos (Hsu, 2017):

· Ingerir datos a escala sin pérdidas (permite transmitir datos de millones de interacciones de usuario o sensores de IoT).

· Analizar los datos en tiempo real.

· Desencadenar acciones basadas en los datos analizados.

· Almacenar los datos a escala de la nube.

· Ejecutar servicios en la nube con un sistema operativo escalable, distribuido y altamente resistente.

SMACK es un modelo de arquitectura pipeline para el procesamiento de datos. Un pipeline[44] de datos es un software que consolida datos de múltiples fuentes y lo pone a disposición para ser utilizado estratégicamente. (Estrada, 2016).

La pila SMACK, creada por Mesosphere[45] en colaboración con Cisco, es una plataforma de arquitectura de Fast Data distribuida y altamente escalable que se compone de una combinación de componentes de código abierto que comparten las mismas características de escalabilidad (McFadin, 2017):

· Spark: Motor rápido y general para el procesamiento de datos distribuido a gran escala.,

36 https://cassandra.apache.org
37 https://www.couchbase.com/
38 https://aws.amazon.com/es/dynamodb/
39 https://hive.apache.org/
40 https://riak.com/products/
41 https://neo4j.com/sandbox/
42 https://www.mongodb.com/es
43 https://www.memsql.com/
44 Se llama pipeline porque cada tecnología contribuye con sus características a una línea de procesamiento como una línea de montaje industrial.
45 https://mesosphere.com/

para cargas de trabajo por lotes y de streaming. El enfoque de Spark es ofrecer procesamiento de datos rápidos, análisis sofisticados. procesamiento de secuencias en tiempo real y capacidad de integración con datos de Hadoop existentes (Estrada, 2016).

· Mesos: Plataforma de gestión de clústeres flexible que proporciona un eficiente aislamiento y uso compartido de recursos entre aplicaciones distribuidas. En MACK Stack, Apache Mesos organiza todos los componentes y administra los recursos. Estrada, 2016).

· Akka: Conjunto de herramientas para crear aplicaciones basadas en mensajes altamente concurrentes, distribuidas y resistentes[46].

· Cassandra: Base de datos NoSql, distribuida, descentralizada, elásticamente escalable, de alta disponibilidad, tolerante a fallas, armoniosamente consistente y orientada a columnas. En la pila SMACK, Akka, Spark y Kafka pueden usar Cassandra para conservar los datos como una capa de datos. Además, Cassandra puede manejar datos operativos y se puede utilizar para devolver datos a la capa de aplicación (Carpenter, 2020).

· Kafka: Sistema de mensajería de publicación-suscripción distribuida de alto rendimiento y baja latencia diseñado para manejar fuentes de datos en tiempo real. En SMACK, Apache Kafka se erige como el punto de ingestión para la capa de datos, que ingieren datos de diferentes fuentes y fluye a través del pipeline hasta el siguiente punto de la pila (McFadin, 2017).

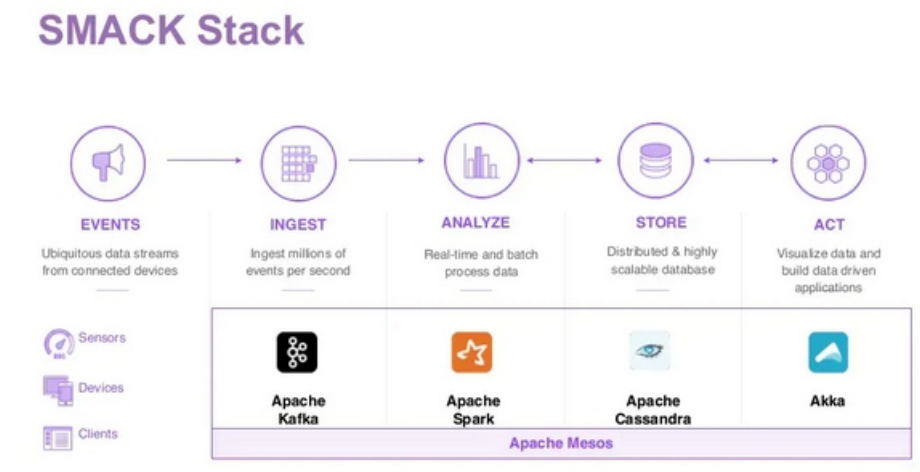

Figura 3. La arquitectura pipeline SMACK (Schad, 2017)

