COMPUTAÇÃO ESCALÁVEL

Frameworks para Processamento de Dados em Larga Escala - Kafka

ROTEIRO

- Introdução
- O que é Apache Kafka?
- Arquitetura
- Pontos de Entrada
- Integração de dados e processamento
- Usos do Kafka

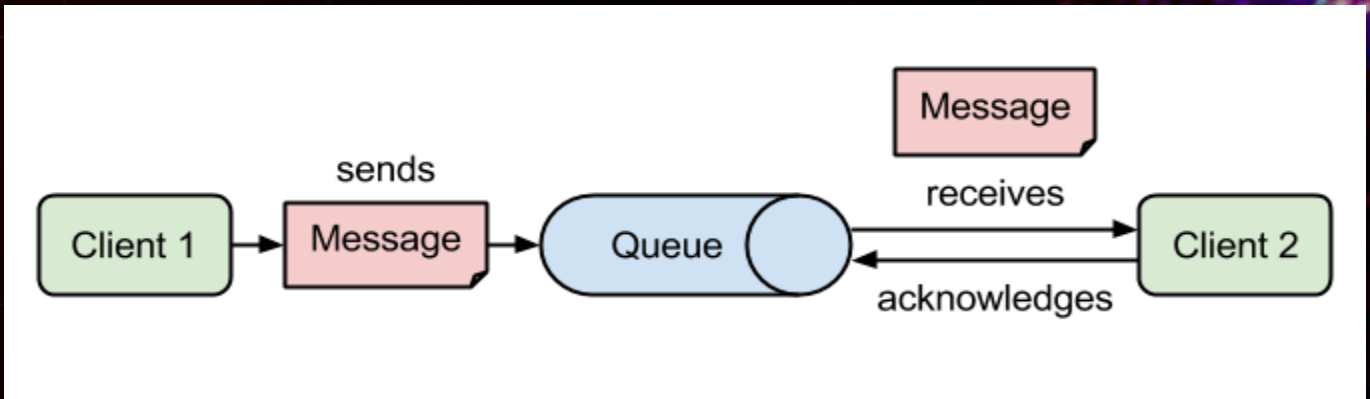
INTRODUÇÃO

Há dois tipos de sistema de mensagens:

- **Sistema ponto a ponto**
- **Sistema Publicar-Assinar**

Sistema ponto a ponto

- **As mensagens são persistidas em uma fila, mas uma mensagem específica pode ser consumida por no máximo apenas um consumidor. Depois que um consumidor lê uma mensagem na fila, ela desaparece dessa fila**

