Why Are We Near Real-Time?

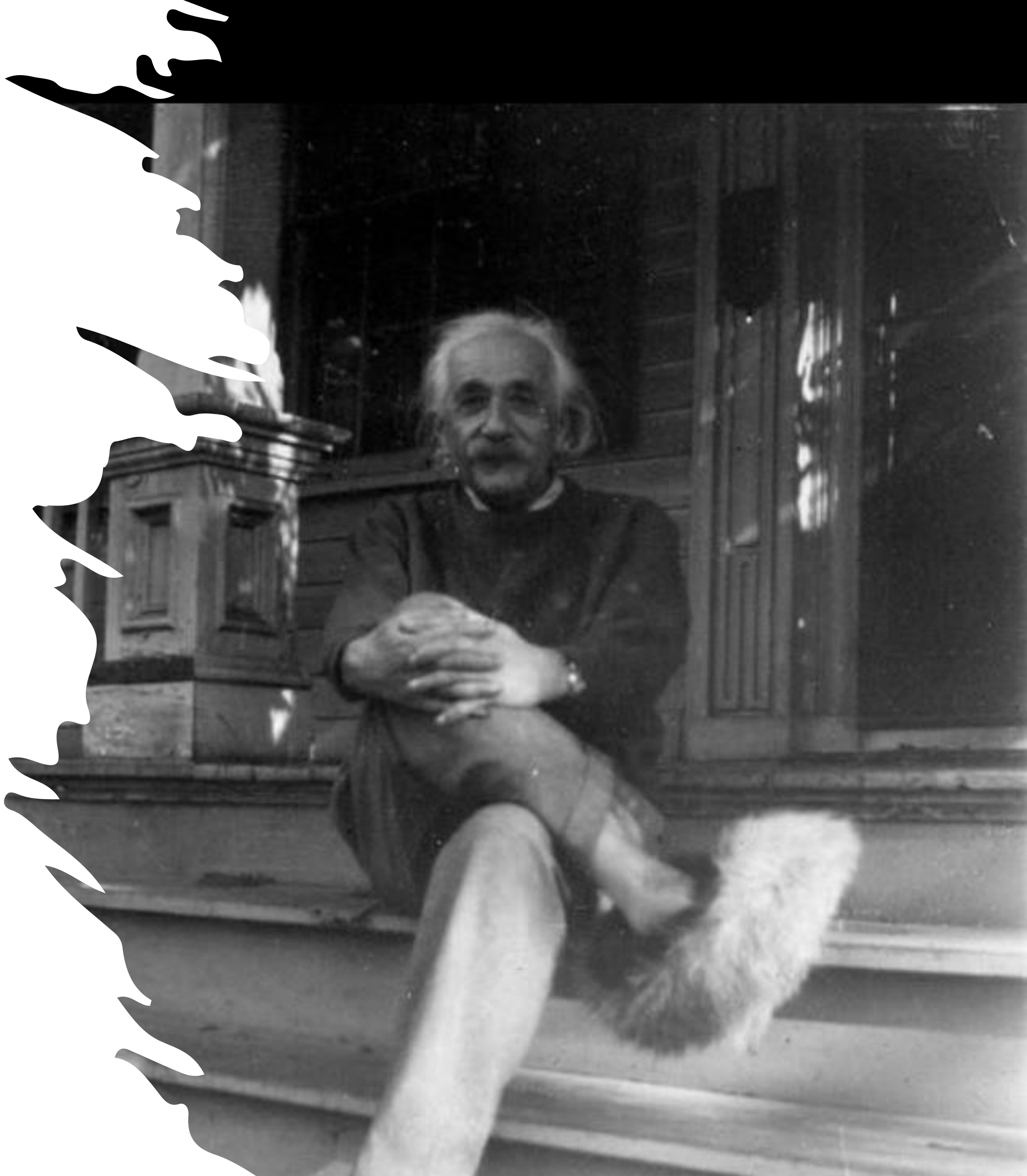
Kristo Raun

Introduction to Near Real-Time Data Analytics

August 2022

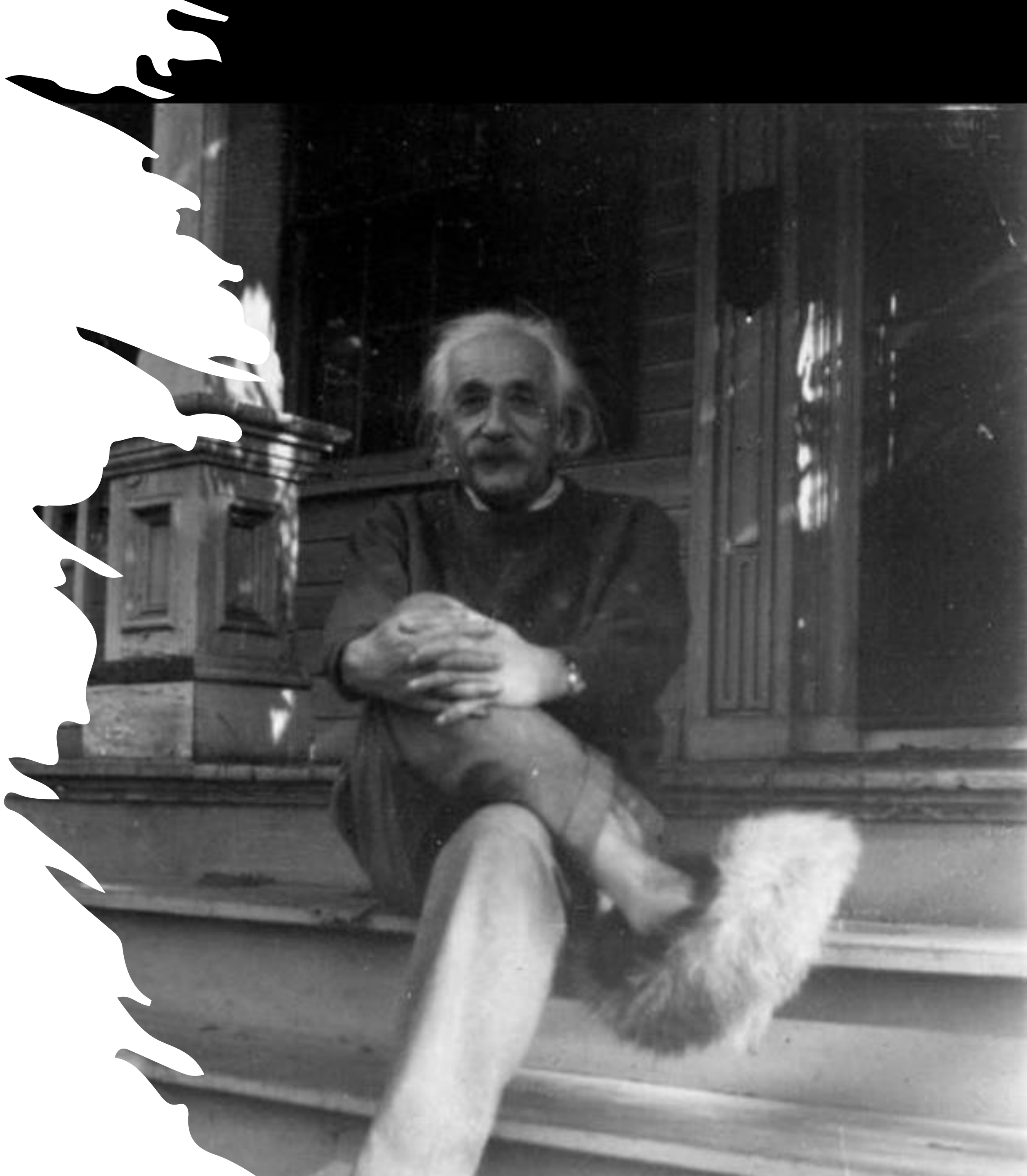
UNIVERSITY OF TARTU

Everything is relative



Everything is relative

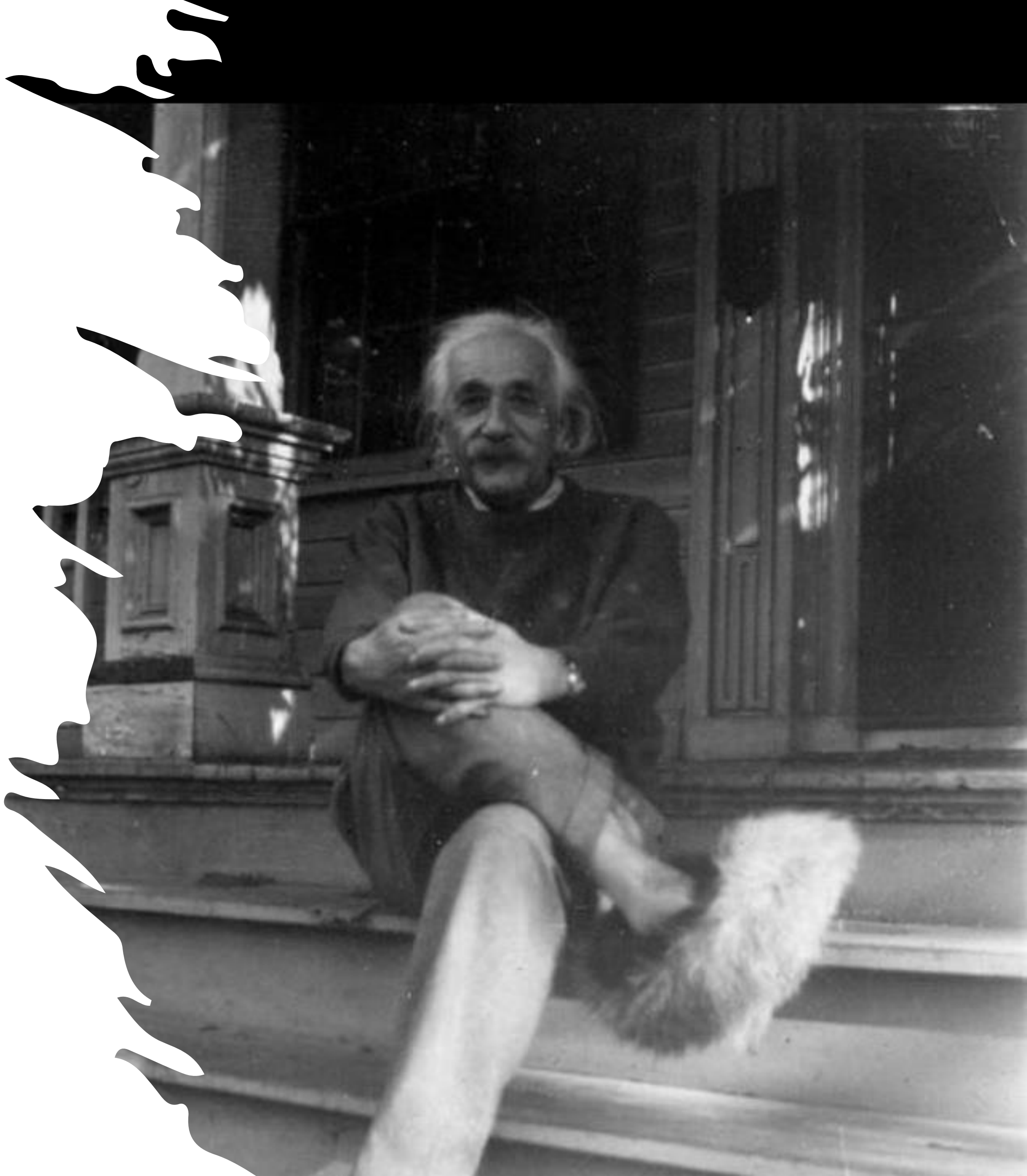
When you sit with a nice girl for two hours you think it’s only a minute.



Everything is relative

When you sit with a nice girl for two hours
you think it’s only a minute.

When you receive streaming data for a minute
you think it’s two hours.