Cada uno de los componentes mencionados desempeña un rol único y bien diferenciado dentro de la arquitectura, pero en combinación, interactúan e integran para simplificar y poner en marcha las tareas rápidamente, brindando así una infraestructura de back-end completa que se adapta a las cargas de trabajo más exigentes de la actualidad.

[46] https://akka.io/.

Guía de recomendaciones de cómo adoptar soluciones de datos rápidos con éxito

Los datos rápidos son importantes. Son el combustible para la inteligencia en tiempo real, para los procesos de negocios digitales de próxima generación y las experiencias de los clientes. Las demandas de infraestructura de hardware y software se expandirán exponencialmente debido a la naturaleza en tiempo real de los servicios digitales; la intensidad de cálculo de los algoritmos de análisis y aprendizaje automático; y el volumen creciente de datos que deben almacenarse. Para satisfacer las demandas de datos rápidos, las organizaciones deben implementar soluciones de Fast Data que ingieran y analicen datos a velocidades más rápidas.

En esta sección se formularán una serie de recomendaciones de cómo los líderes pueden impulsar iniciativas de Fast Data. Si bien, el cambio a un modelo de datos rápidos es ciertamente una decisión tecnológica y arquitectónica, es igualmente una consideración de estrategia empresarial, que al mismo tiempo crea nuevas oportunidades volviendo obsoletos los procesos existentes. La implementación exitosa de una estrategia de datos rápidos requerirá liderazgo, enfoque y comprensión de las siguientes mejores prácticas:

1. Cambiar de una mentalidad de "más datos" a una de "datos rápidos".

Hacer uso de los datos lo antes posible puede abrir oportunidades, pero requiere una mentalidad diferente. La mayoría de las organizaciones están atrapadas en un ciclo de "más datos es mejor" que prioriza la recopilación de datos sobre la acción de datos. Los líderes deben pensar críticamente sobre sus razones para recopilar más y más datos y reevaluar las decisiones en función de la inmediatez. Se deberían plantear lo siguiente: "Si supiera lo que está sucediendo en este mismo instante, ¿qué haría de manera distinta?". Este cuestionamiento los ayudaría a volver a focalizarse en la información más relevante, lo que ahorra tiempo y dinero en la recopilación de más cantidad de datos. Cabe destacar que no hay que decidir entre una u otra opción; ya que algunos casos de uso pueden beneficiarse de información histórica, mientras que otros son más adecuados para una acción inmediata.

2. Evaluar cuidadosamente las tecnologías requeridas para admitir datos rápidos.

El cambio a un modelo Fast Data requiere sistemas actualizados que manejen el procesamiento de datos en tiempo real, y añadir esta infraestructura no debe generar una complejidad indebida. La complejidad nunca es bienvenida en ambientes de Tecnologías de Información, pero es particularmente enemiga de los datos rápidos, ya que ralentiza los sistemas y el procesamiento de datos y dificulta que los equipos internos identifiquen y aprovechen las nuevas oportunidades de datos. Idealmente, los datos rápidos se integran en el enfoque general de gestión de datos de la organización en lugar de agregarse como otra pila o silo de tecnología. Como parte de una estrategia de gestión de datos holística, las

empresas tienen la oportunidad de considerar la posibilidad de cambiar las inversiones de los repositorios históricos (como lagos de datos, almacenes, etc.) a las tecnologías que les permitan conectar y procesar datos en movimiento. Por ejemplo, si se considera el análisis de datos, en lugar de utilizar solo modelos predictivos para extraer hasta el último rendimiento marginal de un lago de datos masivo, una organización podría, en cambio, enfocarse en un análisis rápido y una acción inmediata.

