DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

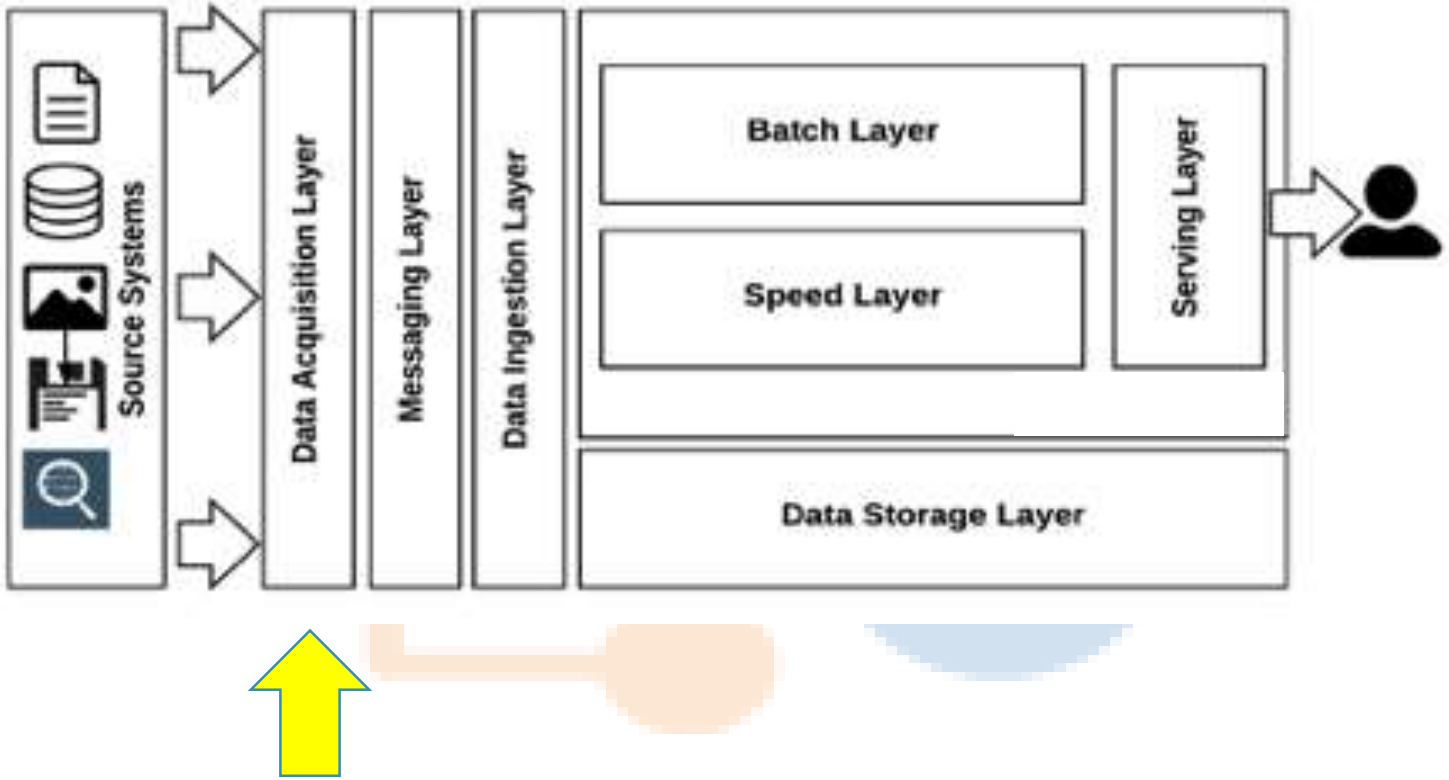
Data Lake
Aquisição de Dados em Streaming

Data Lake

Aquisição de Dados em Streaming

Aquisição de Dados em Batch (Cap05)

Aquisição de Dados em Streaming (Cap06)



Data Lake

Aquisição de Dados em Streaming

Aquisição de Dados em Batch (Cap05)

Aquisição de Dados em Streaming (Cap06)

Data Lake

Aquisição de Dados em Streaming

Não iremos instalar o Hadoop e nem criar outro cluster. Vamos utilizar o cluster criado nos capítulos anteriores.

As atividades práticas serão realizadas na nuvem AWS, mas você pode usar o cluster criado com as máquinas virtuais no VirtualBox, se preferir.

Contexto no Data Lake
Aquisição de Dados

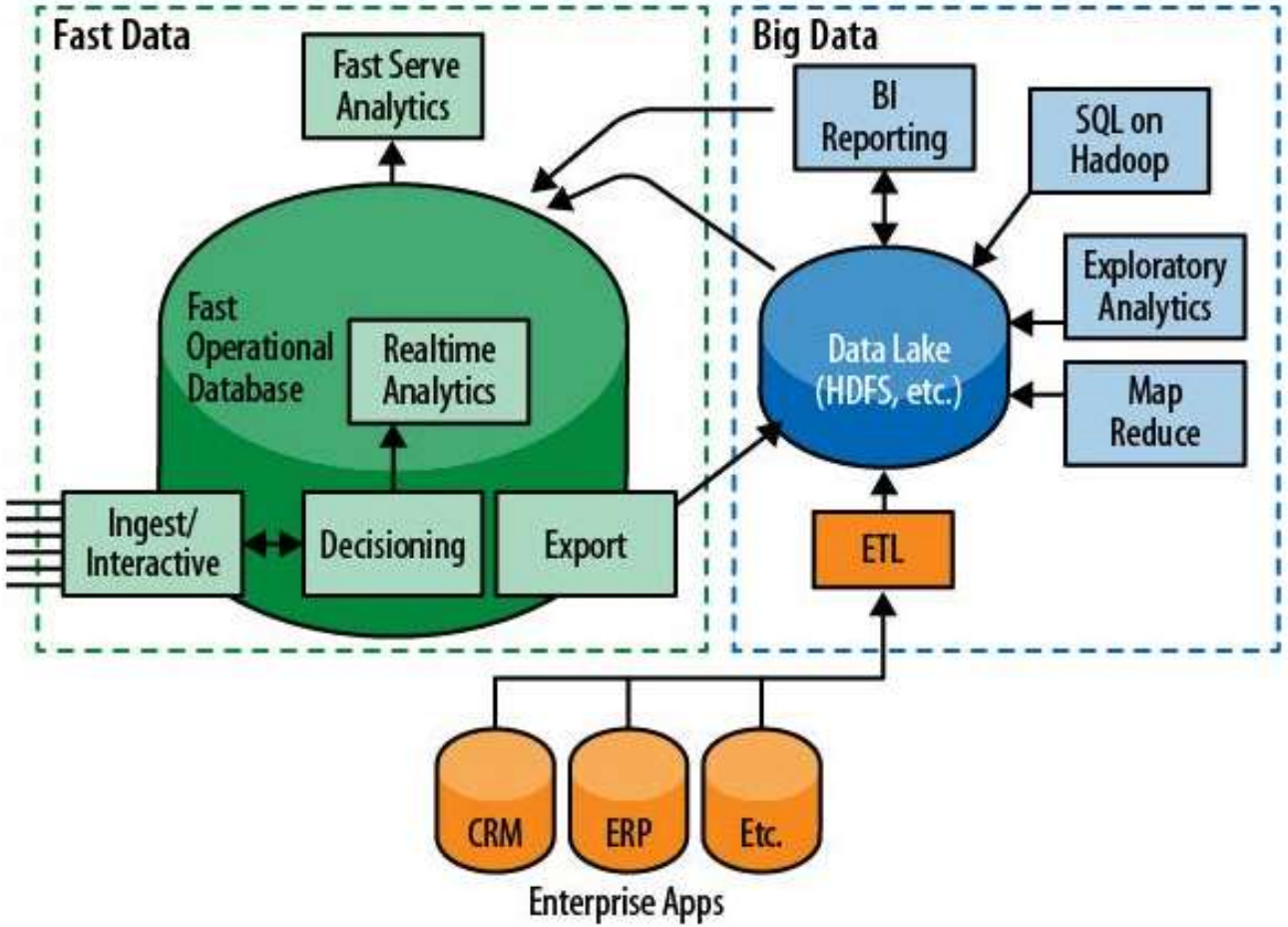
Contexto no Data Lake
Aquisição de Dados

O Big Data é definido pelos seus 4 Vs:
Volume, Variedade, Velocidade e Veracidade

A aquisição de dados em batch trata essencialmente do volume, enquanto a aquisição de dados em streaming trata da velocidade.

