DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Data Lake
Camada de Mensagens

Data Lake
Camada de Mensagens

Qual o principal propósito de um Data Lake?

Tratar grandes conjuntos de dados, estruturados ou não estruturados, em batch ou em streaming!

Data Lake
Camada de Mensagens

Neste capítulo estudaremos uma das principais tecnologias de um Data Lake, a camada de mensagens,

cujo principal objetivo é tratar fluxos de dados gerados em tempo real das mais variadas fontes.

Data Lake

Camada de Mensagens

Data Lake
Camada de Mensagens

Parte 1

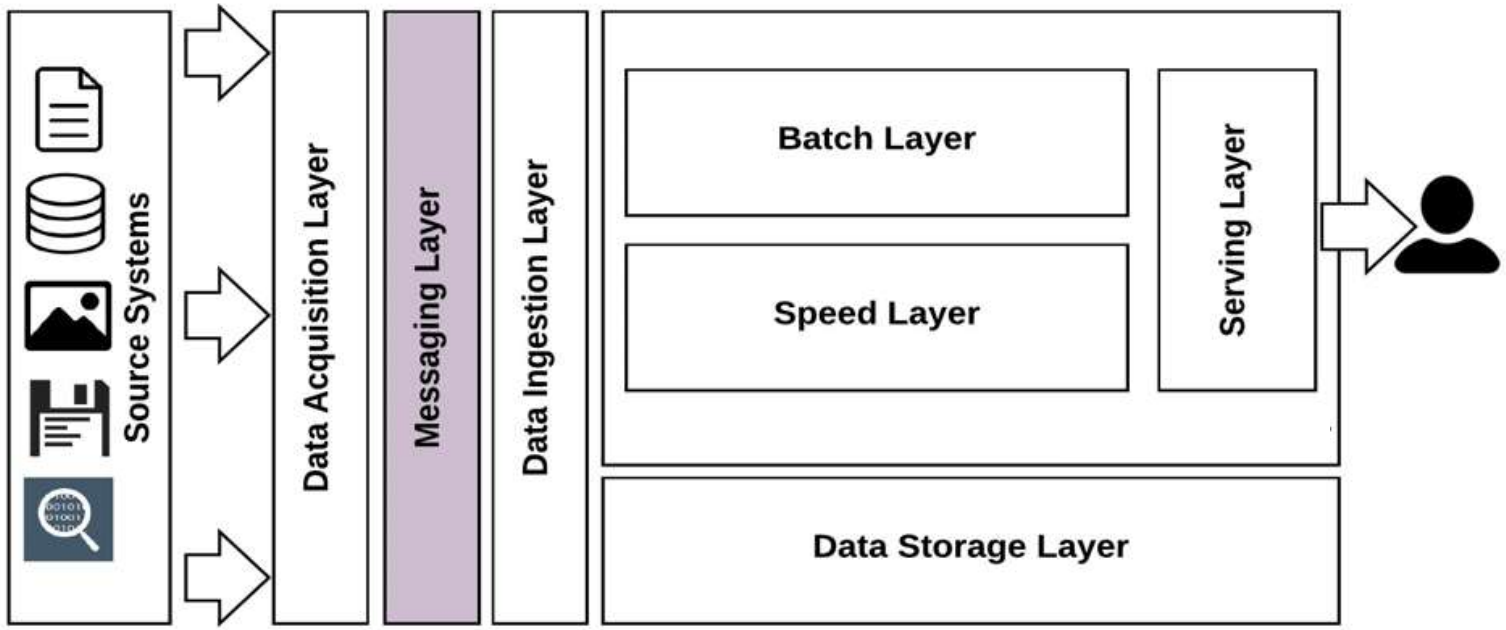
Compreender o que é serviço de mensageria, sua importância na infraestrutura de Big Data e como funciona o Apache Kafka.

Parte 2

Integração do Apache Kafka ao Data Lake.

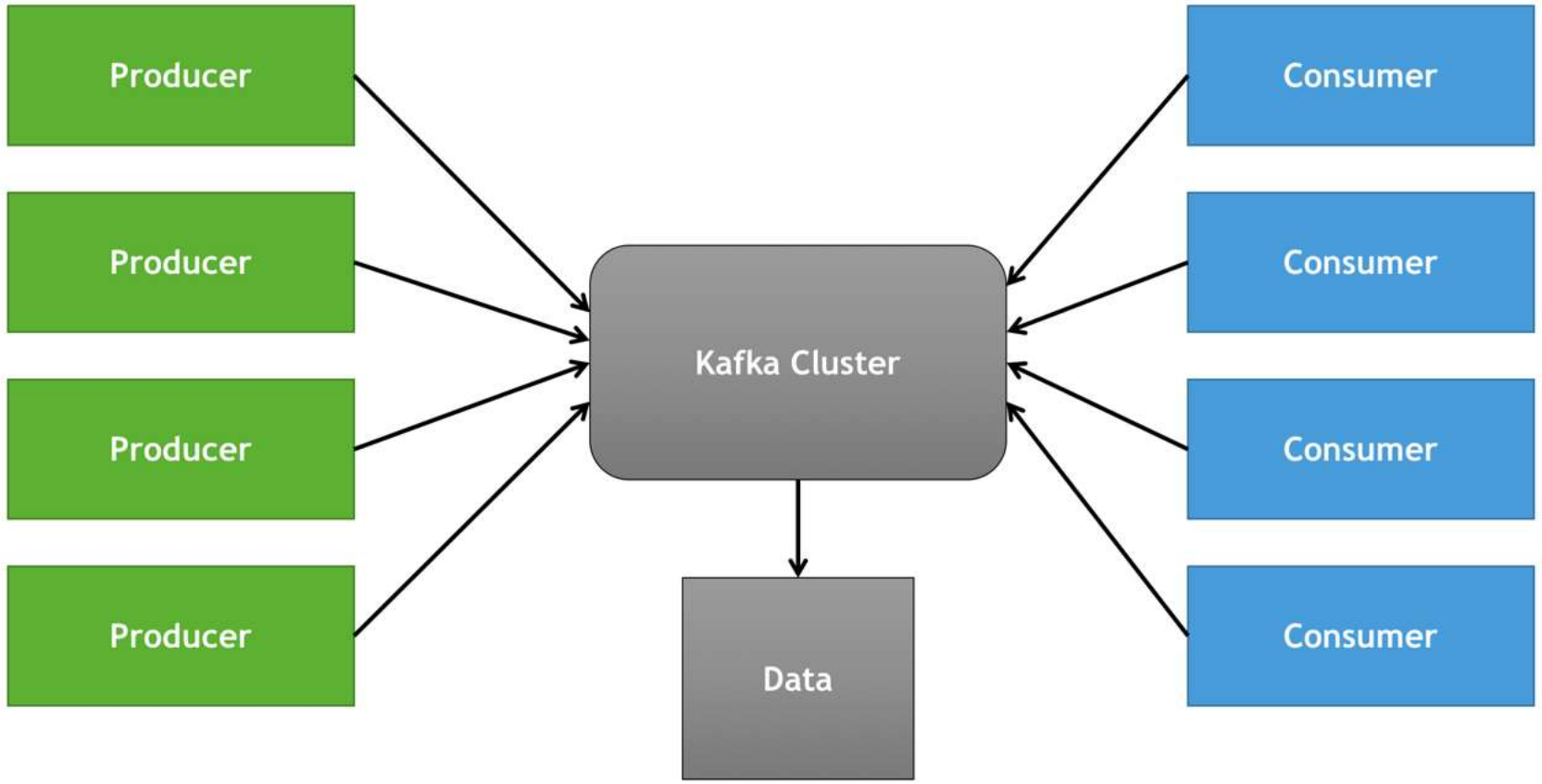
Contexto no Data Lake
Camada de Mensagens

Contexto no Data Lake
Camada de Mensagens

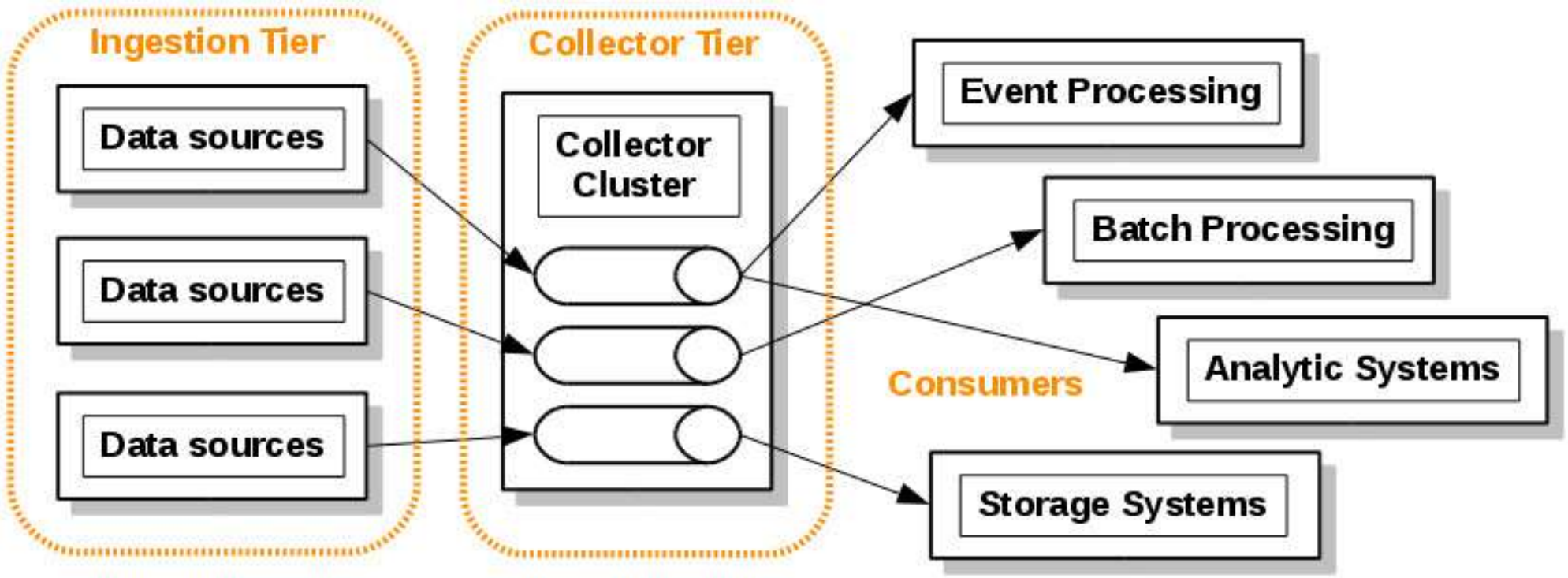


Qual a Função da Camada de Mensagens?

