**Big Data**

**Felipe Meza Obando**

**Tarea 4.**

**Josue Arrieta & Felipe Cubero**

**FlumeJava(Apache Flume):**

Apache Flume es una herramienta de ingestión de datos eficiente, distribuida, confiable y tolerante a fallas. Facilita la transmisión de grandes volúmenes de archivos de registro desde varias fuentes (como servidores web) al Sistema de archivos distribuidos de Hadoop (HDFS), bases de datos distribuidas, como HBase en HDFS, o incluso destinos como Elasticsearch a velocidades casi en tiempo real. Además de transmitir datos de registro, Flume también puede transmitir datos de eventos generados desde fuentes web como Twitter, Facebook y Kafka Brokers. HDFS son las siglas de Hadoop Distributed File System. HDFS es una herramienta desarrollada por Apache para almacenar y procesar grandes volúmenes de datos no estructurados en una plataforma distribuida. Facebook, Yahoo y LinkedIn son algunas de las compañías que confían en Hadoop para su gestión de datos. Los componentes de la arquitectura de trasmisión de datos son: las fuentes de generación de datos, el agente de Flume y el destino o destino. El agente Flume está formado por la fuente Flume, el canal y el sumidero. La fuente Flume recoge archivos de registro de fuentes generadoras de datos como servidores web y Twitter y los envía al canal. El componente sumidero de Flume garantiza que los datos que recibe se sincronicen con el destino, que puede ser HDFS, una base de datos como HBase en HDFS o una herramienta de análisis como Spark.

**Flink Streaming**

Flink Streaming permite analizar fuentes de datos continuas además de archivos estáticos. Utiliza el motor Flink canalizado para procesar flujos de datos en tiempo real y ofrece una nueva API que incluye la definición de ventanas flexibles. Las características de Flink incluyen soporte para procesamiento de flujo y lote, administración de estado sofisticado, semántica de procesamiento en tiempo real y garantías de consistencia exactas para el estado. Además, Flink se puede implementar en varios proveedores de recursos como YARN, Apache Mesos y Kubernetes, pero también como clúster independiente en hardware. Flink puede escalar a miles de núcleos y terabytes de estado de aplicación, ofrece alto rendimiento y baja latencia, y alimenta algunas de las aplicaciones de procesamiento de flujo más exigentes del mundo. Algunos ejemplos de aplicación son los siguientes: Aplicaciones basadas en eventos, aplicaciones de análisis de datos, aplicaciones de canalización de datos y aplicaciones basadas en eventos.

**Beam Streaming (Mixer)**

Mixer es una plataforma de transmisión de videojuegos on line con sede en Seattle. El servicio se lanzó oficialmente el 5 de enero de 2016 como Beam. El servicio tiene un énfasis en la interactividad, con baja latencia de transmisión y una plataforma para permitir a los espectadores realizar acciones que pueden influir en una transmisión. El servicio fue adquirido por Microsoft en agosto de 2016, después de lo cual pasó a llamarse Mixer en 2017 y comenzó a integrarse en la división Xbox de Microsoft. Utiliza un protocolo de transmisión de baja latencia conocido como FTL ("Faster Than Light"), lo que lo hace más apropiado para la interactividad en tiempo real entre un transmisor y sus espectadores. Además, los espectadores pueden usar botones para interactuar con la transmisión (como votar, activar efectos o influir en el juego).

**DataFlow**

Dataflow es un servicio analítico de transmisión totalmente administrado que minimiza la latencia, el tiempo de procesamiento y el costo a través del escalado automático y el procesamiento por lotes. Con su enfoque sin servidor para el aprovisionamiento y gestión de recursos, tiene acceso a una capacidad prácticamente ilimitada para resolver desafíos de procesamiento de datos. Algunas características son: aprovisionamiento y gestión automatizados de recursos de procesamiento, autoescalado horizontal de los recursos de los trabajadores para maximizar la utilización de los recursos, transmisión unificada y modelo de programación por lotes, innovación impulsada por la comunidad OSS con Apache Beam SDK y un procesamiento confiable. Un ejemplo de uso es el análisis de transmisión de Google Cloud, esta herramienta hace que los datos sean más organizados, útiles y accesibles desde el instante en que se generan.