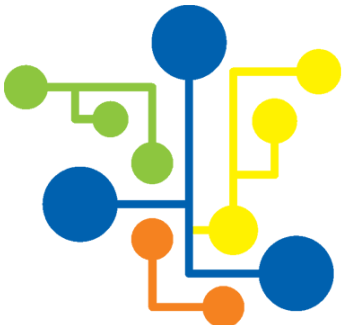

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Flume Sink

Semelhante ao Source, o Sink (ou em português, coletor) é gerenciado pelo *Sink Runner*, que gerencia o modelo de thread e execução. Diferentemente de um Source, no entanto, um coletor é Polling-based e pesquisa o canal em busca de eventos.

O coletor é o componente que gera uma saída (de acordo com o tipo de saída necessário) do agente para um destino externo (fora do Flume). Os coletores também participam do gerenciamento de transações e, quando a saída de um coletor é bem-sucedida, uma confirmação é passada de volta para o canal, que então retira o evento do mecanismo de persistência (o gerenciamento de transações será abordado em detalhes mais adiante). Há uma variedade de Sinks Built-in disponíveis no Flume:

- HDFS: grava no HDFS. Atualmente suporta a gravação de arquivos de texto e sequência (em formato compactado também)
- HBase: grava no HBase
- AsyncHBase: grava no HBase de forma assíncrona
- Hive: grava texto ou JSON em tabelas Hive
- Kafka: pode publicar o evento em um tópico do Kafka (usaremos este Sink em nossas atividades práticas daqui a pouco).

A exemplo do Source e Channel, também é possível criar um Sink customizado usando linguagem Java.

Referências:

https://flume.apache.org/FlumeDeveloperGuide.html