



PUC Minas
DIRETORIA DE
EDUCAÇÃO CONTINUADA

Pós Graduação *Lato Sensu*

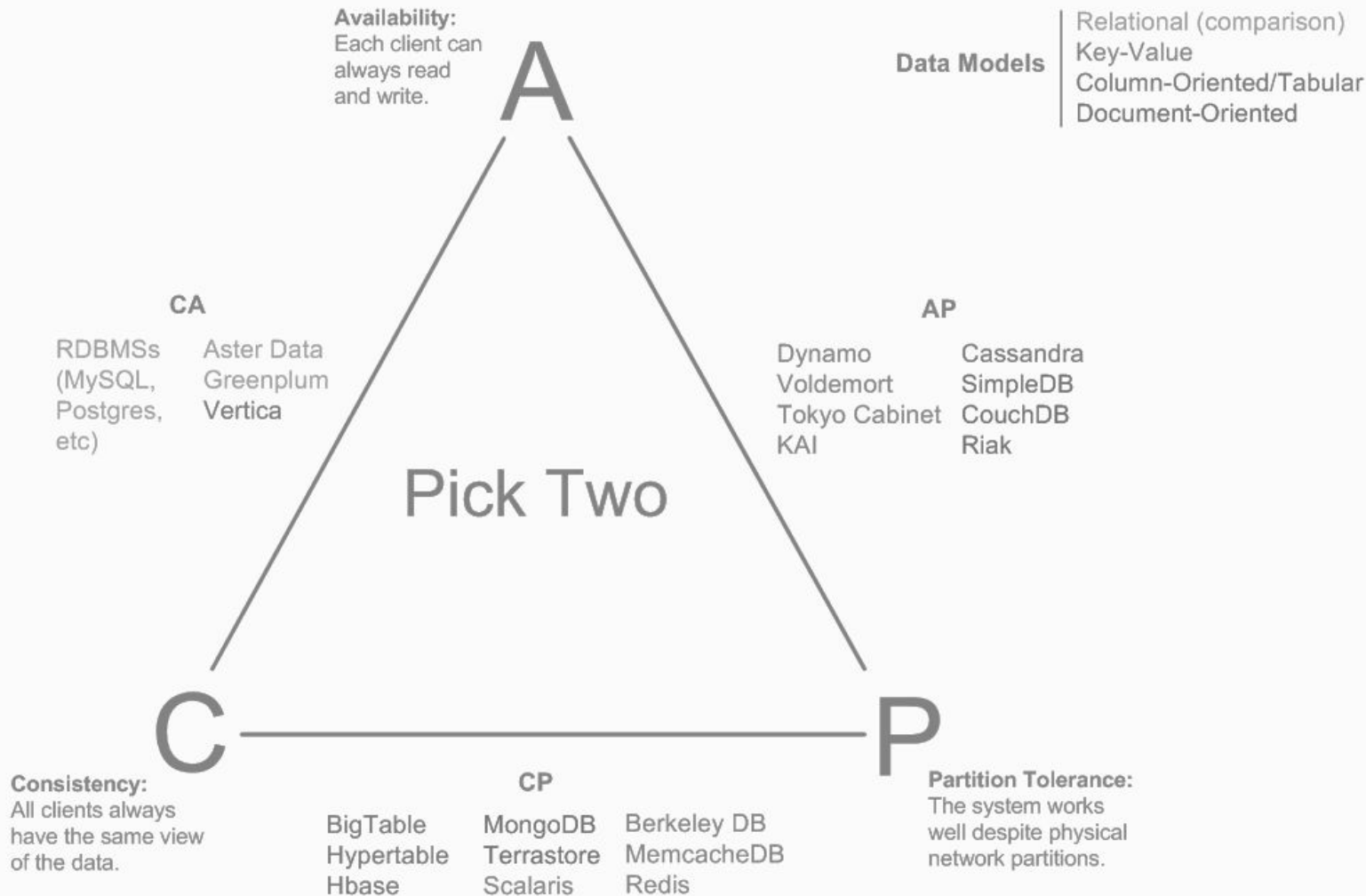
Bancos de dados não
relacionais

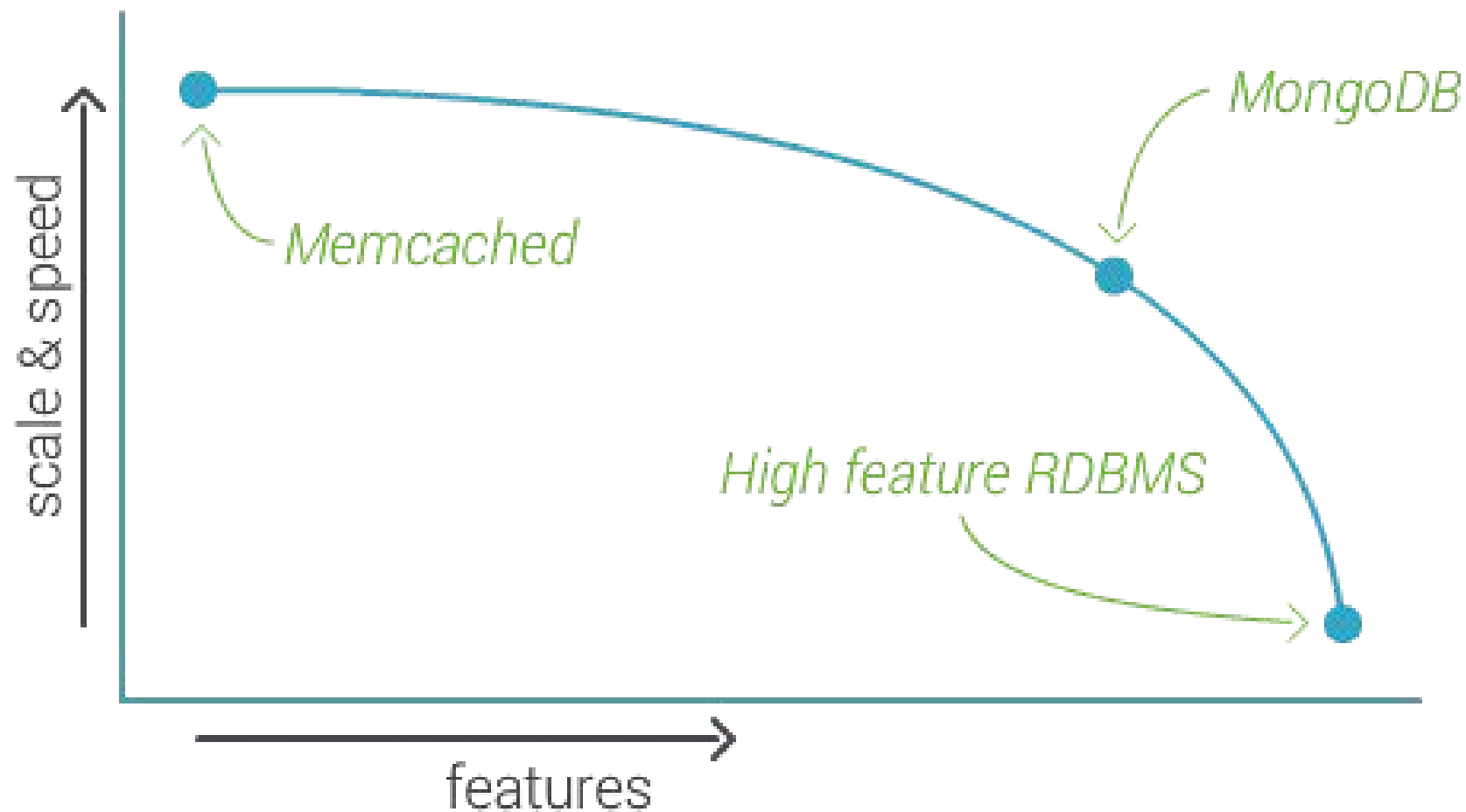
Casos de uso

Trabalho Prático

- Coletar informações de redes sociais ou **importar dados externos** e armazenar ~1M de dados em um banco NoSQL
- Extrair informações do tipo :
 - Termos mais frequentes
 - Volume x dia
 - Volume x hora do dia
- Entregar até o dia **23**/12/2016 e submeter via github.

“Todo banco de dados é ótimo até que você comece a usá-los”



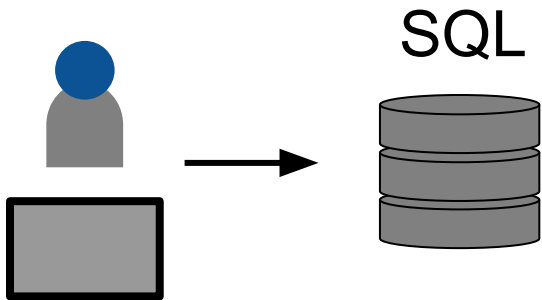


Analytics de baixa latência

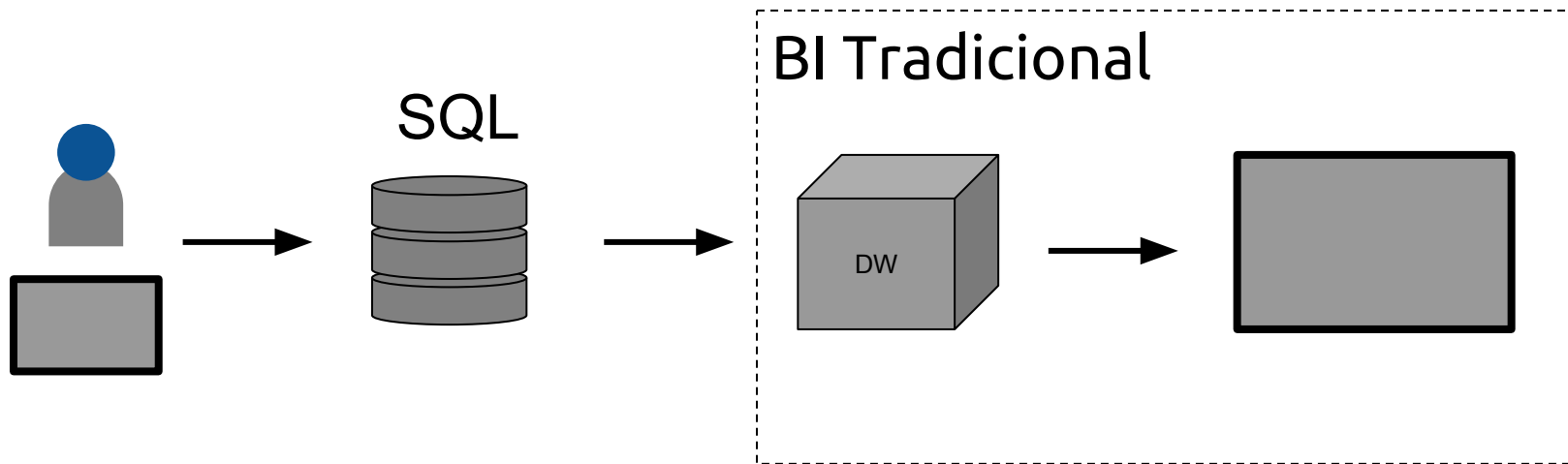
Analytics de baixa latência

- Dashboards realtime;
- Pouca interatividade;
- Acompanhamento de atividades.