About me

About me

2008 – BA start

About me

- 2008 – BA start
- 2011 – working w data

About me

- 2008 – BA start
- 2011 – working w data
- 2015 – MSc start, 1st IT job

About me

- 2008 – BA start
- 2011 – working w data
- 2015 – MSc start, 1st IT job
- 2018-19
  - – 1st child
  - – live in farm
  - – data engineer

About me

- 2008 – BA start
- 2011 – working w data
- 2015 – MSc start, 1st IT job
- 2018-19
  - – 1st child
  - – live in farm
  - – data engineer
- 2021 – PhD start

Agenda for the week

Agenda for the week

Kafka

Agenda for the week

Kafka

Flink

Agenda for the week

Kafka

Flink

Visuals

Agenda for the week

Kafka

Flink

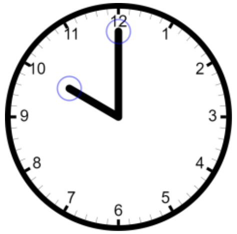
Visuals

Project

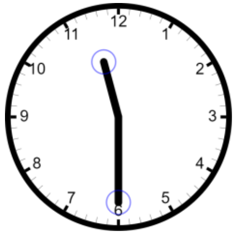
Project

Agenda for today

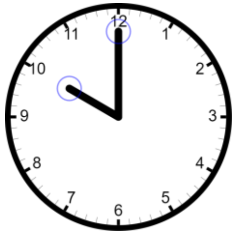
Agenda for today



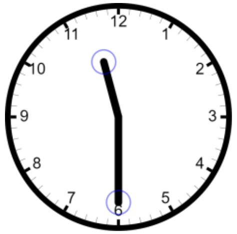
Why are we near real-time?

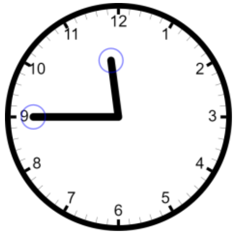


Agenda for today

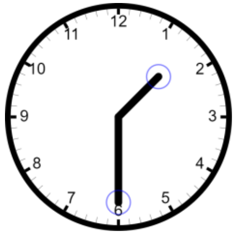


Why are we near real-time?

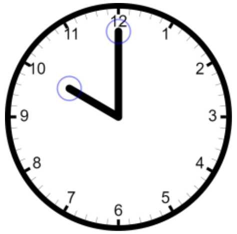




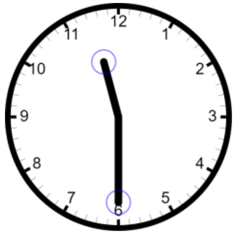
Apache Kafka setup

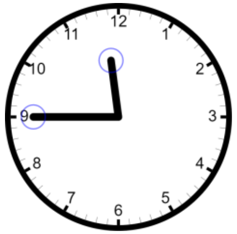


Agenda for today

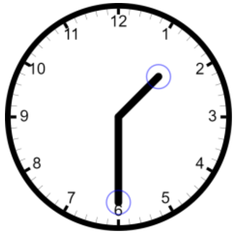


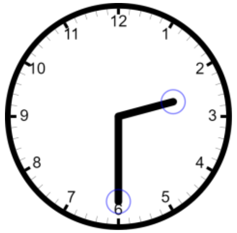
Why are we near real-time?



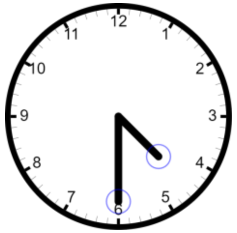


Apache Kafka setup





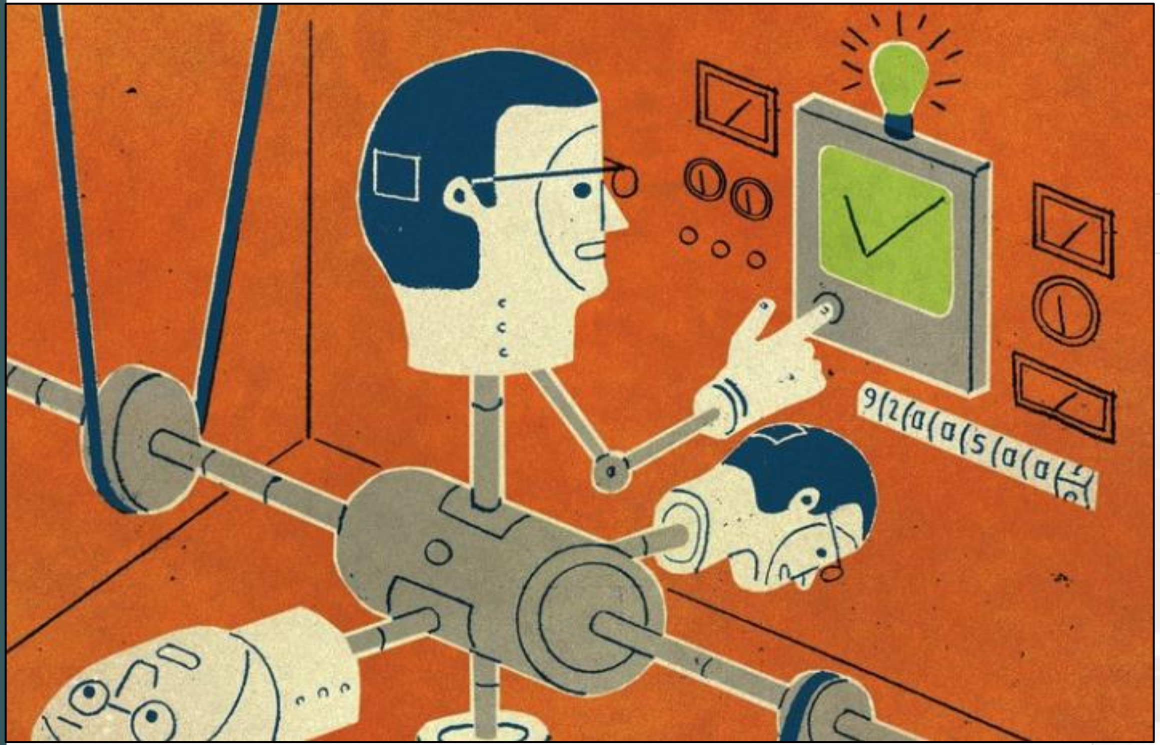
Apache Kafka practice



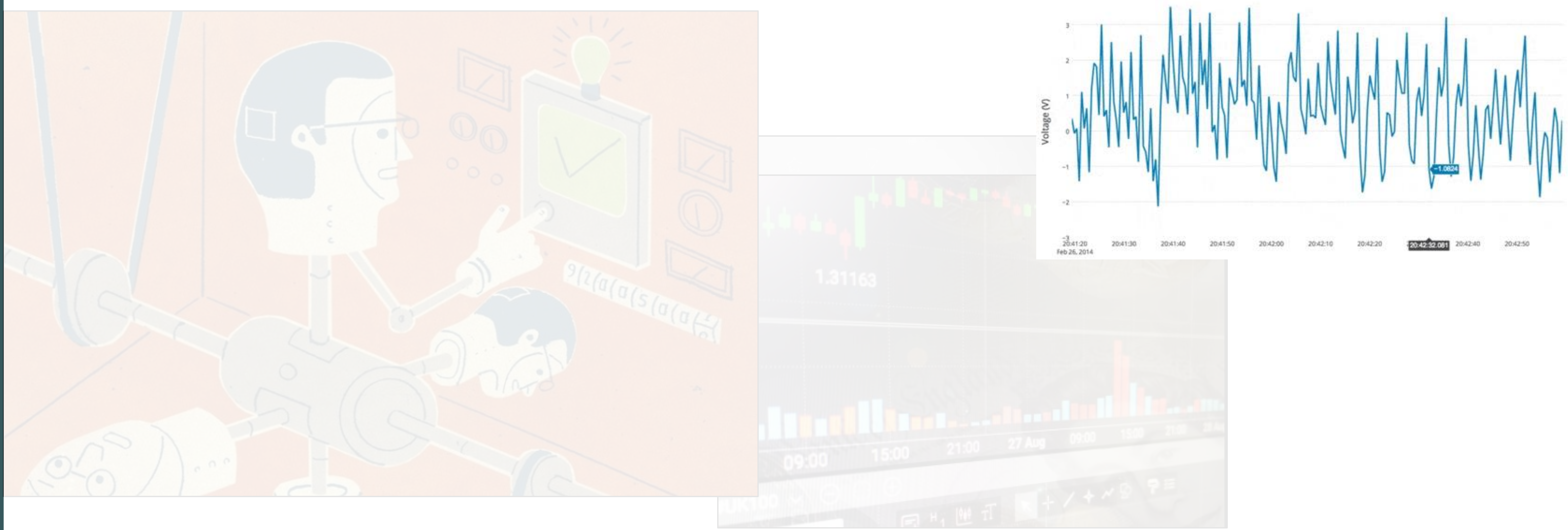
Why?

Why?

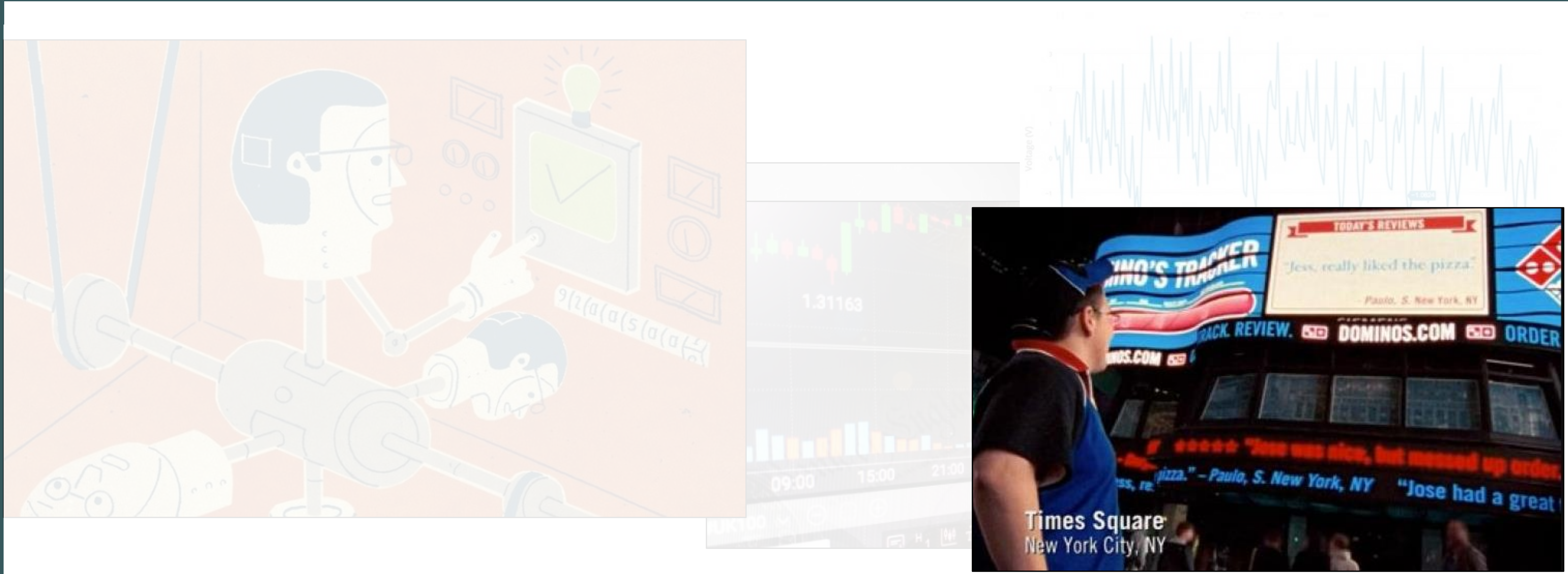
Why?



Why?


1.31163
09:00
15:00
21:00
27 Aug
UK100
Voltage (V)
3
2
1
0
-1
-2
-3
−1.0824
20:41:20
Feb 26, 2014
20:41:30
20:41:40
20:41:50
20:42:00
20:42:10
20:42:20
20:42:32.081
20:42:40
20:42:50

Why?

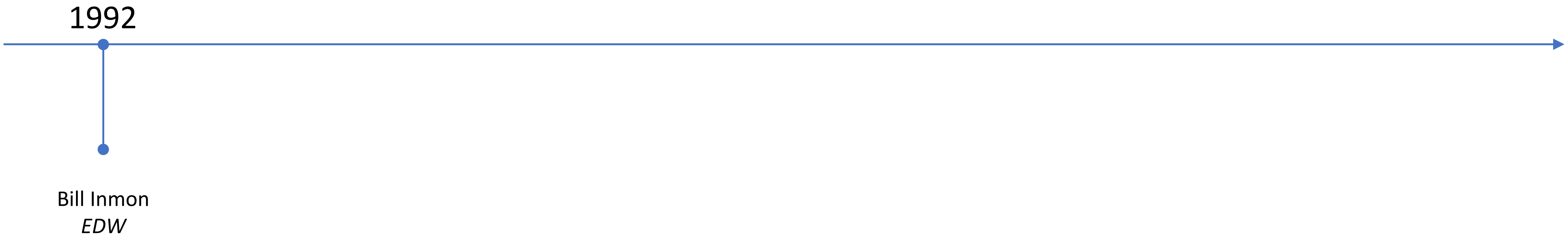

1.31163
09:00
15:00
21:00
Voltage (V)
TODAY'S REVIEWS
"Jess, really liked the pizza."
Paulo, S. New York, NY
INO'S TRACKER
RACK. REVIEW.
DOMINOS.COM
ORDER
pizza." – Paulo, S. New York, NY
"Jose had a great
Times Square
New York City, NY

Why?