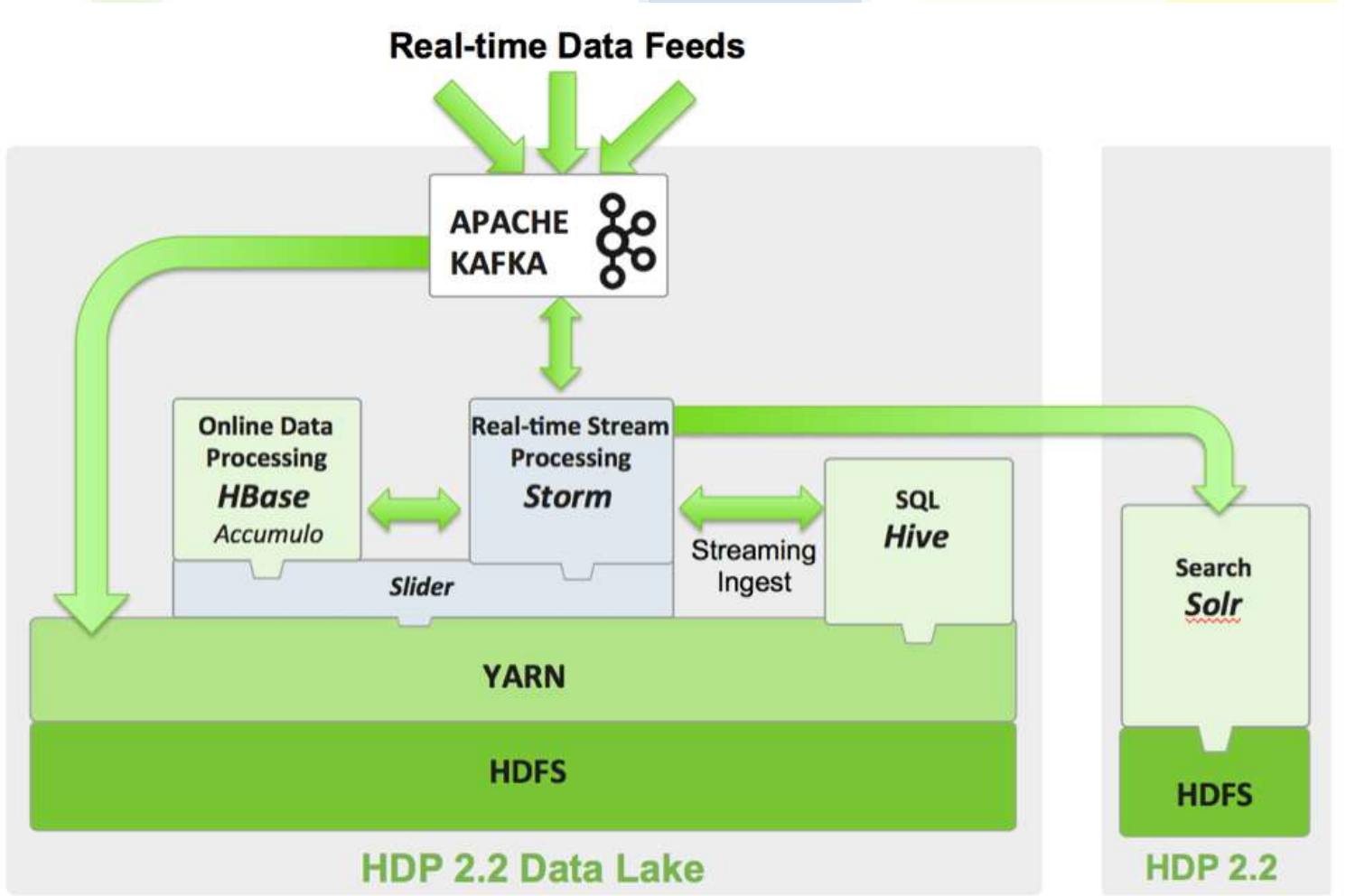
Qual a Função da Camada de Mensagens?



Qual a Função da Camada de Mensagens?

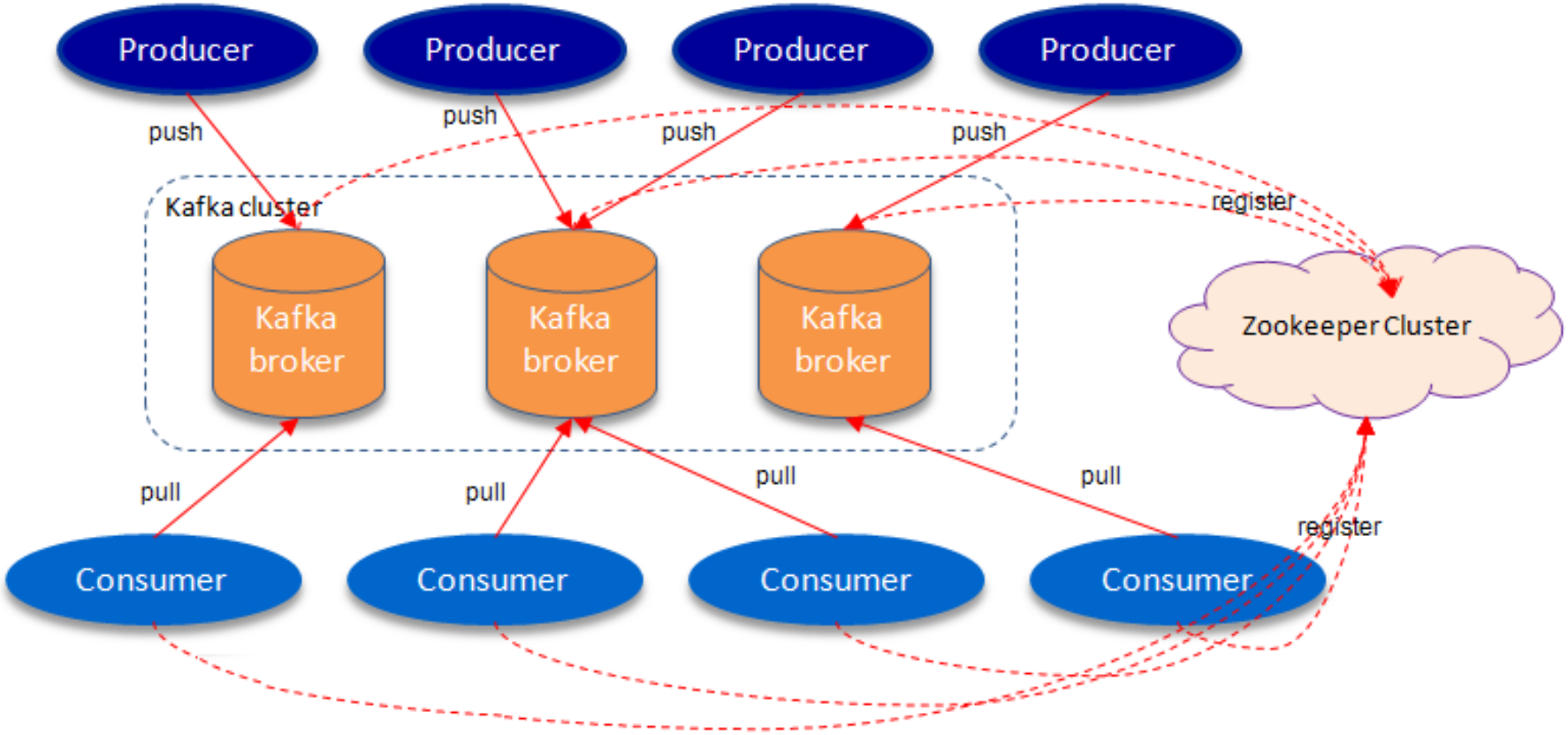


Qual a Função da Camada de Mensagens?

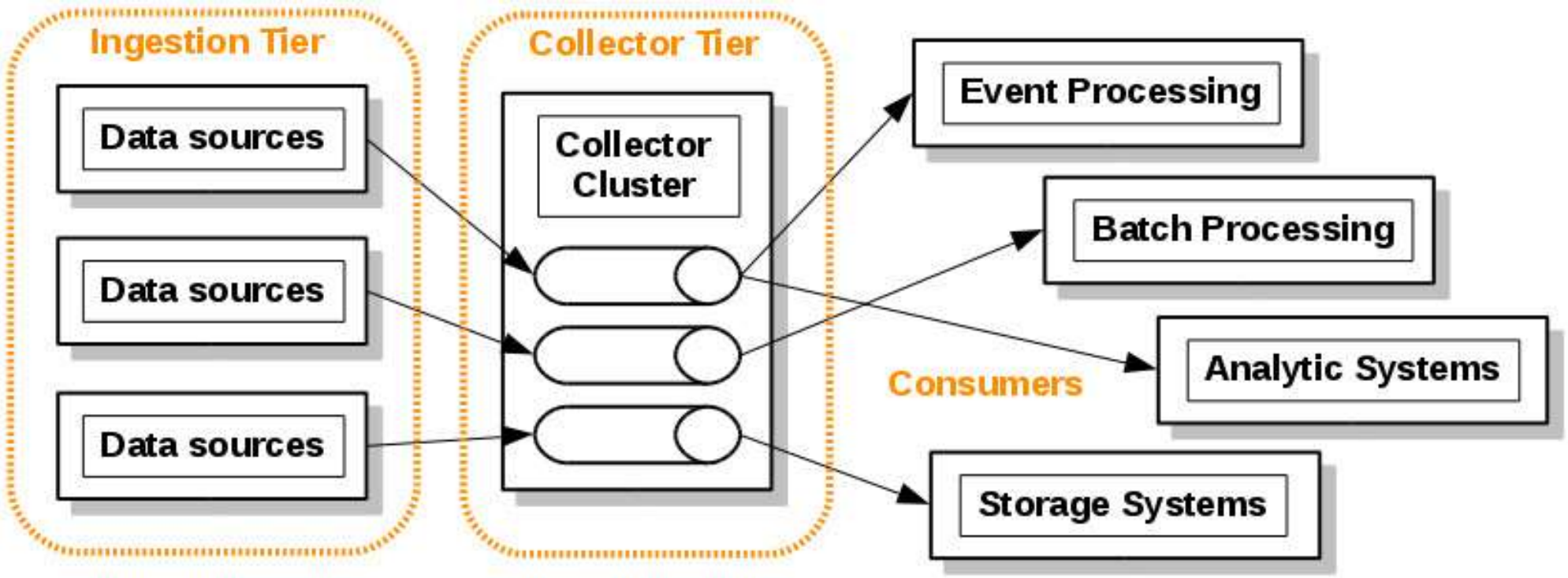


Fonte: Hortonworks

Qual a Função da Camada de Mensagens?