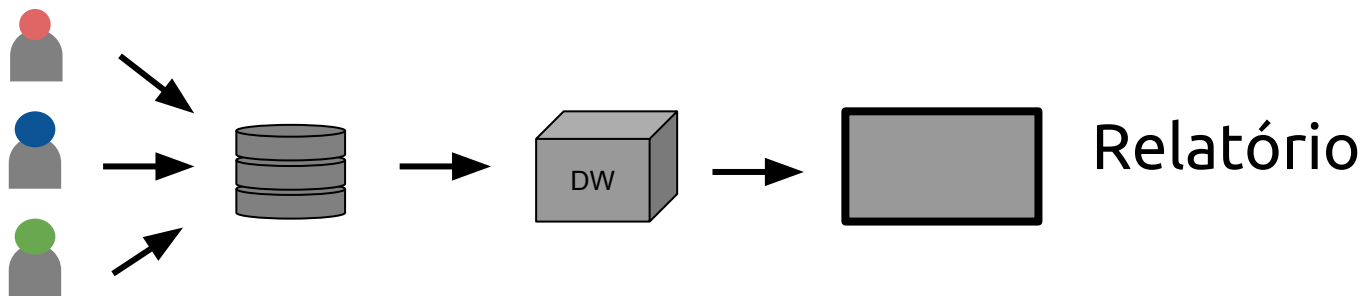
Analytics de baixa latência



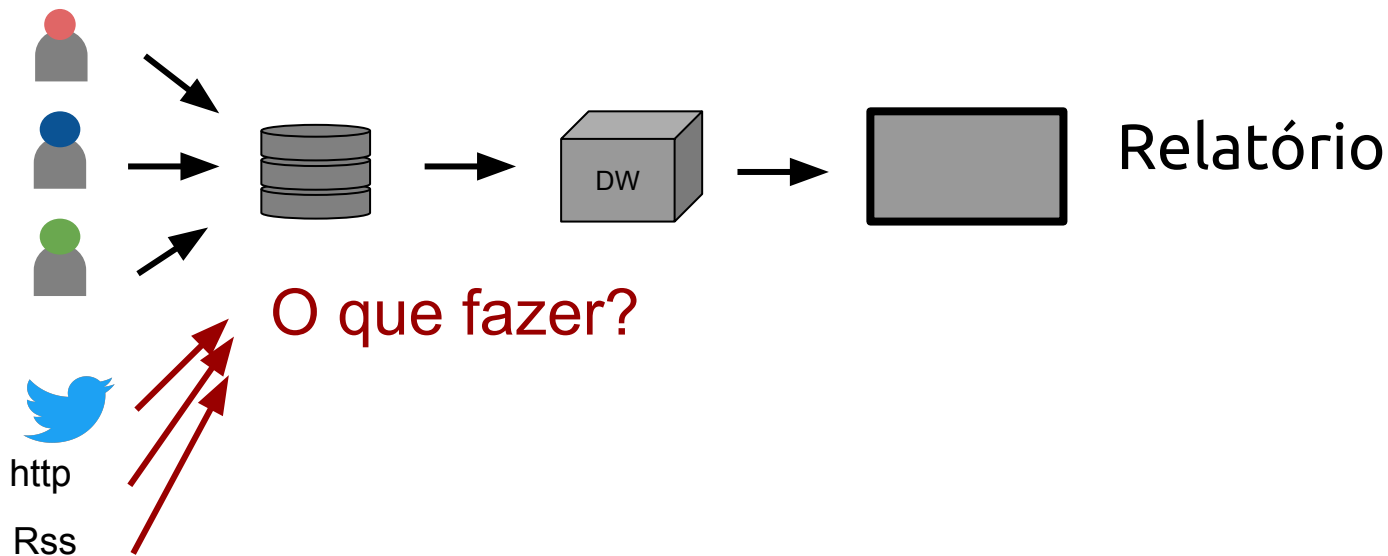
Analytics de baixa latência



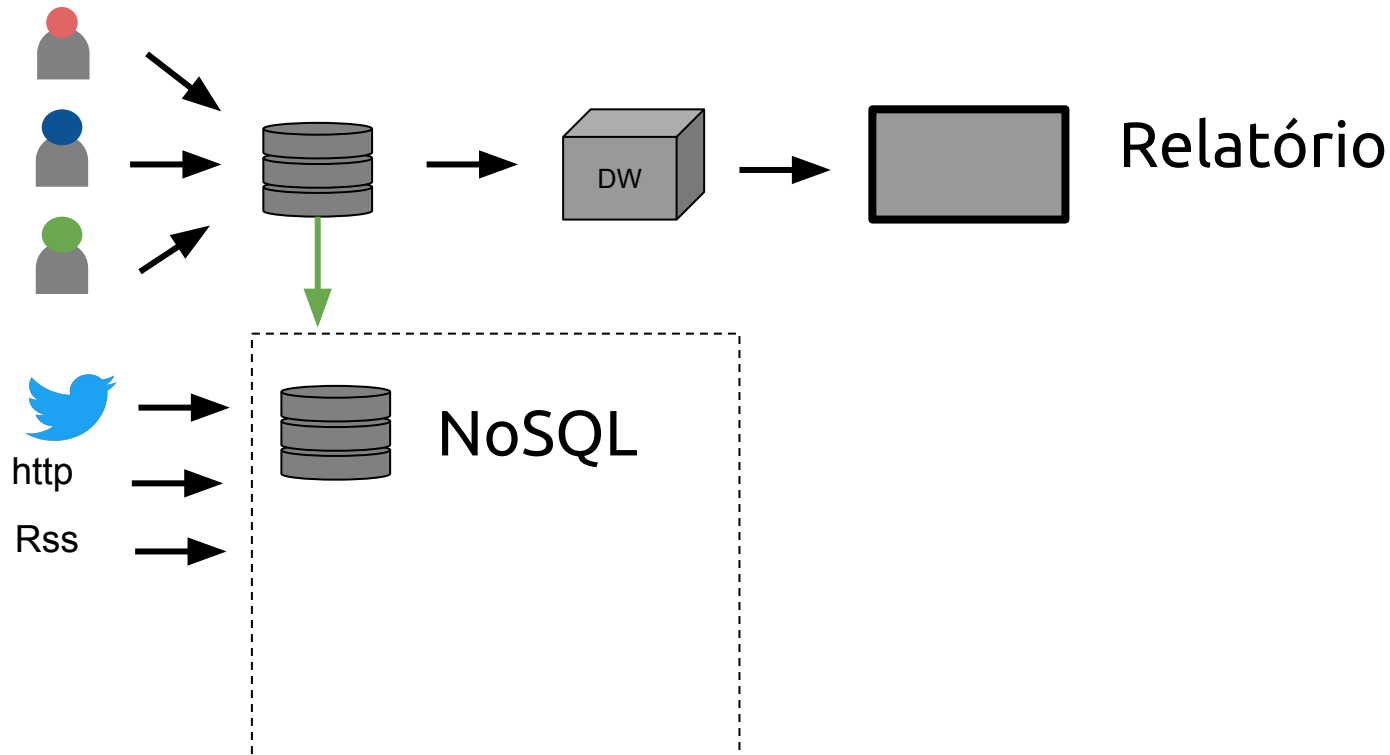
Analytics de baixa latência



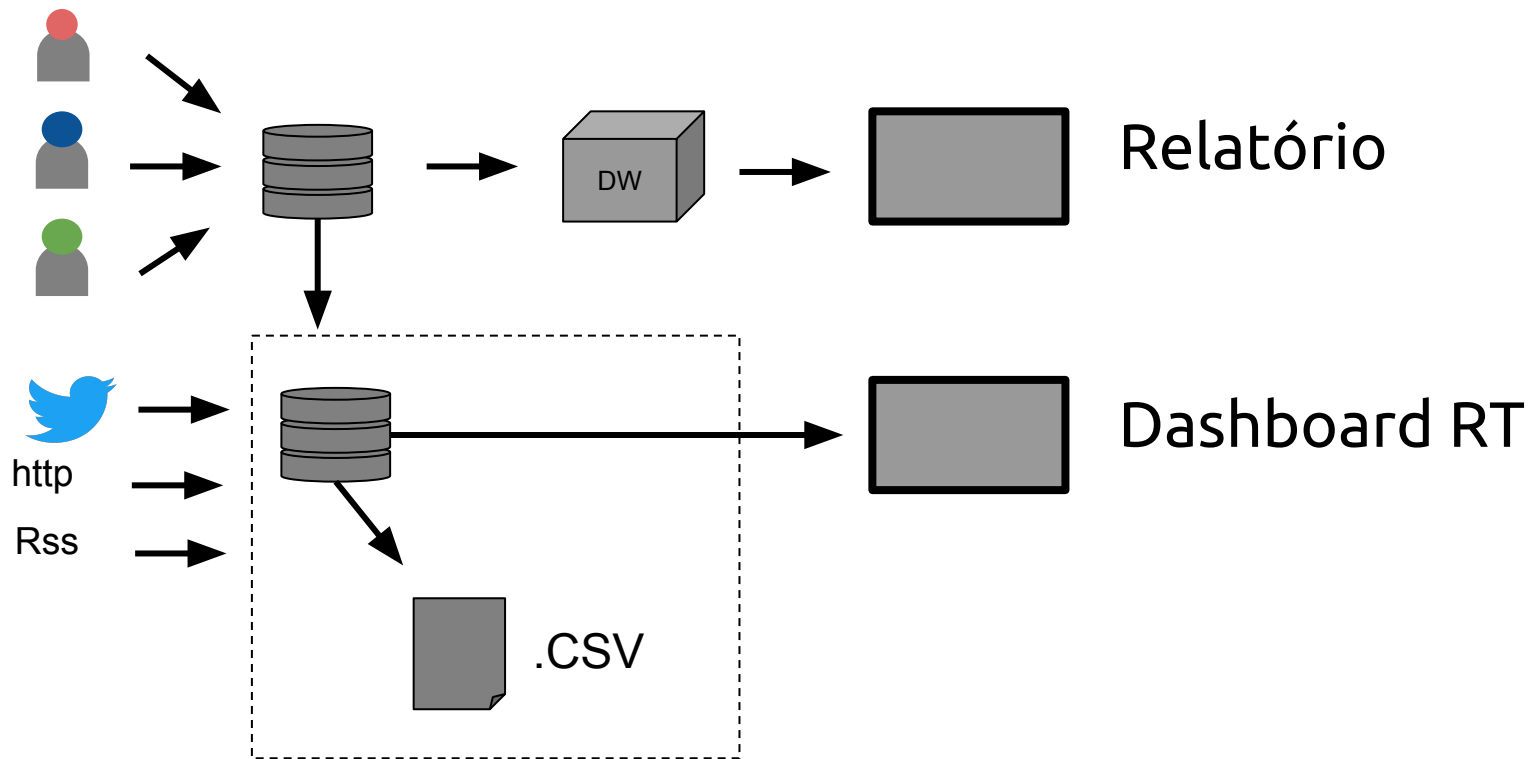
Analytics de baixa latência



Analytics de baixa latência



Analytics de baixa latência

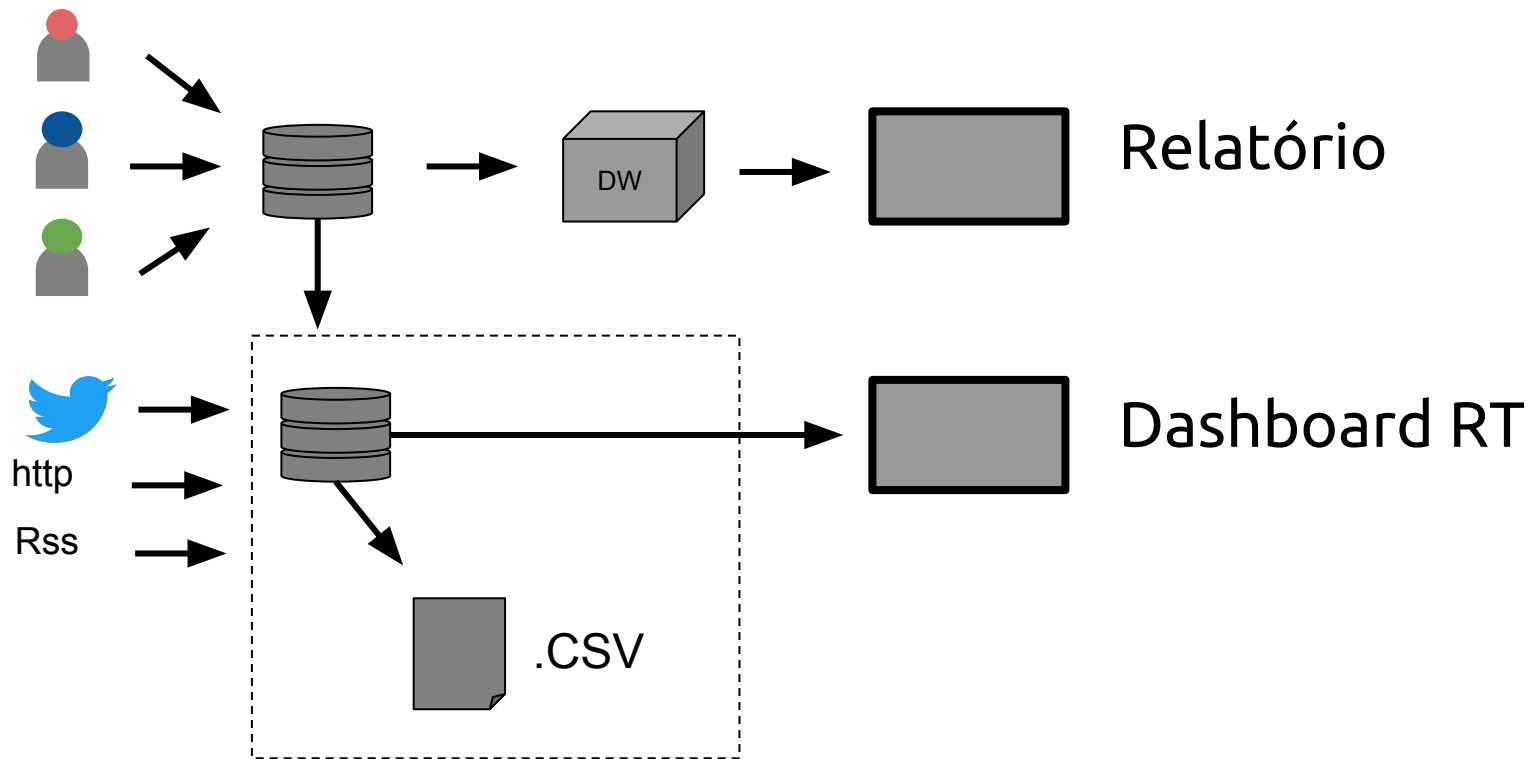


Infraestrutura Big Data

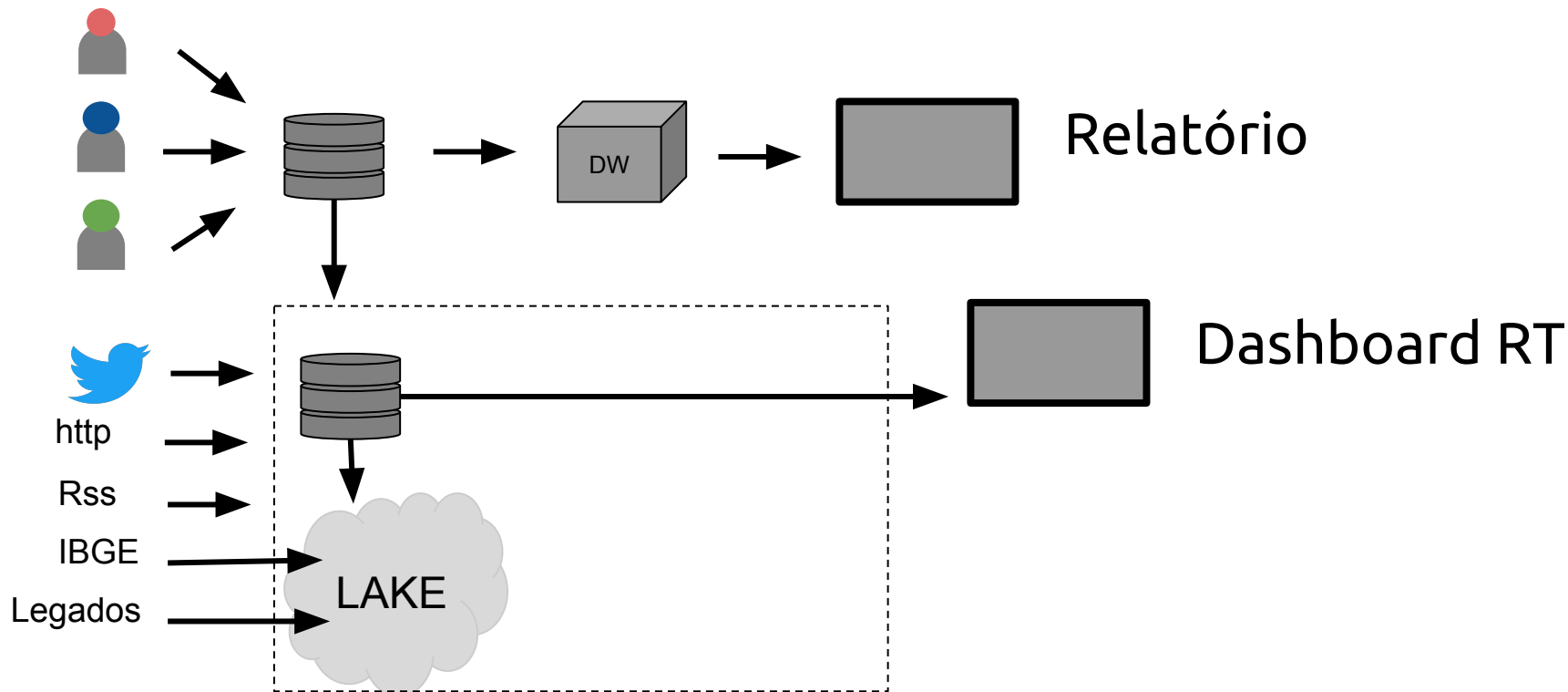
Infraestrutura Big Data

- Variedade de conjunto de banco de dados;
- Processamento de um grande volume de informação;
- Variedade de informações;
- Complexidade de infraestrutura.