3. Garantizar que las nuevas iniciativas de datos rápidos puedan escalar y evolucionar.

Las empresas, en todo momento, tienen que sopesar el tiempo y el dinero invertidos. La escalabilidad en un centro de datos significa que el centro debe crecer en proporción al crecimiento del negocio.

La escalabilidad es un desafío doble con datos rápidos, lo que afecta tanto a las consideraciones técnicas como a los procesos comerciales. En ambos casos, los equipos comienzan con demasiada frecuencia con un entorno de prueba a pequeña escala, donde las demandas son bajas y los procesos están contenidos artificialmente, solo para descubrir más tarde que no pueden escalar para cumplir con los requisitos de volumen y velocidad del mundo real. El liderazgo técnico debe considerar cómo la infraestructura puede escalar hacia arriba y hacia abajo según sea necesario para facilitar una iteración rápida sin interrupciones, así como integrarse con otras iniciativas como dispositivos móviles, IoT, etc.

La escalabilidad vertical implica agregar más capas de procesamiento. La escalabilidad horizontal significa que una vez que una capa tiene más demandas y requiere más infraestructuras, se puede agregar hardware para satisfacer las necesidades de procesamiento. Un requisito moderno es tener escalamiento horizontal con hardware de bajo costo.

La adopción de un enfoque nativo de la nube flexible y resistente ayuda a garantizar que lo que funciona en un entorno de prueba se pueda escalar para dar respuesta a los requisitos organizativos a largo plazo. De manera similar, los líderes empresariales deben evitar crear silos de procesos o iniciativas donde los datos rápidos se limitan sólo a un aspecto del negocio (por ejemplo, la interacción con el cliente) sin tener en consideración cómo los datos deben afectar también otros procesos en la organización (facturación, logística, etc.). Los altos mandos de la organización pueden proporcionar liderazgo en la estrategia general y ayudar a los equipos de ejecución a determinar las áreas de crecimiento potencial que las iniciativas de datos necesitarán integrar y respaldar.

4. Lo importante es la velocidad, no el volumen.

La capacidad de procesar y reaccionar a los datos que se mueven rápidamente en tiempo real separará a las organizaciones ágiles de las lentas. Estos datos valiosos a menudo se originan fuera de una organización, como los teléfonos inteligentes de los clientes o la aplicación en la nube de un socio, pero las organizaciones que aprovechen con éxito esas fuentes de datos y extraigan conocimiento en tiempo real tendrán una ventaja sobre las que

no lo hagan.

Los datos rápidos tienen muchas implicaciones interesantes en todas las industrias. Al centrarse en el hecho de que no es simplemente el volumen de datos lo que importa, sino la rapidez con la que puede actuar sobre ellos, los líderes pueden guiar tanto a sus equipos como a la empresa en general para que se beneficien de sus datos de nuevas formas que les permita brindar nuevos servicios, mejorar la experiencia de usuario, aumentar la eficiencia, lograr operaciones de mayor calidad y, en consecuencia, obtener conocimientos más profundos a partir de sus datos, disfrutar de una mejor calidad y consistencia de los datos, crear una mejor experiencia para el cliente y poder hacer negocios mejor informados.

5. Elegir una solución de datos rápidos que aproveche el código abierto.

La comunidad de código abierto sigue siendo una importante fuente de innovación tecnológica; además, permite evitar dos dependencias: el bloqueo del proveedor y el soporte de entidades externas. La transparencia se garantiza a través de grupos definidos por la comunidad, como Apache Software Foundation[47] o Eclipse Foundation[48], que proporcionan pautas, infraestructura y herramientas para el desarrollo de tecnología sostenible y justa.

Los arquitectos de software deben elegir soluciones de Fast Data que aprovechen las innovaciones de código abierto, pero que estén mejoradas para administrar, escalar y combinar fácilmente con las arquitecturas de aplicaciones y datos ya existentes en la organización.

6. Aprovechar lo mejor de ambos mundos: del Big Data y del Fast Data.

Por supuesto, como ocurre con la mayoría de las cosas en la vida, el debate entre los datos masivos y los datos rápidos no es nada sencillo. Para que las organizaciones tengan éxito en la era moderna, es necesaria una combinación de ambos enfoques. Los macrodatos son extremadamente útiles para encontrar tendencias ocultas en los datos después del hecho, mientras que los datos rápidos son más adecuados para responder a los eventos a medida que ocurren. Efectivamente, los patrones y modelos de inteligencia artificial entrenados que el análisis de Big Data puede desenterrar se pueden ejecutar mediante modelos de datos rápidos para que se utilicen de una manera operativamente relevante.

Al combinar estos dos enfoques, las organizaciones están mejor preparadas para hacer evolucionar sus aplicaciones y procesos a las condiciones del mercado en constante cambio, tanto a medida que se desarrollan las situaciones como después del hecho.

7. Aplicar una mentalidad de prueba y aprendizaje a la construcción de la arquitectura y experimentar con diferentes componentes y conceptos.

Estas prácticas ágiles se han usado en el desarrollo de aplicaciones durante bastante tiempo y recientemente se han trasladado al espacio de los datos. Por ejemplo, en lugar de

[47] https://www.apache.org/
[48] https://www.eclipse.org/

participar en largas discusiones sobre diseños, productos y proveedores óptimos para identificar la opción "perfecta" seguida de aprobaciones presupuestarias prolongadas, los líderes pueden comenzar con presupuestos más pequeños y crear productos mínimos viables o encadenar herramientas de código abierto existente para crear un producto provisional y lanzarlo a producción (utilizando la nube para acelerar) y así poder demostrar su valor antes de expandirse y evolucionar más.

8. Invertir en DataOps.

DataOps[49] (Data Operations) es la orquestación de personas, procesos y tecnología para proporcionar datos de una manera rápida, confiable y lista para aplicarse o enviarse al negocio o a los usuarios requeridos para su implementación en operaciones, apps e inteligencia artificial. La inversión en DataOps puede ayudar a acelerar el diseño, el desarrollo y la implementación de nuevos componentes en la arquitectura de datos con el fin de facilitar, a los equipos de trabajo, la corrección de errores, la adaptación y la adopción de nuevos retos o líneas de negocio, en función de la retroalimentación constante.

Conclusiones

En el mundo digital actual, donde las organizaciones están inundadas de información, una cosa es segura: los datos lo impulsan todo y rápido, esto esta está cambiando la forma en que los líderes empresariales piensan sobre la disrupción y la innovación.

Los datos rápidos significan información en tiempo real, o la capacidad de obtener información a partir de los datos a medida que se generan. Es literalmente, como suceden las cosas.

El término Fast Data ha revolucionado la industria del software y esto se debe a que el tiempo para la comprensión del significado de los datos es cada vez más crítico y juega un papel trascendental en la toma de decisiones inteligente e informada; además de la ventaja comercial obvia que tienen las empresas al tener conocimiento exclusivo de la información sobre el presente, o incluso el futuro.

Un número cada vez mayor de empresas, de una amplia gama de industrias, están evolucionando más allá de la acumulación masiva de datos (Big Data) para centrarse en el "ahora" de los datos, capturando su valor en tiempo real y reaccionando a la información a medida que fluye hacia la organización para competir de manera efectiva.

Comprender la promesa y el valor de los datos rápidos es una necesidad absoluta, pero no es suficiente para garantizar el éxito de las empresas que aún trabajan para implementar iniciativas de Big Data. Tener las herramientas y las habilidades para aprovechar los datos rápidos es fundamental para las empresas de todas las industrias y geografías

Como aporte académico, este trabajo brinda conocimiento general sobre las aplicaciones

[49] https://www.ibm.com/ar-es/analytics/dataops.

