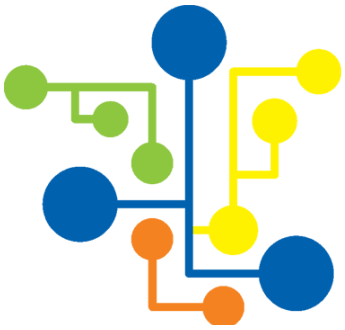

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Flume Source

O agente Flume pode ter várias origens (fontes) de dados, mas é obrigatório ter pelo menos uma fonte para funcionar. A origem é gerenciada pelo *Source Runner*, que controla o aspecto de segmentação e os modelos de execução:

- Orientado a eventos (Event-driven)
- Polling

No modelo de execução orientado a eventos, a origem ouve e consome eventos. No modelo de execução Polling, a fonte mantém uma espécie de *pesquisa* de eventos e, em seguida, lida com essa pesquisa. O evento pode ter uma variedade de conteúdo que satisfaça o esquema do evento (header e payload).

A fonte, em conformidade com o princípio de arquitetura de extensibilidade, funciona na abordagem de plug-in e requer nome e tipo obrigatórios. De acordo com o tipo, a fonte exigirá parâmetros adicionais e, consequentemente, as configurações devem ser definidas para que funcionem corretamente. A fonte pode aceitar um único evento ou um lote de eventos.

Aqui estão as fontes Built-in já disponíveis no Flume e que podem ser usadas através de configuração de arquivos (outras fontes vão requerer desenvolvimento em Java):

Fontes Assíncronas – O cliente que envia os eventos não lida com a falha. Uma vez que o evento é enviado, o cliente esquece (envia o evento e descarta). Alguns dos exemplos são os seguintes:

- **Exec** - executa o comando e ingere a saída como dados. Exemplo de configuração desse tipo de fonte:

  agent.sources.execSource.type = exec

- **Syslog (diretório de spooling)** - analisa o arquivo e ingere os dados. A seguinte configuração é apenas um exemplo de como um diretório de spooling pode ser configurado:

  agent.sources.spool.spoolDir = /data/customerdata

Fontes Síncronas – Depois que o evento é enviado e se a fonte não confirma para o cliente, o cliente poderá lidar com os cenários de falha normalmente. Alguns dos exemplos são:

- **JMS** - são eventos produzidos e manipulados pelo Java Messaging Service (Filas e Tópicos). A configuração de amostra para se conectar a uma fila é a seguinte:

  agent.sources.jms.type = jms
  agent.sources.jms.providerURL = tcp://datalakeserver:61616
  agent.sources.jms.destinationName = customerData
  agent.sources.jms.destinationType = customerDataQueue

- **HTTP** - inicia um servidor da Web para manipular a API REST. Esta é uma configuração de exemplo:

  agent.sources.http.port-8080

Se o seu caso de uso exigir uma fonte especial, uma fonte customizada poderá ser escrita implementando a interface de origem. Um exemplo de como essa fonte personalizada (classe com.nomepacote.CustomSource) pode ser configurada para um agente ag1 é a seguinte:

ag1. fontes = src1
ag1. canais = ch1
ag1. sources.src1.type = com.nomepacote.CustomSource
ag1. sources.src1.canais = ch1

Fontes customizadas (ou seja, que não estejam disponíveis no Flume no formato Built-in) requerem o desenvolvimento em linguagem Java.

Referências:

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