FIAP – FACULDADE DE INFORMÁTICA E ADMINISTRAÇÃO PAULISTA

BRUNO BIANCCHI – RM 84351

LUIS HENRIQUE CALDAS ALTERO – RM 88670

PEDRO GUILHERME POLLONI BARRETO - RM 88964

VITOR LAMPRECHT – RM 86691

PESQUISA SOBRE SOLUÇÃO DE EXTRAÇÃO DE DADOS

São Paulo

2021

Alternativas Para Carga de Dados

Carga de Dados com Flume

Flume é uma ferramenta do apache hadoop que coleta uma grande quantidade de dados e os envia para o HDFS (Hadoop Distributed File System), sendo geralmente usado para dados não estruturados. Um agente Flume roda em uma Java Virtual Machine e possui 3 componentes:

- Source: responsável pela entrada de dados;
- Channel: armazena os dados que passam do source para o sink (parecido com uma fila de espera).
- Sink: responsável por enviar os dados ao destino/ saída.

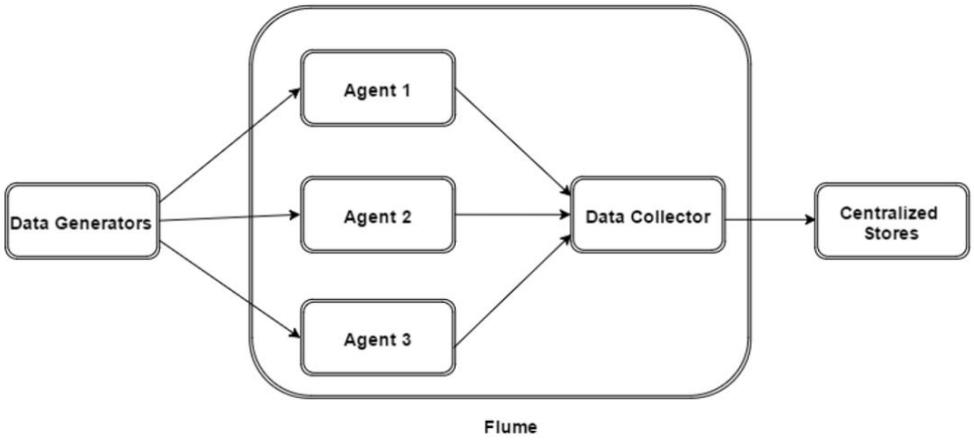

Figura 1 - Arquitetura do Flume
Fonte: https://www.codigofluente.com.br/wp-content/uploads/2018/05/Flume-Arquitetura.png

Os pontos de interesse no Flume são a sua capacidade de transportar dados não relacionais para o HDFS (Hadoop Distributed File System) de forma simples e automatizada. Lembrando que seu uso não se limita apenas ao HDFS, sendo possível enviar também dados para um arquivo ou banco de dados e entre outros. O Apache Flume é uma ferramenta flexível que tem um sistema distribuído, confiável e disponível, tornando-a uma ferramenta simples e poderosa.

Carga de Dados com Kafka

O Kafka é uma plataforma que faz o movimento de uma quantidade imensa de dados para entregá-los a vários destinos desejados ao mesmo tempo, sendo esses dados relacionais ou não. De maneira simplificada, o Kafka faz a intermediação entre a coleta de dados e sua entrega para aplicação que os consumirá, como visto na imagem a seguir:

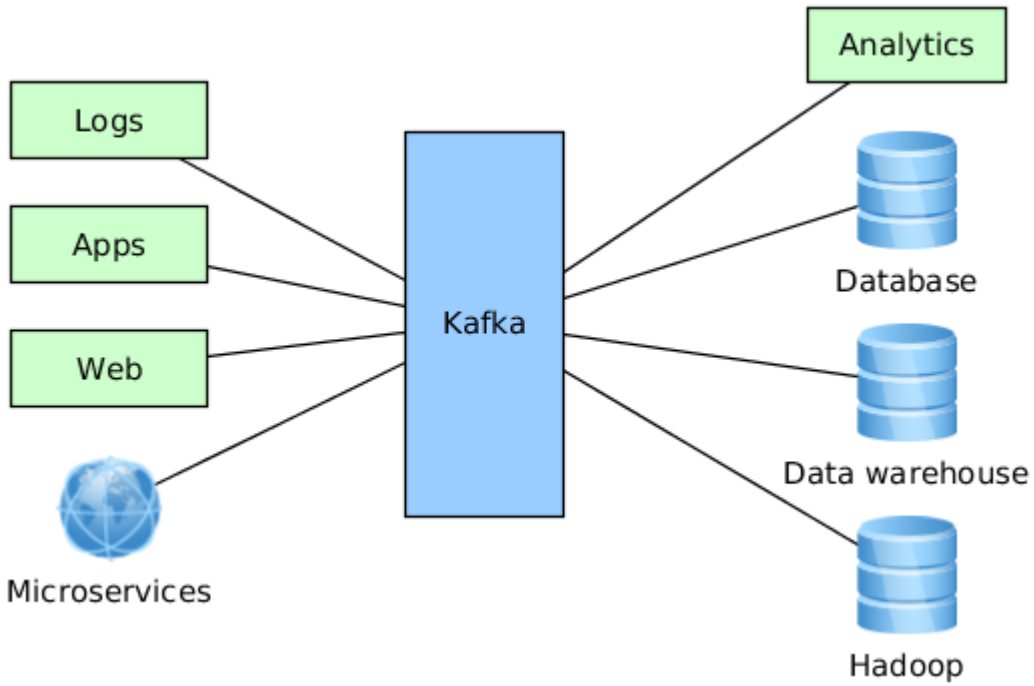

Figura 2 – Coleta de Dados Kafka
Fonte: https://blog-geek-midia.s3.amazonaws.com/wp content/uploads/2020/09/30200240/image3.png

Sua arquitetura funciona a base de producers, consumer e clusters.

- Producer: responsável por publicar as mensagens no cluster
- Cluster: responsável por gravar as mensagens recebidas do producer nos brokers (servidores que gerenciam diversos tópicos)
- Consumer: receptor das mensagens do Kafka

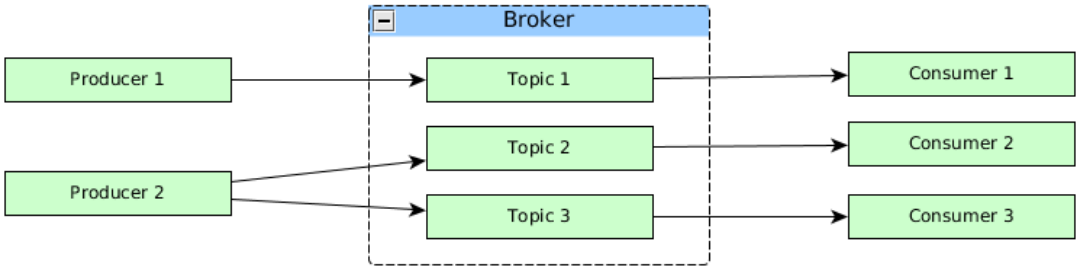

Figura 3 - Arquitetura Kafka
Fonte: https://blog-geek-midia.s3.amazonaws.com/wp-content/uploads/2020/09/30200333/image1.png

Os pontos de interesse no Kafka são suas características (escalabilidade, distribuição, ordenação e alta disponibilidade.) e a sua capacidade de transportar dados não relacionais para o HDFS (Hadoop Distributed File System) e para outro sistema qualquer de forma rápida, organizada e eficaz. Sendo organizado pelo apache zookeeper que é responsável por armazenar os metadados do cluster (em caso de falha, o zookeeper elege um substituto e recupera a operação em questão).

REFERÊNCIAS

https://blog.geekhunter.com.br/apache-kafka/

https://www.redhat.com/pt-br/topics/integration/what-is-apache-kafka

https://medium.com/@gabrielqueiroz/o-que-%C3%A9-esse-tal-de-apache-kafka-a8f447cac028

https://www.codigofluente.com.br/ingestao-de-dados-com-o-flume/

https://imasters.com.br/desenvolvimento/introducao-ao-apache-flume

https://www.dezyre.com/article/sqoop-vs-flume-battle-of-the-hadoop-etl-tools-/176