DOMINOS.COM
ORDER
New York, NY
"Jose had a great
Times Square
New York City, NY

Why?

Timeline

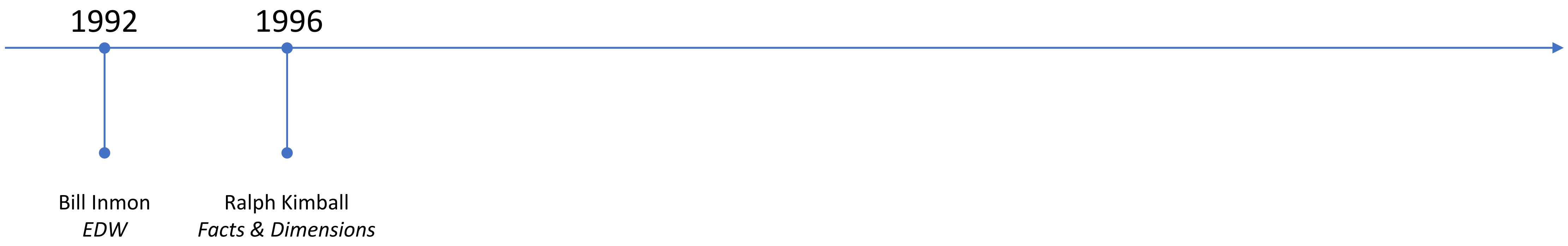


1992

Bill Inmon
EDW

Batch processing, data warehouses

Timeline



1992

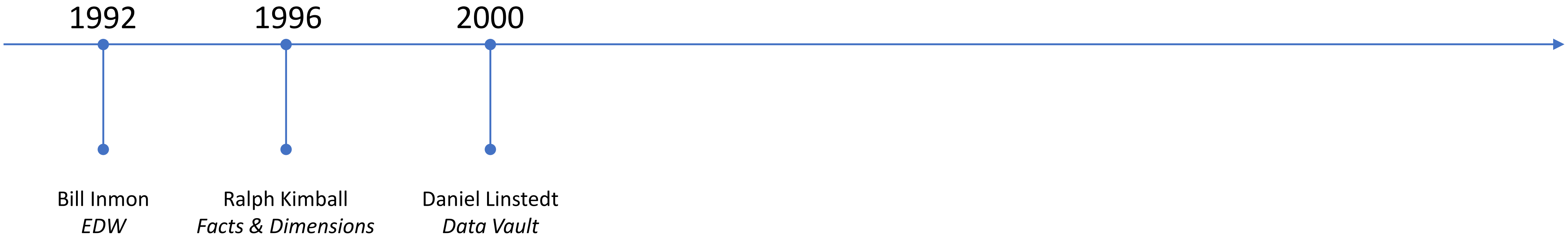
Bill Inmon
EDW

1996

Ralph Kimball
Facts & Dimensions

Batch processing, data warehouses

Timeline



1992

1996

2000

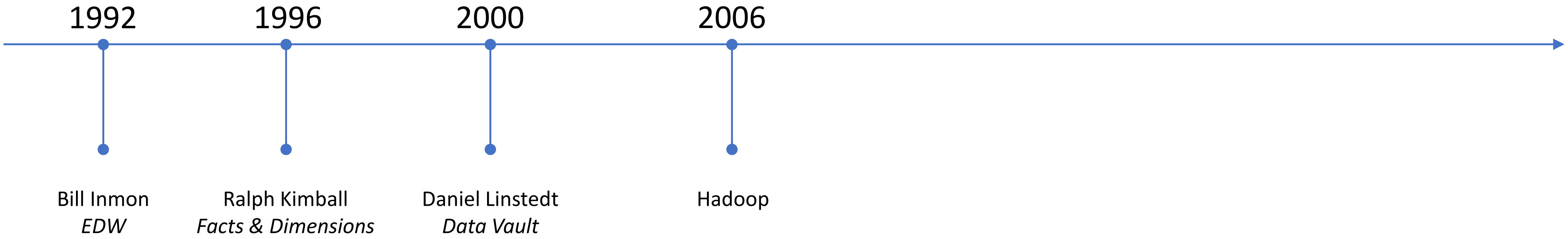
Bill Inmon
EDW

Ralph Kimball
Facts & Dimensions

Daniel Linstedt
Data Vault

Batch processing, data warehouses

Timeline



1992
Bill Inmon
EDW

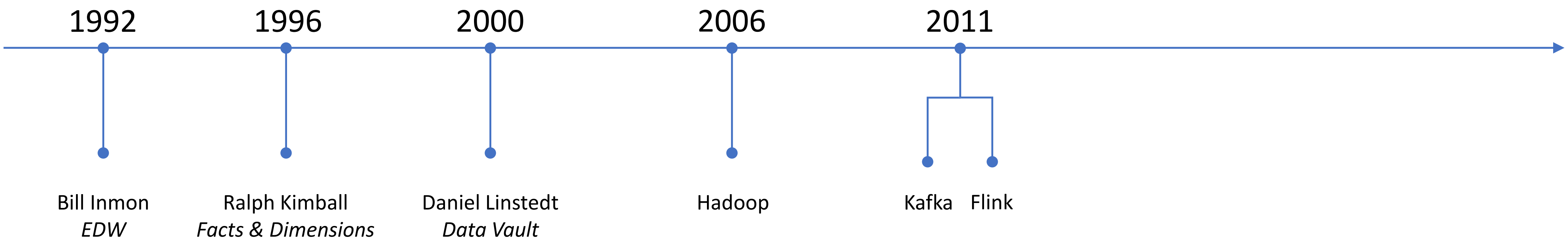
1996
Ralph Kimball
Facts & Dimensions

2000
Daniel Linstedt
Data Vault

2006
Hadoop

Batch processing, data warehouses ……… data lakes

Timeline



1992 Bill Inmon *EDW*

1996 Ralph Kimball *Facts & Dimensions*

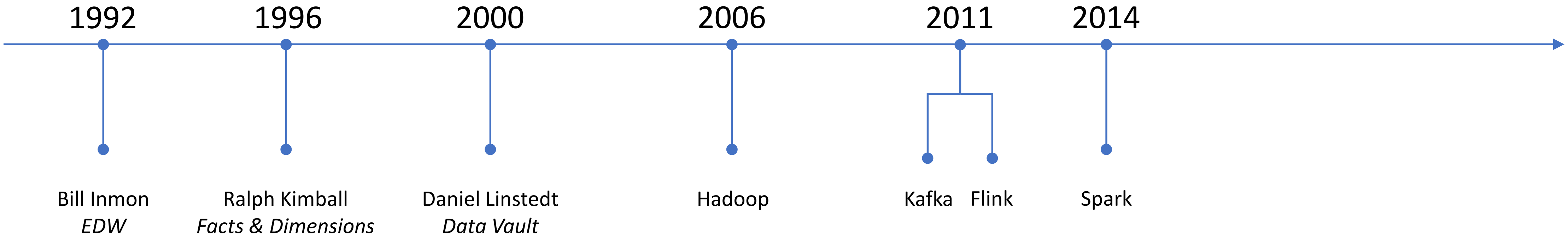
2000 Daniel Linstedt *Data Vault*

2006 Hadoop

2011 Kafka Flink

Batch processing, data warehouses ……… data lakes ……… stream processing

Timeline



1992

Bill Inmon
EDW

1996

Ralph Kimball
Facts & Dimensions

2000

Daniel Linstedt
Data Vault

2006

Hadoop

2011

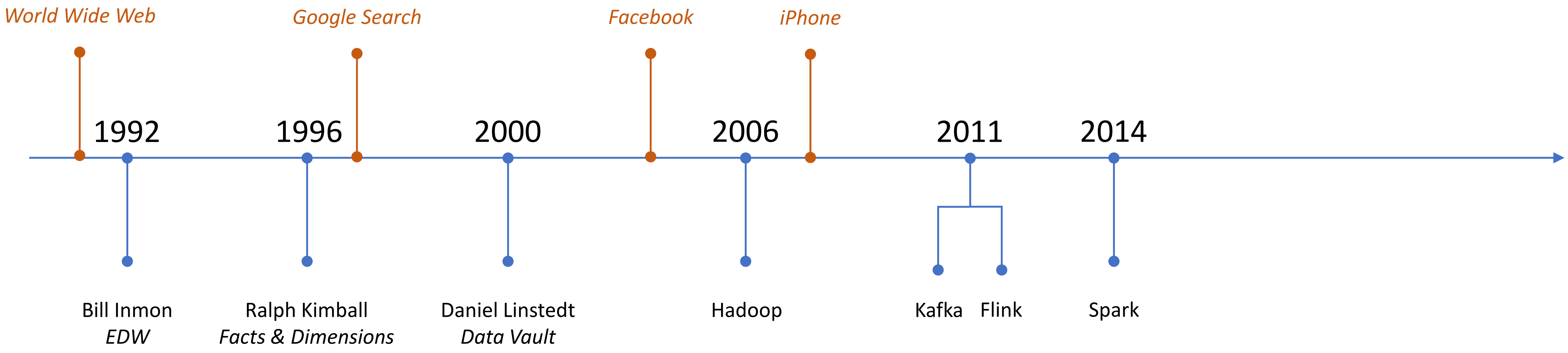
Kafka Flink

2014

Spark

Batch processing, data warehouses ……… data lakes ……… stream processing

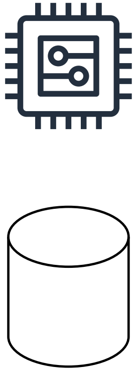
Timeline



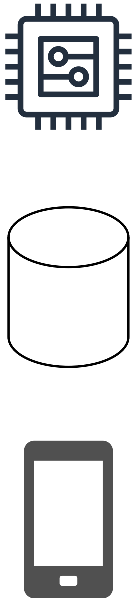
Batch processing, data warehouses ……… data lakes ……… stream processing

