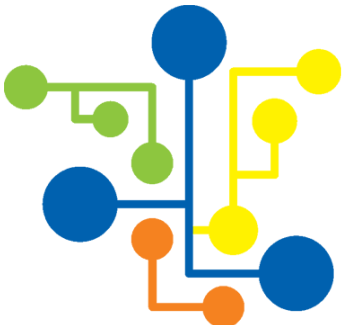

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Flume Channel

Um Canal (Channel) é um mecanismo usado pelo agente Flume para transferir dados da origem para o coletor. Os eventos são persistidos no canal e até serem entregues/removidos por um coletor, eles residem no canal. Essa persistência no canal permite que o coletor repita a tentativa para cada evento, caso haja uma falha enquanto os dados persistem no armazenamento real (HDFS).

Os canais podem ser categorizados em dois:

- **In-memory**: Os eventos estão disponíveis até que o componente do canal esteja ativo. Filas na memória no canal permitem o menor tempo de latência para processamento porque os eventos são persistidos na memória.

- **Durável**: Mesmo depois que o componente estiver inativo, o evento persistente estará disponível e, quando o componente ficar online, esses eventos serão processados. Existem 3 opções para este tipo de canal:

    - Arquivo (WAL ou Write-Ahead Log): O tipo de canal mais usado. Requer normalmente uma configuração de disco do tipo RAID, SAN ou similar (mais detalhes abaixo).

    - JDBC: Um canal suportado pelo RDBMS adequado que fornece conformidade com ACID.

    - Kafka: armazenado em cluster Kafka.

Existe outro canal especial chamado Spillable Memory Channel, que armazena dados na memória e no disco. Quando a capacidade de memória está cheia, os demais eventos são armazenados em disco (canal de arquivo incorporado ou *embedded file channel*), muito similar ao funcionamento do Apache Spark.

O aspecto de confiabilidade depende do tipo de canal que está sendo configurado. Os canais também cuidam da solicitação de eventos e também ajudam na garantia de transação para o agente.

Considerações Sobre Infraestrutura de Disco (Armazenamento)

O RAID (Redundant Array of Independent Disks) é uma tecnologia de virtualização de armazenamento de dados que combina vários componentes de unidade de disco físico em uma única unidade lógica para fins de redundância de dados, melhoria de desempenho ou ambos.

Uma SAN (Storage Area Network) é uma rede que fornece acesso a armazenamento de dados em nível de bloco consolidado. As SANs são usadas principalmente para aprimorar dispositivos de armazenamento, como matrizes de disco, acessíveis a servidores para que os dispositivos apareçam no sistema operacional como dispositivos conectados localmente.

Em ciência da computação, o WAL (Write-Ahead Log) é uma família de técnicas para fornecer atomicidade e durabilidade (duas das propriedades ACID) em sistemas de banco de dados. Em um sistema que usa o WAL, todas as modificações são gravadas em um log antes de serem aplicadas. Geralmente, as informações de refazer e desfazer uma operação são armazenadas no log.

Em ciência da computação, o ACID (Atomicidade, Consistência, Isolamento, Durabilidade) é um conjunto de propriedades das transações do banco de dados. No contexto de bancos de dados, uma única operação lógica nos dados é chamada de transação.

Canal Personalizado

O aspecto *pluggable* do Flume pode ser usado para escrever um canal personalizado de acordo com sua necessidade, satisfazendo o caso de uso. Para isso, a classe deve ser escrita implementando a interface do canal. Uma configuração de amostra de um canal customizado para um agente ag1 para uma classe de canal customizada com.nomepacote.CustomChannel seria a seguinte:

```
ag1.canais = ch1
ag1. channels.ch1.type = com.nomepacote.CustomChannel
```

Referências:

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