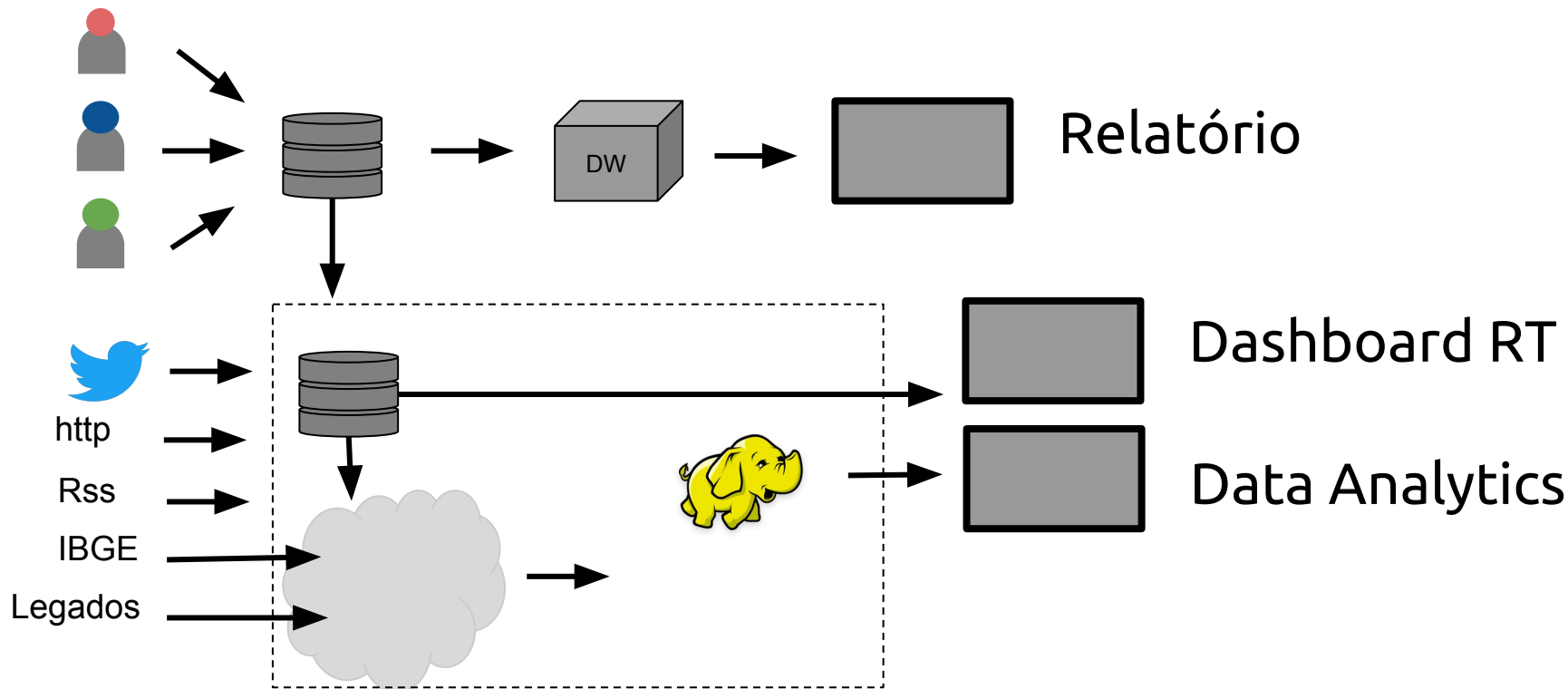
Infraestrutura Big Data



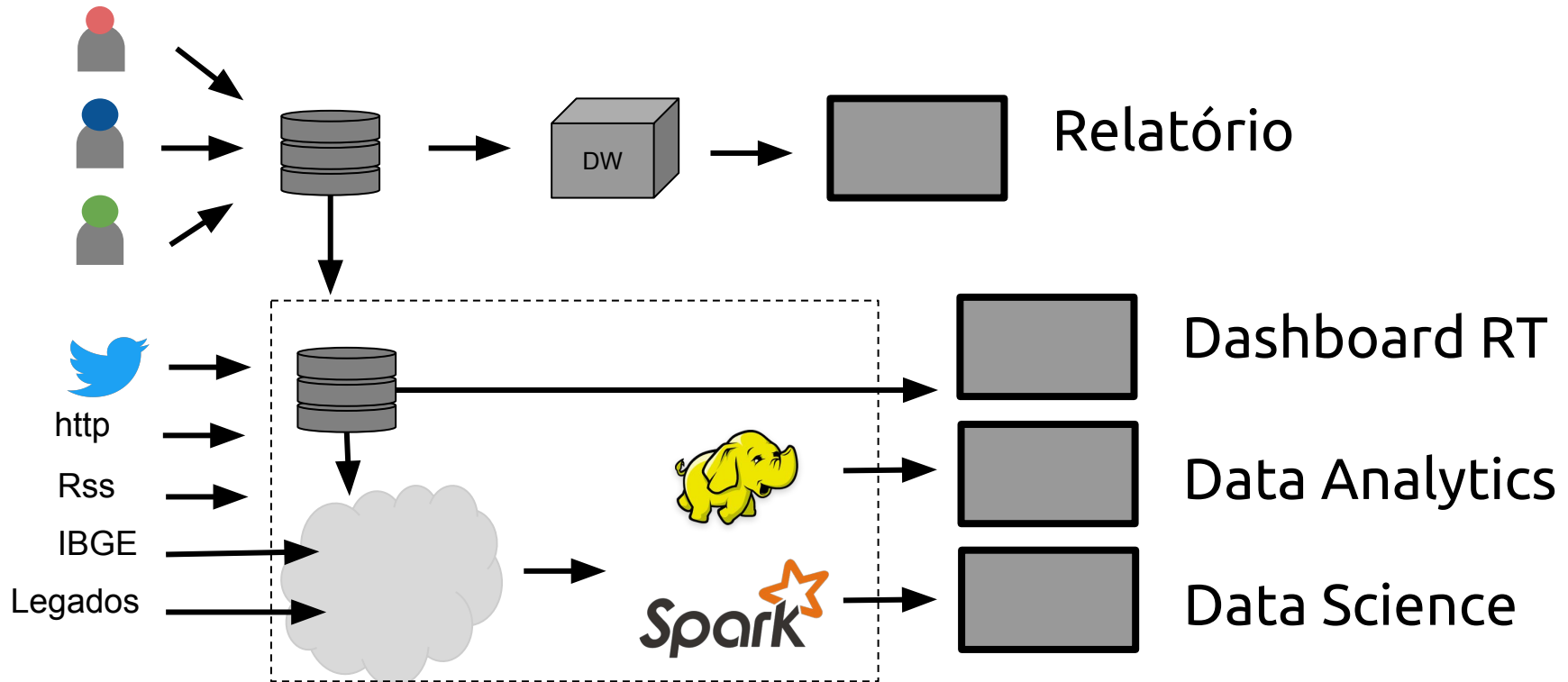
Infraestrutura Big Data



Infraestrutura Big Data



Infraestrutura Big Data



Por que não fazer tudo em um só lugar?

- SQL
 - Inserção rápida dos dados;
 - Poluir aplicação.
- NoSQL
 - Necessidade das transações;
- Hadoop
 - Overhead de processamento;

Banco de dados as a Service

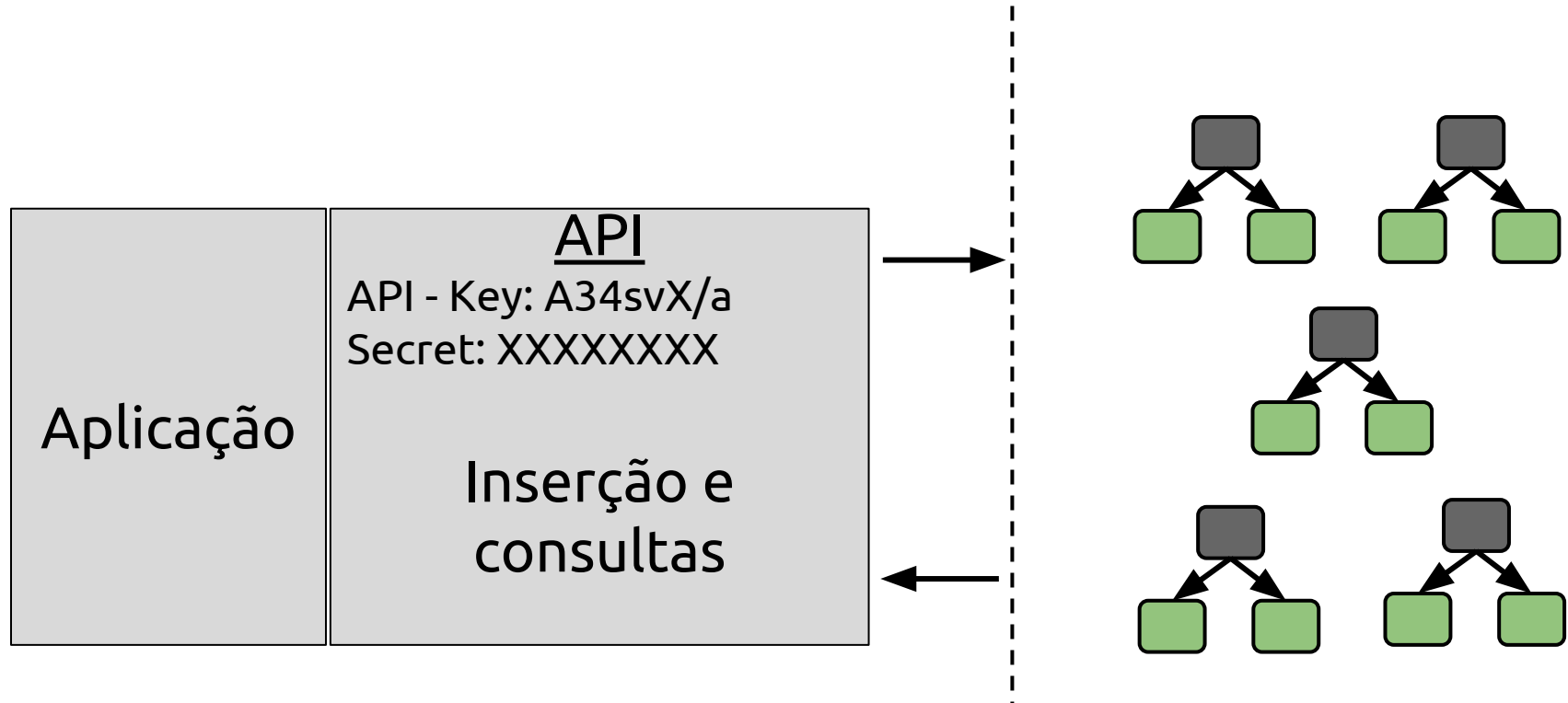
Data Base as a Service

- Projetos com poucos recursos;
- Sem necessidade de dev-ops;
- Aplicações pequenas;
- Pay as you use:
 - Transações
 - Quantidade de dados

Data Base as a Service

- Azure Table Service
- DynamoDB
- BigTable
- Amazon Redshift

Data Base as a Service



Create Table

Cancel 



PRIMARY KEY

PROVISIONED
THROUGHPUT CAPACITY

THROUGHPUT ALARMS
(optional)

Table Name:

Table will be created in eu-west-1 region

Primary Key:

DynamoDB is a schema-less database. You only need to tell us your primary key attribute(s).

Primary Key Type: ☒ Hash ☐ Hash and Range

☒ String ☐ Number

Hash Attribute Name:



Choose a hash attribute that ensures that your workload is evenly distributed across hash keys.

For example, "Customer ID" is a good hash key, while "Game ID" would be a bad choice if most of your traffic relates to a few popular games.