Role

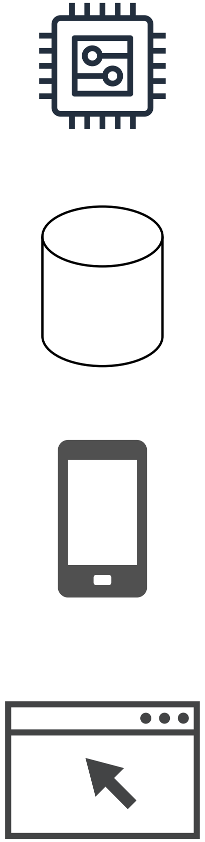
Role



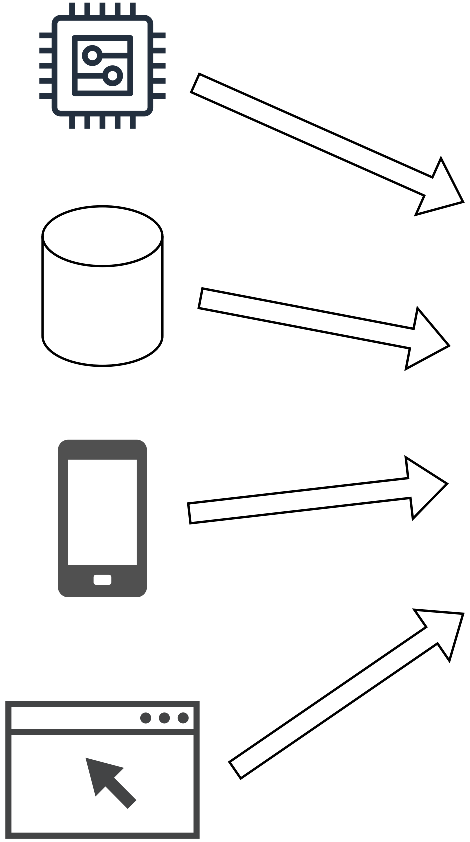
Role



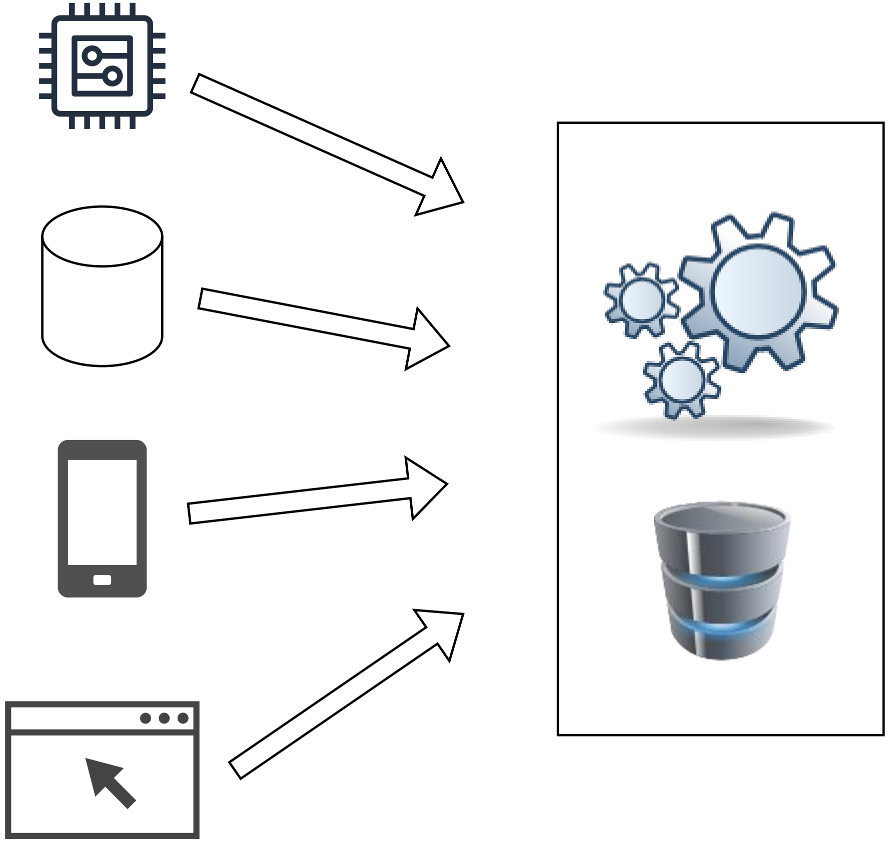
Role



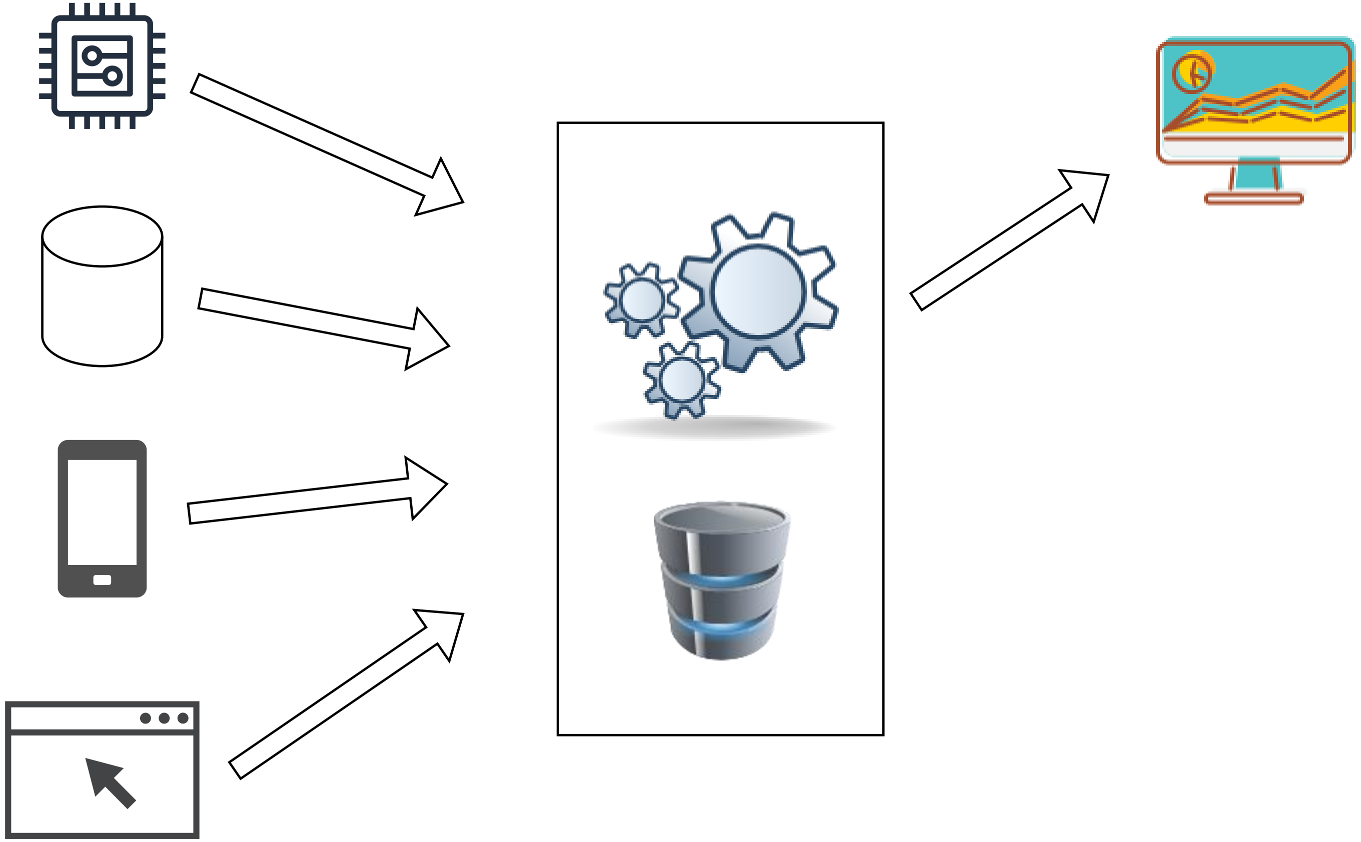
Role



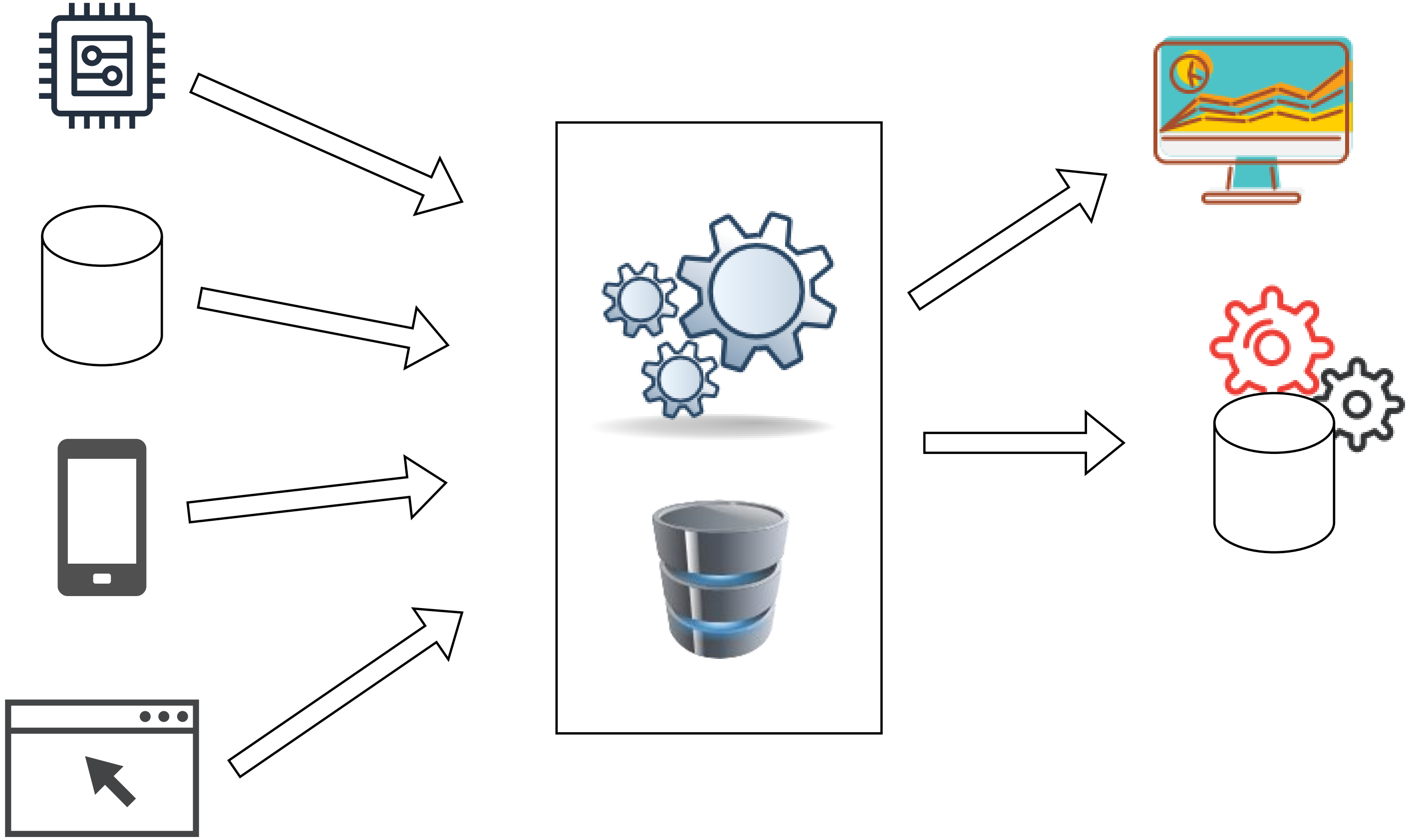
Role



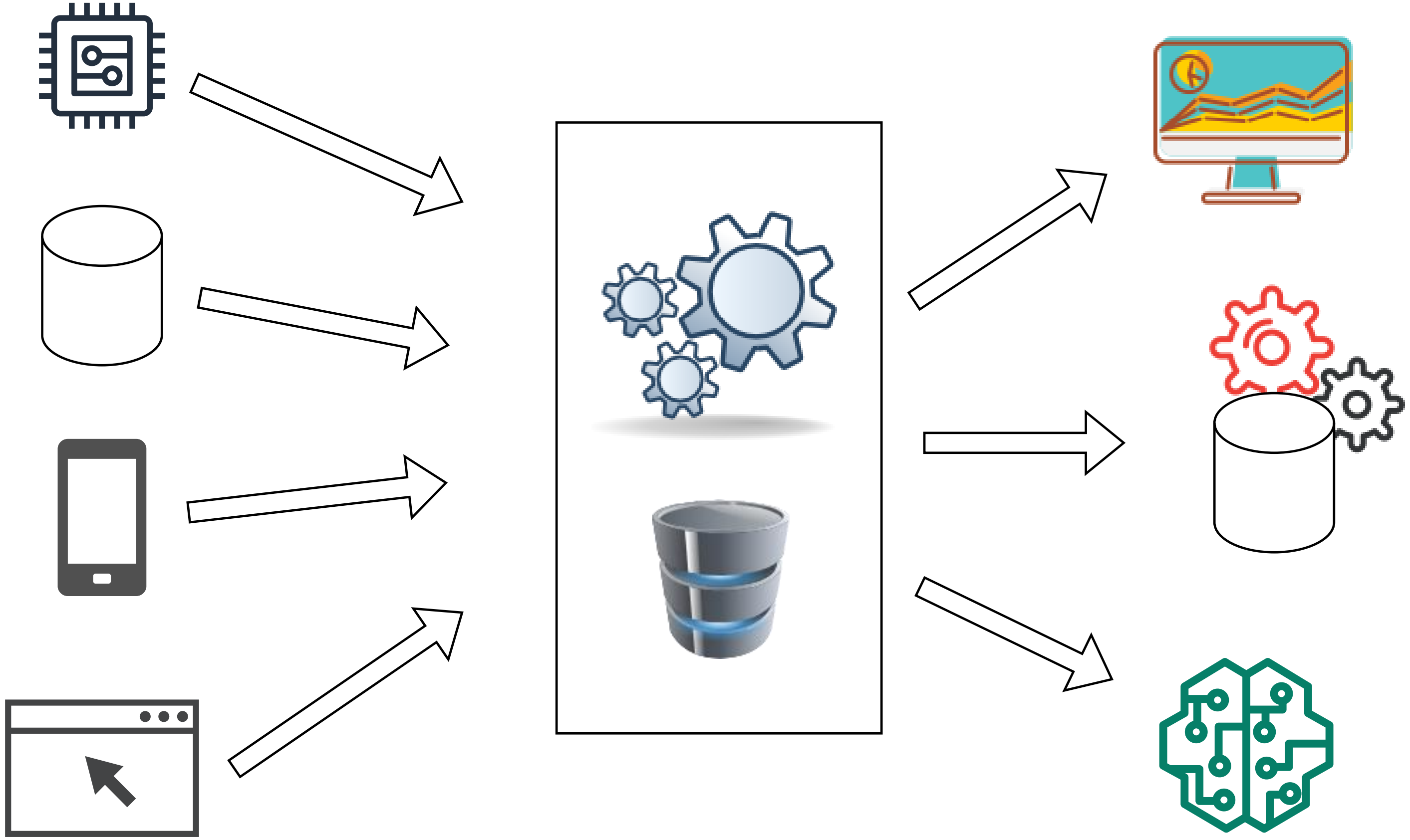
Role



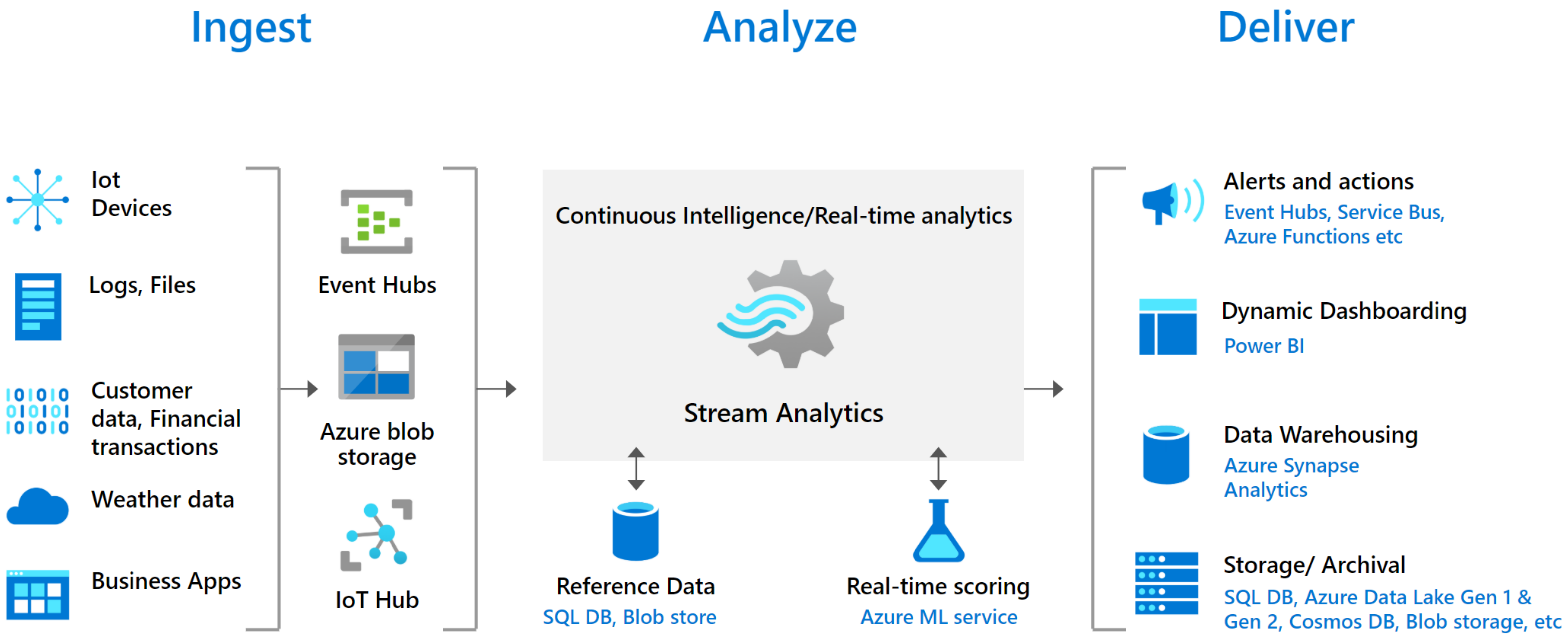
Role



Role



Data architecture



Ingest

Analyze

Deliver

Iot Devices

Logs, Files

Customer data, Financial transactions

Weather data

Business Apps

Event Hubs

Azure blob storage

IoT Hub

Continuous Intelligence/Real-time analytics

Stream Analytics

Reference Data
SQL DB, Blob store

Real-time scoring
Azure ML service

Alerts and actions
Event Hubs, Service Bus, Azure Functions etc

Dynamic Dashboarding
Power BI

Data Warehousing
Azure Synapse Analytics

Storage/ Archival
SQL DB, Azure Data Lake Gen 1 & Gen 2, Cosmos DB, Blob storage, etc

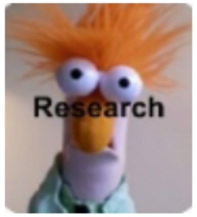
Source: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

Accepted Latency