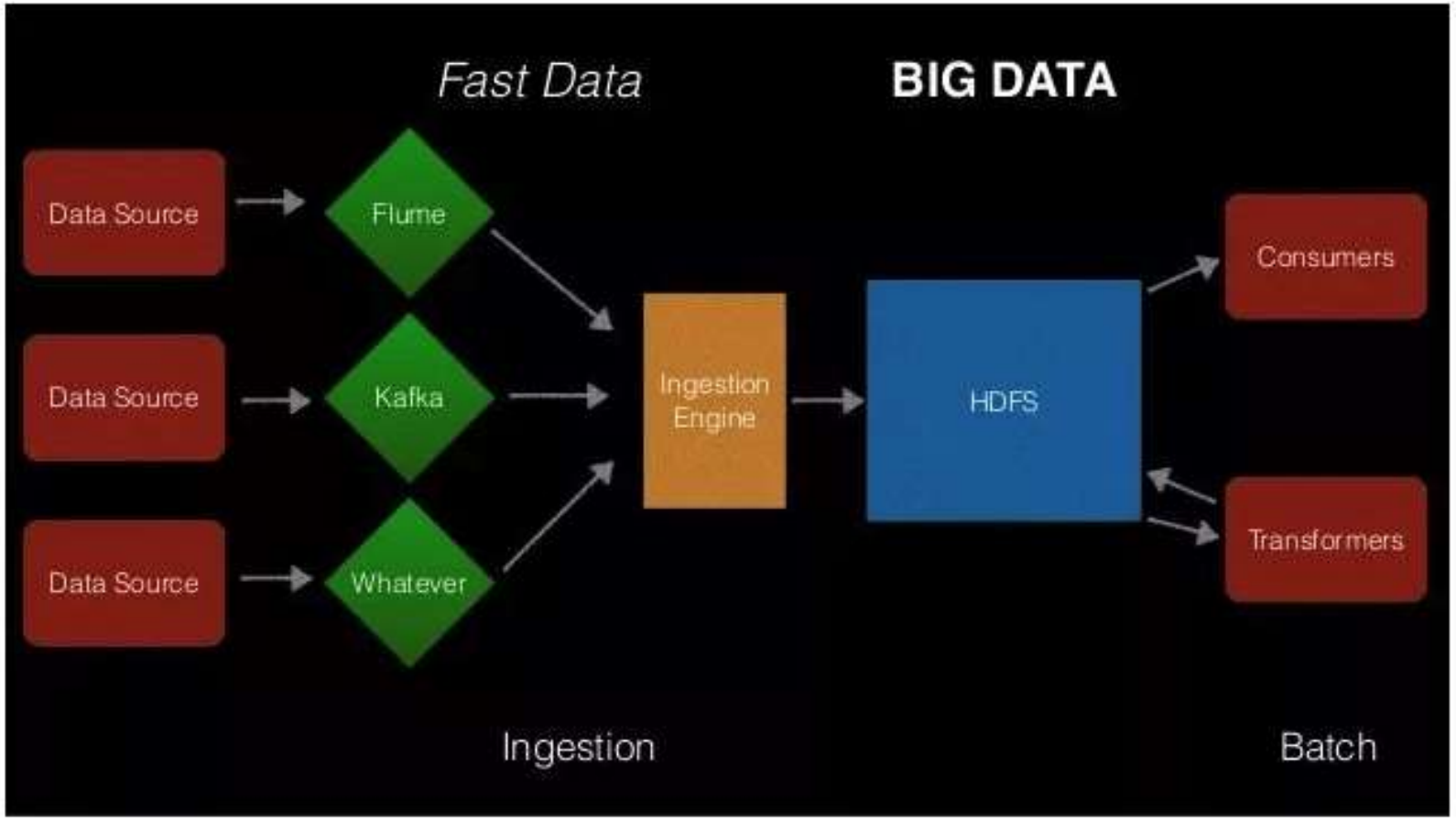
Contexto no Data Lake

Aquisição de Dados



Contexto no Data Lake

Aquisição de Dados



O Que é Streaming de Dados?

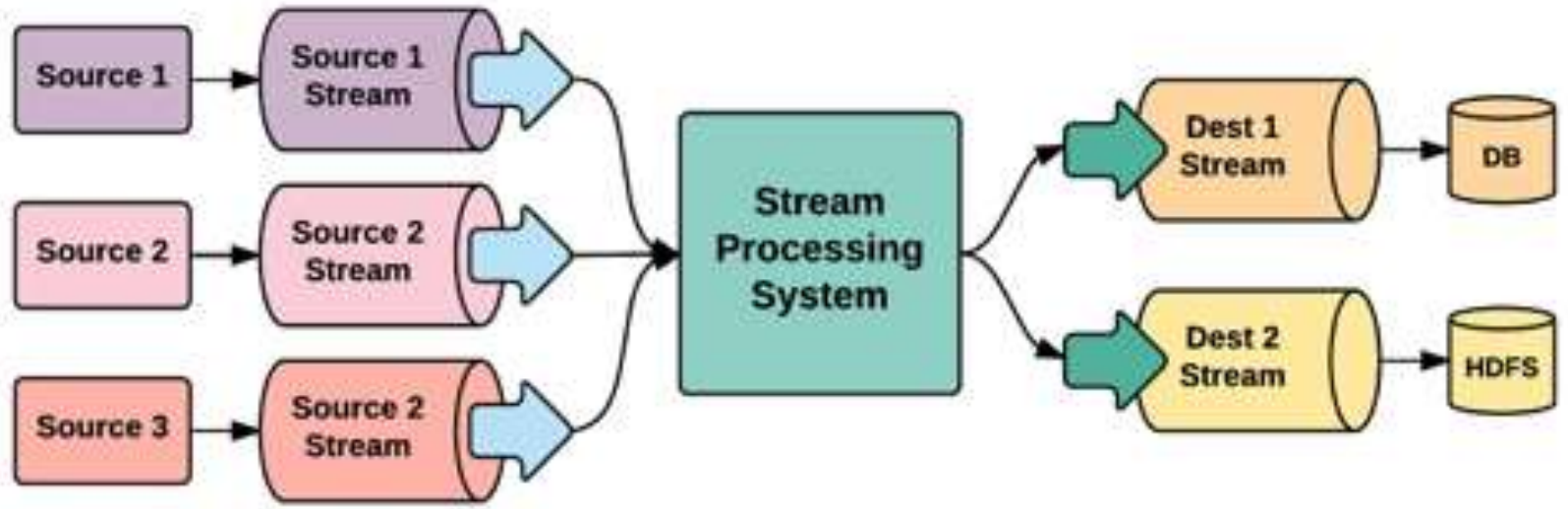
O Que é Streaming de Dados?

Dados em streaming são dados gerados continuamente por milhares de fontes de dados, que geralmente enviam os registros de dados simultaneamente, em tamanhos pequenos (na ordem de kilobytes).

O Que é Streaming de Dados?