[Learn more about choosing your primary key](#)

Cancel

Continue 

Help 

Streaming de alta resiliência

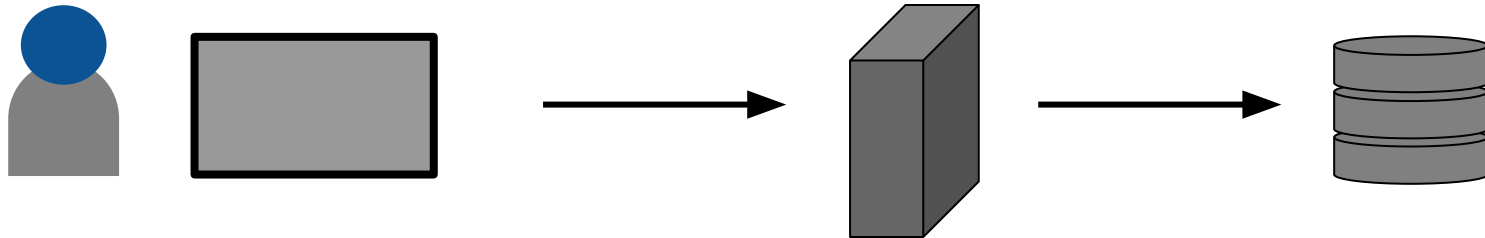
Streaming de alta resiliência

- Fluxo **contínuo** de informação
- Inserção de uma **grande** quantidade de dados
- Aplicações **resistentes** a falhas

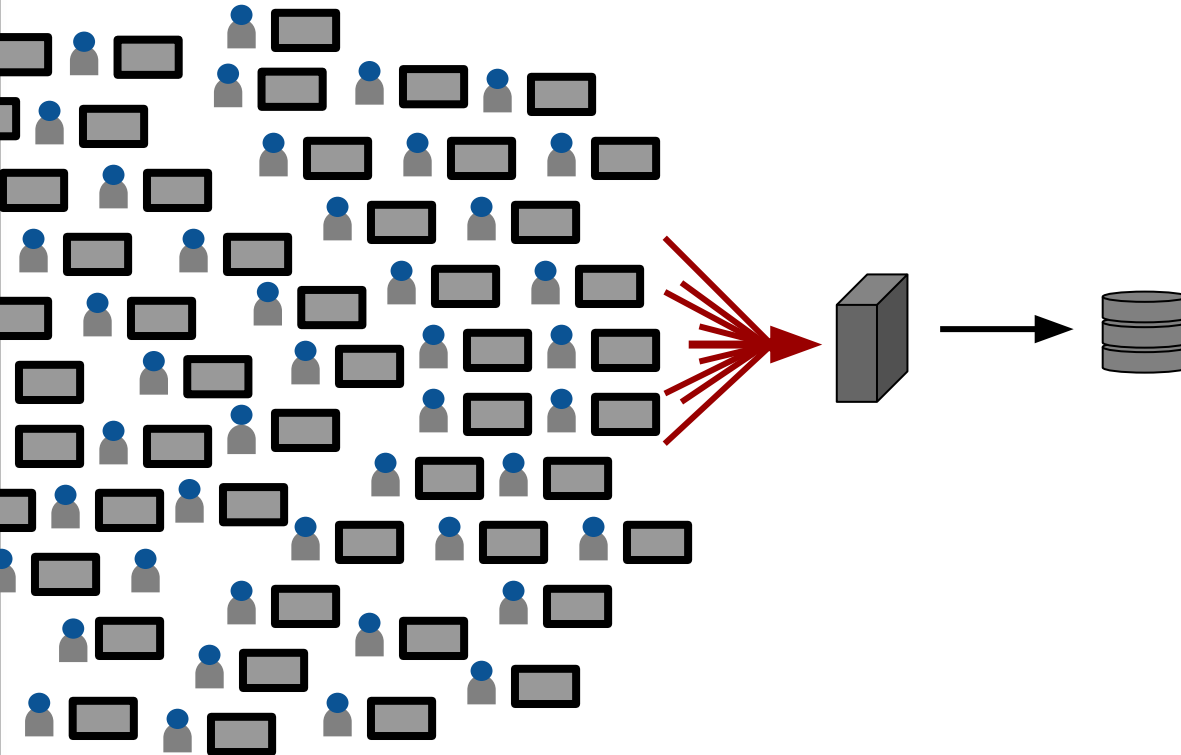
Casos:

- IoT
- Redes sociais
- Coleta de grande volumes em geral

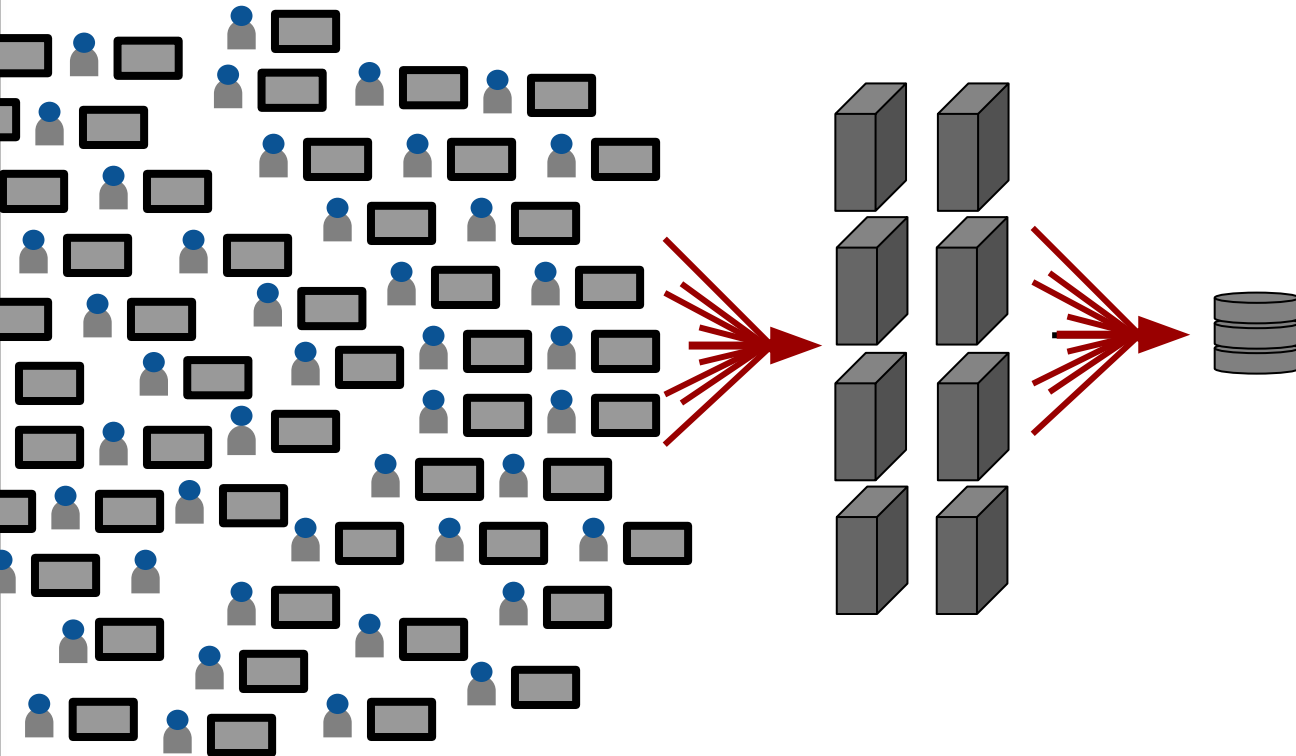
Streaming de alta resiliência



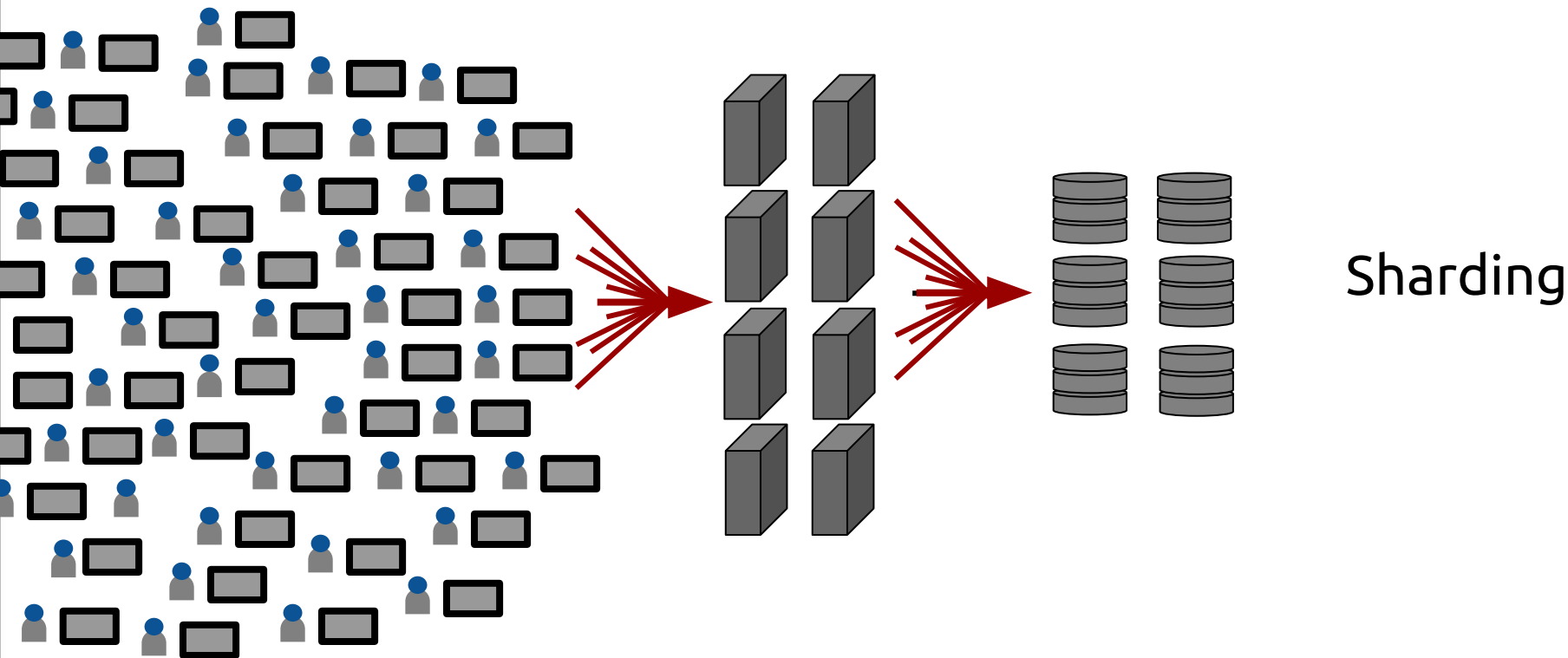
Streaming de alta resiliência



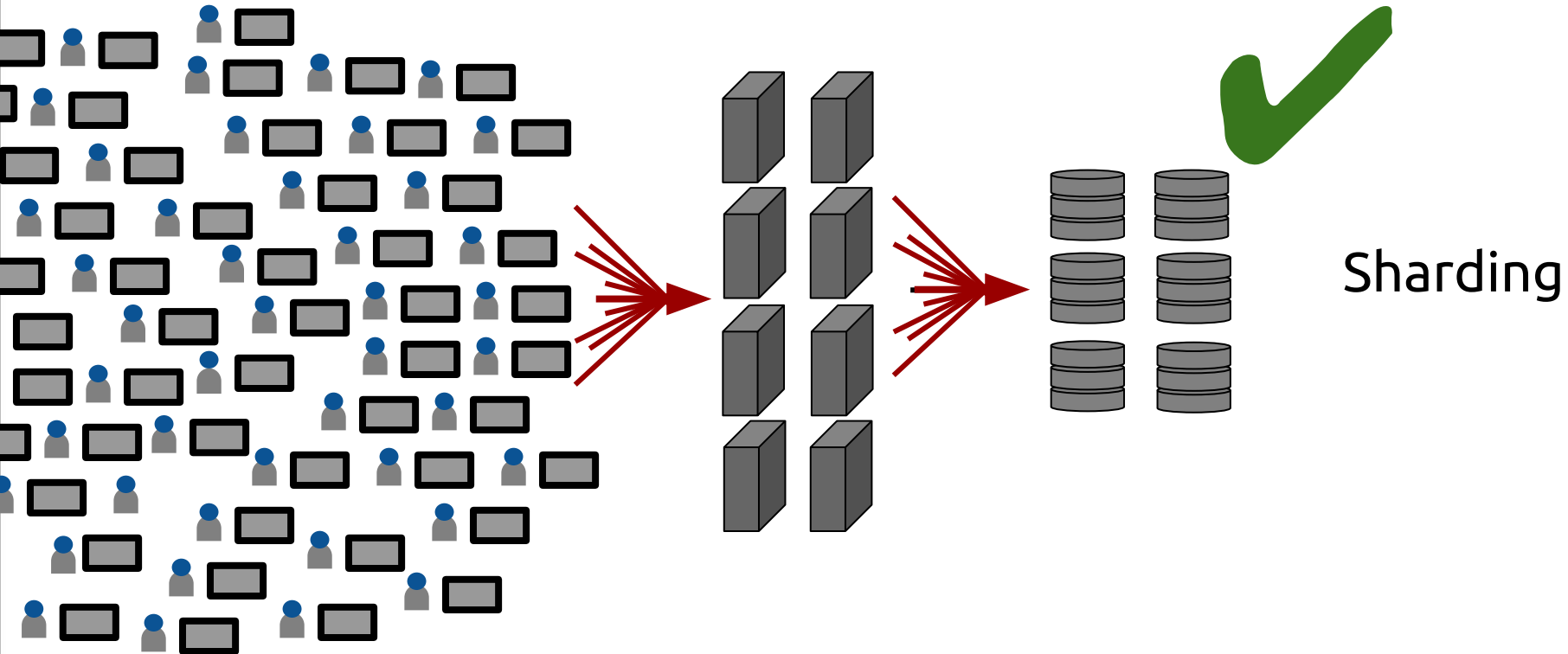
Streaming de alta resiliência



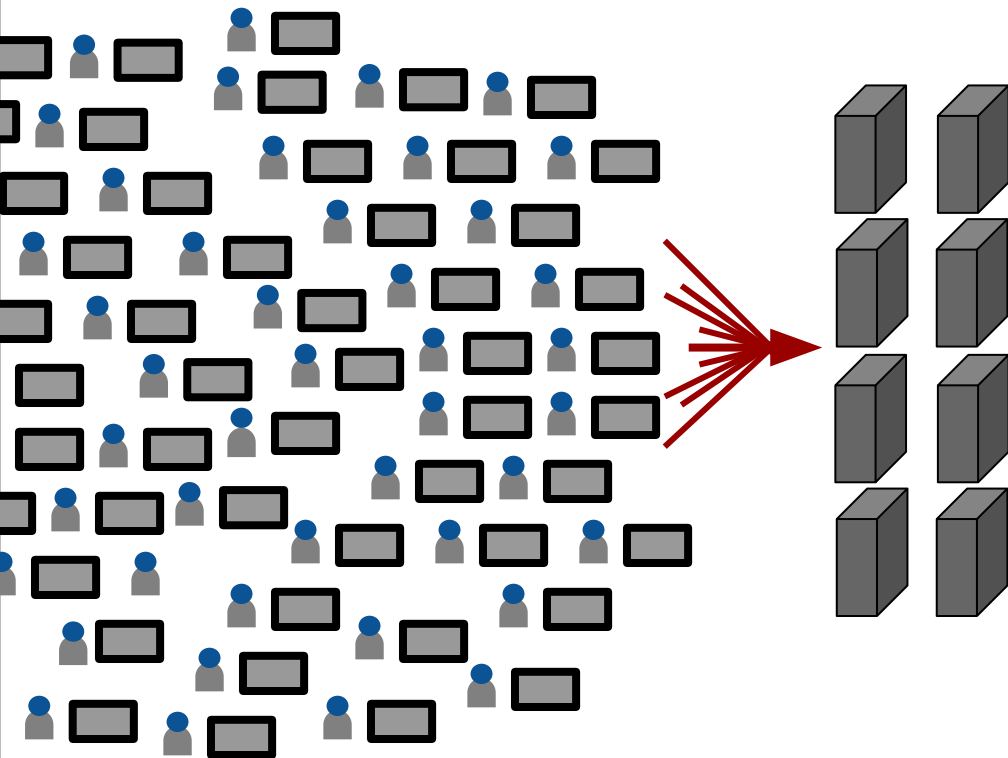
Streaming de alta resiliência



Streaming de alta resiliência



Streaming de alta resiliência



Como utilizar
tecnologias NoSQL para
auxiliar a construção de
sistemas resilientes