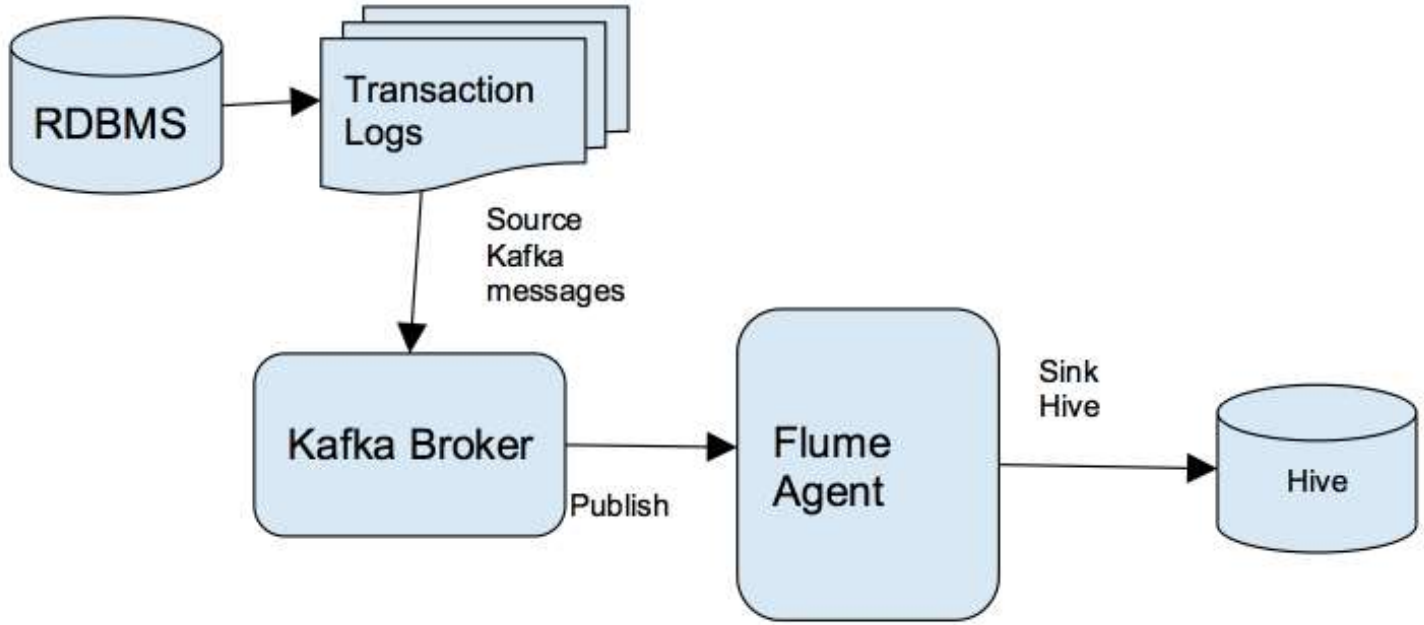


Qual a Função da Camada de Mensagens?



Qual a Função da Camada de Mensagens?

- Uma das principais funções dessa camada é a capacidade de dissociar a origem (produtor) e o destino (consumidor).



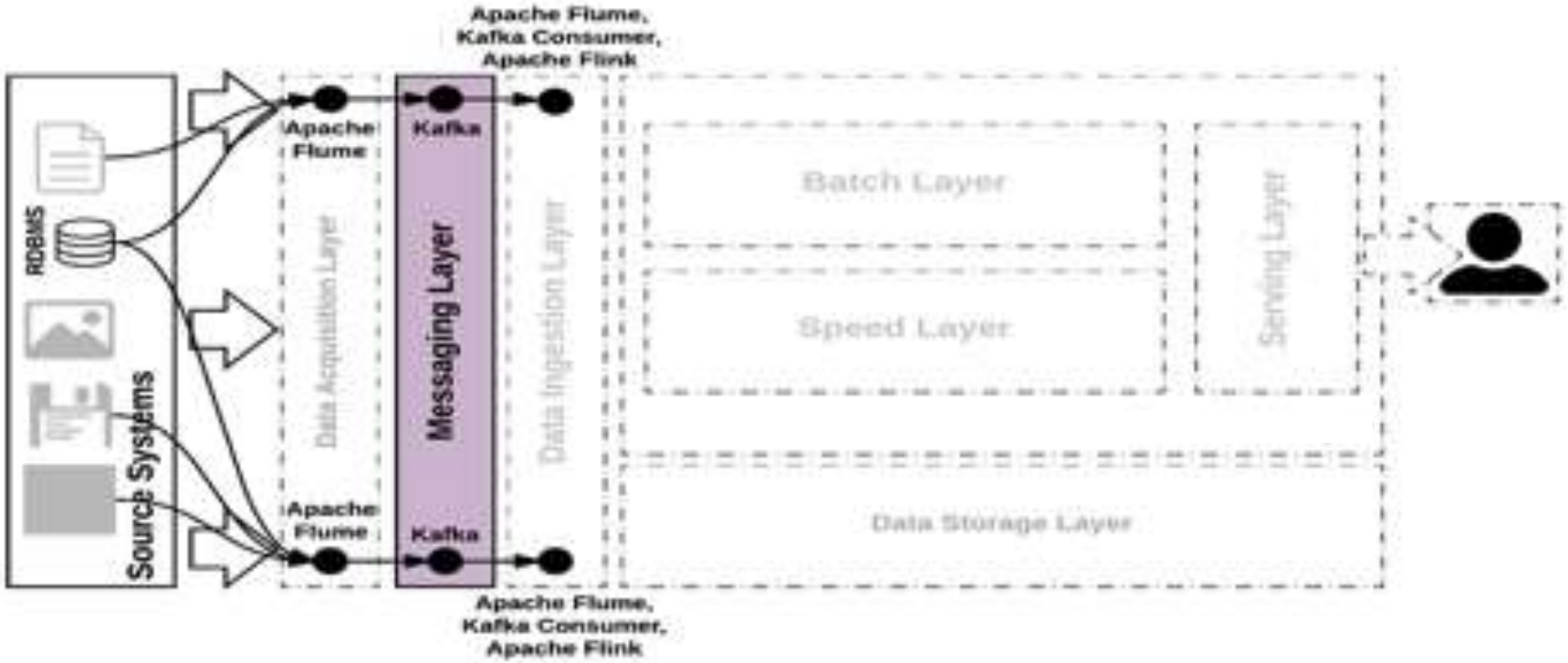
Qual a Função da Camada de Mensagens?

- Uma das principais funções dessa camada é a capacidade de dissociar a origem (produtor) e o destino (consumidor).
- Capacidade de manipular mensagens de alta velocidade na ordem de centenas de megabytes por segundo de cada nó do servidor de aplicativos.
- Capacidade de lidar com grandes volumes de dados da ordem de terabytes a petabytes.
- Capacidade de lidar com mensagens com latência muito baixa sob requisitos extremos de taxa de transferência.
- Capacidade de garantir a entrega de mensagens (durabilidade) de forma ordenada.

Qual a Função da Camada de Mensagens?

- Capacidade de fornecer a mesma mensagem a vários consumidores. Por exemplo: fornecer mensagens para a camada Speed Layer (que vai processar os dados em tempo real) e Data Storage Layer (que vai armazenar os dados) ao mesmo tempo.
- Capacidade de análise de dados para derivar estatísticas operacionais.
- Capacidade de agregar dados provenientes de várias fontes e fazer algumas análises.
- Obviamente, alto desempenho com menos requisitos de hardware (sim, na verdade, isso é um requisito).
- Capacidade de executar recursos mínimos de enriquecimento e transformação.

Qual a Função da Camada de Mensagens?



O Que é o Apache Kafka?

O Que é o Apache Kafka?

O Que é o Apache Kafka?

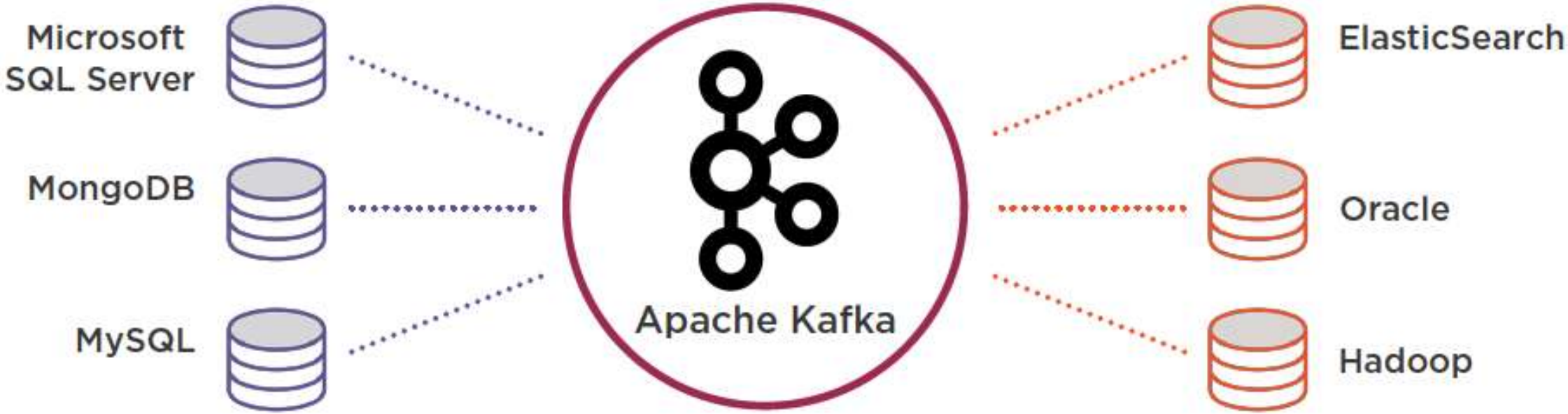
O Apache Kafka é uma plataforma open-source de processamento de streaming de dados, mantida pela Apache Software Foundation, escrita em Scala e Java. O projeto tem como objetivo fornecer uma plataforma unificada de alta produtividade e baixa latência para o tratamento de feeds de dados em tempo real.

Sua camada de armazenamento é essencialmente uma "fila de mensagens massivamente escalável arquitetada como um log de transações distribuídas", tornando altamente valioso para infraestruturas corporativas de processamento de streaming de dados.

O Que é o Apache Kafka?

O Apache Kafka é um MOM - Middleware Orientado a Mensagens, que suporta o envio e recebimento de mensagens entre sistemas distribuídos.

O Que é o Apache Kafka?



O Que é o Apache Kafka?

High Volume:

- Over 1.4 trillion messages per day
- 175 terabytes per day
- 650 terabytes of messages consumed per day
- Over 433 million users

High Velocity:

- Peak 13 million messages per second
- 2.75 gigabytes per second

High Variety:

- Multiple RDBMS (Oracle, MySQL, etc.)
- Multiple NoSQL (Espresso, Voldemort)
- Hadoop, Spark, etc.

