Tarea 4: Investigación Streaming
Curso: BigData
Programa: Ciencia de Datos

Fundatec
2020

Alonso Nuñez Sanchez

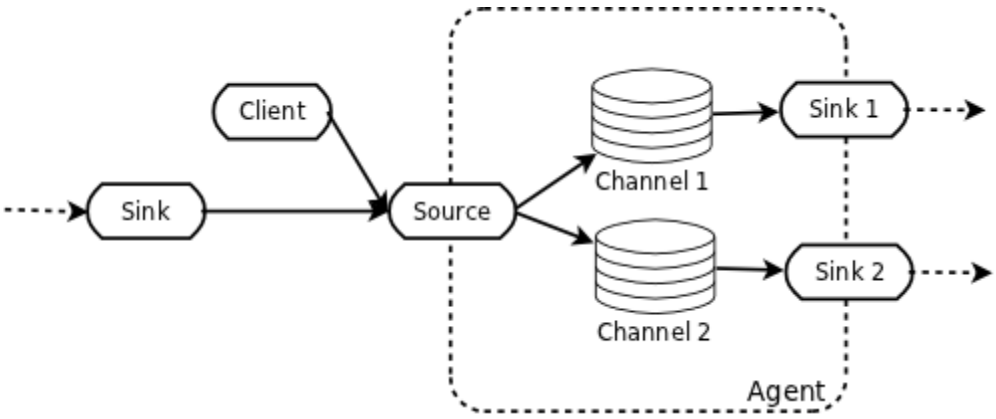
Flume

Flume, es una herramienta que forma parte de Hadoop, cuyo objetivo es el manejo de grandes volúmenes de datos desde diferentes fuentes hacia un repositorio de datos centralizada.

Arquitectura:

- Se basa en flujos de streaming
- Tiene mecanismos para asegurar una alta disponibilidad, tales como asegurar la entrega, failover y recuperación.
- Los datos viajan a través de agentes desde un punto de origen hasta un destina final

Elementos:



Evento:
Datos a transportar desde la fuente hacia el destino

Flujo:
Movimiento de eventos

Cliente:
Elemento que opera en la fuente (punto de origen) y entrega los datos a un agente

Agente:
Proceso que agrupa diferentes componentes (fuente, canales, sinks). Es una unidad de procesamiento dentro del flujo, recibe, almacena y envía eventos.

Fuente:
Recibe eventos y los entrega a los canales

Canal:
Es un almacenamiento de eventos, entregadas por una fuente. El evento permanece en el canal hasta que un sink lo toma, lo elimina del canal y lo entrega al siguiente agente

Sink:
Elimina eventos de los canales, y los retransmite al siguiente agente o al destino final

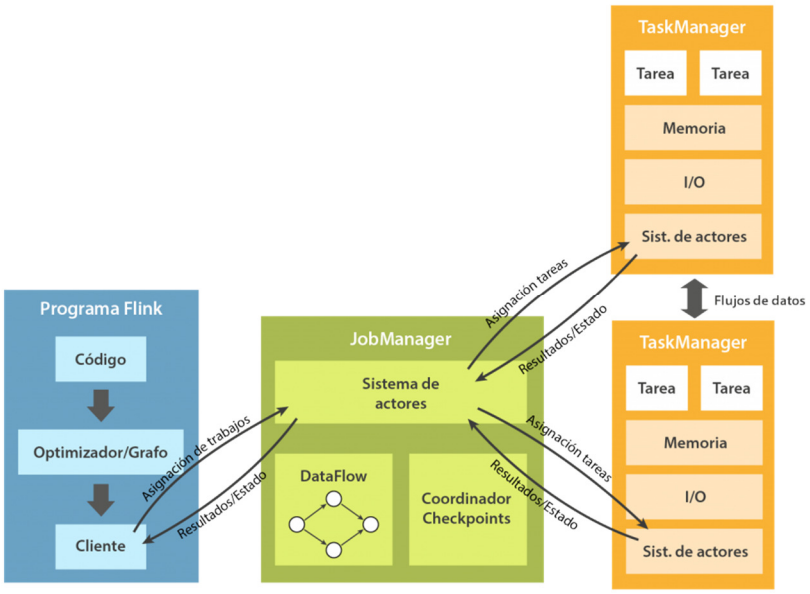
Flink

Flink es un framework de Apache, para el manejo de flujos de datos, ya sea tipo batch o streaming