Sistemas de filas

- Armazenamento temporário de dados
- Compartilhamento de informação

Principais Sistemas:

- **Kafka**
- **RabbitMQ**
- ActiveMQ
- Kestrel

Chave-Valor (Key-Value)

0001



A

0002



B

0003



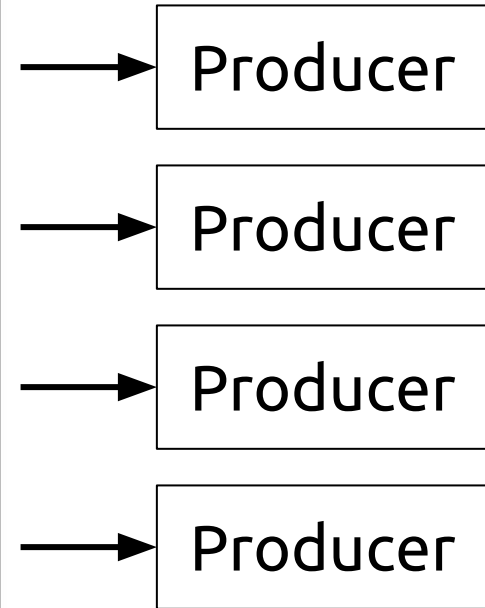
C

0004

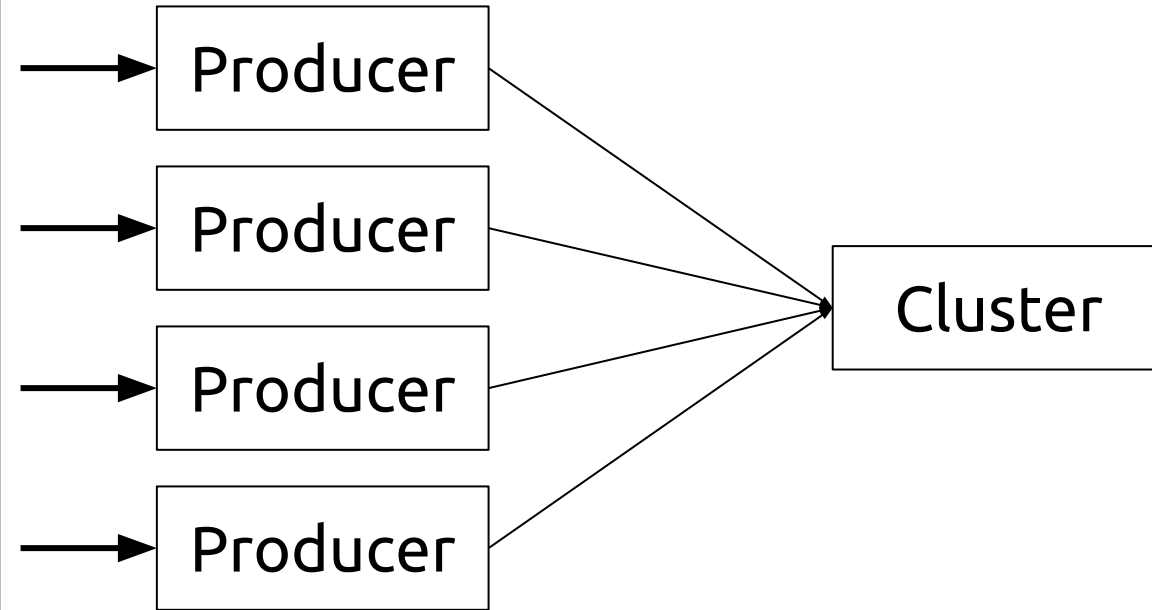


D

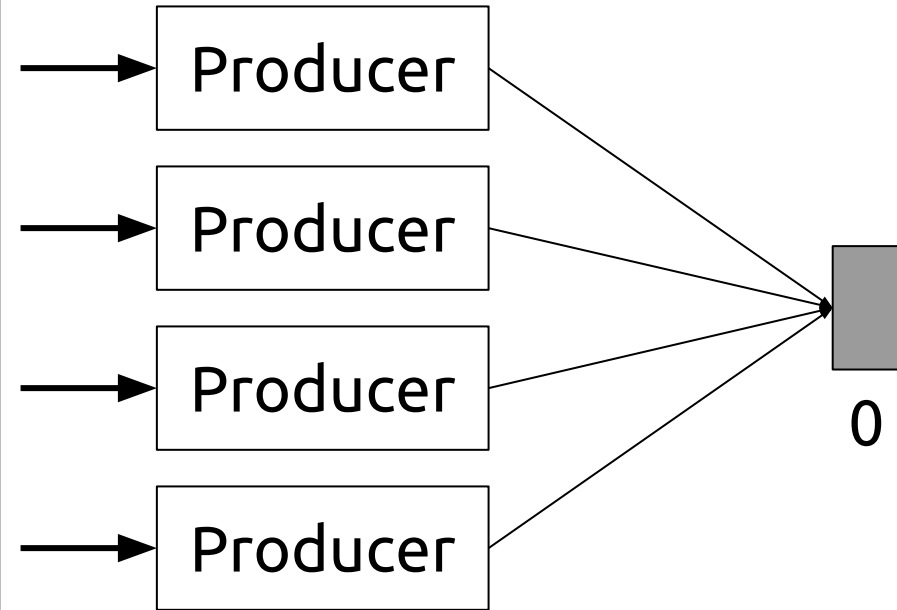
Sistema de filas



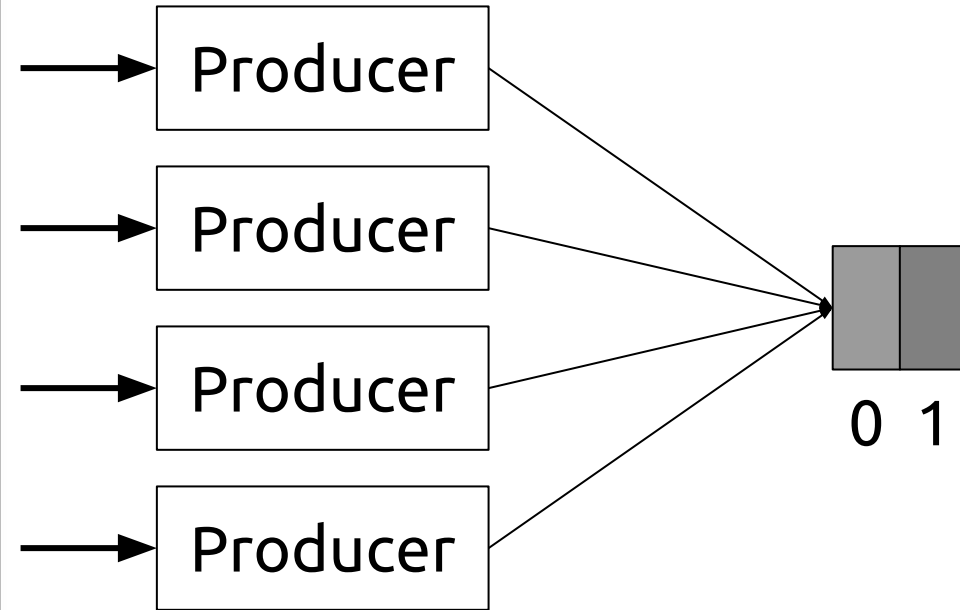
Sistema de filas



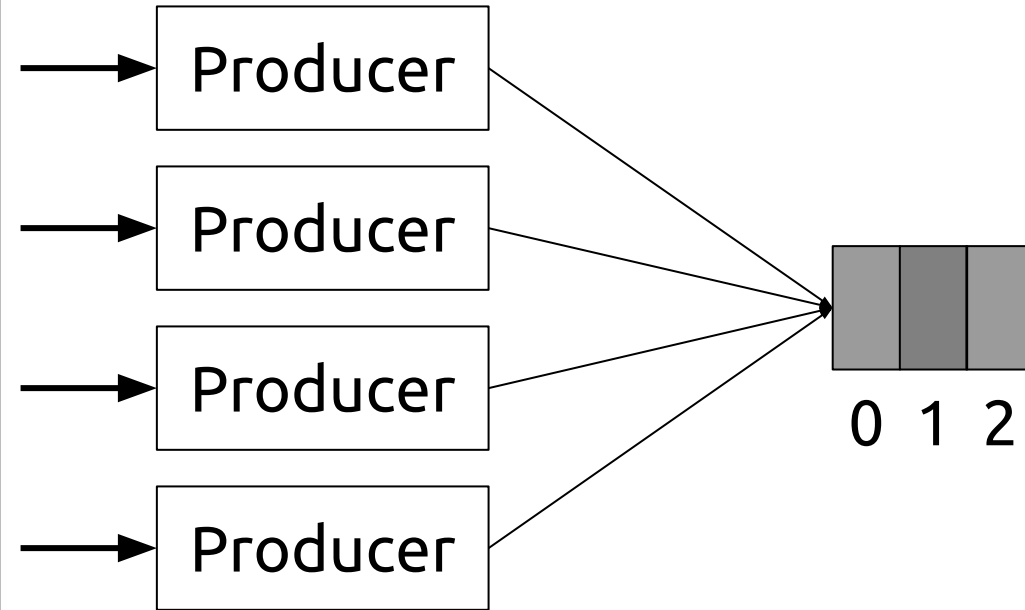
Sistema de filas



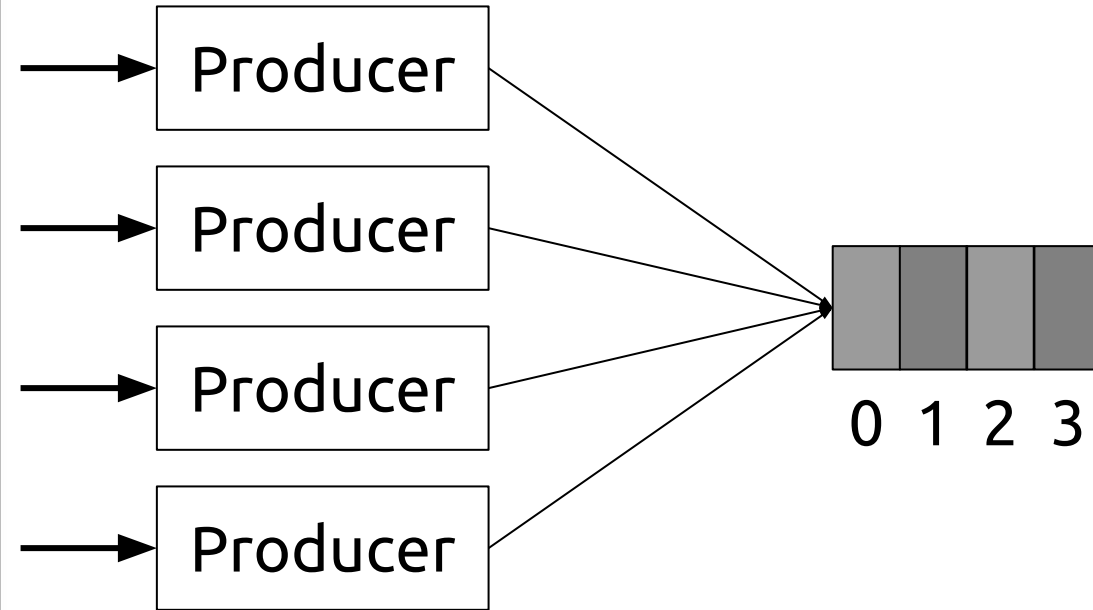
Sistema de filas



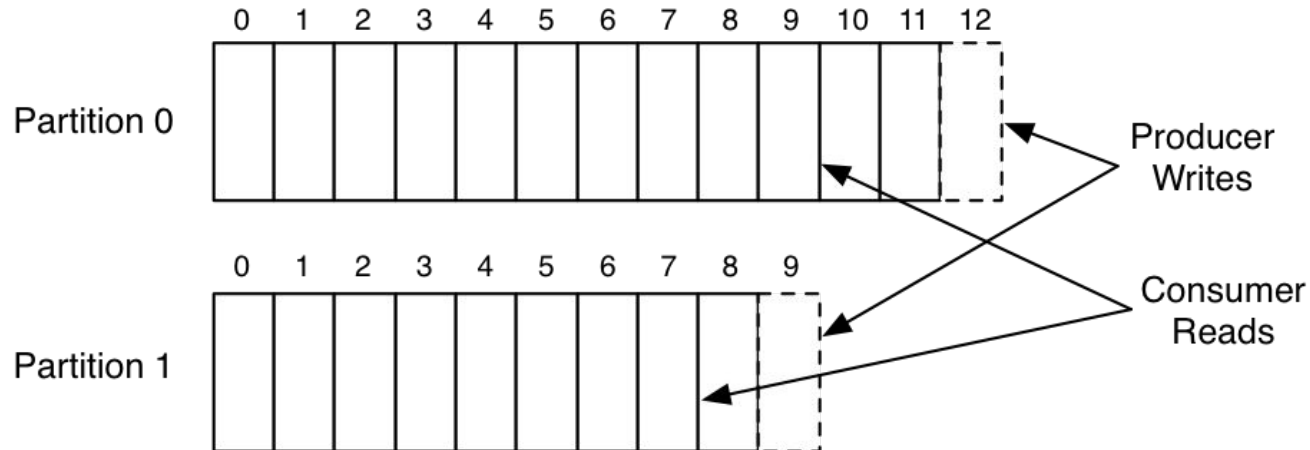
Sistema de filas



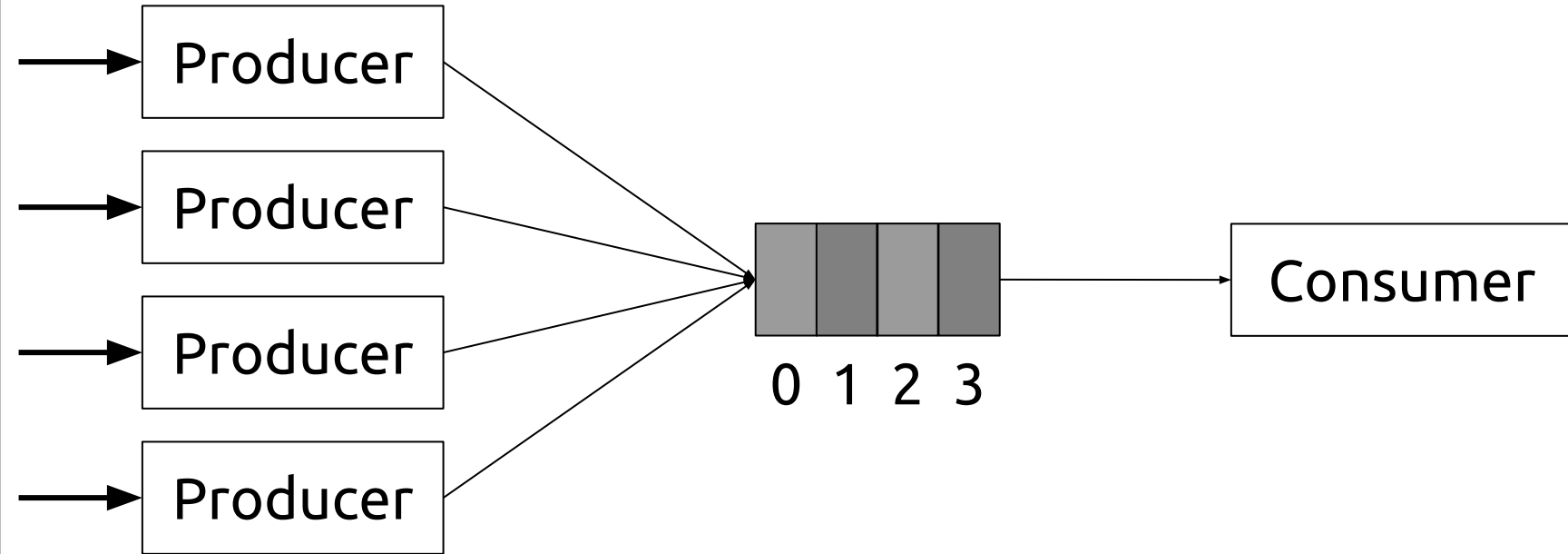
Sistema de filas



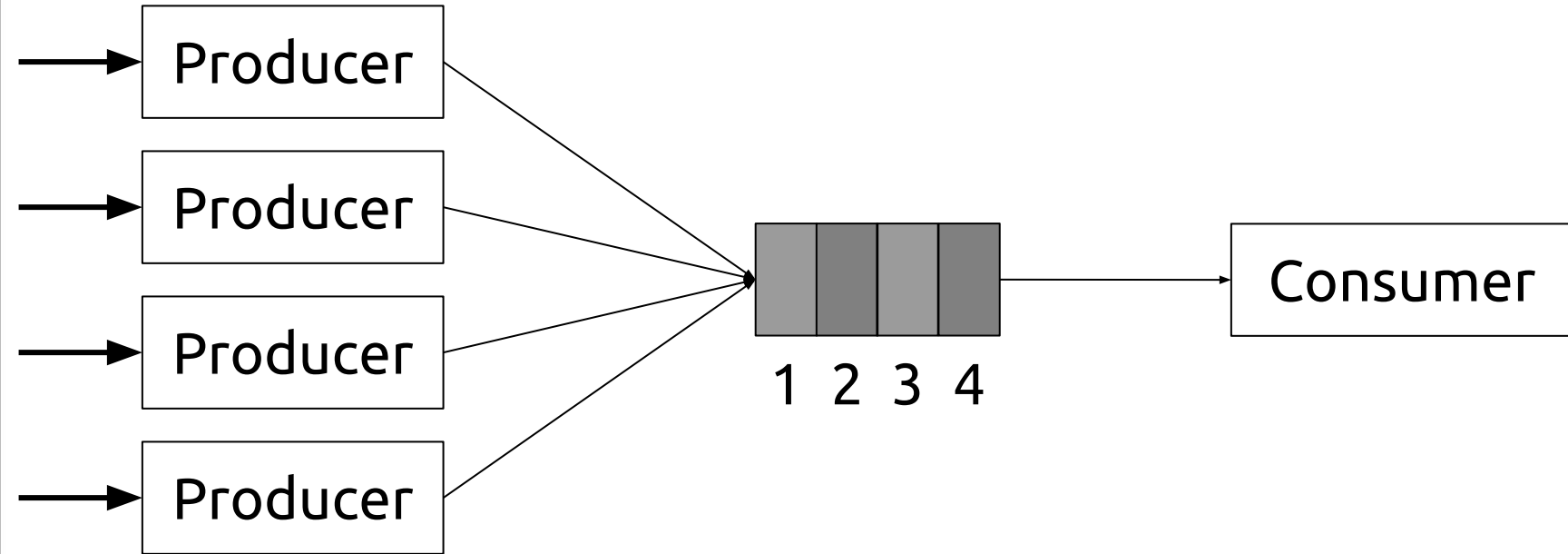
Sistema de filas



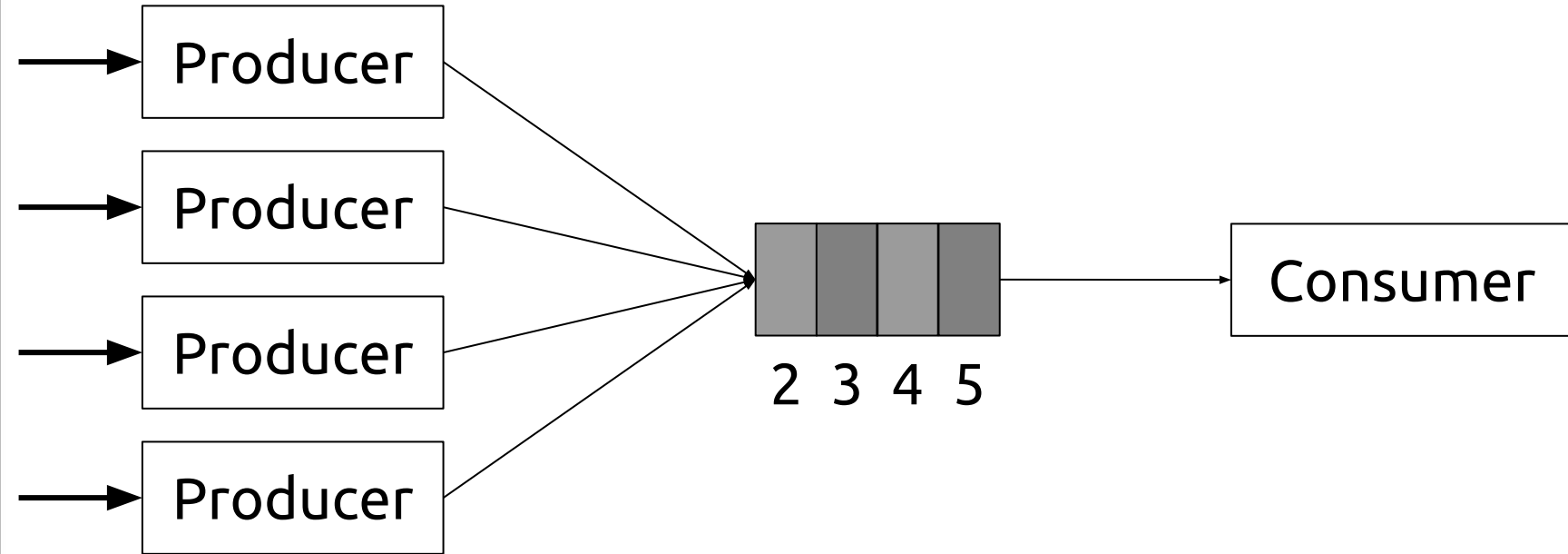
Sistema de filas



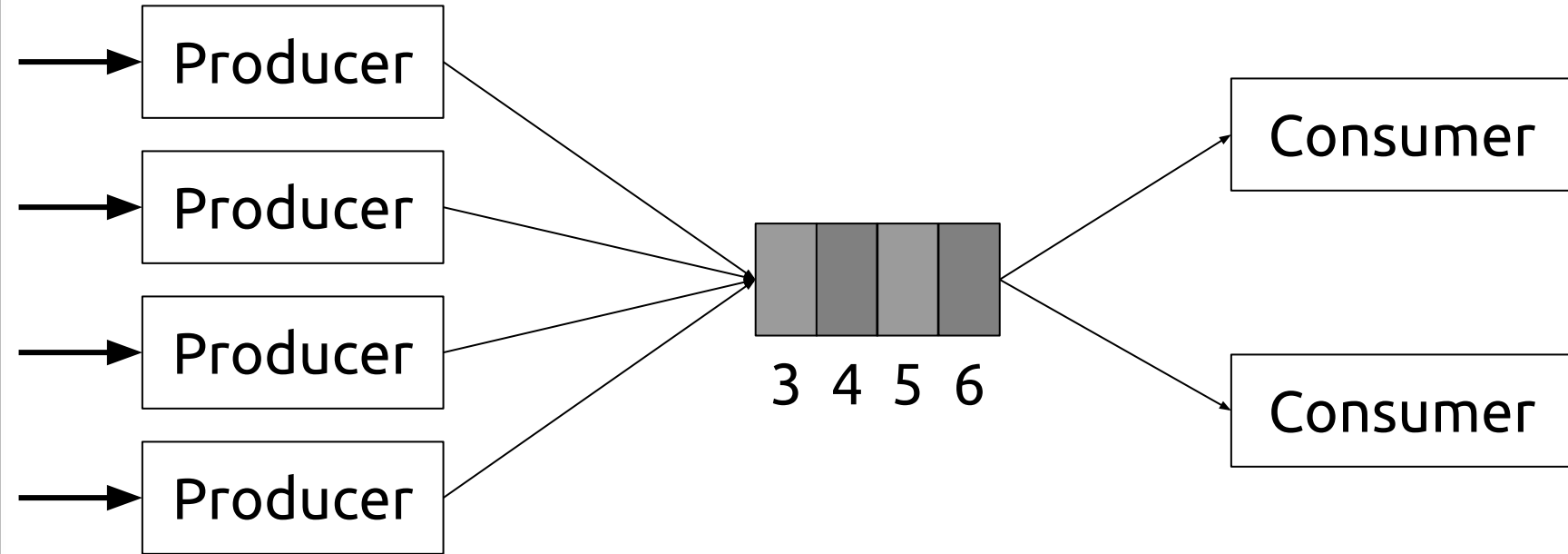
Sistema de filas



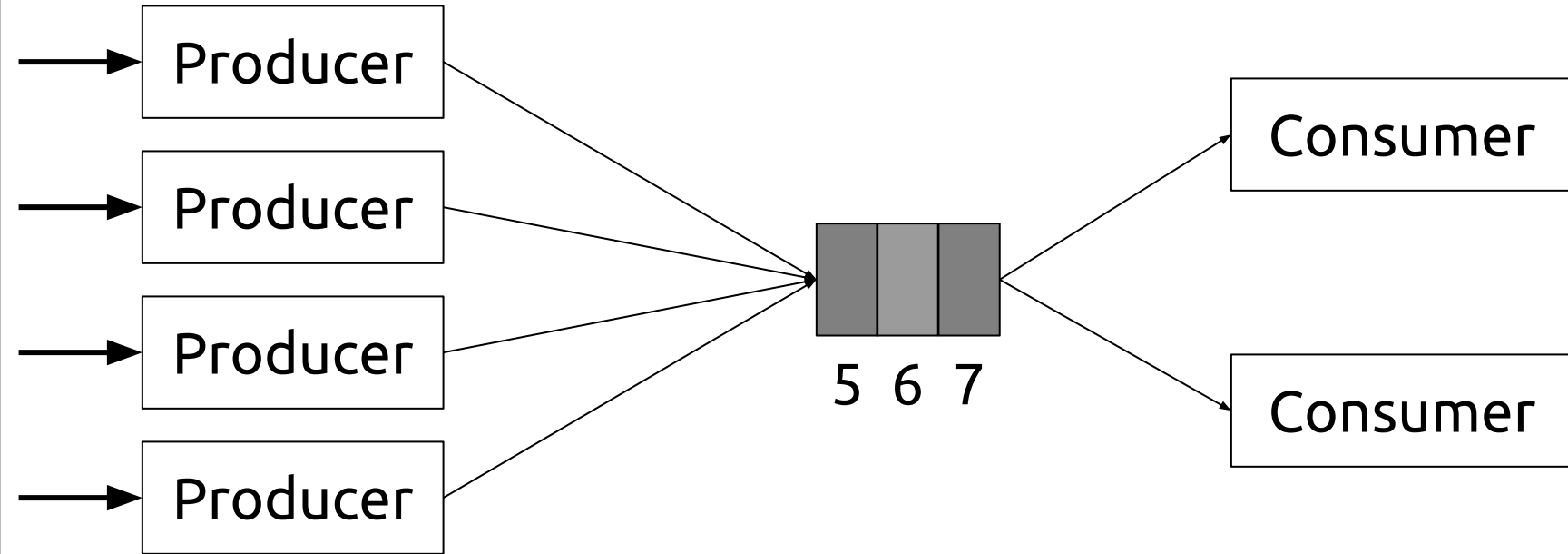
Sistema de filas



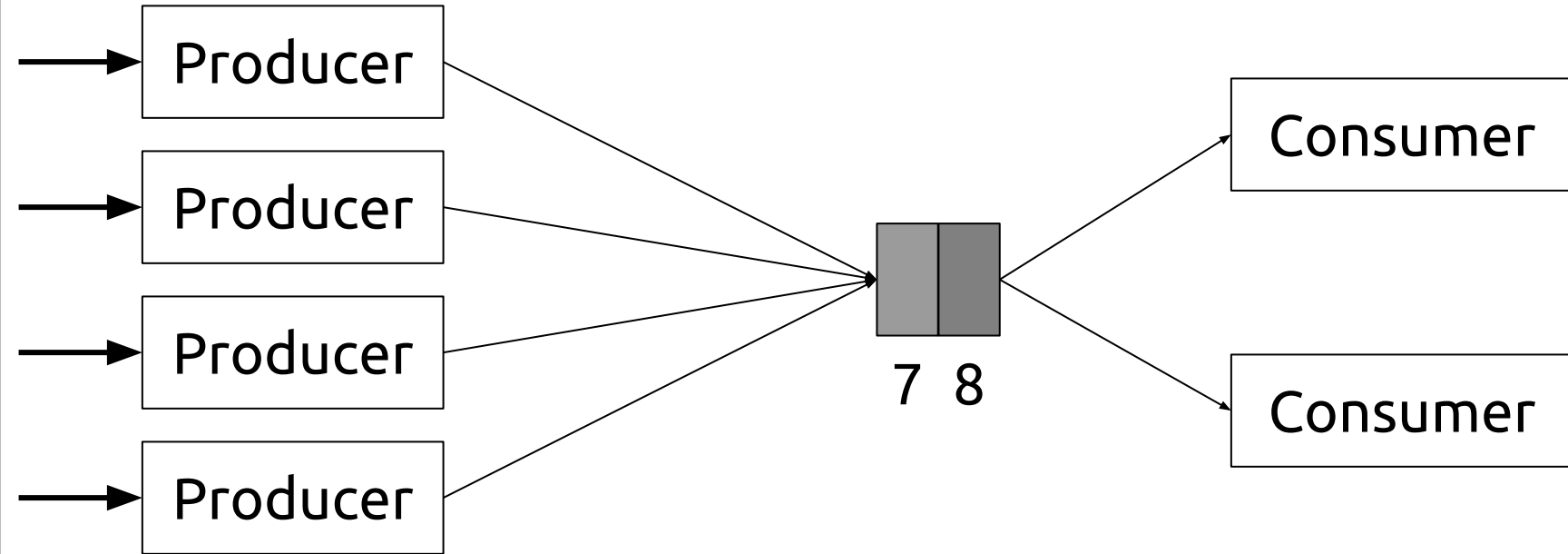