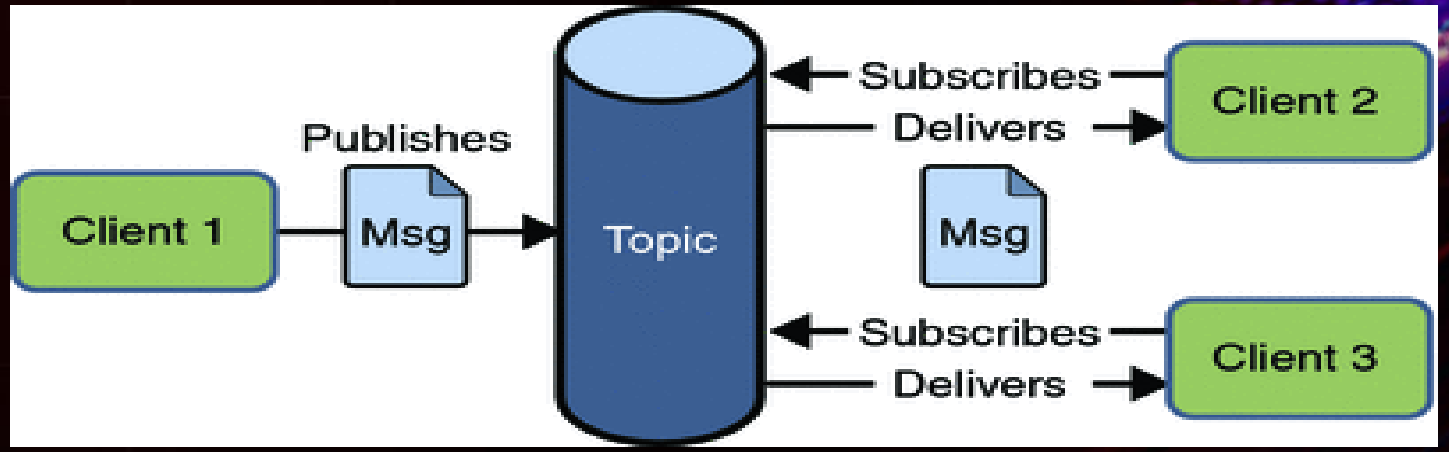


INTRODUÇÃO

Sistema Publicar-Assinar

- As mensagens são persistidas em um tópico. Ao contrário do sistema ponto a ponto, os consumidores podem se inscrever em um ou mais tópicos e consumir todas as mensagens desse tópico

- Neste sistema, os produtores de mensagens são chamados de publicadores e os consumidores de mensagens são chamados de assinantes