Arquitectura

Está diseñado para trabajar con grandes volúmenes de datos y diversas arquitecturas específicas tales como Hadoop YARN, Apache Mesos y Kubernetes, pero también como un clúster independiente (stand-alone) lo que permite integrarse con diferentes entornos.

Trabaja con procesos distribuidos (varios task manager) gestionados por un coordinador (job manager)



Características:

- Baja latencia (resultados en milisegundos).
- Alto throughput (millones de eventos por segundo).
- Consistencia (resultado correcto en caso de errores).
- Tolerancia a fallos a través de un sistema de snapshots distribuidos.
- Eventos desordenados (procesamiento de eventos en función de un tiempo asociado).
- Sistema de ventanas de streaming muy flexible.
- Un único sistema para procesar batch y streaming.
- APIs intuitivas multilenguaje (Scala, Python y Java)

Beam

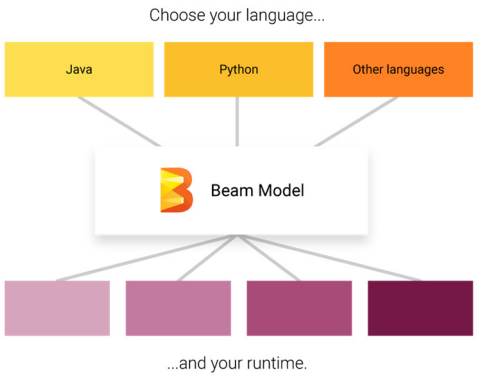
Beam es un framework para procesamiento de grandes volúmenes de datos
Puede manejar datos tipo batch y flujos en tiempo real, sin tener que cambiar de código, o de librerías

Se puede ejecutar en diferentes lenguajes, usando el SDK correspondiente, y así definir un pipeline para el procesamiento de datos
SDKs disponibles:

- Java
- Python
- Go
- Scala

Un programa Beam escrito en cualquiera de los lenguajes mencionados, puede ser ejecutado en diferentes “Beam Pipeline Runners”:

- Apache Apex
- Apache Flink
- Apache Gearpump
- Apache Samza
- Apache Spark
- Google Cloud Dataflow
- Hazelcast Jet



Conceptos:

Pipeline:
Flujos de procesamiento que encapsulan las tareas o el trabajo de inicio a fin

PCollection:
es una abstracción de conjunto de datos distribuida utilizada para para transferir datos entre PTransforms. Es un contenedor de datos

PTransform:
es un proceso que recibe datos de entrada (PCollection de entrada) y genera datos de salida (PCollection de salida).

Google Cloud Dataflow

- Dataflow es un Servicio de Google, para procesar datos
- Está basado en Beam
- Se puede decir que es uno de los “Beam Pipeline Runners” mencionados anteriormente, con la ventaja de ser un servicio administrado con todas las bondades ofrecidas por los servicios de Google
- Se puede escalar hacia arriba o hacia abajo según los volúmenes de datos que se requiera procesar, de esta forma, solo se paga por lo que se usa
- Además, al formar parte de la plataforma/ecosistema Cloud de Google, se puede integrar con otros servicios o herramientas, por ejemplo hacer uso de repositorios prácticamente ilimitados para recibir entradas o generar salidas de pipelines Beam

Ejemplo de cómo integrar Dataflow con otras herramientas más allá del procesamiento: