¿Qué es una arquitectura de Big Data?

Una arquitectura de big data se refiere a la estructura organizativa y técnica que permite la captura, almacenamiento, procesamiento y análisis eficientes de grandes volúmenes de datos. Estas arquitecturas están diseñadas para abordar los desafíos específicos asociados con la gestión de grandes cantidades de datos, así como para aprovechar las oportunidades que ofrece el análisis de esos datos.

Características claves de una arquitectura de Big Data

Fuentes de datos:
Incluye la identificación y conexión con diversas fuentes de datos como bases de datos transaccionales, registros de aplicaciones, sensores, datos de redes sociales, archivos de registro, entre otros.

Ingesta de datos:
Se refiere al proceso de recolección y carga de datos desde diversas fuentes hacia el sistema de almacenamiento de big data. Puede incluir técnicas como la ingesta en tiempo real y por lotes.

Almacenamiento de datos:
Involucra el diseño de un sistema de almacenamiento escalable y distribuido para manejar grandes volúmenes de datos. Puede incluir tecnologías como Hadoop Distributed File System (HDFS), sistemas de almacenamiento en la nube, bases de datos NoSQL, entre otros.

Procesamiento de datos:
Incluye un potente procesamiento de datos para análisis, consulta y extracción de datos. Esto puede incluir marcos como Apache Spark, Apache Flink y técnicas de procesamiento por lotes o en tiempo real.

Nivel de servicio:
Proporciona servicios adicionales que facilitan el uso y procesamiento de datos. Esto puede incluir servicios de catálogo, metadatos, control de versiones, etc.

Seguridad y gestión:
Incluye mecanismos de seguridad de datos, control de acceso, cifrado y cumplimiento normativo. También aborda aspectos de gobernanza, como la calidad de los datos y la integridad.

Visualización y análisis

Implica la capa de presentación, donde los usuarios pueden interactuar con los datos a través de herramientas de visualización y análisis. Esto facilita la toma de decisiones informadas basadas en los resultados de los análisis.

Escalabilidad y tolerancia a fallos
Se refiere a la capacidad de la arquitectura para escalar horizontalmente para manejar cargas de trabajo crecientes y para mantener la operación incluso en caso de fallos en algunos componentes.

Integración con tecnologías existentes
Considera la integración de la arquitectura de big data con las tecnologías y sistemas existentes en la organización para garantizar una implementación coherente.

Optimización de costos
El objetivo es optimizar los costos de almacenamiento, procesamiento y operación de grandes cantidades de datos, tomando en cuenta oportunidades como el uso de servicios en la nube, el aprovisionamiento eficiente de recursos y la gestión de costos operativos.

Arquitectura Lambda vs Kappa

Existen varias arquitecturas de big data, cada una diseñada para abordar diferentes necesidades y desafíos en el manejo y análisis de grandes volúmenes de datos. A continuación, comentaremos las dos principales.

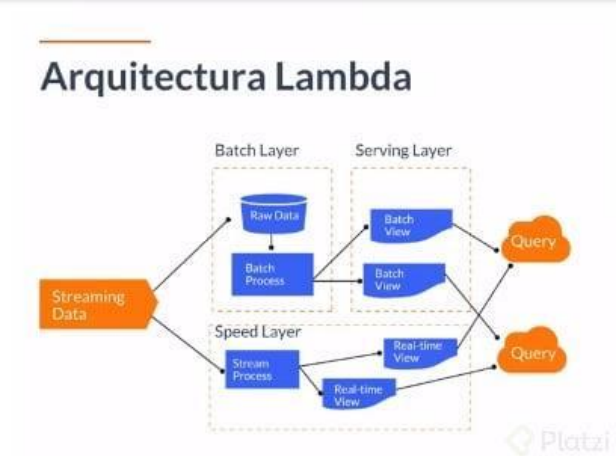
¿Qué es Lambda?

La arquitectura Lambda apareció en el año 2012. Su creador, Nathan Marz, pretendía generar un sistema fuerte que tolerase los fallos, que fuera escalable y con baja latencia.

Su principal característica es la utilización de distintas capas en el procesamiento batch y el streaming, pero, en general, se caracteriza por lo siguiente:

- La información nueva que se recoge se envía a la capa de batch y a la capa de streaming.
- Dentro de la capa batch se produce la gestión de la información en crudo (sin modificar) y los datos nuevos se añaden a los ya existentes. Tras esto se hace un tratamiento a través de un proceso batch cuyo resultado serán las denominadas Batch Views, que se utilizarán en la capa que sirve los datos para ofrecer la información ya transformada al exterior.
- La capa que sirve los datos indexa las Batch Views generadas anteriormente para que puedan ser consultadas con baja latencia.
- La capa de streaming compensa la alta latencia de las escrituras que ocurre en la serving layer y únicamente tiene en cuenta los datos nuevos.
- Por último, la respuesta a las consultas realizadas se produce combinando los resultados de las Batch Views y de las vistas en tiempo, generadas en el anterior paso.

Esquema:



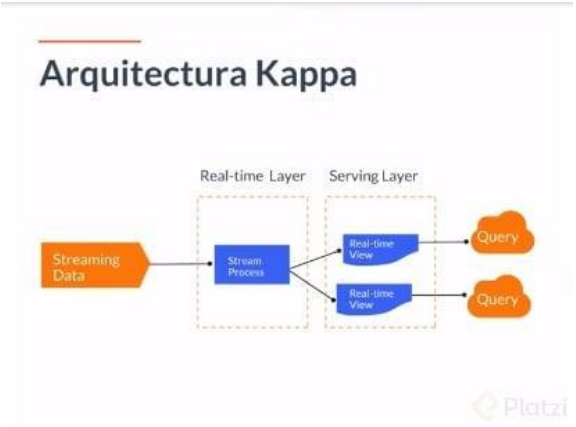
¿En qué consiste Kappa?

Kappa, por su parte, se atribuye a Jay Kreps. Fue introducida en el 2014 y su creador señaló y solucionó algunos puntos débiles de Lambda, tales como eliminación de la capa batch, dejando únicamente la capa de streaming.

Esta evolución de Kappa se basa principalmente en una simplificación de Lambda, donde todo el procesamiento se lleva a cabo en una sola capa denominada de tiempo real o Real-time Layer.

Además, con Kappa no se modifican los datos de partida y existe un solo flujo de procesamiento, provocando que el código, el mantenimiento y la actualización del sistema se reduzcan también.

Esquema:



¿Cuáles son las diferencias entre ambas arquitecturas?

La principal diferencia entre ambas reside en los flujos de tratamiento de datos que intervienen, siendo Kappa mucho más sencilla que su predecesora. Además, Kappa, por su parte, permite construir su sistema de transmisión y procesamiento por lotes en una sola tecnología y ofrece una infraestructura más moderna, con un procesamiento bien distribuido.

Existen otras arquitecturas de Big Data como la de microservicios, grafos, entre otras.

Más información

https://platzi.com/blog/arquitectura-para-big-data-cloud/