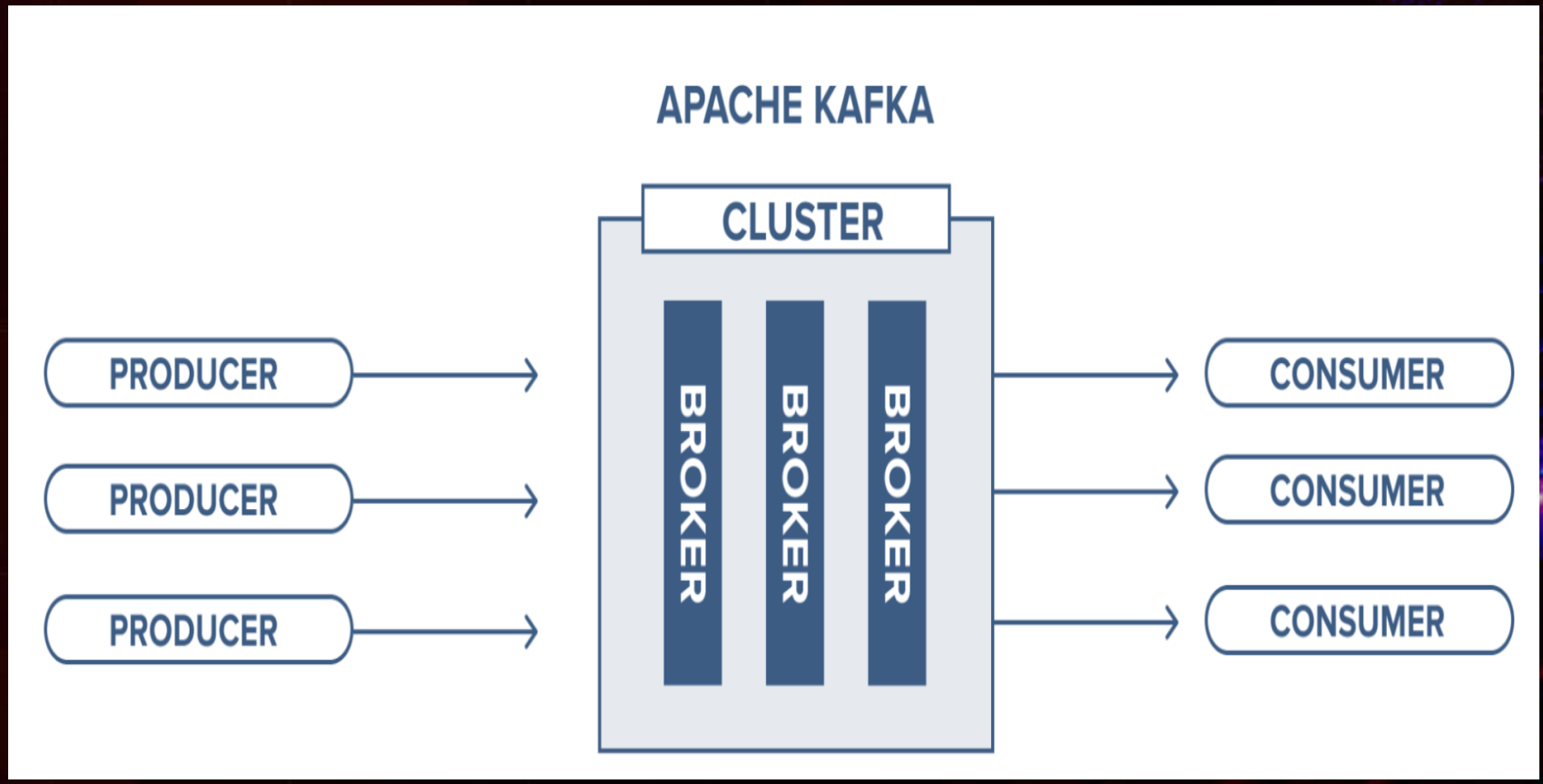


O QUE É APACHE KAFKA?