Time Dimension v/s SLA

SLA is order of Hours / Day

Batch

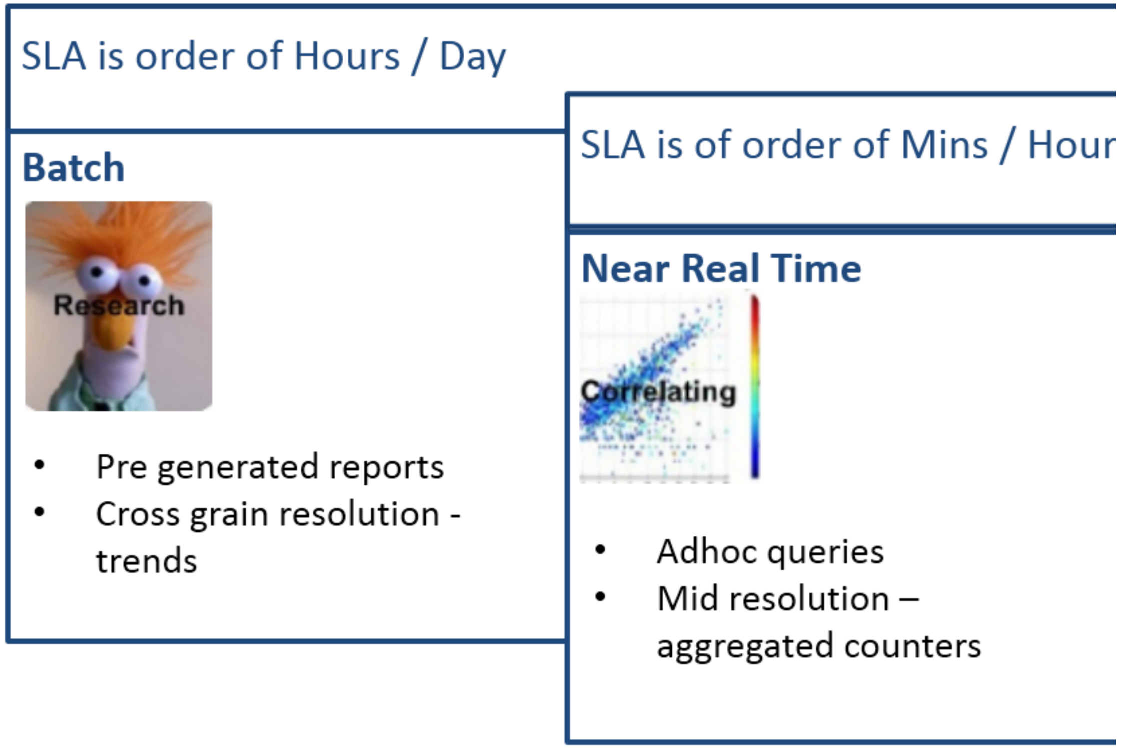


- Pre generated reports
- Cross grain resolution - trends

Source: Practical Real-time Data Processing and Analytics (2017)

Accepted Latency

Time Dimension v/s SLA



SLA is order of Hours / Day

Batch

- Pre generated reports
- Cross grain resolution - trends

SLA is of order of Mins / Hour

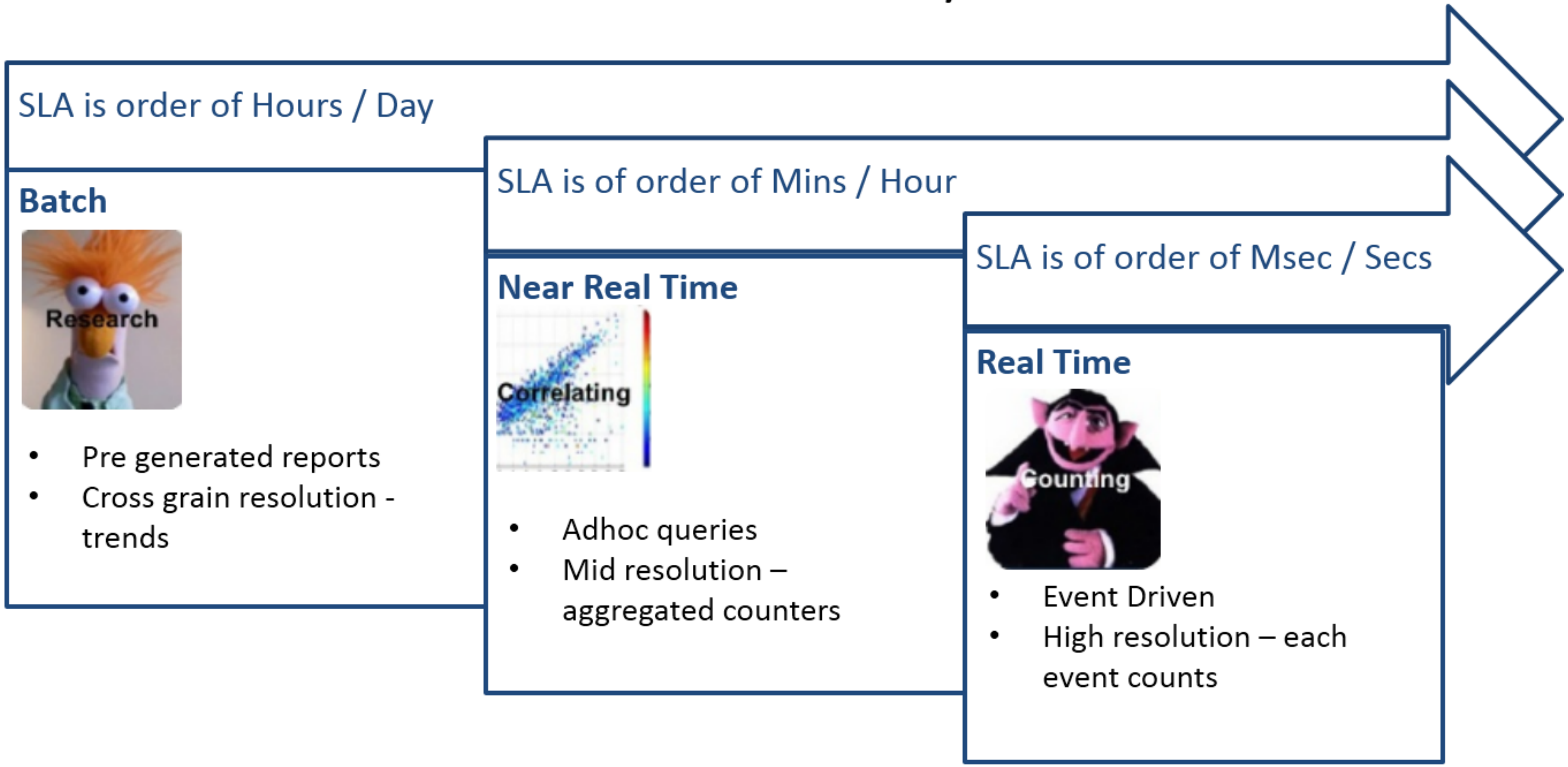
Near Real Time

- Adhoc queries
- Mid resolution – aggregated counters

Source: Practical Real-time Data Processing and Analytics (2017)

Accepted Latency

Time Dimension v/s SLA



SLA is order of Hours / Day

Batch

- Pre generated reports
- Cross grain resolution - trends

SLA is of order of Mins / Hour

Near Real Time

- Adhoc queries
- Mid resolution – aggregated counters

SLA is of order of Msec / Secs

Real Time

- Event Driven
- High resolution – each event counts

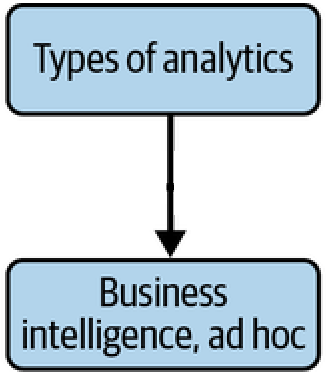
Source: Practical Real-time Data Processing and Analytics (2017)

