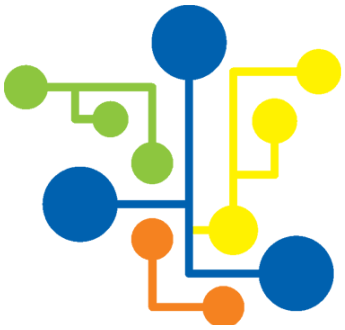


Data Science Academy

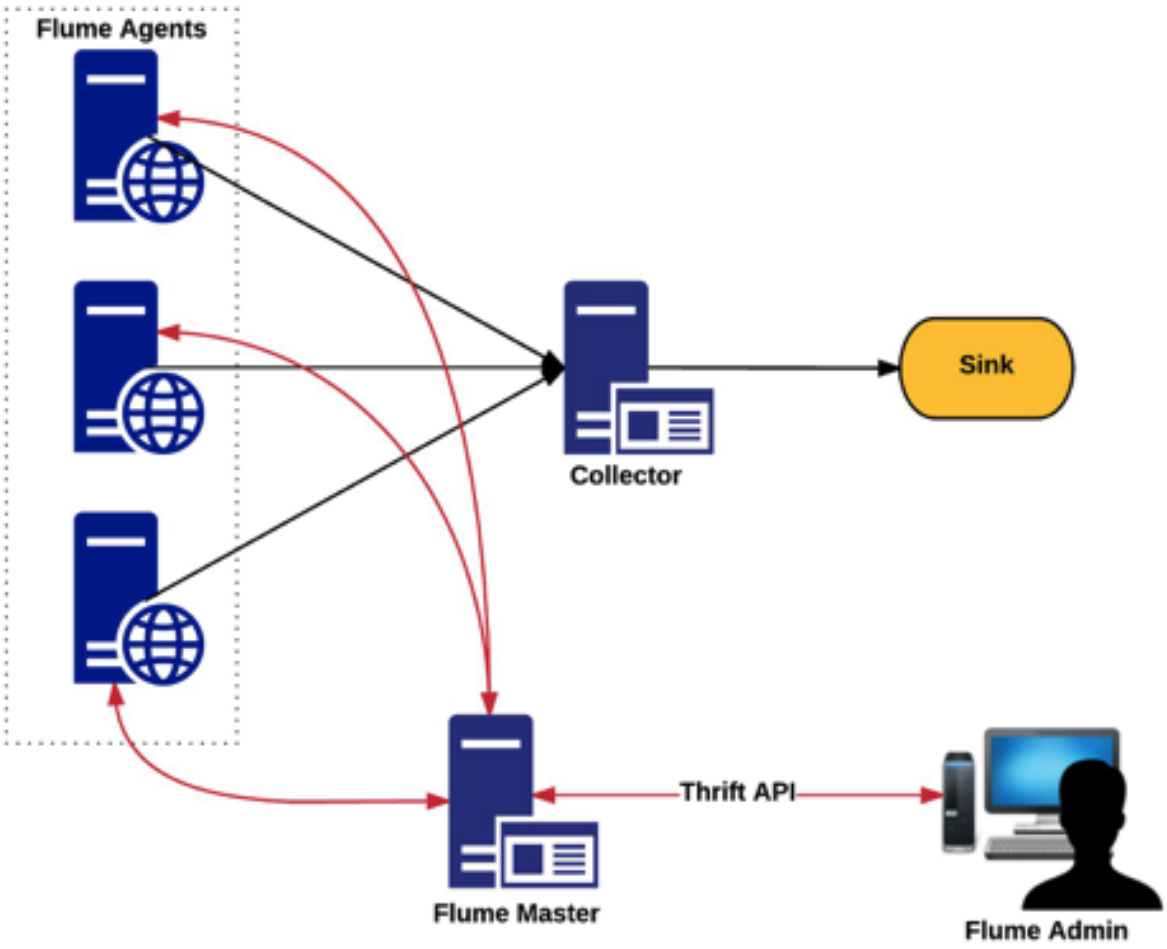
www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Flume
Arquitetura Avançada

A arquitetura Flume em um ambiente corporativo terá mais um componente importante: o Flume Master, cuja principal responsabilidade é servir como autoridade centralizada para todas as configurações de todos os nós na arquitetura geral.

Cada máquina participante da arquitetura é denominada nó. Cada nó depende do mestre para recuperar as configurações que determinam como o Flume deve executar suas ações. Toda a topologia Flume pode ser configurada ou reconfigurada dinamicamente enviando comandos relevantes usando a API Thrift para o mestre Flume. A figura abaixo mostra uma visão geral da arquitetura Flume, que demonstra o mestre controlando a configuração e ditando o funcionamento geral e a topologia:



A confiabilidade é uma das bases sobre a qual a arquitetura Flume foi projetada. Para atingir este nível de confiabilidade, o Flume fornece ao usuário níveis de confiabilidade configuráveis. Eles são classificados da seguinte forma:

End-to-end: Quando esse nível de confiabilidade é definido, o evento enviado para o Flume certamente chegará ao outro extremo, desde que o agente de origem esteja ativo. Para atingir esse nível de confiabilidade, o agente ao receber um evento armazena/grava no disco no WAL (Write Ahead Log). Quando o evento atinge o endpoint definido e o reconhecimento (acknowledgement) é enviado até o agente de origem, o evento gravado é apagado. Este nível pode suportar a falha de qualquer componente após o agente de origem. Mas cuidado! Quanto maior a confiabilidade, menor é a escalabilidade e isso se enquadra no nível mais alto de confiabilidade oferecido pelo Flume.

Store on Failure: Quando esse nível de confiabilidade é definido, o evento ao passar por agentes diferentes (saltos), o agente de origem do evento armazenará/gravará no disco somente se o agente para o qual o evento foi enviado falhar. O agente principal apenas grava no disco o detalhe do evento, se não houver confirmação do próximo salto no agente. Este é um nível de confiabilidade mais prático, mas se houver falhas silenciosas, os eventos podem ser perdidos para sempre.

Best-effort: Este nível de confiabilidade é o mais fraco e o mais leve no qual o evento é enviado para o próximo salto sem gravar no disco e não depende de nenhum reconhecimento ou falha que retorne do próximo agente para o qual o evento foi enviado.

Escolha o nível de confiabilidade correto que seu caso de uso exige e tenha sempre em mente que quanto maior a confiabilidade, menor a escalabilidade e maior o custo de manutenção.

Vejamos em mais detalhes o funcionamento do Flume e sua matéria-prima, o streaming de dados.