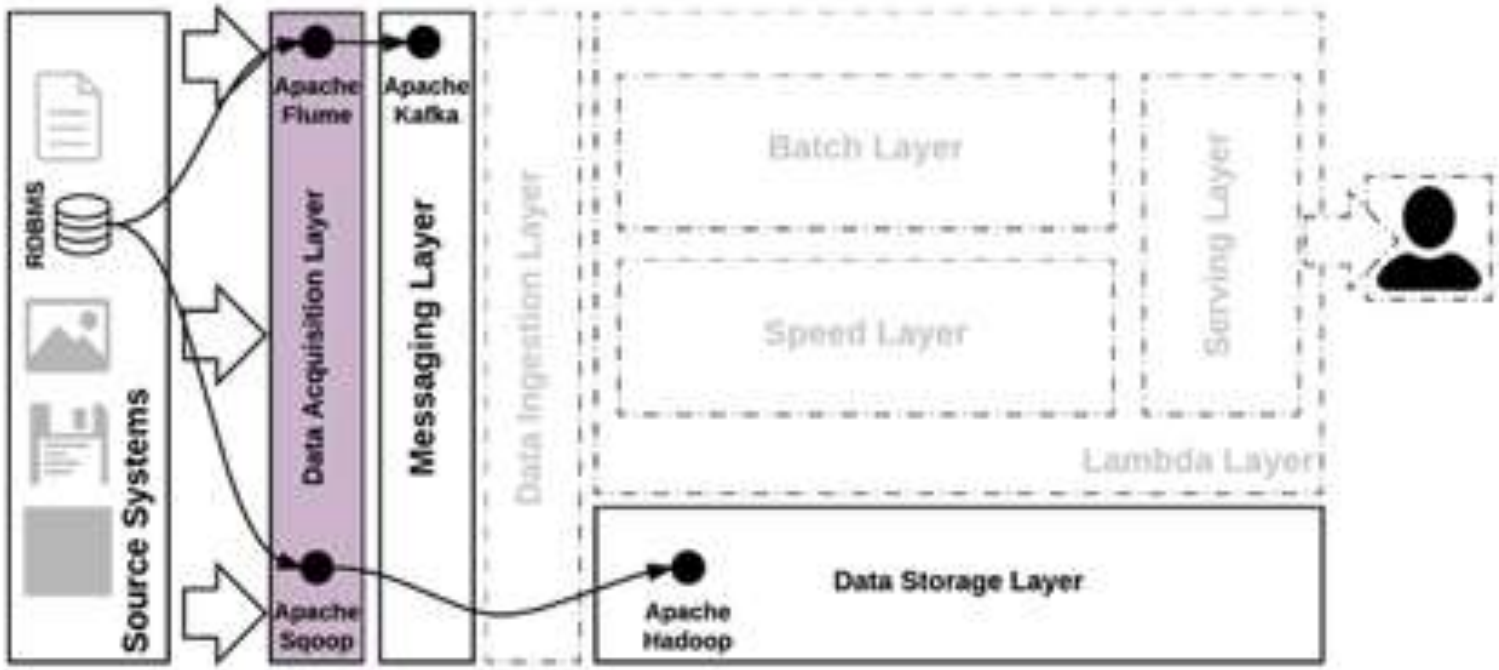
- Dados gerados por plataformas de redes sociais.
- Dados de arquivos de logs de servidores (web, e-mail, aplicações).
- Dados gerados pelo comportamento do usuário em um website (cliques, impressão de páginas).
- Dados de sensores e plataformas IoT.
- Cotações de ações e pregões financeiros.
- Serviços geoespaciais.

O Que é Streaming de Dados?

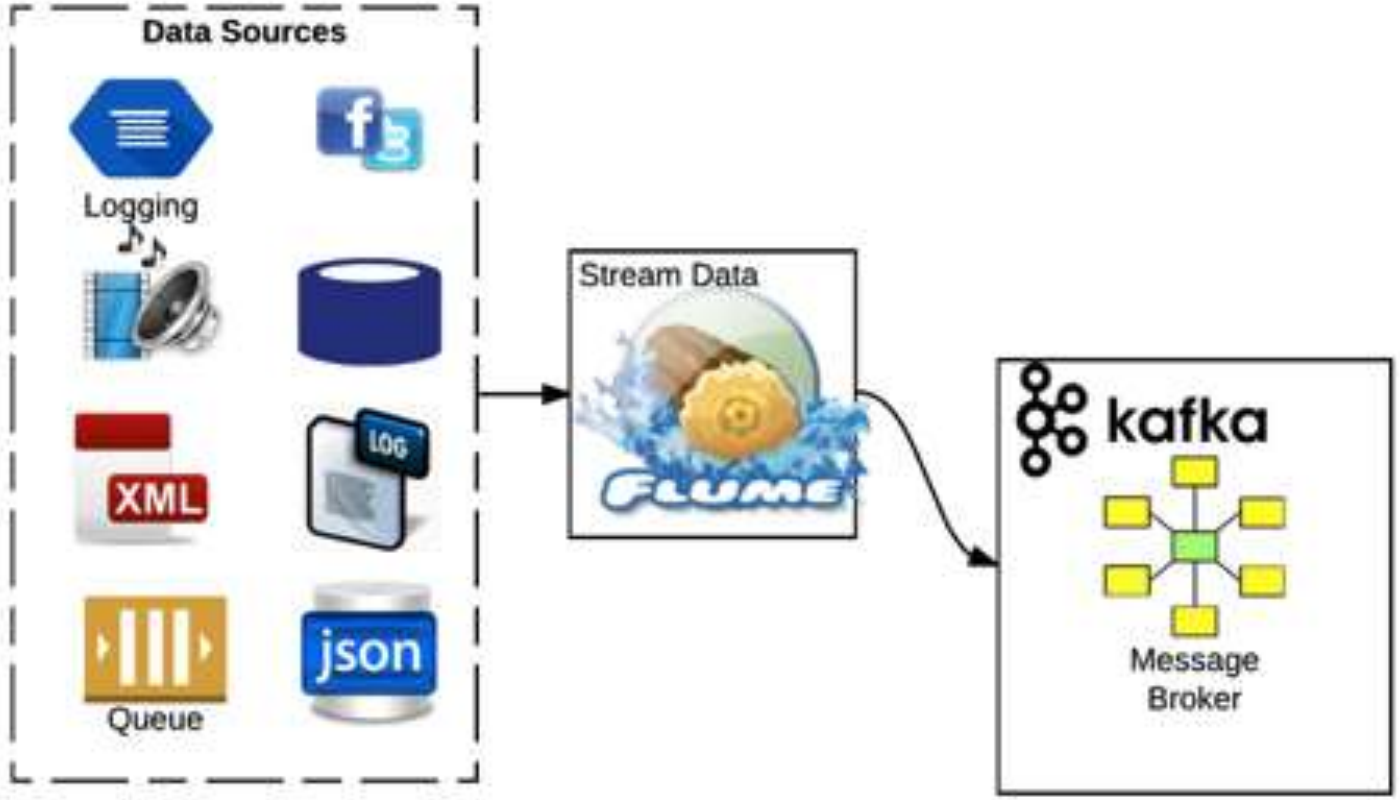


Como Vamos Processar Dados de Streaming?

Como Vamos Processar Dados de Streaming?



Como Vamos Processar Dados de Streaming?