Analytics

Types of analytics

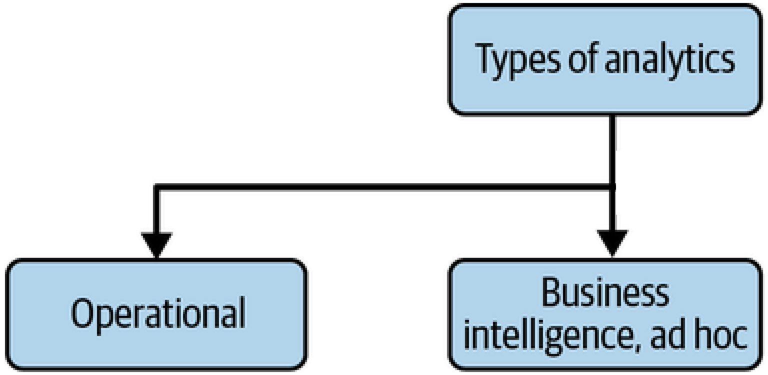
Source: Fundamentals of Data Engineering (2022)

Analytics



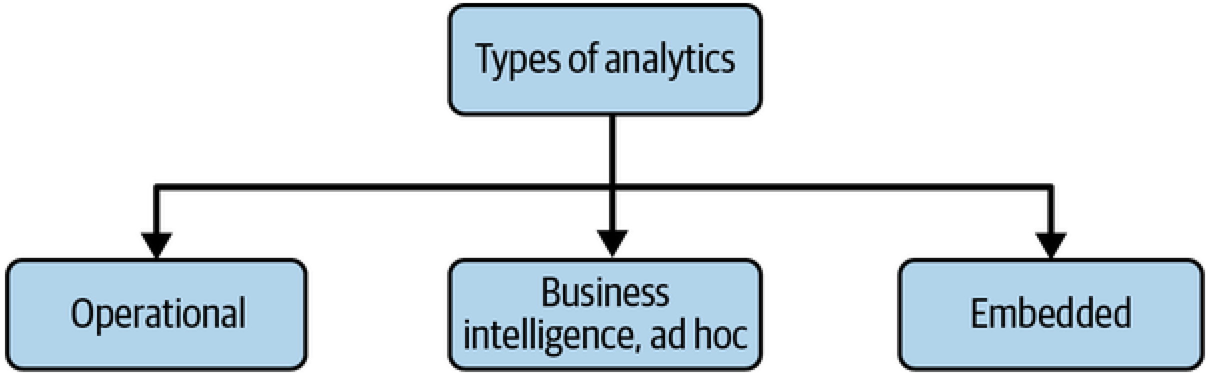
Source: Fundamentals of Data Engineering (2022)

Analytics



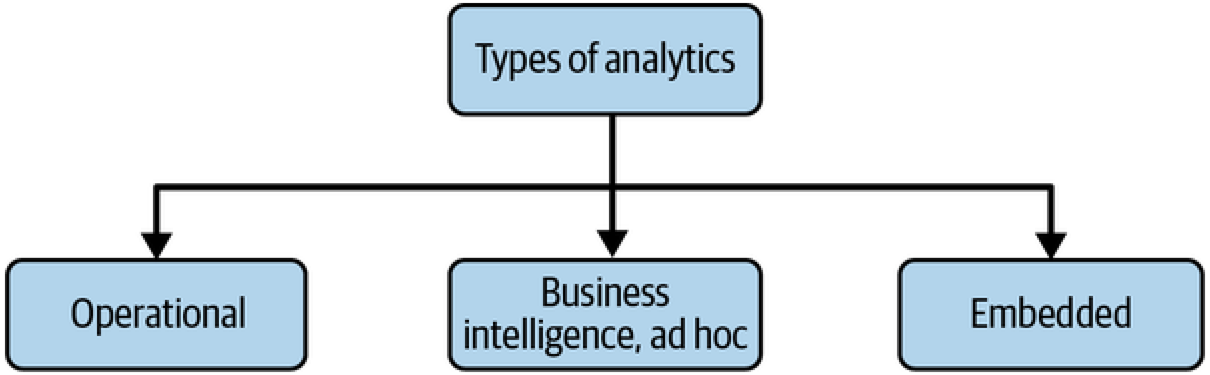
Source: Fundamentals of Data Engineering (2022)

Analytics



Source: Fundamentals of Data Engineering (2022)

Analytics



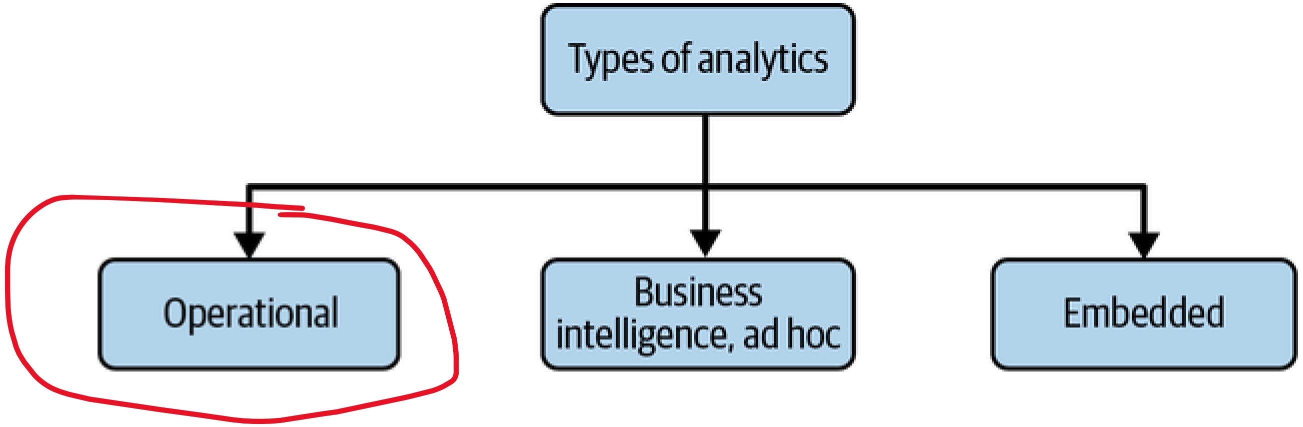
Live view

Past and present of business

Multi-tenant (outbound BI)

Source: Fundamentals of Data Engineering (2022)

Analytics



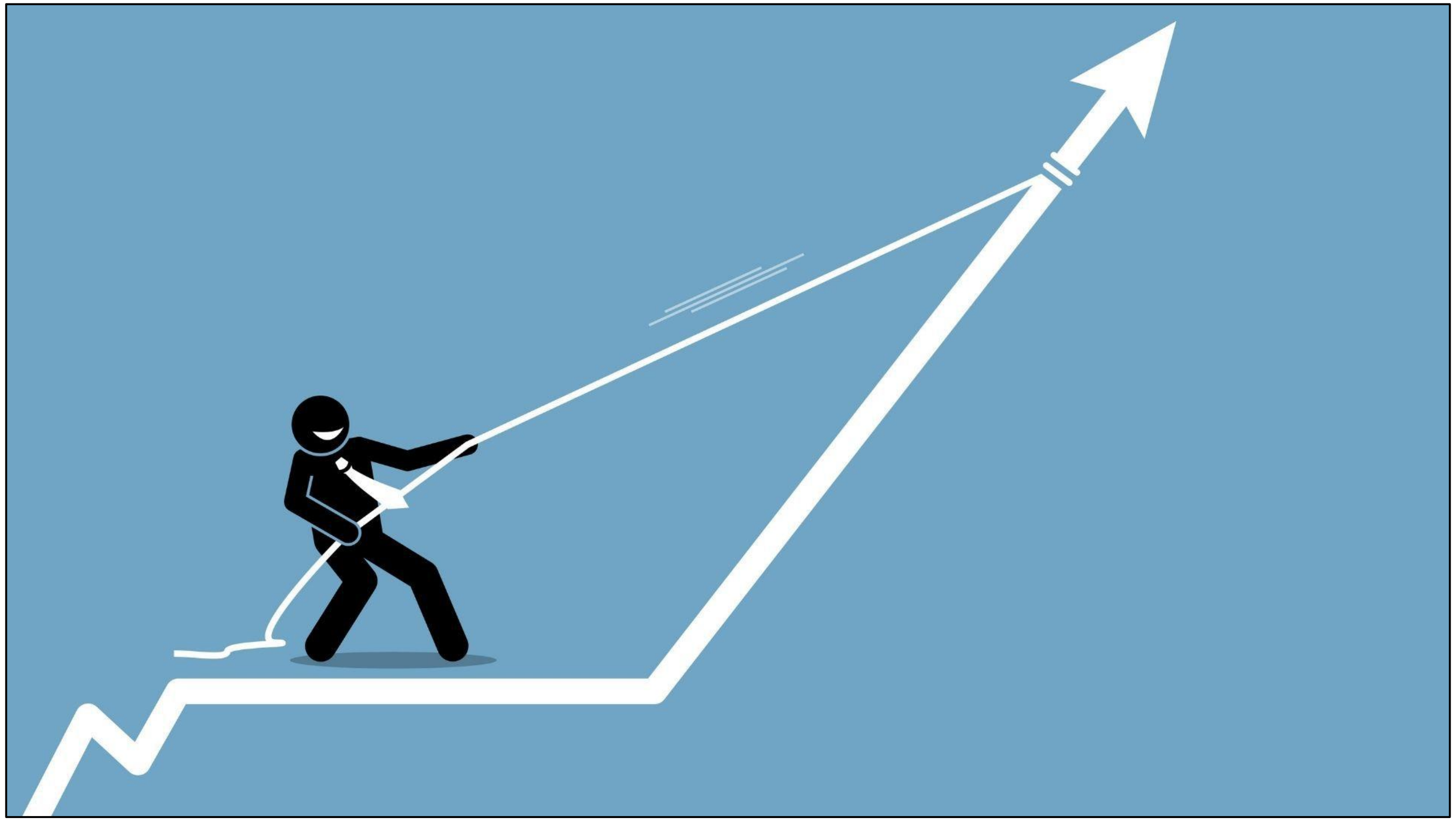
Live view

Past and present of business

Multi-tenant (outbound BI)

Source: Fundamentals of Data Engineering (2022)

Challenges



Scalability

Challenges

Ordering

Challenges

Consistency

Challenges

Fault Tolerance

Recap

- Data generation boom
- Data processing shift
  - From batch to streaming
- We need a live view
  - Operational analytics
  - Near real-time (seconds/minutes latency)
- Challenges
  - Scalability – Ordering – Consistency – Fault Tolerance

Not discussed but important

- Reverse ETL
- Machine Learning
- DataOps
- Security
- Data Management

Kafka

Messaging

Messaging

Messaging

- Hello, would you like to hear a TCP joke?

Messaging

- Hello, would you like to hear a TCP joke?
- Hello, yes, I’d like to hear a TCP joke.

Messaging

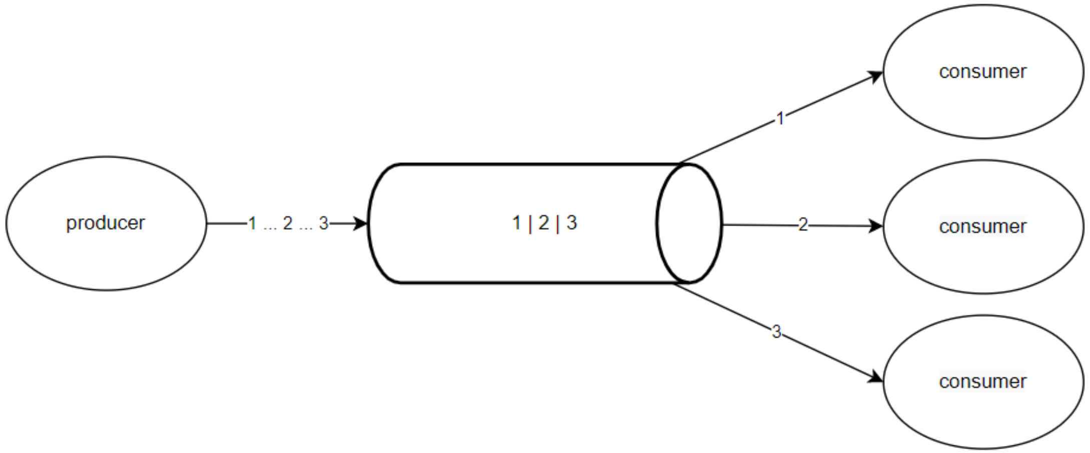
- Hello, would you like to hear a TCP joke?
- Hello, yes, I’d like to hear a TCP joke.
- OK, I’ll tell you a TCP joke.