Fast Data; los bloques funcionales de la arquitectura y sus características claves; y las tecnologías líderes que implementan estas funciones. Además, se presentan ejemplos de aplicaciones de datos rápidos que están siendo exitosas en el mercado y un conjunto de buenas prácticas para impulsar iniciativas de Fast Data correctamente. Con el fin de brindar a la organización moderna la capacidad de desarrollar, implementar y operar aplicaciones que brinden información en tiempo real y acciones inmediatas aumentando su ventaja competitiva y agilidad para reaccionar a los nuevos desafíos del mercado.

Bibliografía

Amazon (s.f.) "¿Qué son los datos de streaming?" Disponible en: https://aws.amazon.com/es/streaming-data/

Anadiotis, George (2017). "Streaming hot: Real-time big data architecture matters"(https://www.zdnet.com/article/streaming-hot-real-time-big-data-architecture-matters/)

Anuff, Ed (2021). "Fast Data. D It's Not Your Grandfather's Operational Data". Disponible en: https://www.cio.com/article/3614604/fast-data.html

Carpenter Jeff (2020). "Cassandra The Definitive Guide". 3rd Edition. O'Reilly, Inc.

Estrada Raul (2016)."Fast Data Processing Systems with SMACK Stack". Packt Publishing Ltd.

Estrada, Raul (2018). "From big data to fast data. Designing application architectures for real-time decisions. "O'Reilly and Mesosphere. Disponible en: https://www.oreilly.com/ideas/from-big-data-to-fast-data

Hsu Edward (2017). The SMACK Stack is the New LAMP Stack, 2017, available at https://d2iq.com/blog/smack-stack-new-lamp-stack.

Holst, Arne(2021). "Cantidad de datos creados, consumidos y almacenados 2010-2025". Disponible en: https://www.statista.com/statistics/871513/worldwide-data-created/.

Forrester A. Consulting (2018). "Don't Get Caught Waiting On Fast Data". Paper Commissioned By IBM.

Jarr, Scott (2015). "Fast Data and the New Enterprise Data Architecture". O'Reilly Media.

Maas Gerard y Kontopoulos Stavros (2018). "Designing fast data application architectures. O'Reilly Media.

Kleppmann, Martin (2017). "Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainablesystems".O'Reilly Media Inc.

Lopez, Roberto (2020). "Arquitecturas Fast Data para el procesado masivo de datos en tiempo real". GFT. Disponible en: https://blog.gft.com/es/2020/09/28/arquitecturas-fast-data-para-el-procesado-masivo-de-datos-en-tiempo-real/.

McFadin Patrick (2017). "The SMACK stack. A new architecture for today's datarich modern applications". O'Reilly Media. Disponible en https://www.oreilly.com/radar/the-smack-stack

Piekos, J. (2015). Tree Fast Data Application Patterns, available at http://highscalability.com/blog/2015/4/13/three-fast-data-application-patterns.html

Perera, Srinath (2018). “What Is Stream Processing? A Gentle Introduction”. Disponible en: https://dzone.com/articles/what-is-stream-processing-a-gentle-introduction?fromrel=true

Schreiner, Vladimir (2018). “Understanding Stream Processing: Fast Processing of Infinite and Big Data”. Disponible en: https://dzone.com/refcardz/understanding-stream-processing

Schad, Jorg (2017). “Smack Stack and Beyond. Building Fast Data Pipelines”. Mesosphere, Inc.

Tripathi, Pankaj (2018). “Data Streaming for Big Data”. Disponible en: https://www.digitalvidya.com/blog/data-streaming-big-data/

Tyler Akidau(2015). “The world beyond batch: Streaming 101”. O'Reilly Media. Disponible en:https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101

Wampler, Dean (2015). “Fast Data: Big Data Evolved”. White paper.

Wampler, Dean (2016). “Fast Data Architectures for Streaming Applications”. Ebook. O'Reilly Media. Disponible en: https://www.lightbend.com/blog/fast-data-architectures-for-streaming-applications-free-oreilly-mini-book-by-dean-wampler