O Que é o Apache Kafka?

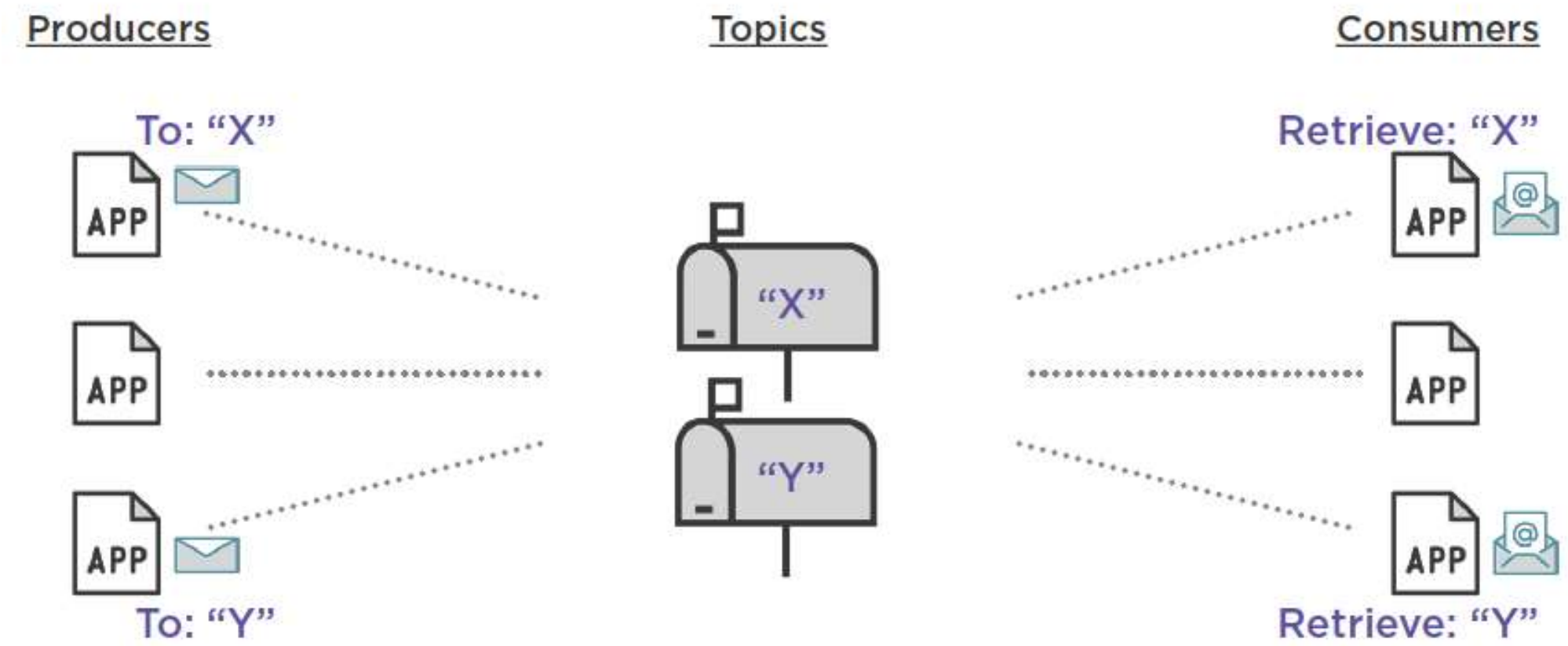
Apache Kafka foi originalmente desenvolvido pelo LinkedIn, e foi posteriormente open-sourced no início de 2011.

Em novembro de 2014, Jun Rao, Jay Kreps e Neha Narkhede, que haviam trabalhado no desenvolvimento do Kafka no LinkedIn, criaram uma nova empresa chamada Confluent com foco no Kafka. De acordo com um post do Quora de 2014, Kreps escolheu nomear o software em homenagem ao escritor alemão Franz Kafka porque ele é "um sistema otimizado para escrita", e ele gostava do trabalho de Kafka.

