ARQUITECTURAS BIG DATA, APACHE SPARK Y KAFKA

Actividad de transferencia

Definiendo Tecnologías

En los proyectos de Big Data y fundamentalmente en aquellos proyectos que son de Big Data Streaming, es decir, de tiempo casi-real, hemos visto que existen arquitecturas por un lado y tecnologías por otro que permiten de alguna forma dar respuesta eficiente a estos proyectos.

Entre las arquitecturas vimos la arquitectura Lambda y la Arquitectura Kappa. Entre las tecnologías, vimos Apache Spark y Apache Kafka. No obstante, es frecuente que en los proyectos de big data se combinen diversas tecnologías.

La actividad de esta semana consiste en armar un **cuadro comparativo**, para que puedas conocer otras tecnologías también muy utilizadas y, sobre todo, para revisar algunas diferencias sutiles que hay entre ellas.

Vamos a comparar las siguientes tecnologías:

- Apache Storm
- Apache Spark
- Apache Samza

Aplicación	¿Para qué se utiliza?	Algunas ventajas	Proyectos ideales para implementar esta tecnología	Ejemplo de empresa que lo utiliza
Apache Storm	Procesamiento de datos en tiempo real y en flujos continuos.	• Baja latencia. • Alto rendimiento	• Análisis de redes sociales en	Twitter

		• Escalabilidad horizontal • Arquitectura basada en topologías flexibles	tiempo real • Monitoreo de sistemas y detección de fraudes • Procesamiento de logs y clics	
Apache Spark	Procesamiento en tiempo real (con Spark Streaming), por lotes y análisis avanzado de grandes volúmenes de datos.	• Soporte para múltiples lenguajes (java, scala, python y R) • Alto rendimiento con procesamiento en memoria • API unificada para batch y streaming	• Machine learning con grandes volúmenes de datos • Procesamiento ETL • Análisis de logs, métricas y comportamiento de usuarios	Netflix, eBay
Apache Samza	Procesamiento de flujos en tiempo real, especialmente con integración nativa a Kafka y YARN.	• Integración nativa con Apache Kafka • Alta tolerancia a fallos • Procesamiento local	• Aplicaciones de monitoreo en tiempo real • Dashboards de	LinkedIn

		(near data processing)	eventos en vivo • Enriquecimiento de datos en flujos Kafka	

Fuente de datos consultada:

- Apache Storm Official Documentation: https://storm.apache.org/
- Apache Spark Official Documentation: https://spark.apache.org/
- Apache Samza Official Documentation: https://samza.apache.org/
- Karau, H., & Warren, R. (2015). *High Performance Spark*. O'Reilly Media.
- Gulisano, V., et al. (2012). *StreamCloud: An elastic and scalable data streaming system*. IEEE Transactions on Parallel and Distributed Systems.
- Zaharia, M. et al. (2016). *Apache Spark: a unified engine for big data processing*. Communications of the ACM.
- Toshniwal, A. et al. (2014). *Storm@Twitter: Scaling distributed stream processing*. VLDB.
- Noghabi, S.A. et al. (2017). *Samza: stateful scalable stream processing at LinkedIn*. Proceedings of the VLDB Endowment.
- Netflix Tech Blog: https://netflixtechblog.com/

La entrega será un documento WORD o PDF con el cuadro comparativo. Debes de estar lo más completo que puedas y, sobre todo, debes colocar las referencias de dónde has obtenido la información.

NOTA: Recuerda utilizar recursos como Google Académico, para acceder a papers o documentos académicos que puedan tener validez para educación.