Sistema de filas



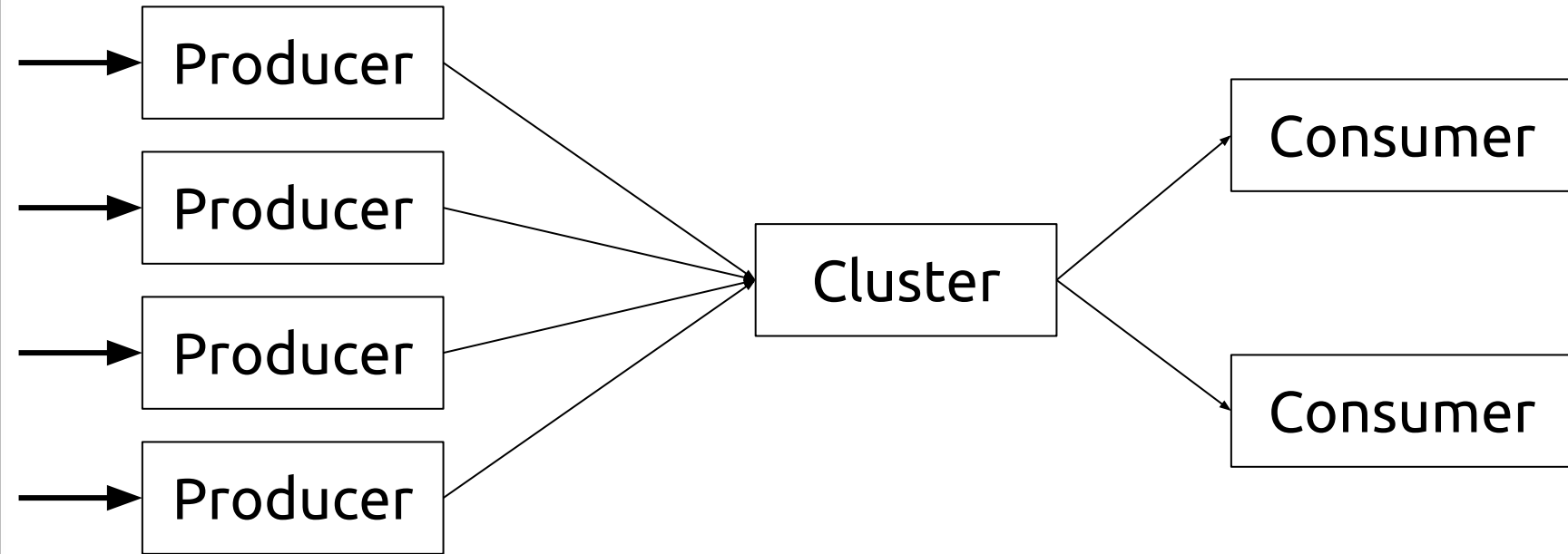
Sistema de filas



Sistema de filas



Sistema de filas









Queues

▼ All queues (6)

Pagination

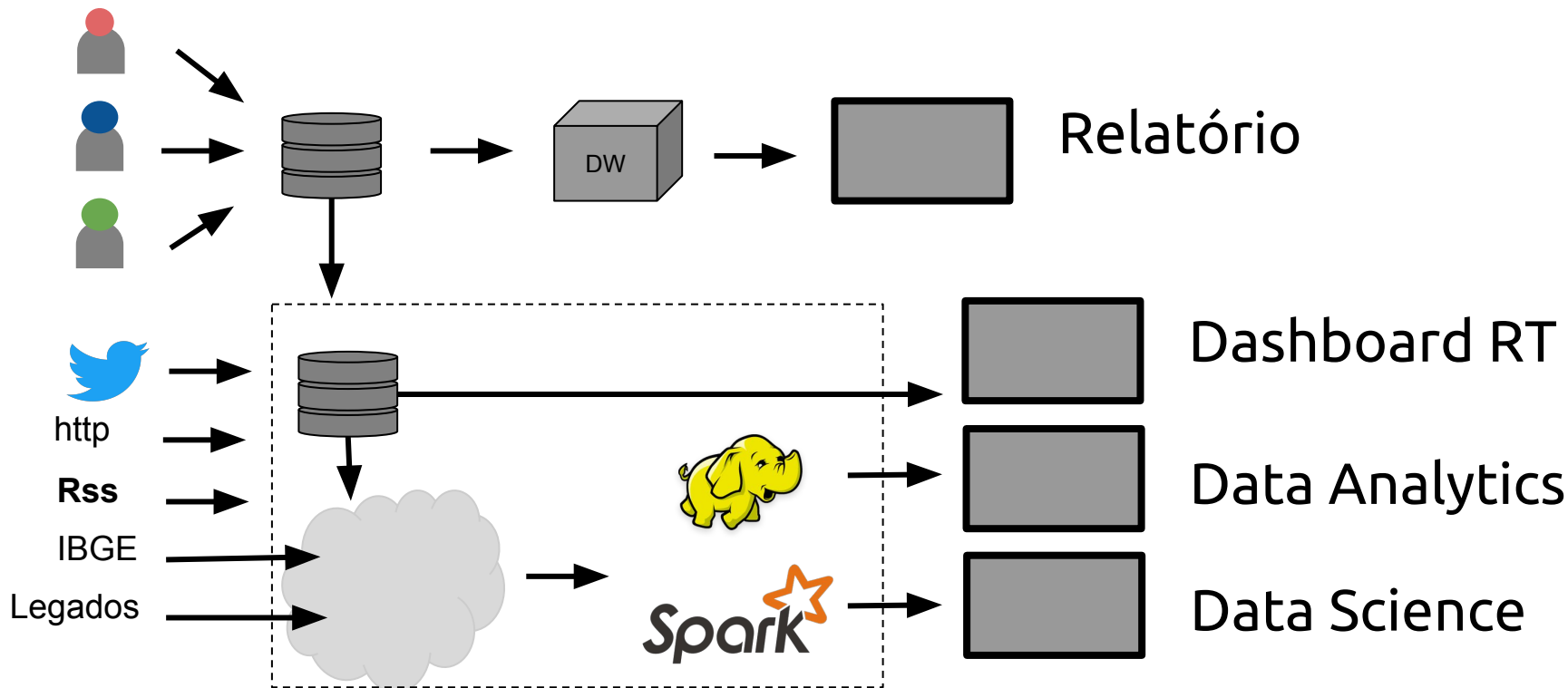
Page **1** ▼ of 1 - Filter: ☐ Regex (??)

Overview			Messages			Message rates			+/-
Name	Features	State	Ready	Unacked	Total	Incoming	deliver / get	ack	
events	D	 running	71	0	71	0.00/s	0.00/s	0.00/s	
events_update	D	 running	0	6,546	6,546	0.00/s	89/s	0.00/s	
posts	D	 idle	0	0	0				
similarity	D	 running	8,809	9,897	18,706	0.00/s	810/s	0.00/s	
sources	D	 idle	5,776	0	5,776	0.00/s	0.00/s	0.00/s	
users	D	 running	0	0	0				

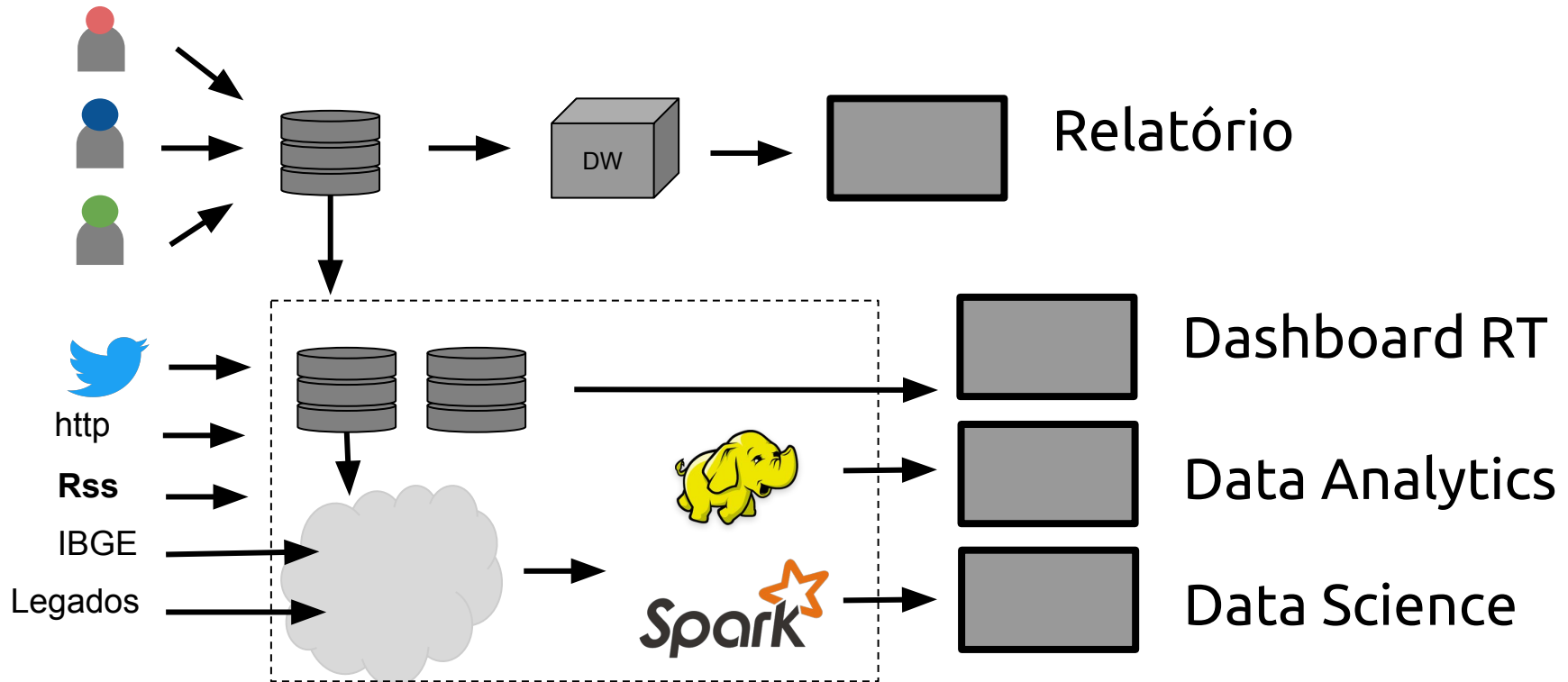
► Add a new queue

Processamento de texto

Processamento de texto



Infraestrutura Big Data



Processamento de texto

- Inserção de uma grande quantidade de dados textual;
- Diferentes tipos de indexação de texto;
- Filtro de informações
 - Remove stop-words, em diversas línguas