Messaging

- Hello, would you like to hear a TCP joke?
- Hello, yes, I’d like to hear a TCP joke.
- OK, I’ll tell you a TCP joke.
- OK, I’ll hear a TCP joke.

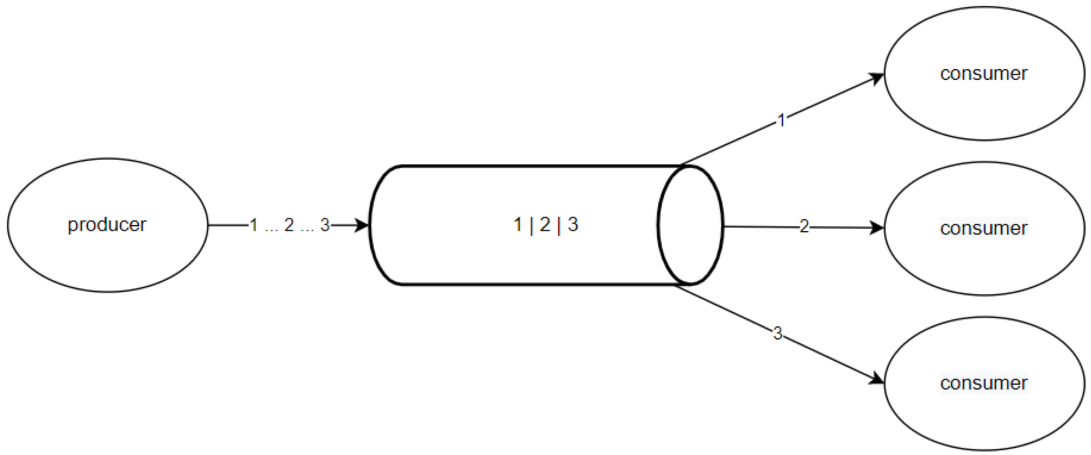
Kafka motivation

Traditional
Message-Queue

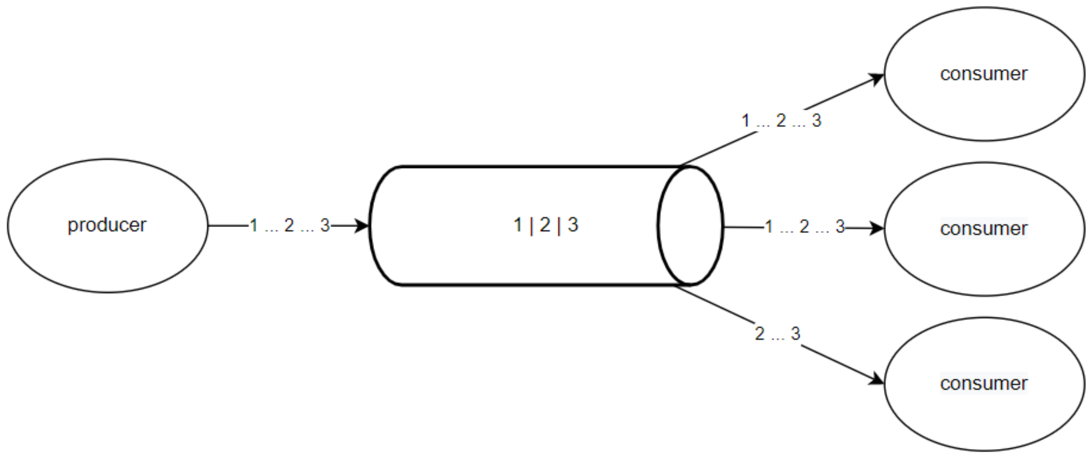


Kafka motivation

Traditional
Message-Queue

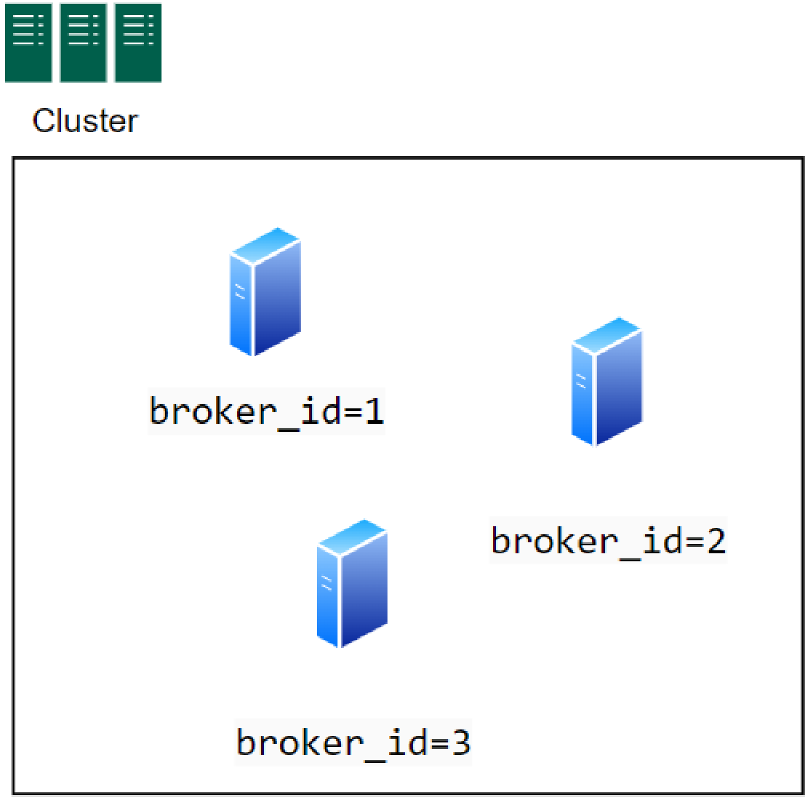


Kafka



- Distributed streaming platform
    - Publish & subscribe
    - Store streams durably
    - Process streams as they occur

Kafka cluster



Kafka topic

The core abstraction Kafka provides for a stream of records is the **topic**.
A topic is a category or feed name to which records are published.