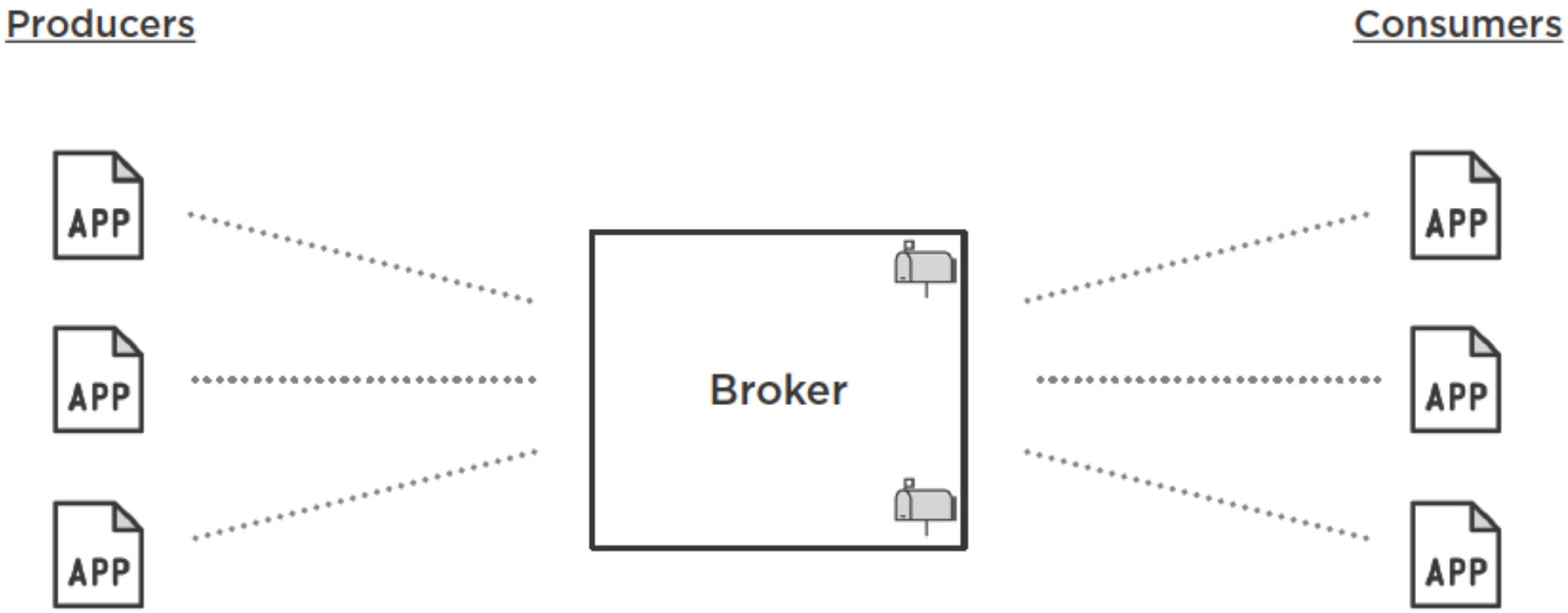
Franz Kafka

Arquitetura do Apache Kafka

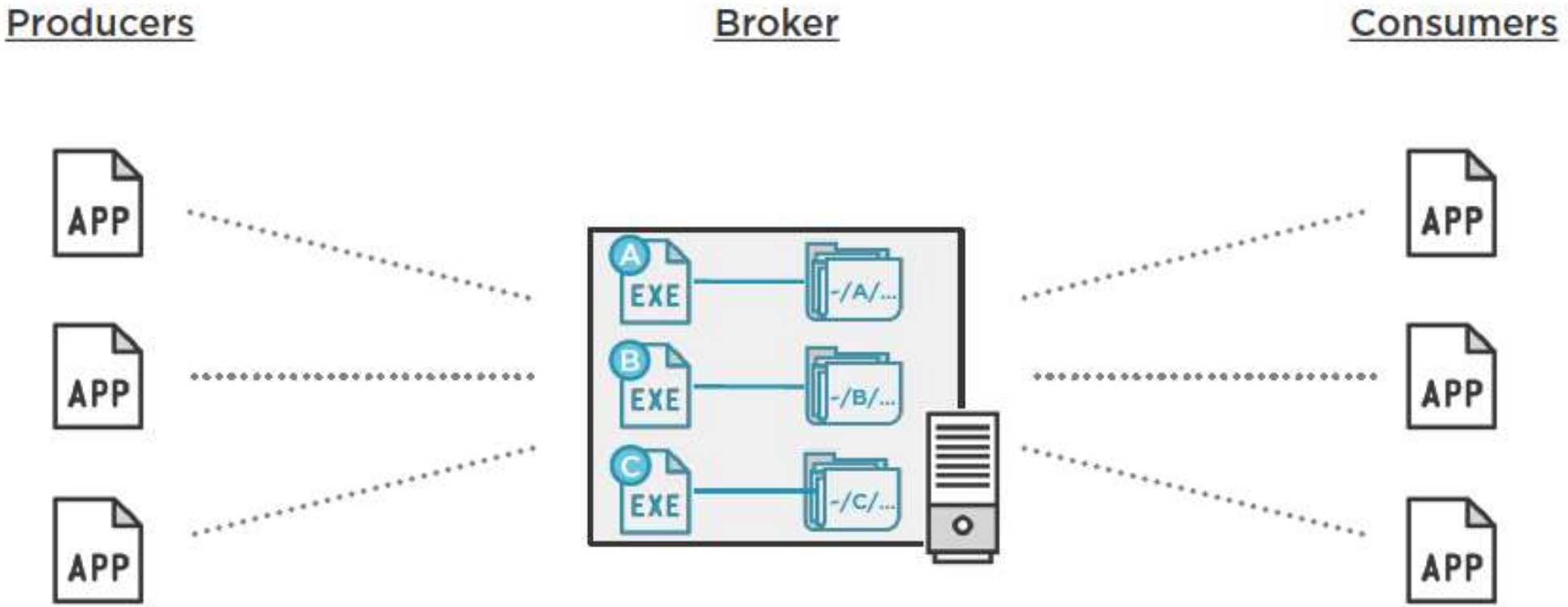
Arquitetura do Apache Kafka



Arquitetura do Apache Kafka



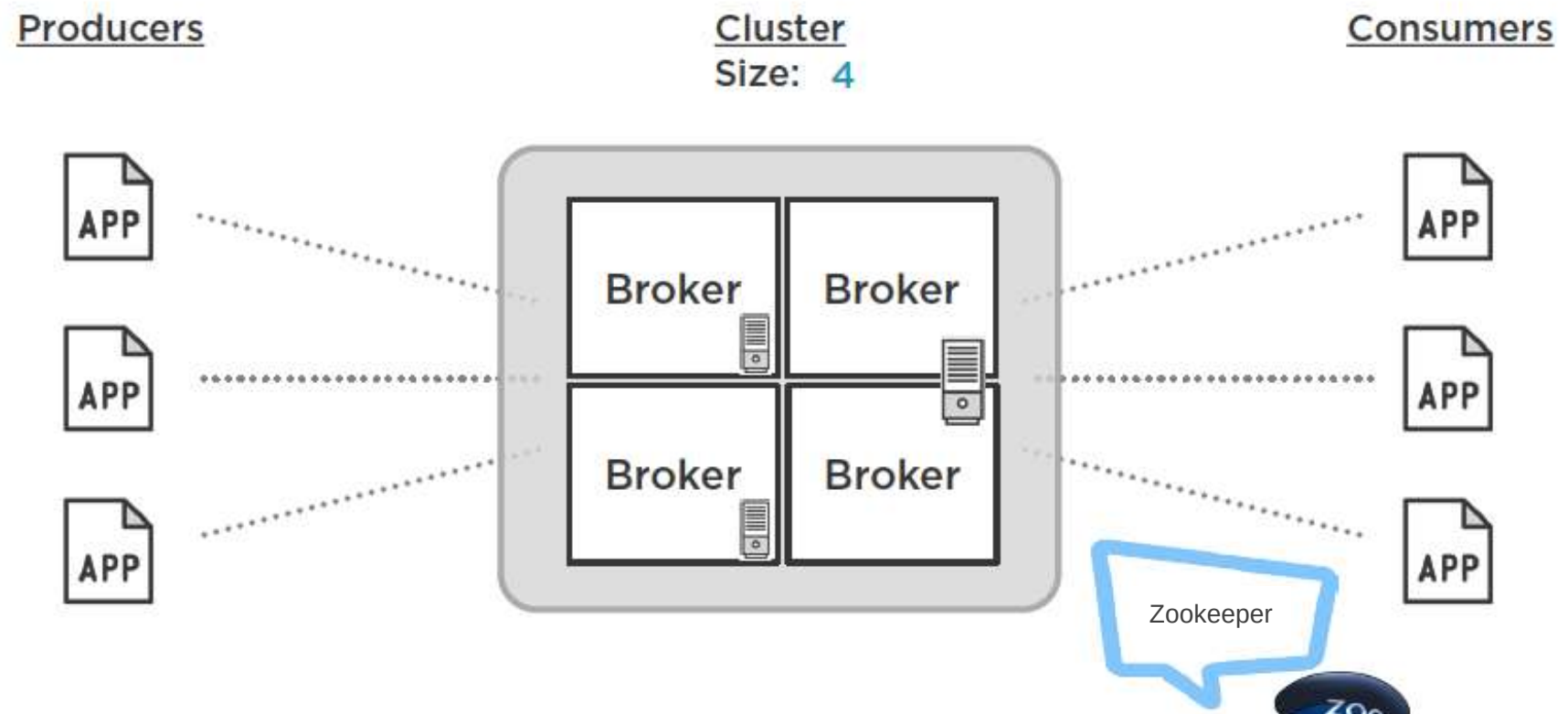
Arquitetura do Apache Kafka



Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Kafka Cluster

Apache Kafka Cluster



Em 2016 o LinkedIn divulgou que seu cluster Kafka tinha 1.400 brokers, para processar 2 Petabytes de dados por semana.

Apache Zookeeper

Apache Zookeeper

O que são **Sistemas Distribuídos**?

- Coleção de recursos que possuem um objetivo ou função específicos.
- Consistem de múltiplos nodes ou workers.
- Um sistema distribuído requer coordenação para garantir consistência e progresso para o objetivo comum.
- Cada node se comunica com os outros nodes através de mensagens.

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Zookeeper

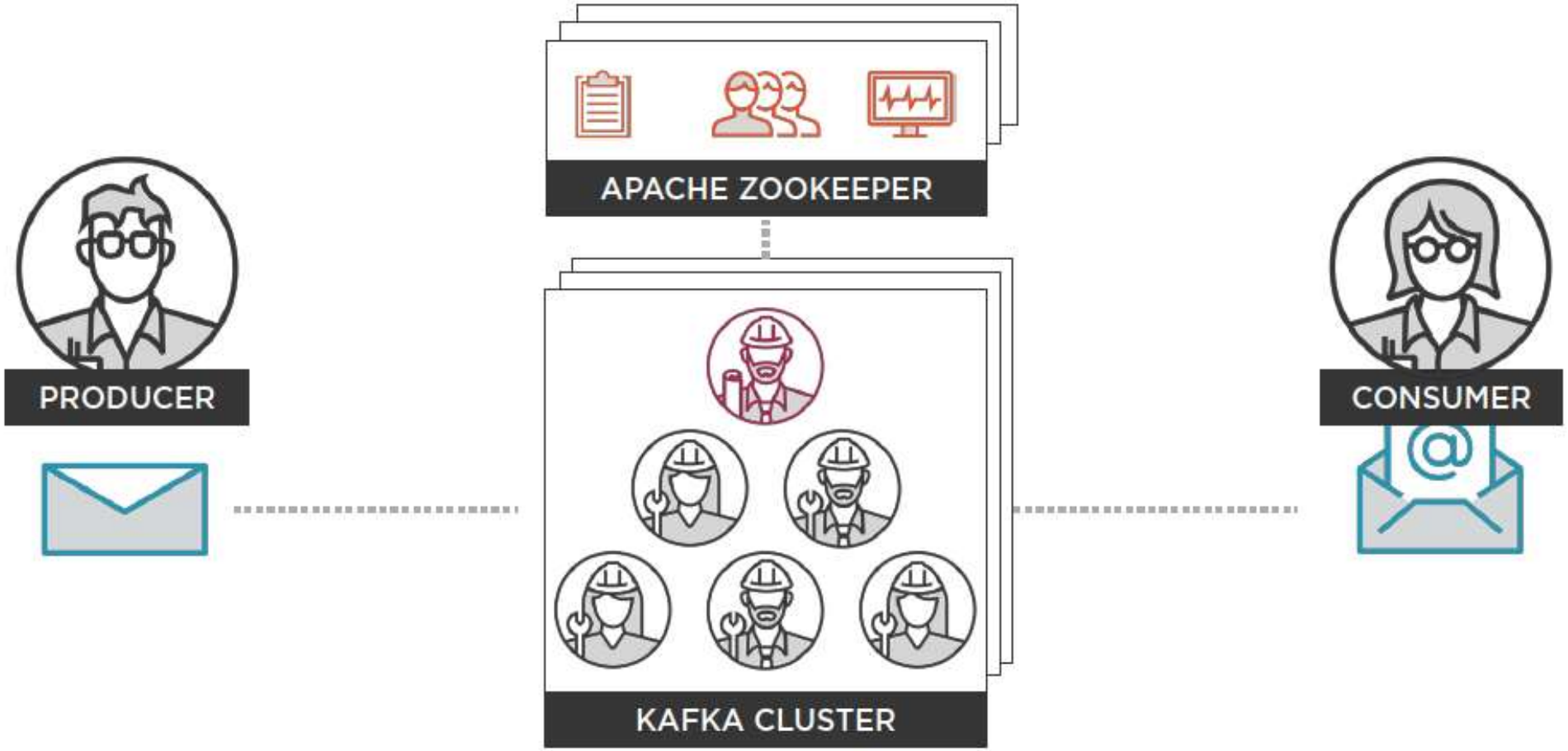
Exemplos de sistemas distribuídos?

Kafka, Hadoop, HBase, Redis, Neo4j

Apache Zookeeper

- Serviço centralizado para manter metadados sobre o cluster de nodes distribuídos, ou seja, gerenciar um sistema distribuído.
- Usado com Hadoop, HBase, Mesos, Solr, Redis e Neo4j e Kafka.
- O Zookeeper é um sistema distribuído por si próprio, consistindo de múltiplos nodes.

Apache Zookeeper



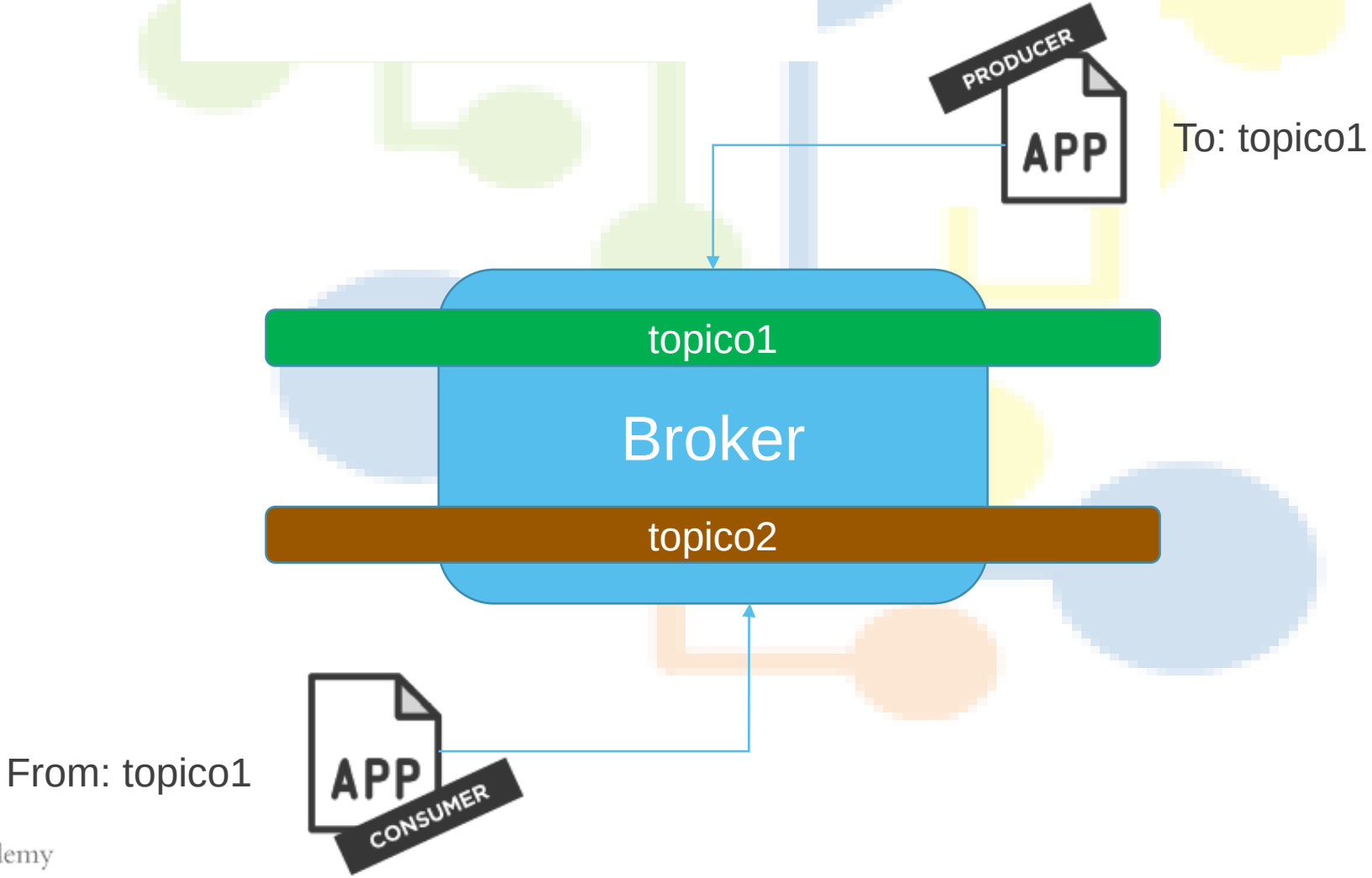
Apache Kafka Topics

Apache Kafka Topics

- Abstração central do Kafka.
- Categoria de Mensagens.
- Entidade Lógica, fisicamente representada por um arquivo de log.