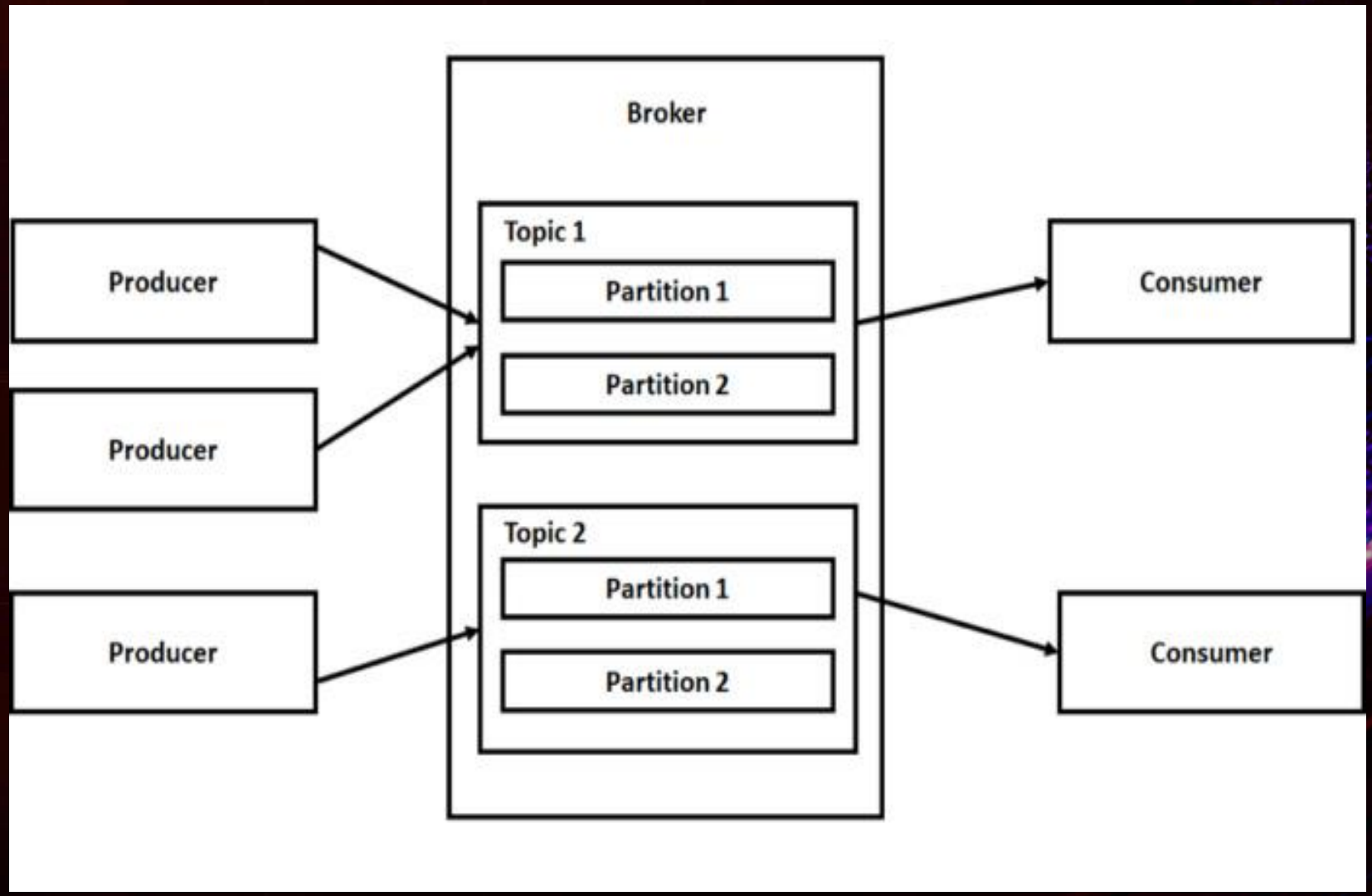
Kafka é uma plataforma de streaming distribuído