 - Identifica tags HTML e aproveita somente o necessário

Processamento de texto



Processamento de texto

- Stemming: Processo de identificação de derivações e inflexões de palavras

“Aula de **bando** de dados não relacionais”

Termo de pesquisa: banco

Processamento de texto

- Tokenização:
“Aula de bando de dados não relacionais”

Tokens: “Aula”, “de”, “bando”, “de”, “dados”, “não”,
“relacionais”

Processamento de texto

- Score de relevância
- Entende padrões
 - URL
 - Emails
 - #hashtags
 - @menções
 - Valores monetários R\$, \$..

Processamento de texto

- Parser:
"Hoje é dia 24/11/2008"

Formato do parser: dd/mm/aaaa

Retorna: Date(2008,11,24)

Processamento de texto

- Procura por texto usando linguagem DSL

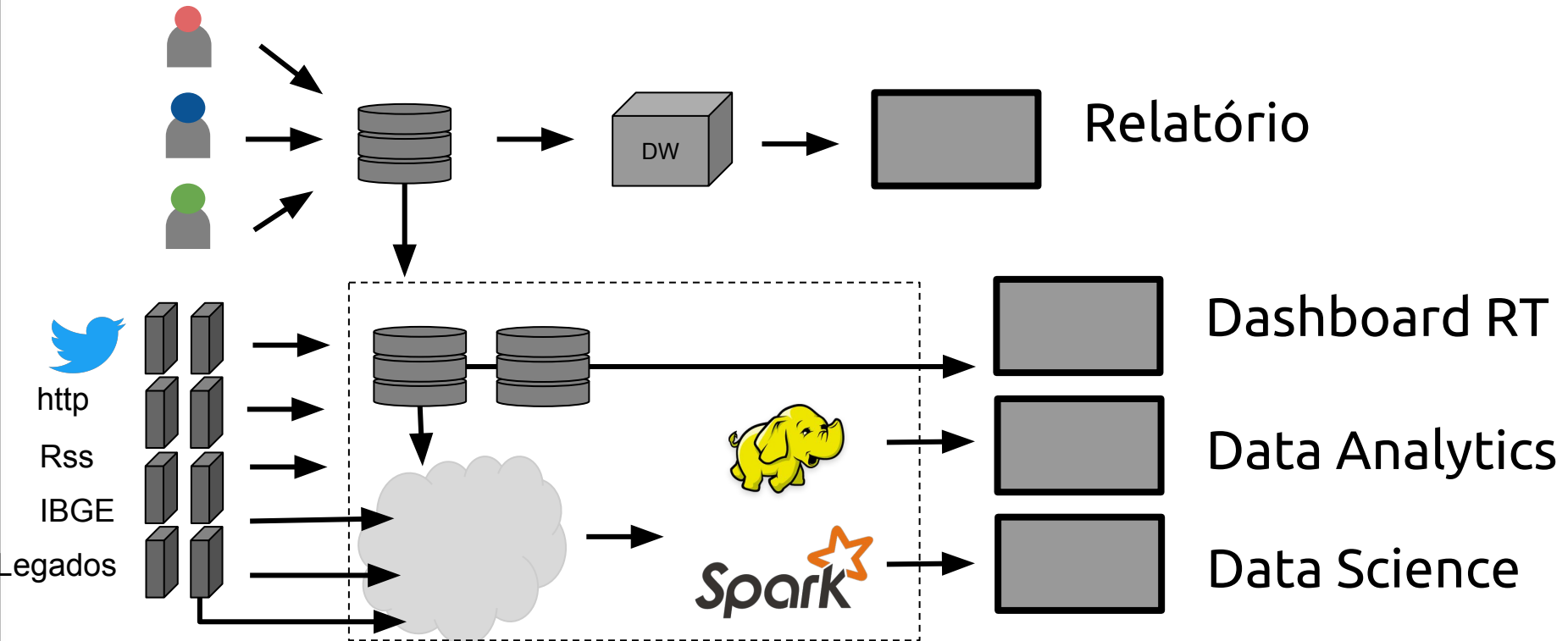
banco AND dados OR professor

Monitoramento de LOG

Monitoramento de LOG

- LOG: Registro de ações realizados por aplicações;
- Muito utilizados em sistemas complexos.