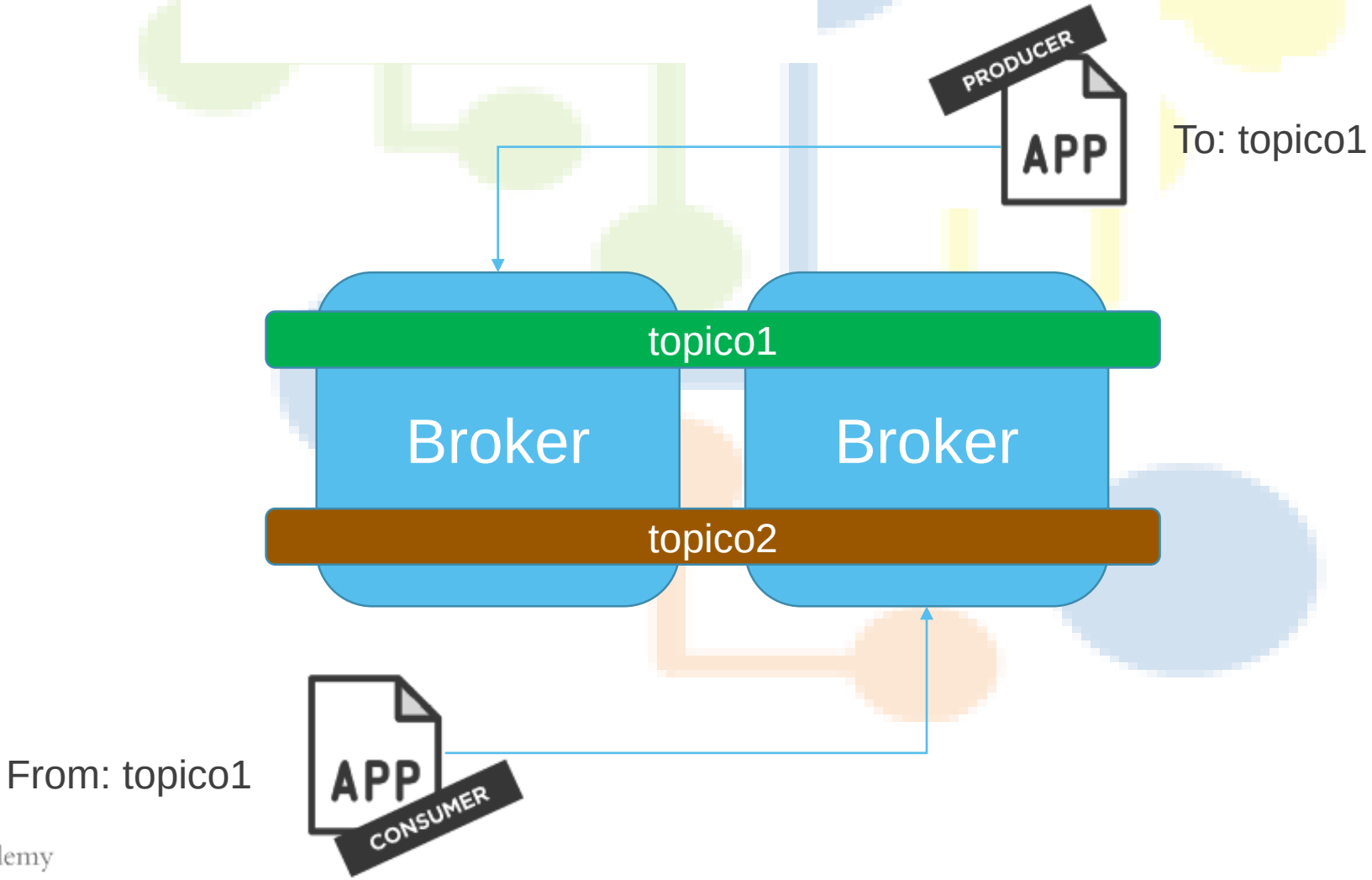
Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Kafka Topics

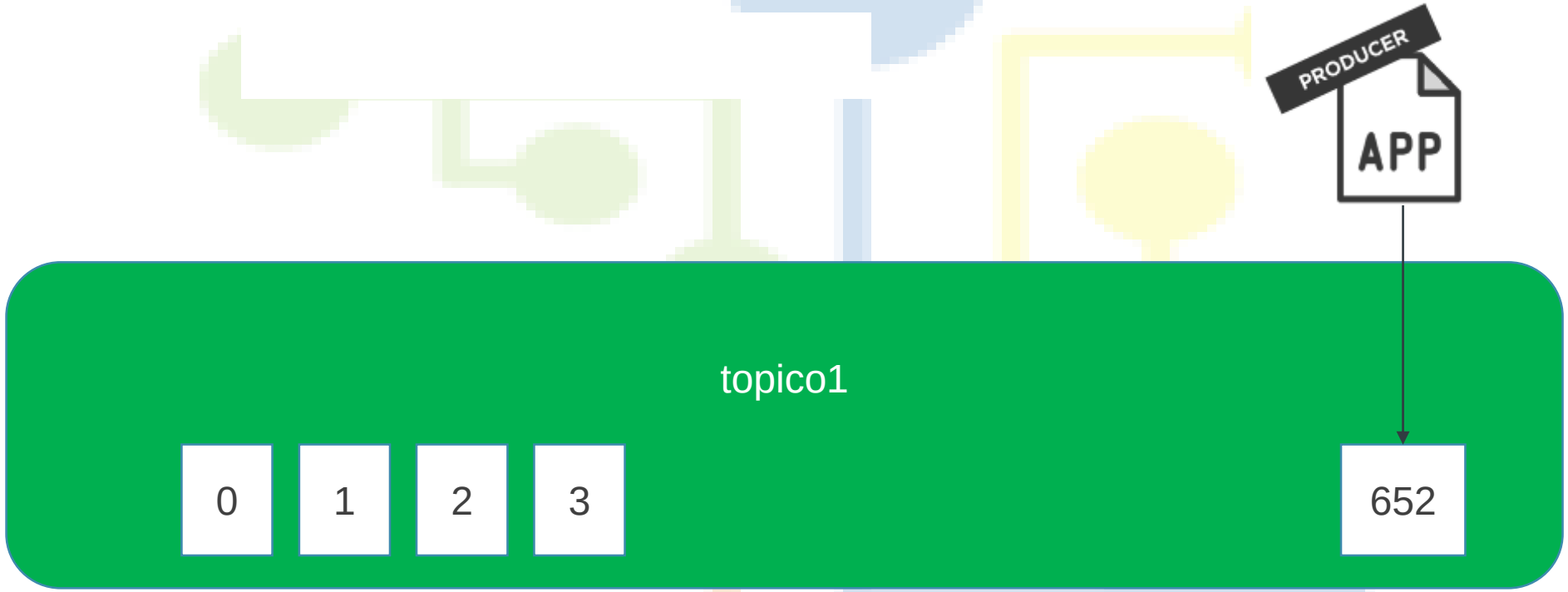


Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Kafka Topics



Apache Kafka Topics

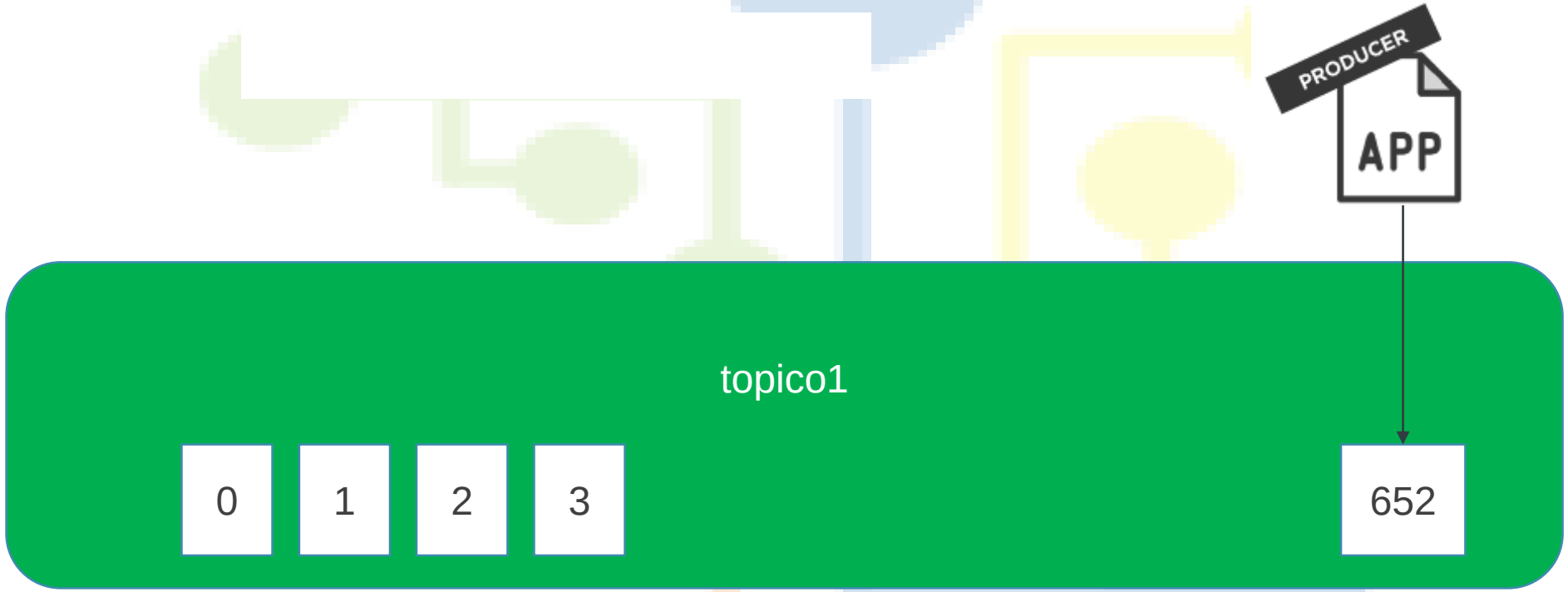


Essa arquitetura é chamada de “Event Sourcing” e o objetivo é manter o estado de aplicações capturando todas as mudanças como uma sequência de eventos imutáveis, ordenando por tempo.