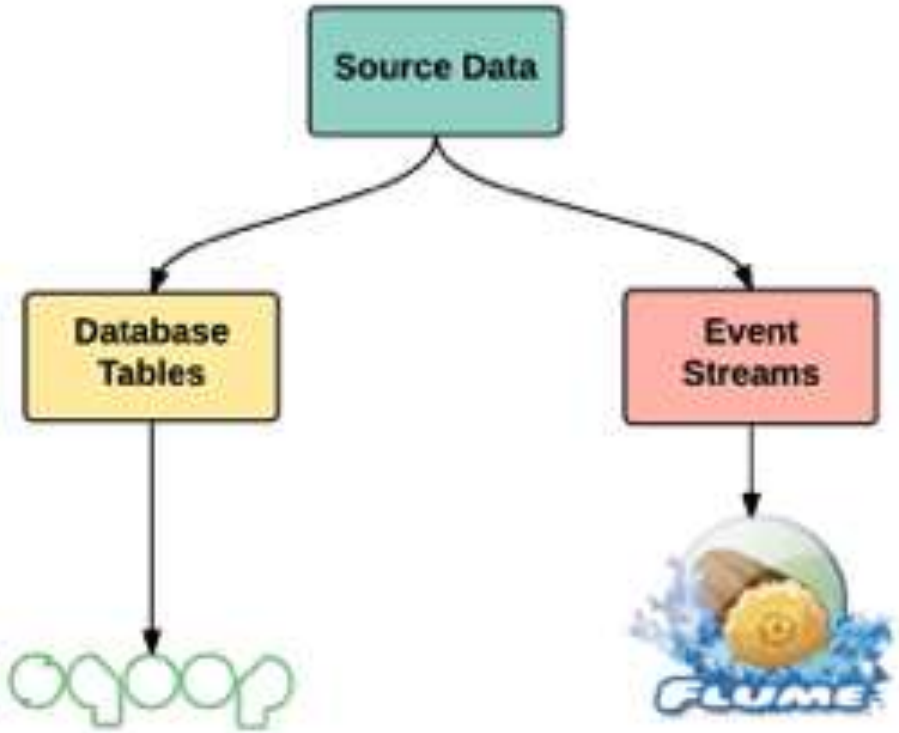


Sqoop x Flume

Sqoop x Flume

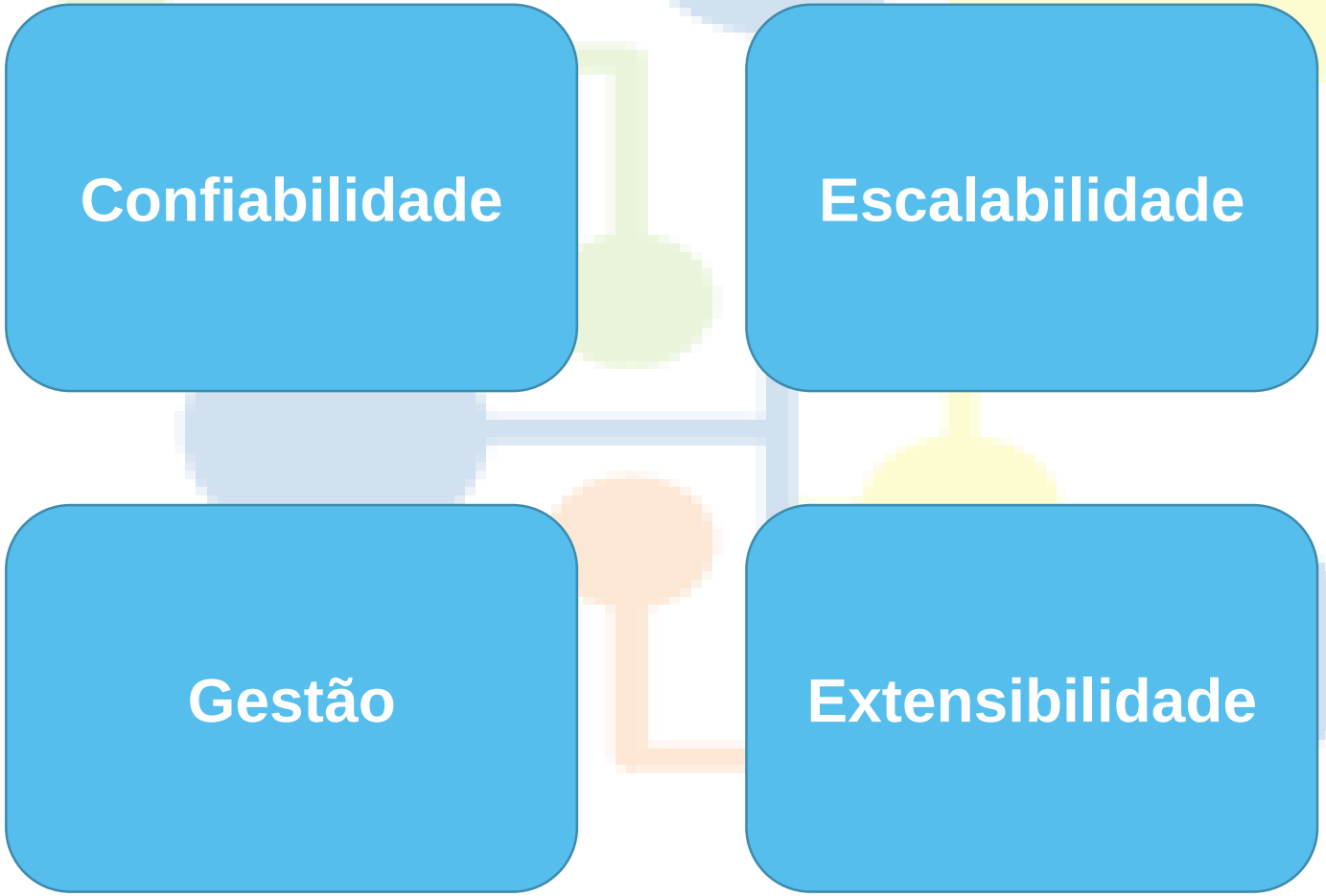
Sqoop	Flume
Dados em Batch	Dados em Streaming
Dados estáticos	Dados em movimento
Para grandes quantidades de dados (de GB a TB)	Para dados gerados em alta velocidade (de KB a MB)
Transfere dados de RDBMS para o HDFS	Transfere streaming de dados de diversas fontes para o HDFS
Coleta dados normalmente já agregados	Permite agregar dados durante a coleta
Pode gerar overhead na fonte de dados	Não gera overhead na fonte de dados

Sqoop x Flume



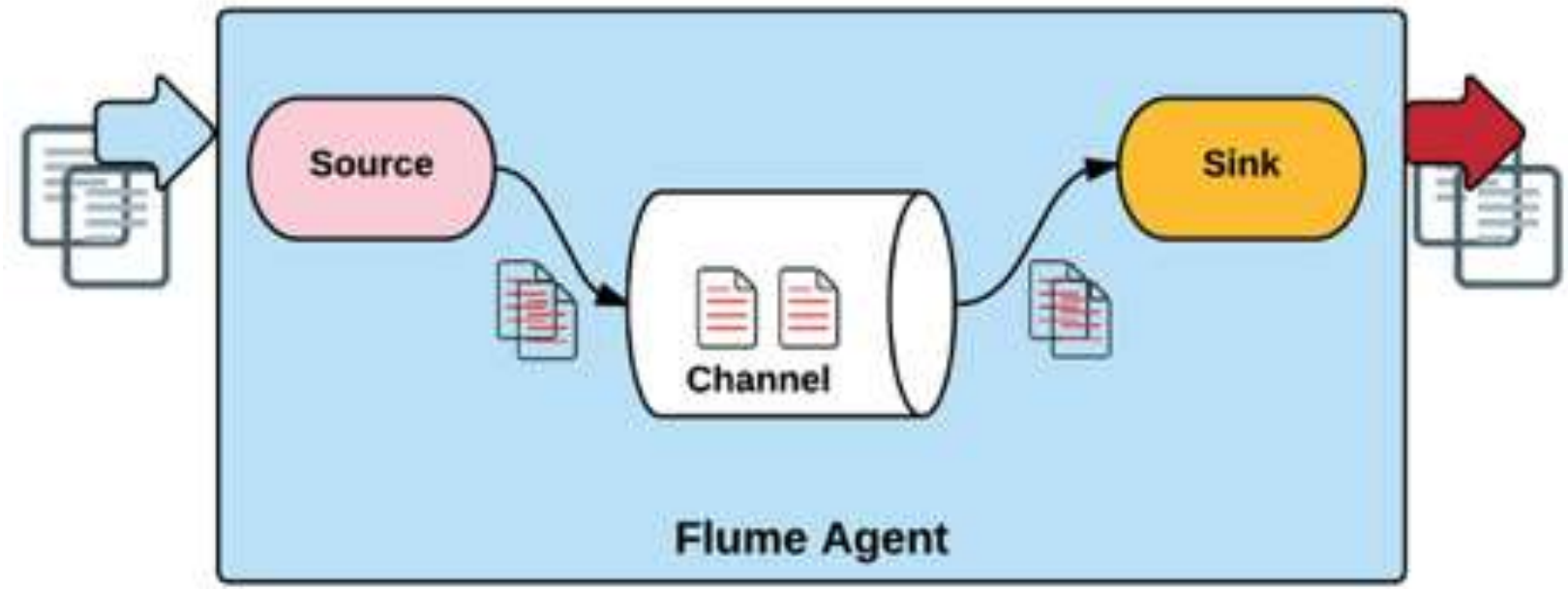
Principais Características do Flume

Principais Características do Flume



Arquitetura Flume

Arquitetura Flume

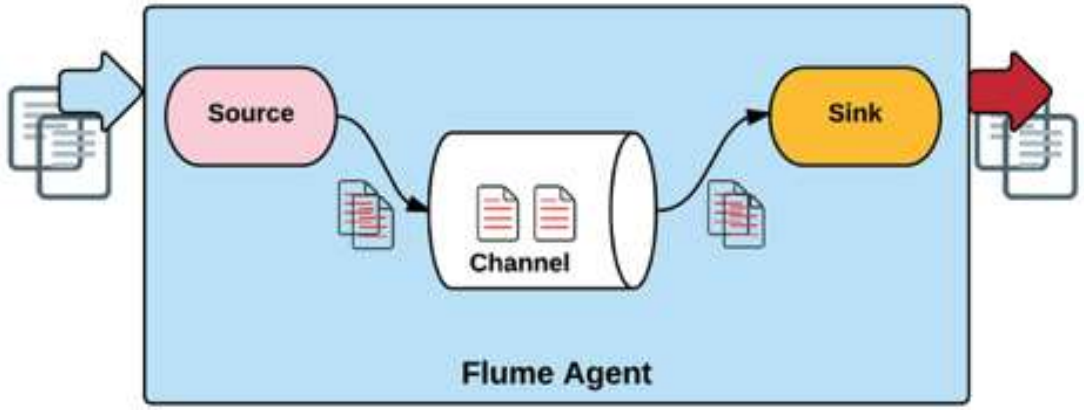


Arquitetura Flume

Source – Responsável por “ouvir” o streaming de dados ou eventos e coloca-los no canal (channel).

Channel – Onde os eventos são armazenados até que sejam consumidos.