Infraestrutura Big Data



NoSQL + Hadoop

Aplicações em infraestrutura BigData

- Modelagem de risco;
- Análises preditivas e em retrospecto;
- Aprendizado de máquina;
- Identificação de padrões frequentes.

Hadoop + NoSQL

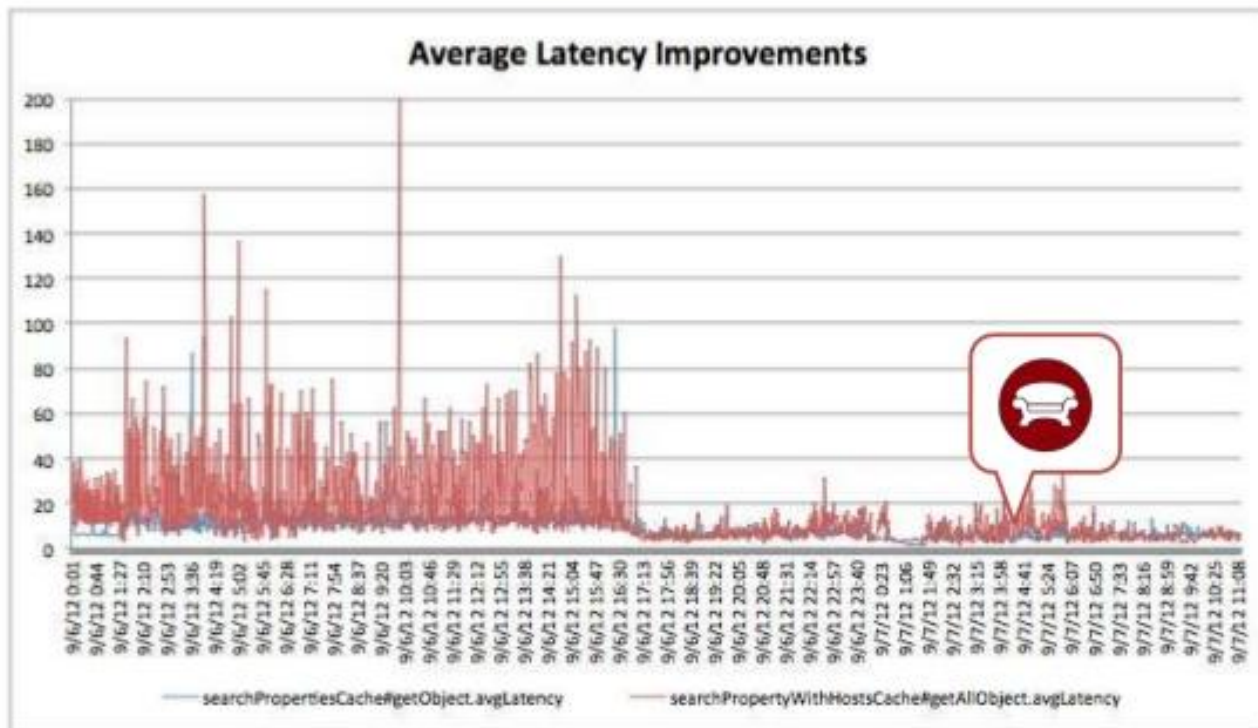
NoSQL



Gerenciamento dos preços e datas de disponibilidade

Hadoop

Estudo de segmentação dos consumidores



Pull data from Couchbase using the rest API into Graphite

Hadoop + BD - Mundo financeiro

NoSQL

Tick data, quants analysis, reference data distribution

Hadoop

Análises de risco, segurança e detecção de fraude.

Hadoop + BD - Logística

NoSQL

Armazenamento de dados de sensores conectados aos veículos

Hadoop

Programa de manutenção preventiva e análise de comportamento de motoristas

Hadoop + BD - Logística

NoSQL

Armazenamento de dados de sensores conectados aos veículos

Hadoop

Programa de manutenção preventiva e análise de comportamento de motoristas

Hadoop + BD - Planos de saúde

NoSQL

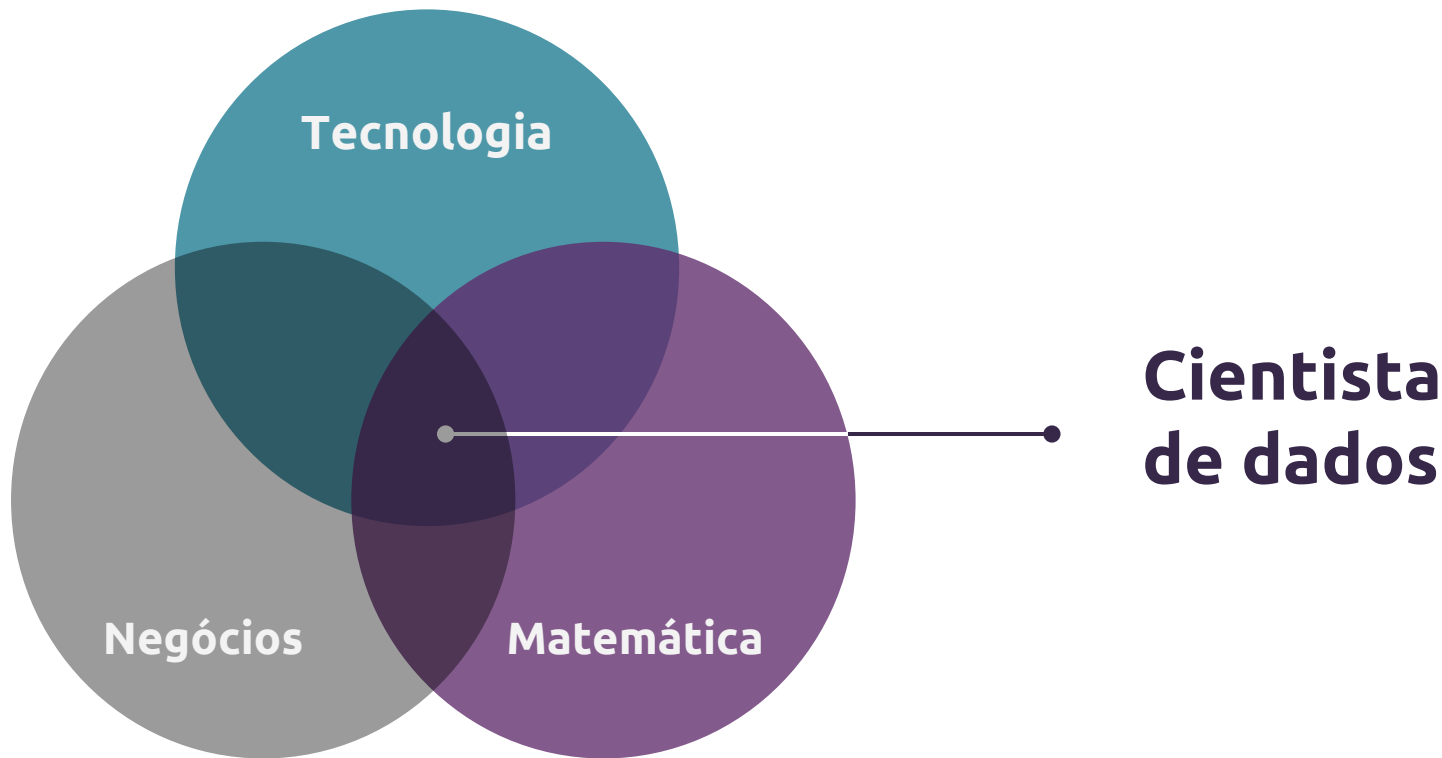
Armazenamento das várias transações de itens de saúde e comportamento diário

Hadoop

Análise predição dos filiados que tem chances de terem grande quantidade de gastos nos próximos meses.

Mensagem para os cientistas de dados

Ciência de dados



Dia a dia do cientista de dados

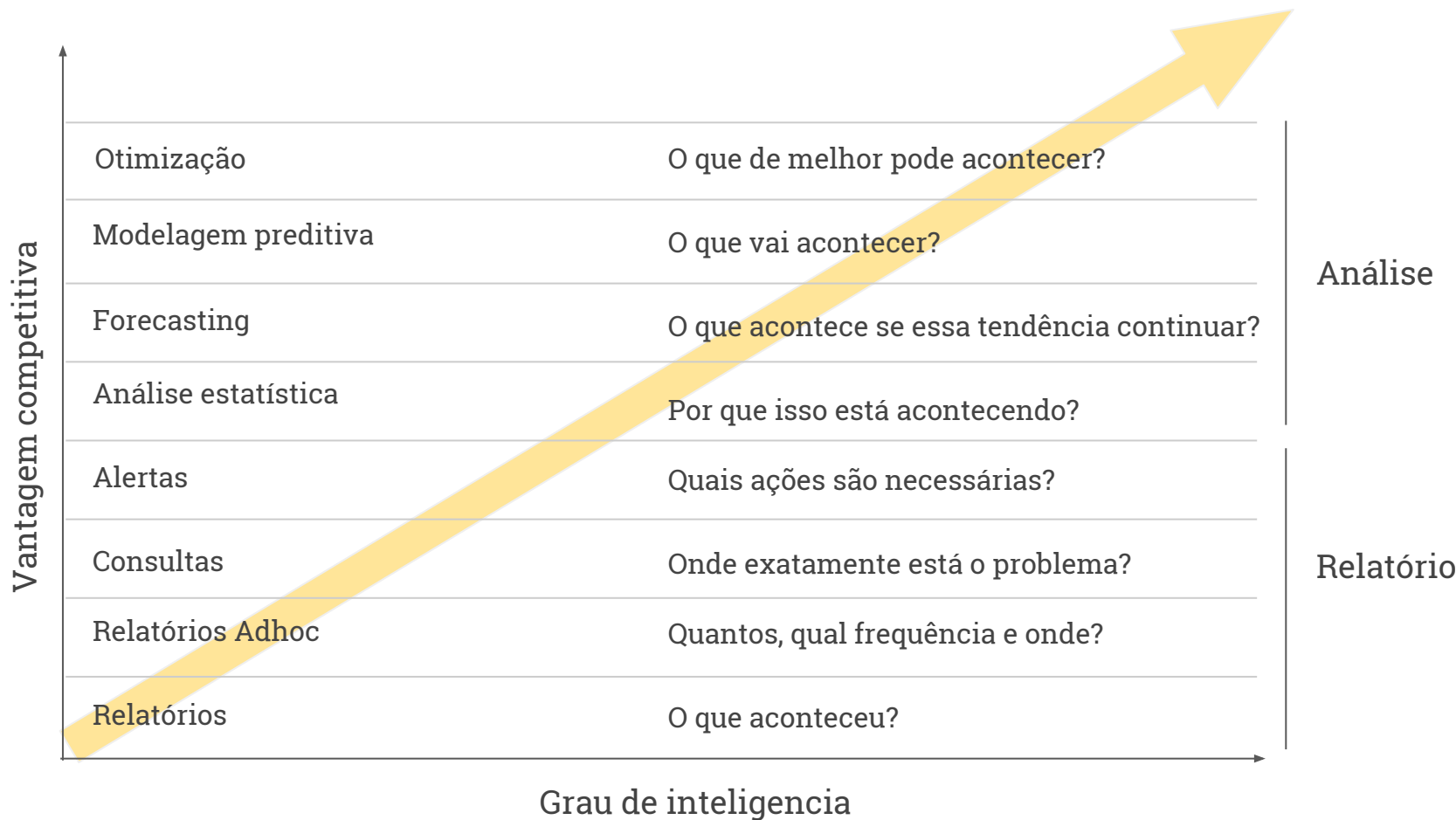
Banco de dados não relacionais:

- Operações e consulta a dados;
- Entender os índices;
- Sempre questionar se a opção que de BD é a melhor para a aplicação em questão.

Ciência de dados

Análise de dados:

- Aprendizado de máquina;
- Validações estatísticas;
- Senso crítico;



Exercícios

Para cada uma das situações dos exercícios escolha a infraestrutura que você acha mais adequada e justifique sua escolha.

Exercício 1

Você e um grupo de amigos da faculdade decidem-se juntar e criar uma empresa de na área de IoT. Todos seus amigos são excelentes programadores porém estão em dúvida como montar a infraestrutura para suportar a grande quantidade de dados gerados pelos sensores da aplicação. O que vocês devem fazer?

Exercício 2

Dentro de sua empresa certamente existem pontos que podem ser adaptados para a

Explique a infraestrutura atual e o que você mudaria para melhorar a eficiência

Pode ser alguma infraestrutura que já lidou no passado :)

Referências

<http://nosql-database.org/>

<https://dzone.com/articles/better-explaining-cap-theorem>

Bases de dados

<http://www.kaggle.com>

<http://www.bigdatabusiness.com.br/6-bases-de-dados-gratuitas-para-mineracao-estudos-e-testes/>

Orbitz:

<http://www.couchbase.com/cn/presentations/couchbase-at-orbitz>