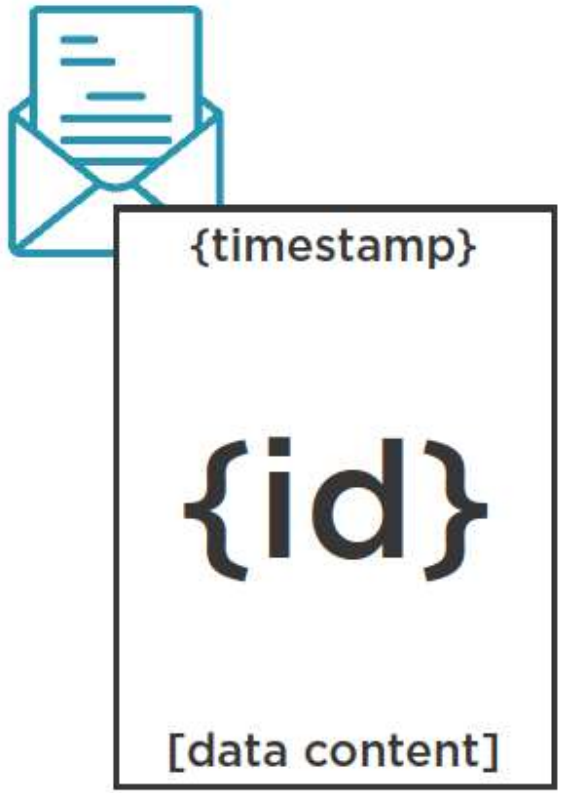
Apache Kafka Message

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Kafka Message



Apache Kafka Message

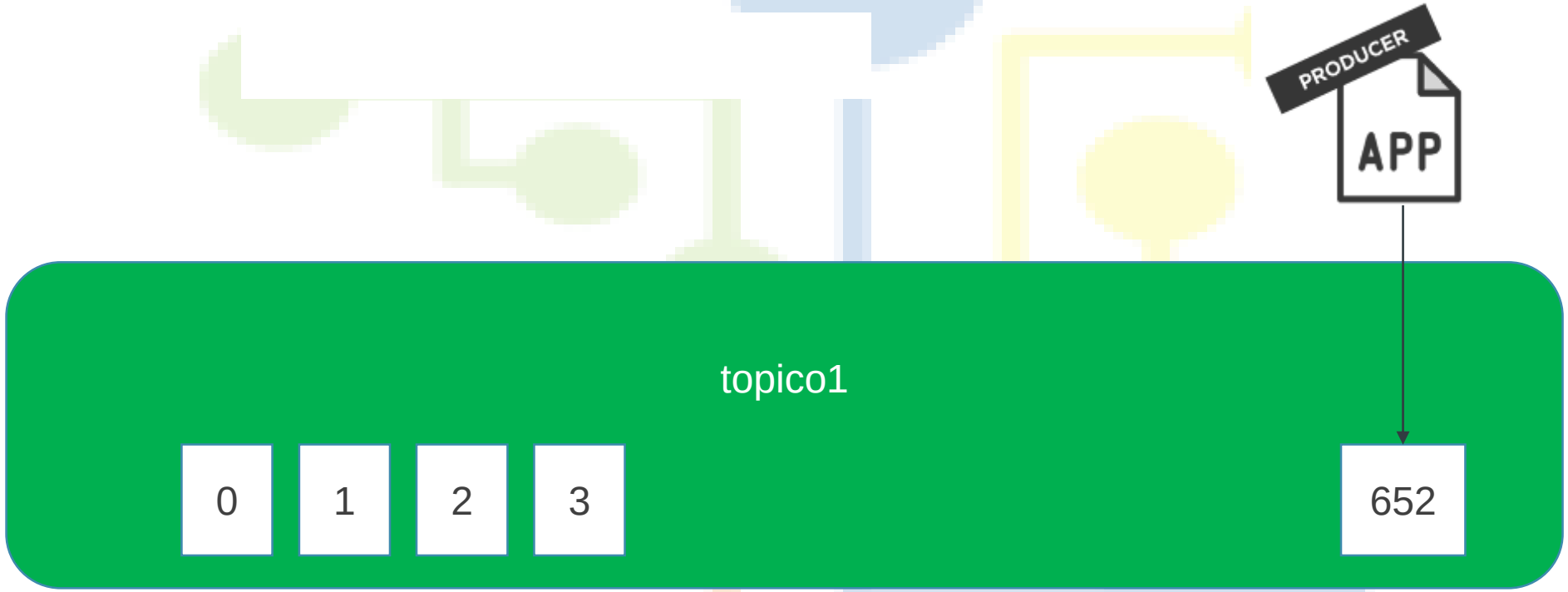


Cada mensagem tem:

- Um timestamp
- Identificador
- Payload (binary) → dados

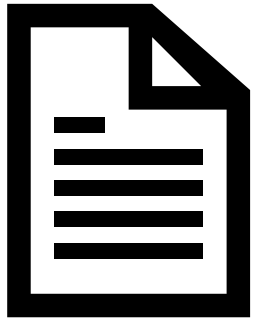
Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Kafka Message



Como o Kafka sabe que uma mensagem já foi lida pelo Consumer?

Apache Kafka Message

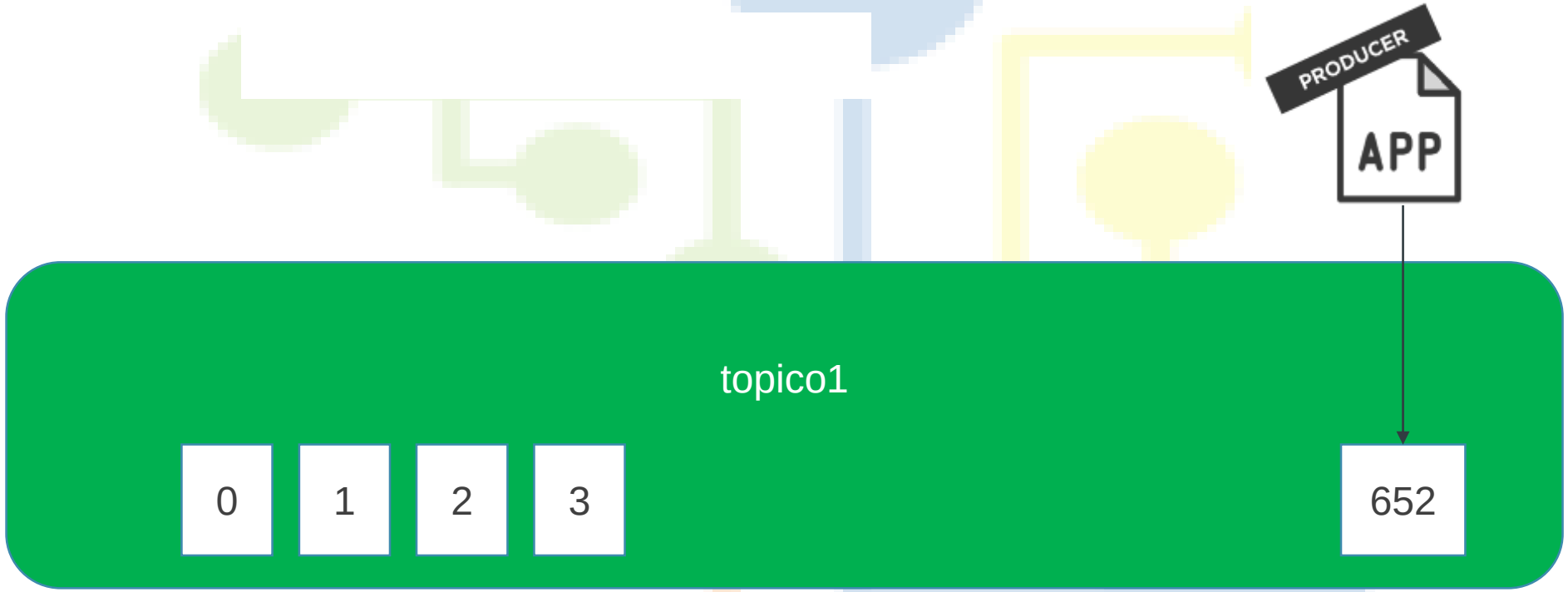


Offset

- Placeholder
- Guarda a posição da última mensagem lida
- É mantido pelo Kafka Consumer
- Corresponder ao identificador da mensagem

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Kafka Message



Por quanto tempo o Kafka retém as mensagens?

Apache Kafka Message

Política de Retenção:

- O Kafka retém todas as mensagens publicadas independentemente do consumo.
- O período de retenção é configurável e o padrão é de 168 horas (7 dias).
- O período de retenção é definido por tópico.
- É preciso considerar restrições de espaço em disco ao definir a política de retenção.

O Que São Commit Logs?

