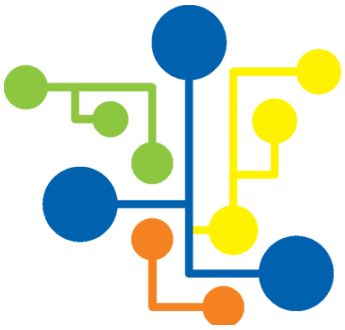


Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Um Pouco de História do Flume

Arquivos de logs são uma fonte básica de informação e ajudam a monitorar a 'saúde' de sistemas, detectando falhas em hardware e serviços. Podem ajudar também na solução de problemas. Embora, de uma forma geral, um log represente registro de eventos, esta é uma ferramenta importante na administração de sistemas.

Com o surgimento de tecnologias e ferramentas para Big Data, a geração e utilização dos logs se tornaram cada vez mais importantes, uma vez que agora é possível 'prever' quando um equipamento precisa de manutenção, o comportamento de um indivíduo em um e-commerce e a geração de recomendações, entre outras possibilidades.

Administrar o volume de logs que é gerado por diversos sistemas/ dispositivos não é uma tarefa fácil. Em 2011, a Cloudera criou o Flume, um sistema distribuído, confiável e disponível para coletar, agregar e mover grandes quantidades de dados de muitas fontes diferentes para um armazenamento de dados centralizado. Em 2012, esse passou a ser um projeto top level na Apache Software Foundation.

Muito semelhante ao Sqoop, Flume também tem duas versões principais:

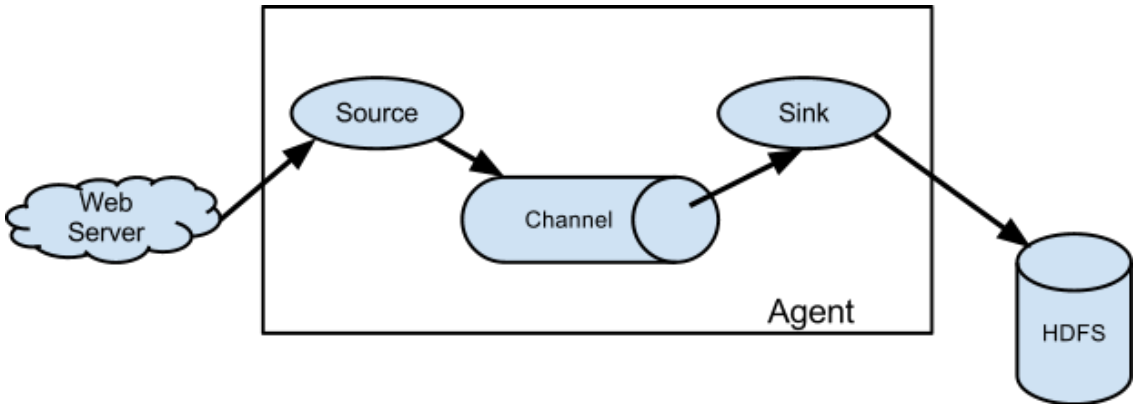
- Flume OG (Old Generation: pre 1.0)
- Flume NG (New Generation)

Como o nome sugere, o Flume OG foi a distribuição inicial do Flume, que foi então submetida a reescrita e refatoração completas dando origem ao Flume NG. O Flume NG é o atual projeto Apache suportado e ativo e essa é a versão que usaremos neste capítulo.

O Apache Flume não se restringe apenas à coleta de logs, e pode ser usado para transportar grandes quantidades de dados, como os gerados em social media, e-mails e qualquer fonte de dados possível.

O objetivo principal do Flume é ingerir dados de eventos no HDFS (Hadoop Distributed File System) de forma simples e automatizada. Porém, seu uso não se limita apenas ao HDFS; é possível enviar também dados para um arquivo ou banco de dados, entre outros.

Abaixo, o modelo de fluxo de dados do flume:



Um agente Flume roda na JVM (Java Virtual Machine) e possui os seguintes componentes:

- **Source**: responsável pela entrada de dados;
- **Channel**: armazena os dados que passam do source para o sink. Seu comportamento é parecido com uma fila;
- **Sink**: responsável por enviar os dados ao destino/ saída. A saída pode ser outro agente Flume.

A configuração de um agente é feita por meio de um arquivo local que tem o formato de um arquivo properties utilizado em Java, muito similar aos arquivos de configuração do HDFS.

Apache Flume é uma ferramenta flexível, podendo ser usada para diversas situações, o que a torna bem poderoso apesar de sua simplicidade.

Referência:

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