Sink – Responsável por consumir os eventos do Channel e enviar para o destino, processando ou persistindo no data store. Se o Sink falhar, continua tentando até obter sucesso.



Arquitetura Flume

Os componentes do Flume podem ser organizados em 3 diferentes topologias:

Pipeline Distribuído

Fan Out

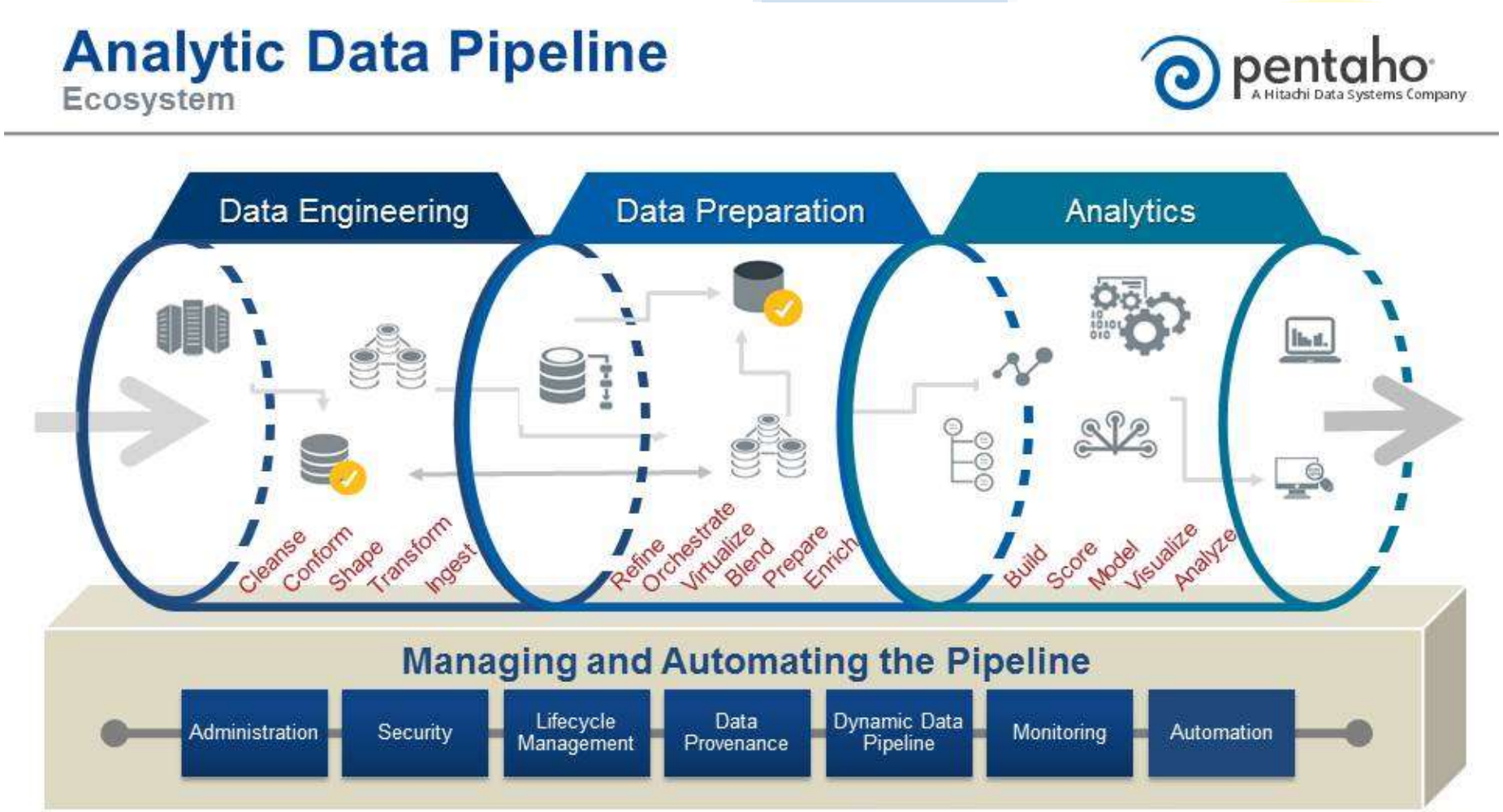
Fan In

Pipeline Distribuído

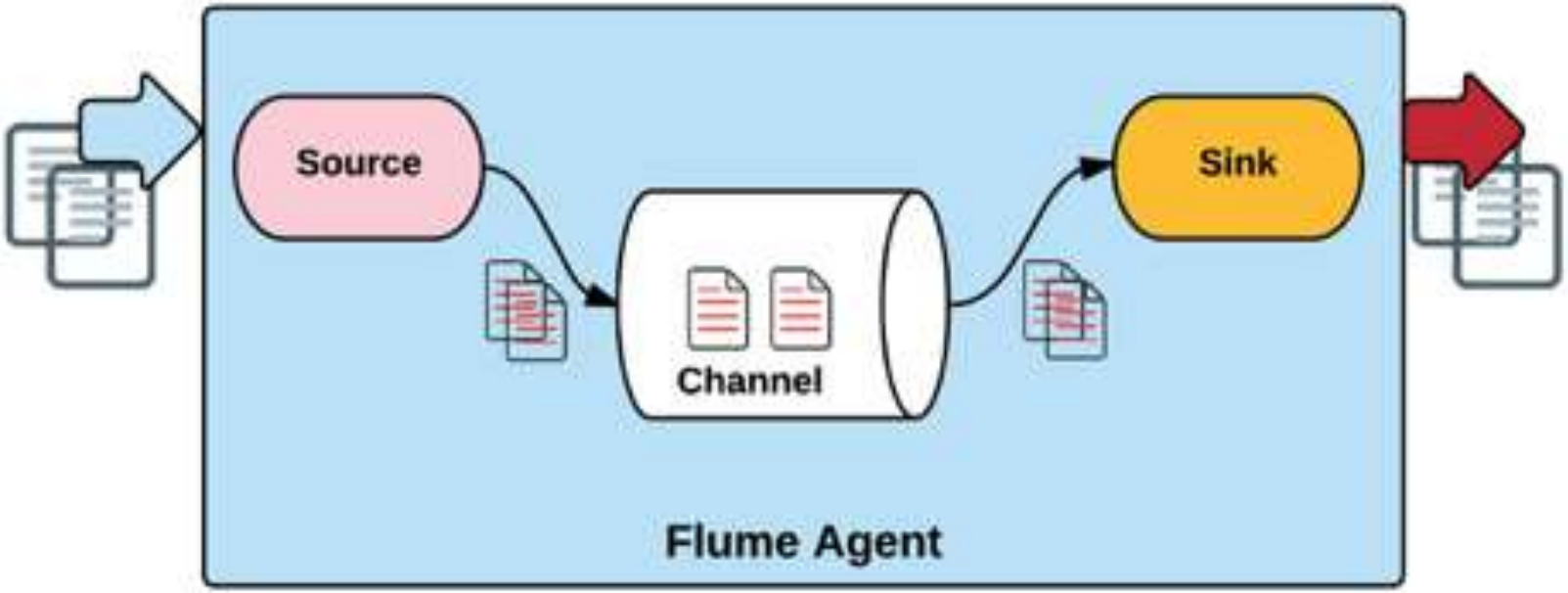
Pipeline Distribuído

Pipeline – Linha de tubos (canos)