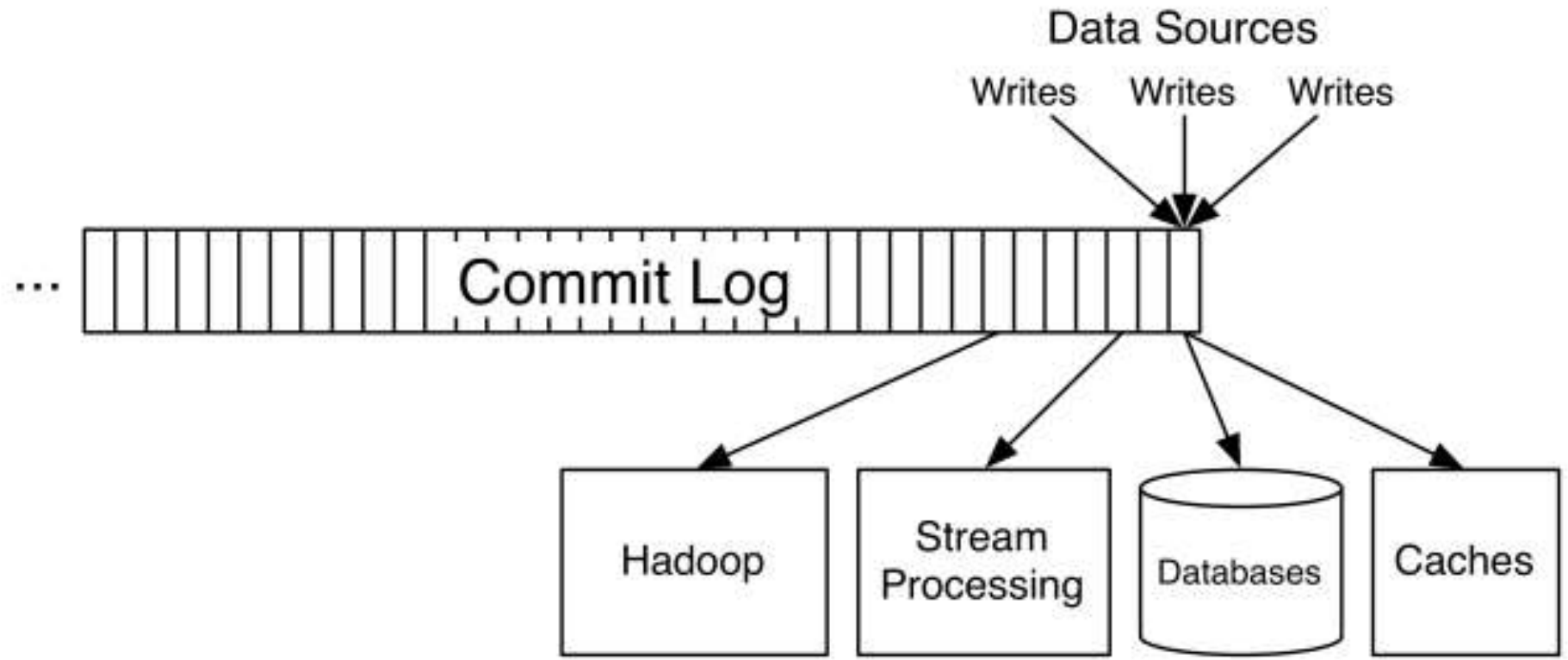
O Que São Commit Logs?

- Registro de transações
- Fisicamente armazenados e mantidos
- Fonte de confiança nos dados
- Ponto de recuperação
- Base para replicação e distribuição

O Que São Commit Logs?

Apache Kafka é um serviço de mensagens do tipo publicação/subscrição repensado como um Distributed Commit Log.

O Que São Commit Logs?



Partições no Apache Kafka

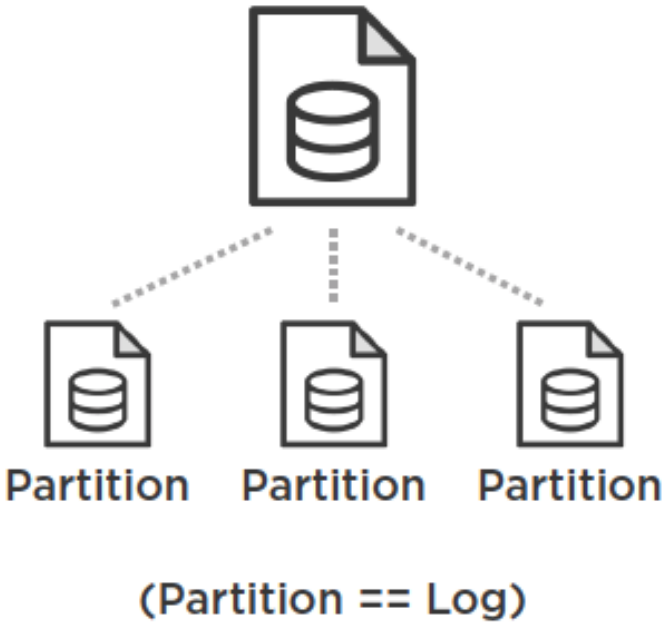
Partições no Apache Kafka

Um tópico é um conceito lógico.

Um tópico é representado por um ou mais arquivos de log.

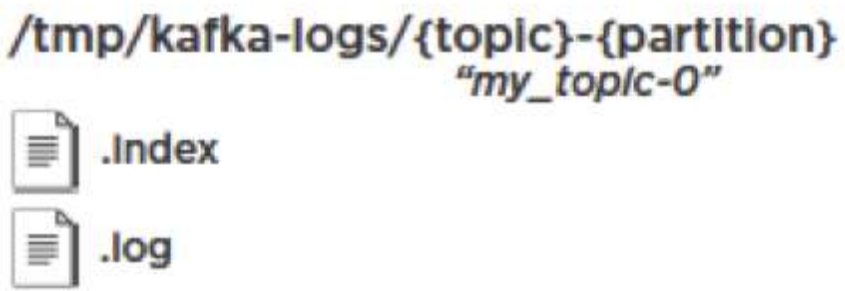
Esses arquivos são chamados de partições.

Partições no Apache Kafka



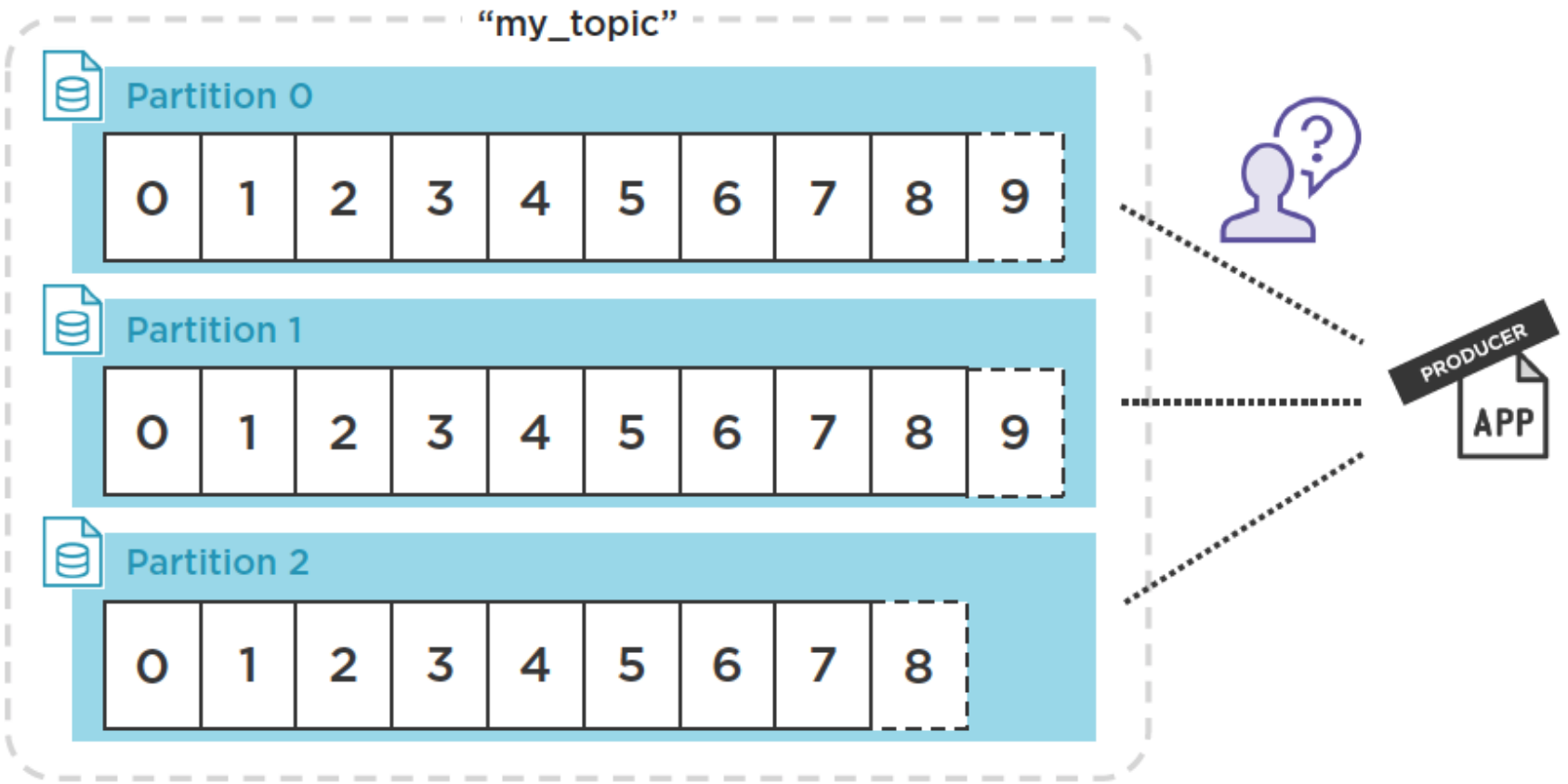
- Cada tópico tem uma ou mais partições e o número de partições é configurável
- Uma partição é a base do Kafka para conseguir:
  - Escalabilidade
  - Tolerância a Falhas
- Cada partição é mantida em um ou mais Brokers

Partições no Apache Kafka



Cada partição tem sua própria pasta dentro do diretório de logs do Kafka.

Partições no Apache Kafka



Partições no Apache Kafka

Em geral, a escalabilidade do Apache Kafka é determinada pelo número de partições sendo gerenciado por múltiplos Brokers.

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Pipeline de Captura e Armazenamento de Dados em Tempo Real

Pipeline de Captura e Armazenamento de Dados em Tempo Real

O Pipeline de Captura e Armazenamento de Dados em Tempo Real com o Apache Kafka pode ser feito da seguinte forma:

- Integrando o Kafka com o Apache Flume
- Integrando o Kafka com o Apache NiFi
- Desenvolvendo Kafka Producer e Consumer em Java e Scala
- Integrando com o StreamSets

Pipeline de Captura e Armazenamento de Dados em Tempo Real

- Criado em 2014 por um ex-engenheiro da Cloudera
- Gratuito
- Operações de DataOps
- Totalmente visual (Drag and Drop)
- Diversos drivers para diferentes fontes e destinos de dados
- Fácil integração com o Kafka
- Dashboards com estatísticas do Pipeline em tempo real

Pipeline de Captura e Armazenamento de Dados em Tempo Real

Dados gerados pelos sensores IoT serão automaticamente levados para armazenamento no Data Lake, através de uma camada de mensagens, que garante entrega dos dados.

Sempre que um novo registro for inserido ou atualizado no bando de dados relacional, o Pipeline levará os dados em tempo real para armazenamento no Data Lake.

Pipeline de Captura e Armazenamento de Dados em Tempo Real

No curso de **Machine e IA em Ambientes Distribuídos**, vamos reproduzir esses Pipelines para processamento dos dados em tempo real e aplicação de modelos de Machine Learning.

Muito Obrigado.

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