- Alta escalabilidade (partição)
- Tolerante a falhas (replicação)
- Permite alto nível de paralelismo e desacoplamento entre produtores de dados e consumidores de dados
- Padrão de fato para armazenamento, acesso e processamento de fluxo de dados quase em tempo real
- Componente crítico da maior parte da plataforma de Big Data e, portanto, do ecossistema Hadoop

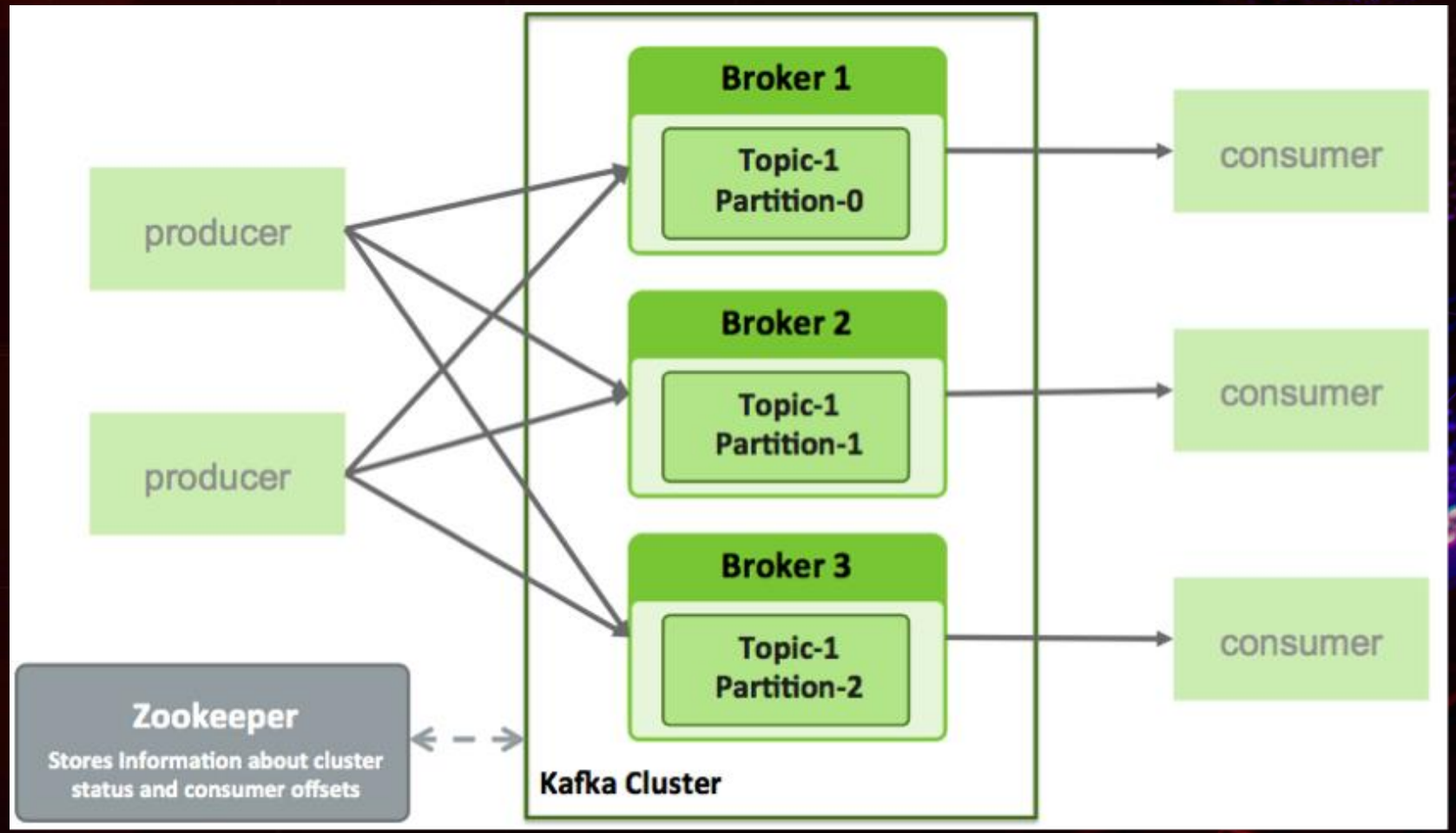
INTRODUÇÃO


APACHE KAFKA
CLUSTER
BROKER
BROKER
BROKER
PRODUCER
PRODUCER
PRODUCER
CONSUMER
CONSUMER
CONSUMER

ARQUITETURA



ARQUITETURA



ARQUITETURA

- Um produtor é uma entidade/aplicação que publica dados em um cluster Kafka, composto por brokers
- Um broker é responsável por receber e armazenar os dados quando um produtor publica
- Um consumidor consome dados de um broker em um deslocamento especificado, ou seja, posição
  - O Kafka usa o zookeeper para facilitar a verificação de integridade, o gerenciamento e a coordenação
    - O zookeeper oferece aos brokers, metadados sobre os processos em execução no sistema

ARQUITETURA

- Um **tópico** é um nome de categoria/feed, no qual os registros são armazenados e publicados. Os tópicos têm partições e ordem garantida por partições

- Todos os **registros** do Kafka são organizados em tópicos. Aplicativos produtores gravam dados em tópicos e aplicativos consumidores leem tópicos

PONTOS DE ENTRADA

Implementação personalizada de produtor e consumidor usando a API do cliente Kafka