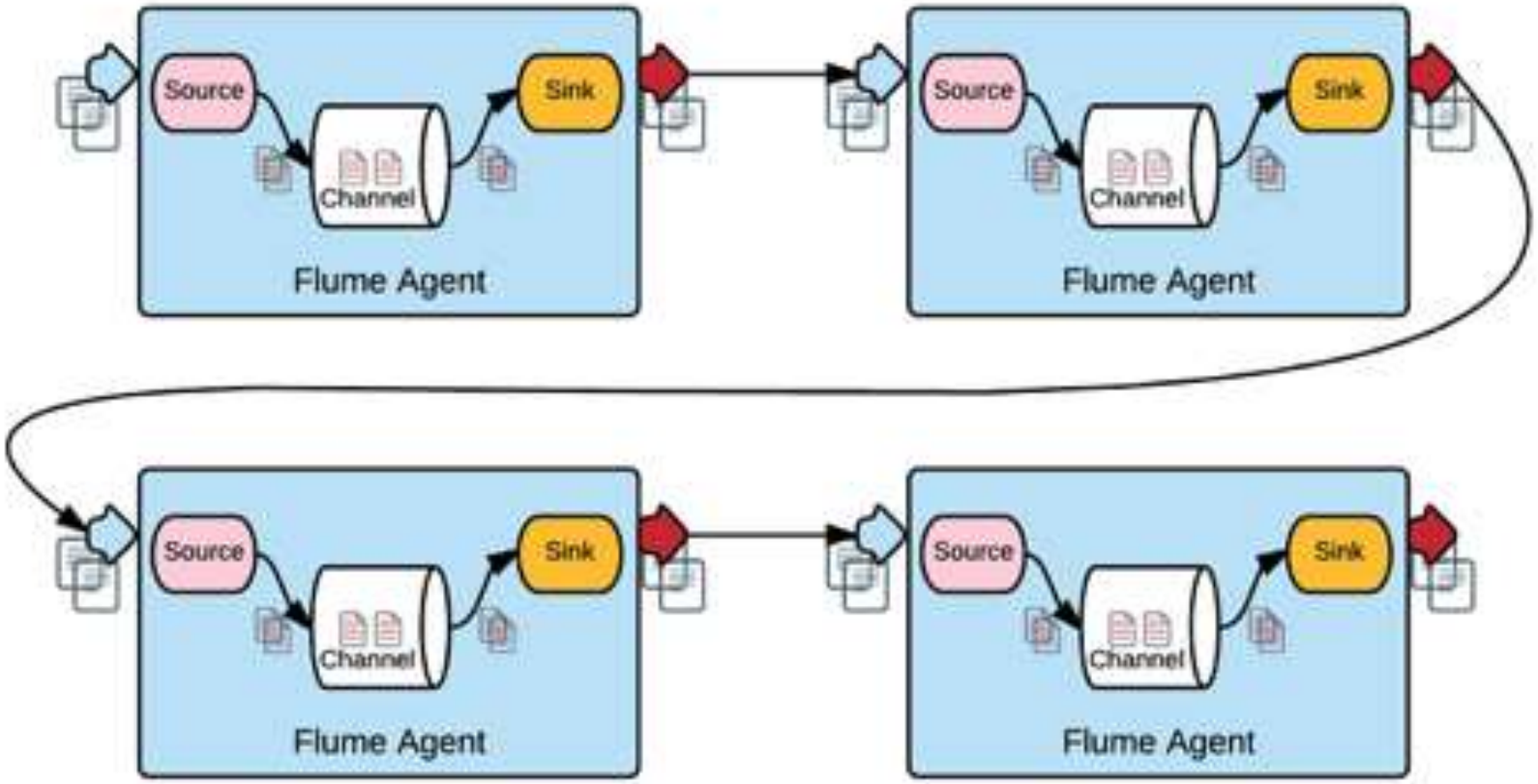
Pipeline Distribuído



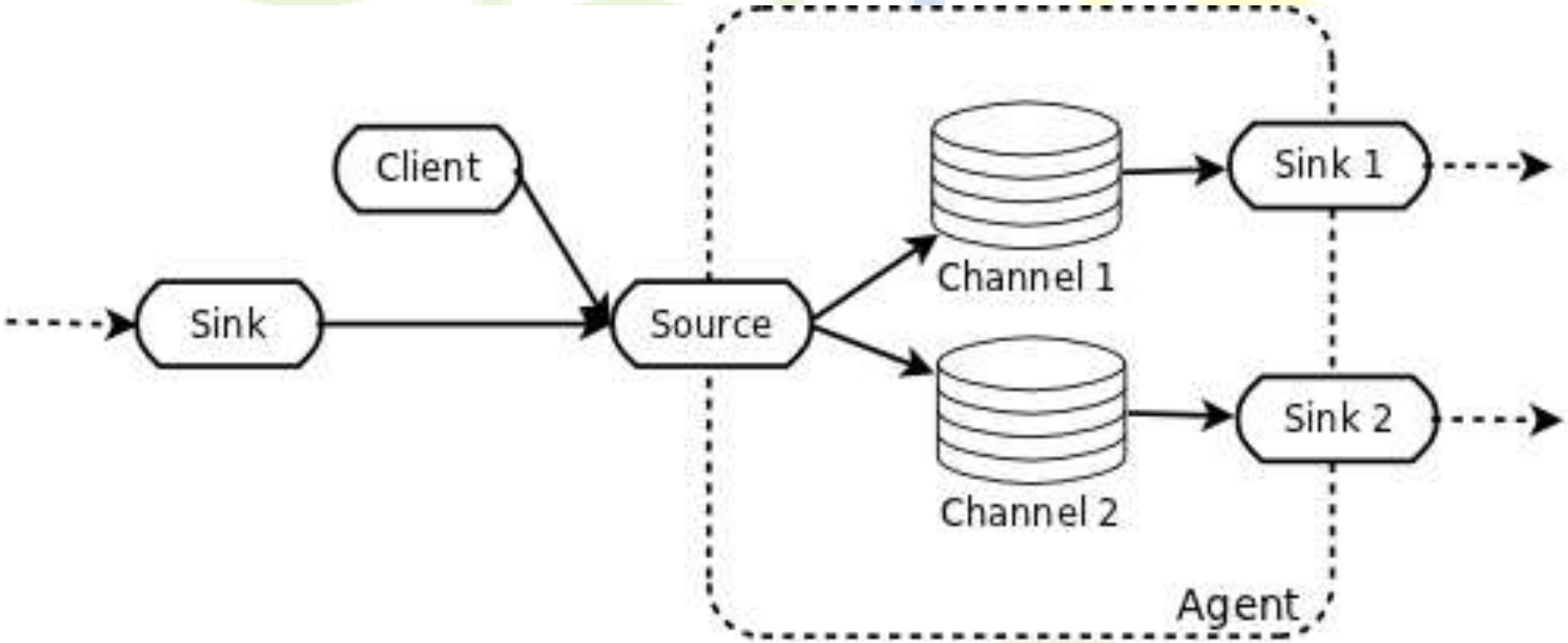
Pipeline Distribuído



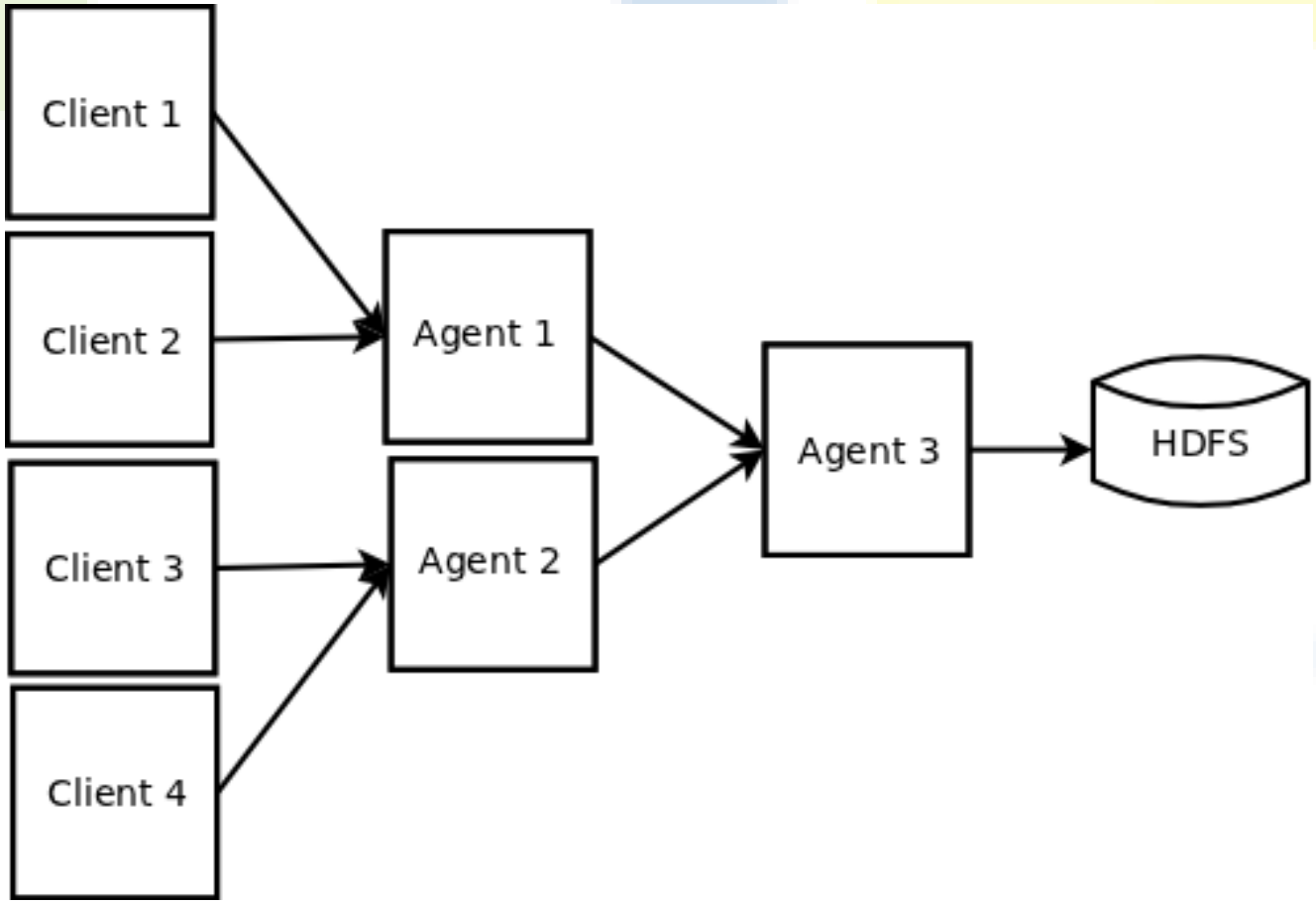
Pipeline Distribuído



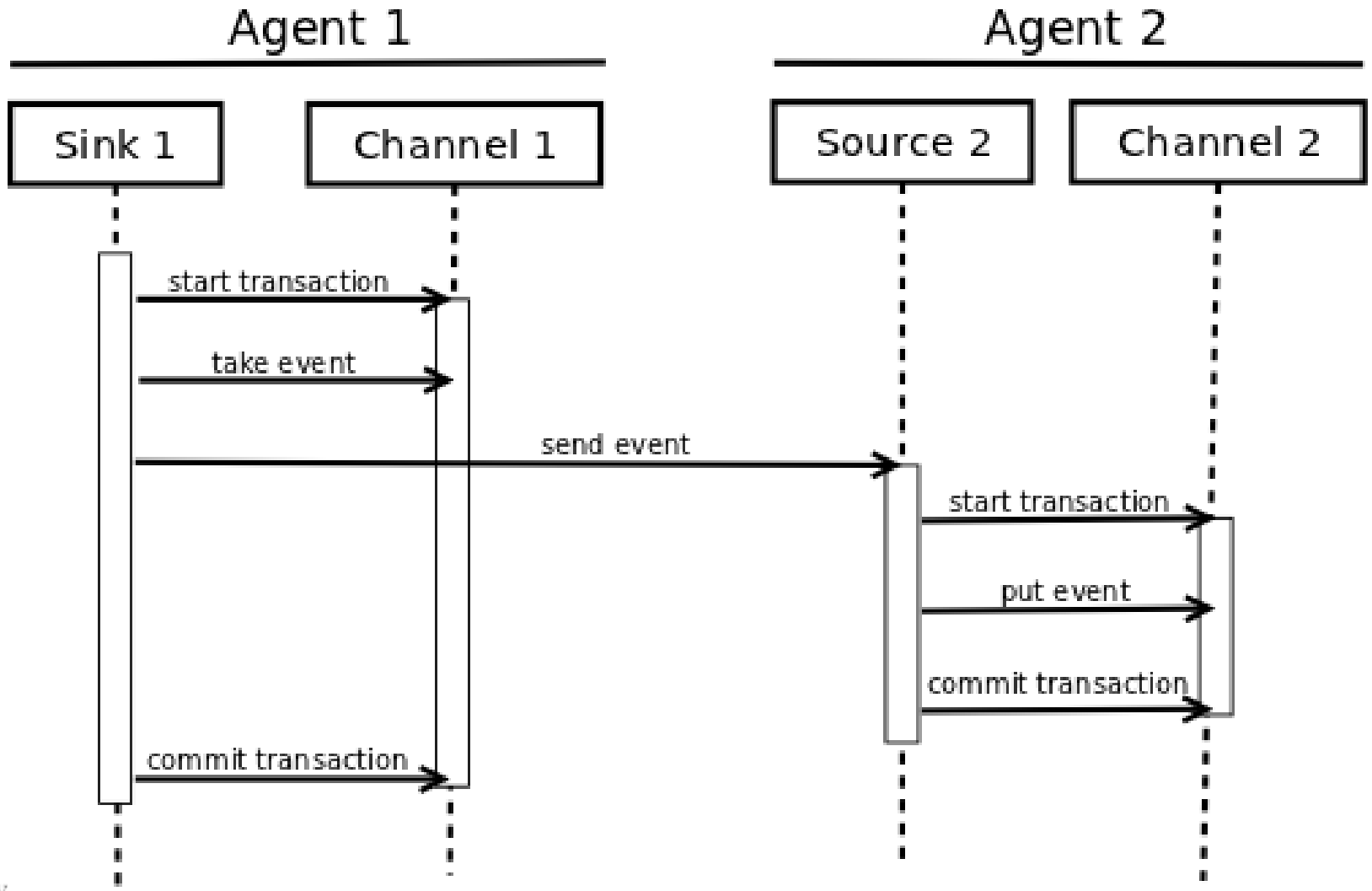
Pipeline Distribuído



Pipeline Distribuído

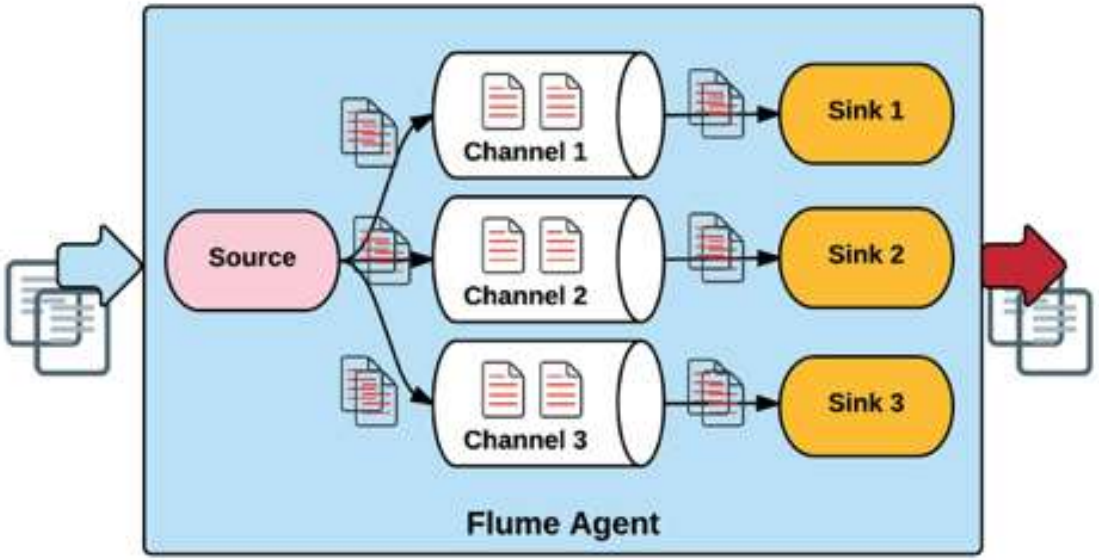


Pipeline Distribuído



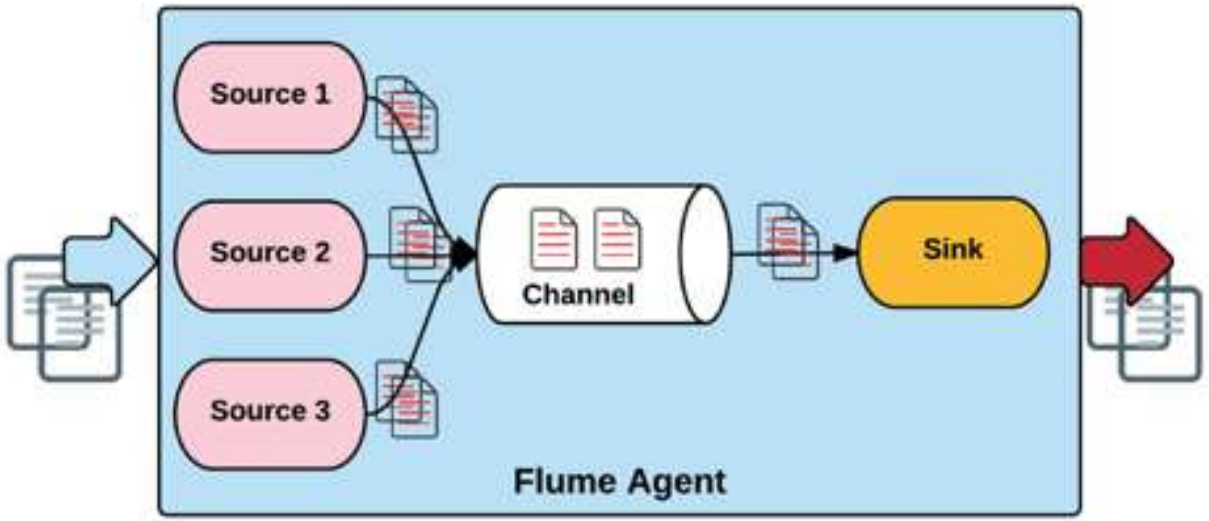
Topologia Fan-Out e Fan-In

Topologia Fan-Out e Fan-In



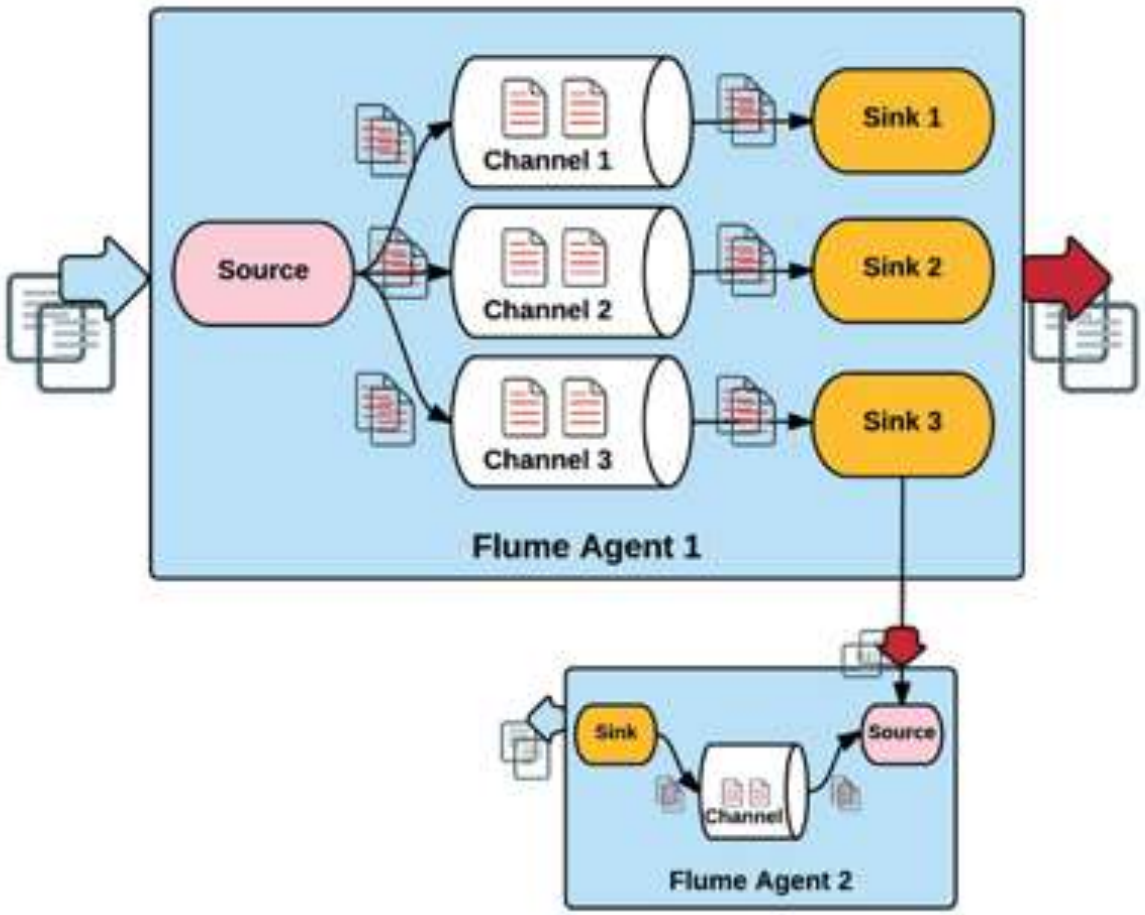
Topologia Fan-Out

Topologia Fan-Out e Fan-In



Topologia Fan-In

Topologia Fan-Out e Fan-In



E você ainda pode criar sua própria topologia, baseada na sua necessidade.

Arquitetura Flume Design de 3 Camadas