- **Java, Scala, C++, Python**

Conectores Kafka

- **LogFile, HDFS, JDBC, ElasticSearch…**

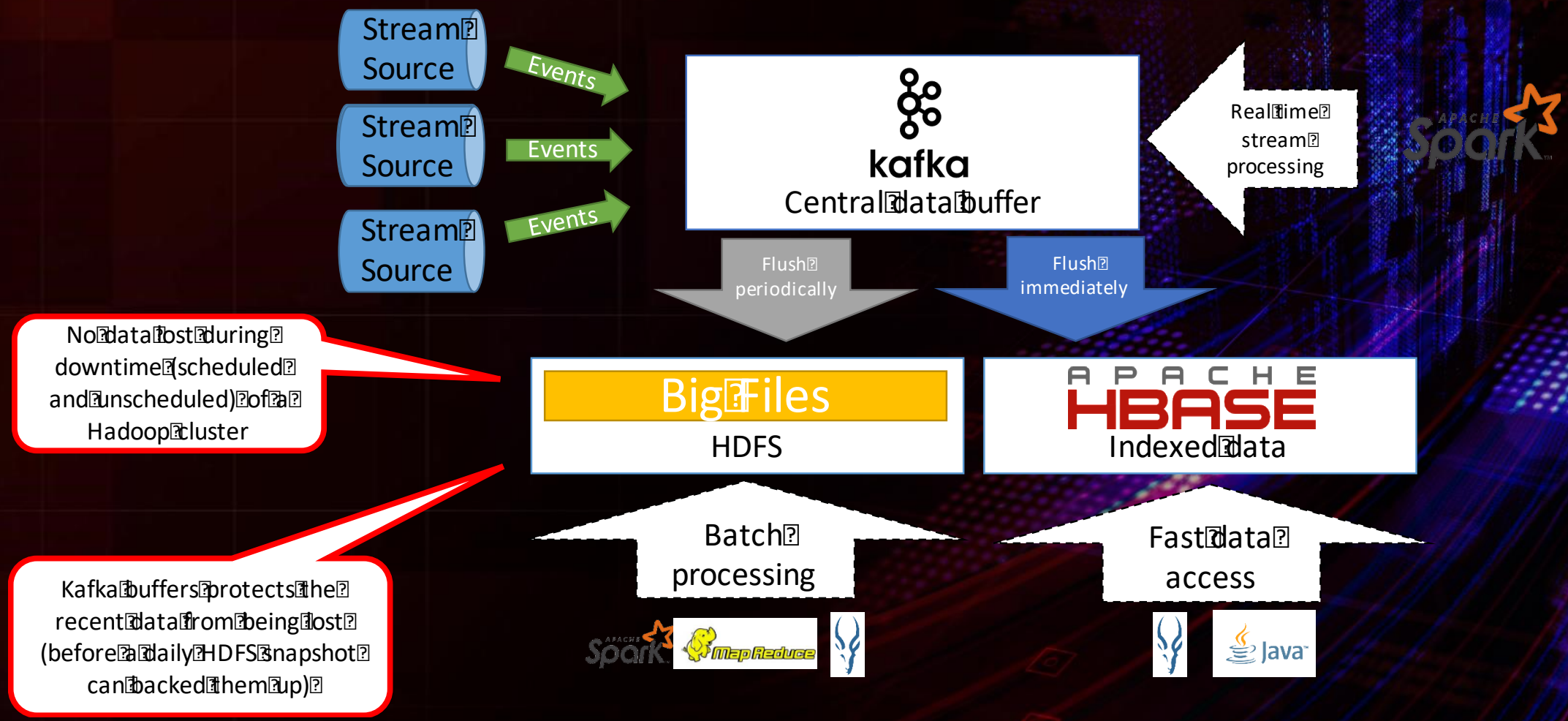
Logstash

- **Fonte e destino**

Outras ferramentas de ingestão ou processamento que suportam Kafka

- **Apache Spark, LinkedIn Gobblin, Apache Storm…**

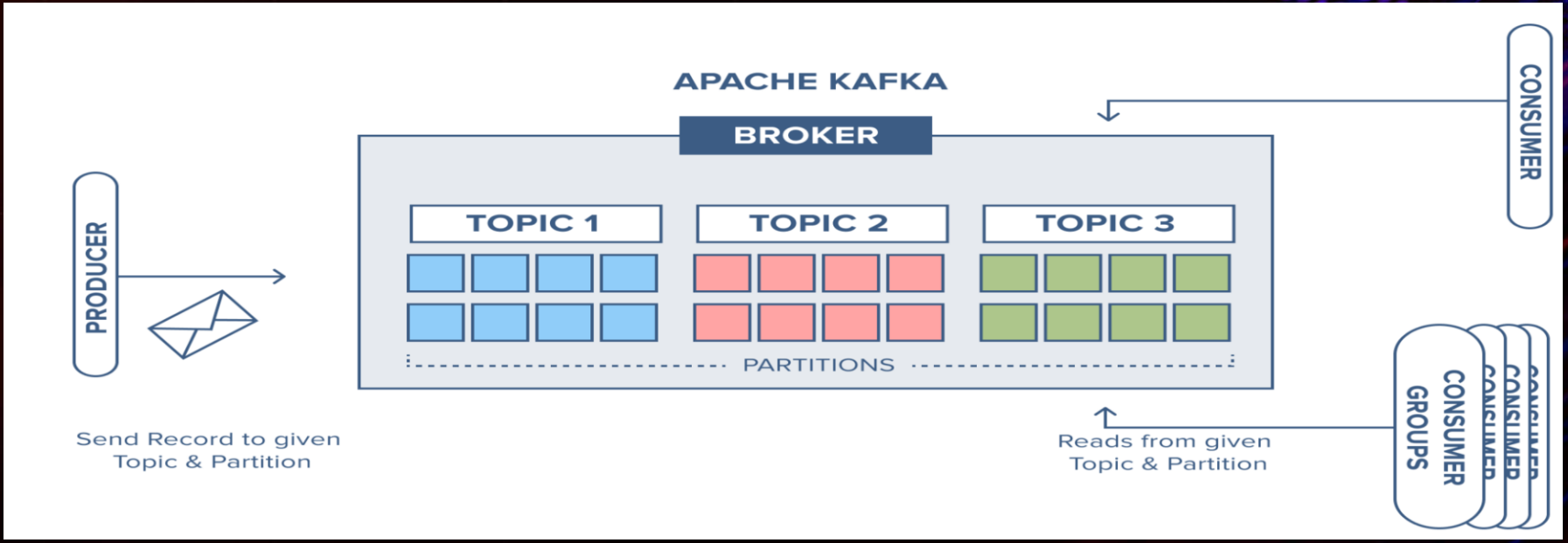
INTEGRAÇÃO DE DADOS E PROCESSAMENTO



REGISTRO DE FLUXO

Considere um broker com três tópicos, no qual cada tópico possui 8 partições

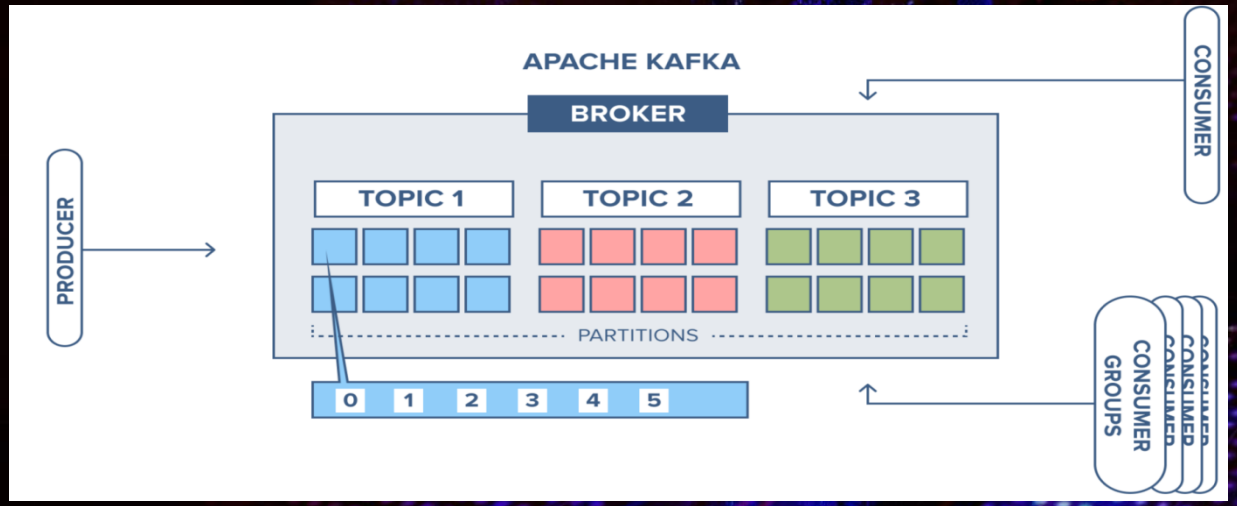
O produtor envia um registro para a partição 1 no tópico 1 e, como a partição está vazia, o registro termina no deslocamento 0



REGISTRO DE FLUXO

O próximo registro é adicionado à partição 1 e acima no deslocamento 1, e o próximo registro no deslocamento 2 e assim por diante

Esse é um log de confirmação e cada registro é anexado ao log e não há como alterar os registros existentes no log (imutável). Esse também é o mesmo deslocamento que o consumidor usa para especificar onde iniciar a leitura



USOS DO KAFKA

- Mensageria
- Métricas
- Agregação de logs
- Registros de atividades em sites e aplicações Web
- Operações de TI
- Internet das Coisas

BIBLIOGRAFIA

1. https://www.redhat.com/pt-br/topics/integration/what-is-apache-kafka#:~:text=O%20Apache%20Kafka%20%C3%A9%20uma,entreg%C3%A1%2Dlos%20a%20%C3%A1rios%20clientes
2. https://vepo.medium.com/entendendo-o-kafka-bf64169e421f
3. https://www.infoq.com/br/articles/real-time-api-kafka/
4. https://www.infoq.com/br/articles/apache-kafka-licoes/