Arquitetura Flume - Design de 3 Camadas

Camada Agente

Camada onde o Agente Flume está localizado

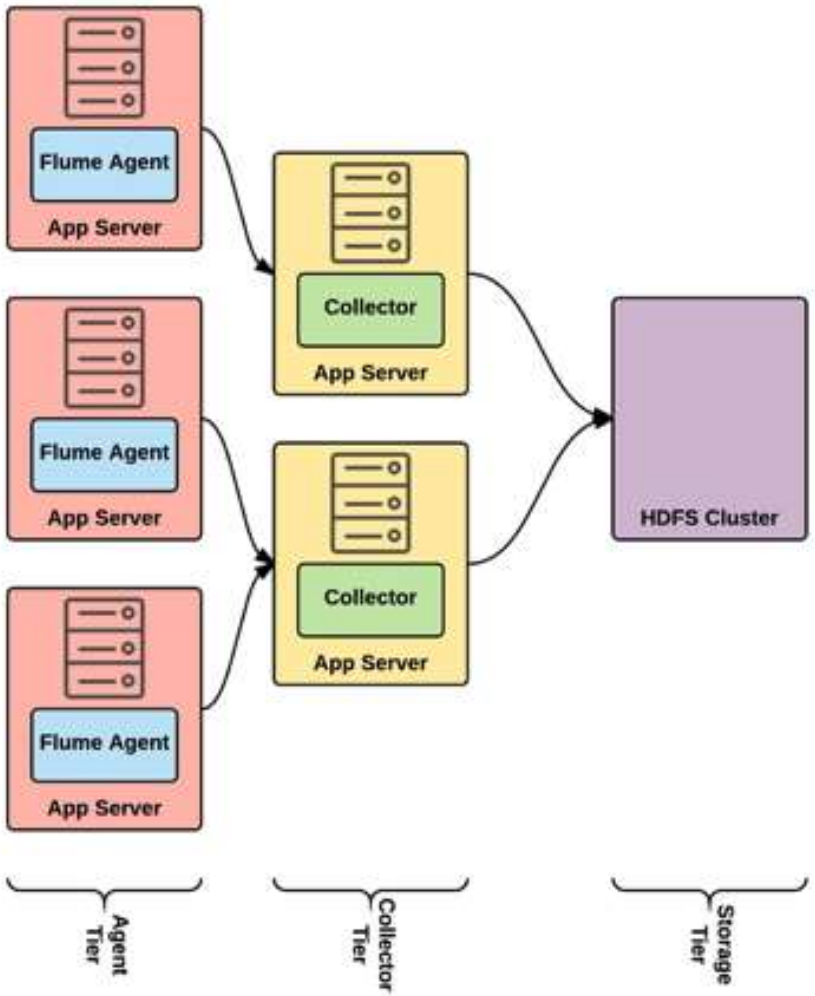
Camada Coletor

Camada onde os dados do agente são coletados

Camada Storage

Camada onde os dados dos coletores são armazenados

Arquitetura Flume - Design de 3 Camadas



Implementação física do Flume

O Que é Apache NiFi?

O Que é Apache NiFi?

Apache NiFi suporta “roteamento” de dados como grafos direcionados, de forma escalável e simples, para movimentação e transformação de dados.

O Que é Apache NiFi?

Então Apache NiFi é uma solução de ETL?

Sim, um ETL Turbinado e gratuito!

O Que é Apache NiFi?

Principais características do NiFi:

- Permite automatizar o fluxo de dados entre sistemas
- Interface Drag and Drop
- Foco na configuração dos “Processors”
- Escalável em um cluster de computadores
- Entrega garantida de dados

O Que é Apache NiFi?

Quando usar o NiFi?

- Necessidade de um sistema seguro para transferência de dados entre sistemas
- Entrega de dados da fonte para plataformas analíticas
- Processamento e transformação dos dados durante a movimentação (conversão, parsing, limpeza, etc...)

O Que é Apache NiFi?

Quando **NÃO** usar o NiFi?

- Computação distribuída (nesse caso use o Apache Spark)
- Processamento complexo de eventos (nesse caso use o Kafka/Flume/Flink)
- Operações de agregação e joins (nesse caso use Sqoop)

Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.