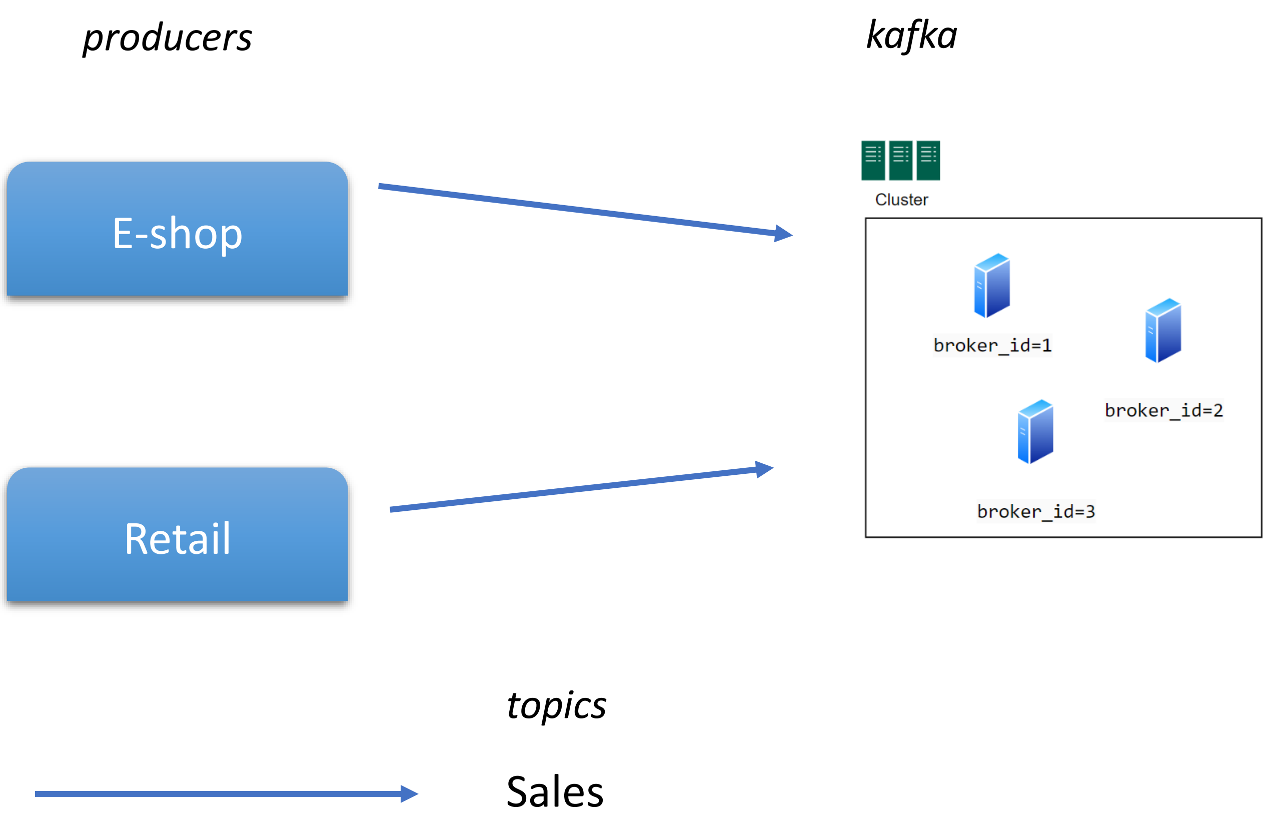
Kafka topic

producers

E-shop

Retail

Kafka topic



producers

kafka

E-shop

Retail

Cluster

broker_id=1

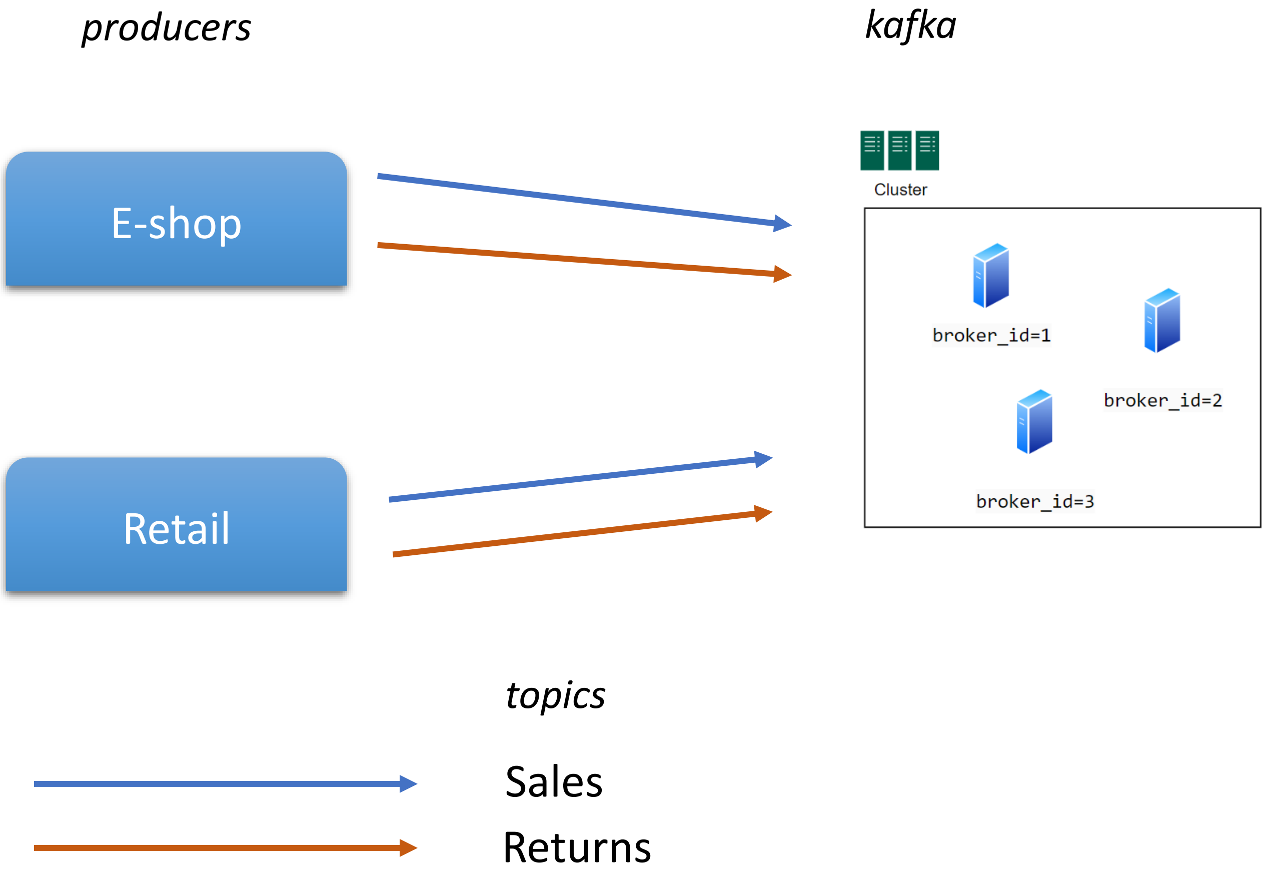
broker_id=2

broker_id=3

topics

Sales

Kafka topic



producers

kafka

E-shop

Retail

Cluster

broker_id=1

broker_id=2

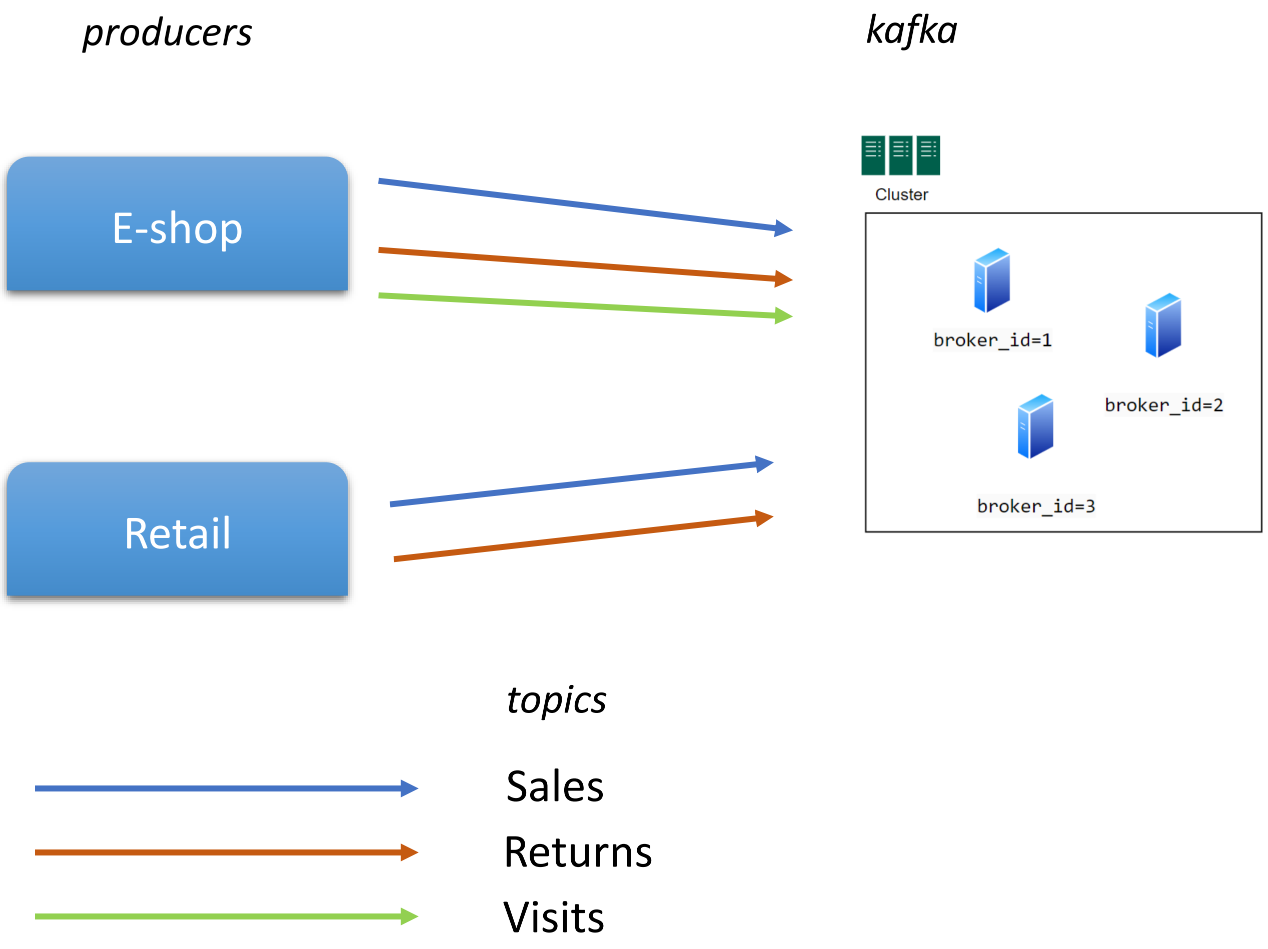
broker_id=3

topics

Sales

Returns

Kafka topic



producers

kafka

E-shop

Retail

Cluster

broker_id=1

broker_id=2

broker_id=3

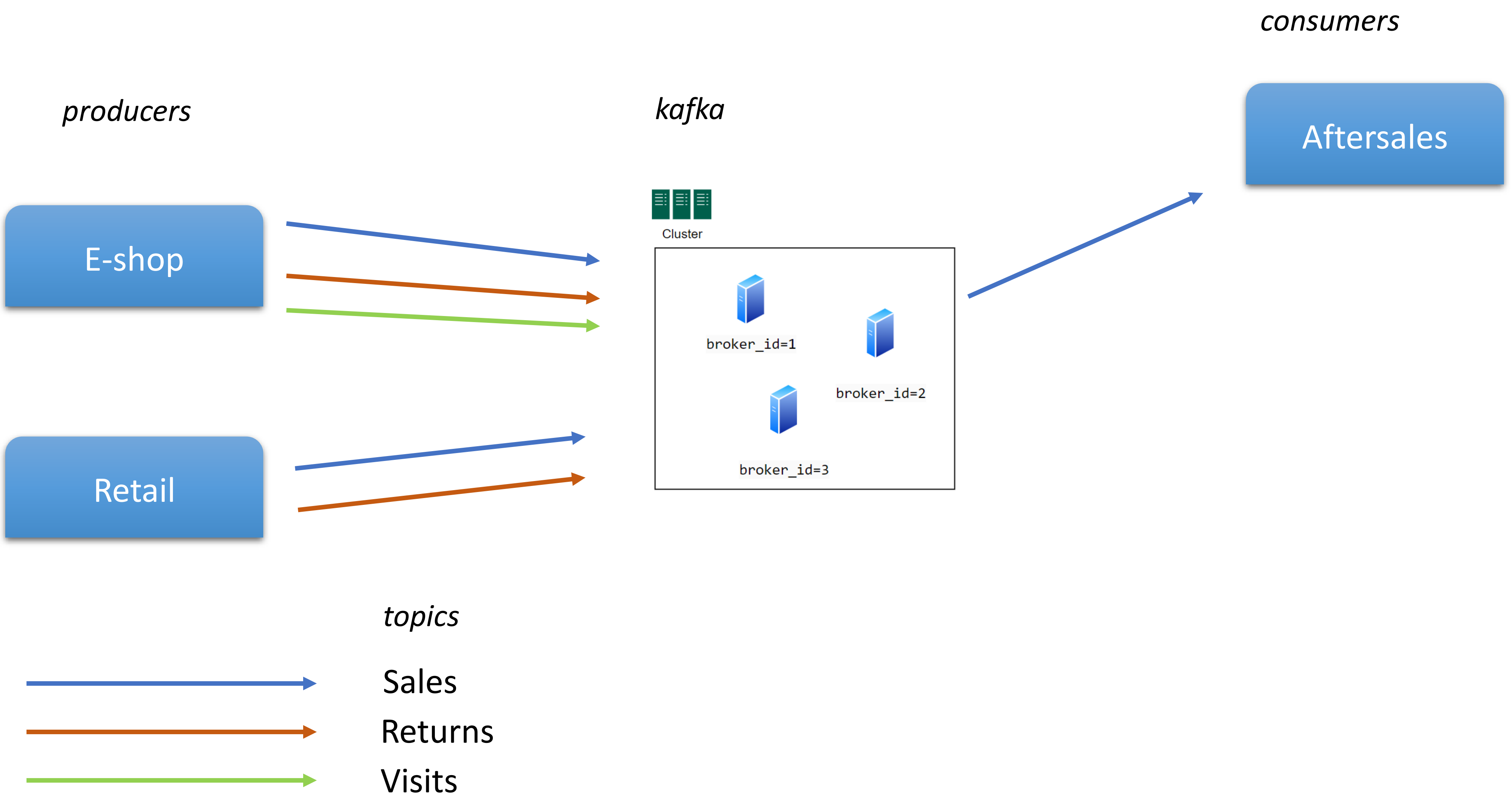
topics

Sales

Returns

Visits

Kafka topic



consumers

producers

kafka

Aftersales

E-shop

Cluster

broker_id=1

broker_id=2

broker_id=3

Retail

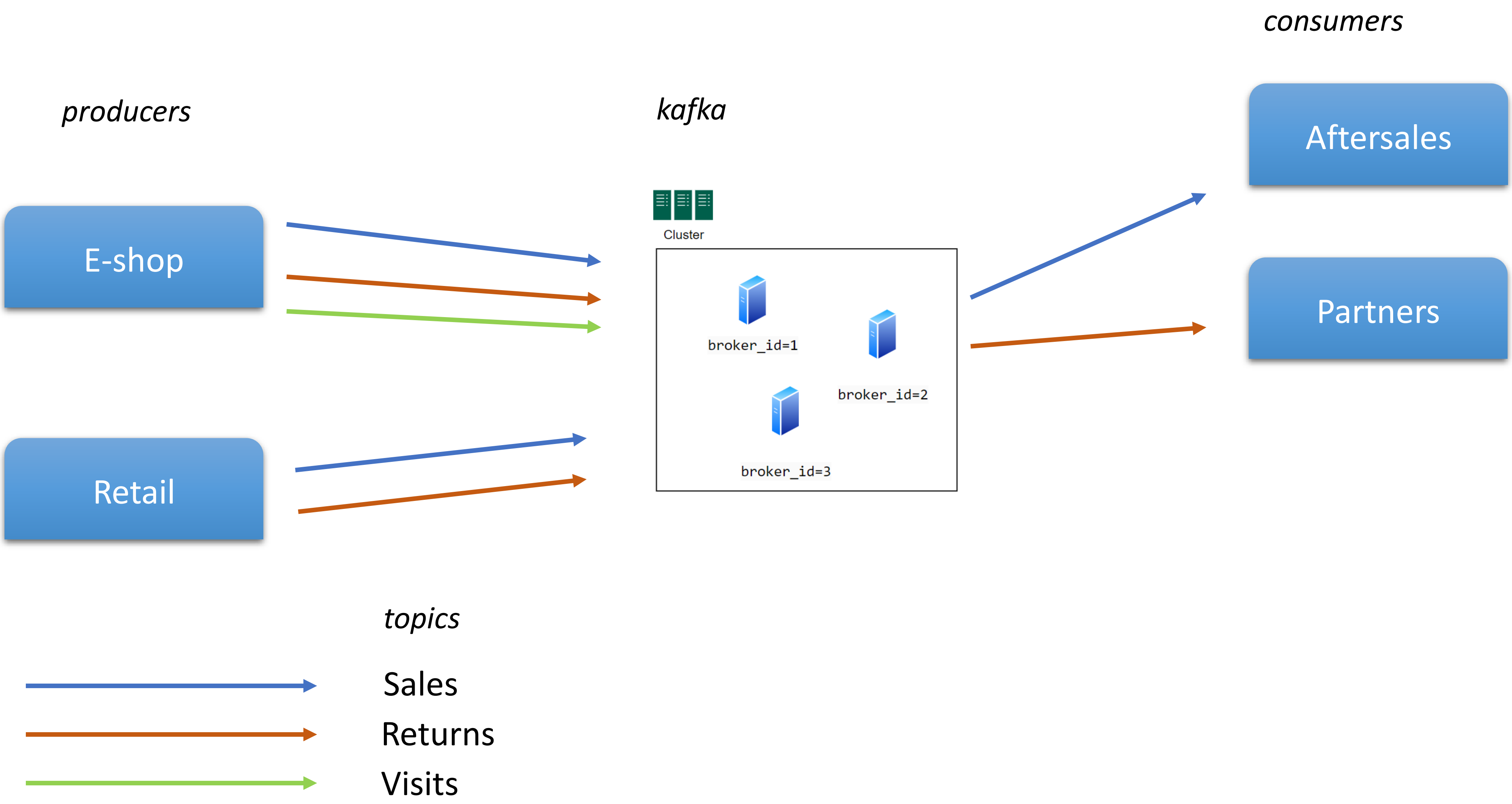
topics

Sales

Returns

Visits

Kafka topic



producers

E-shop

Retail

kafka

consumers

Aftersales

Partners

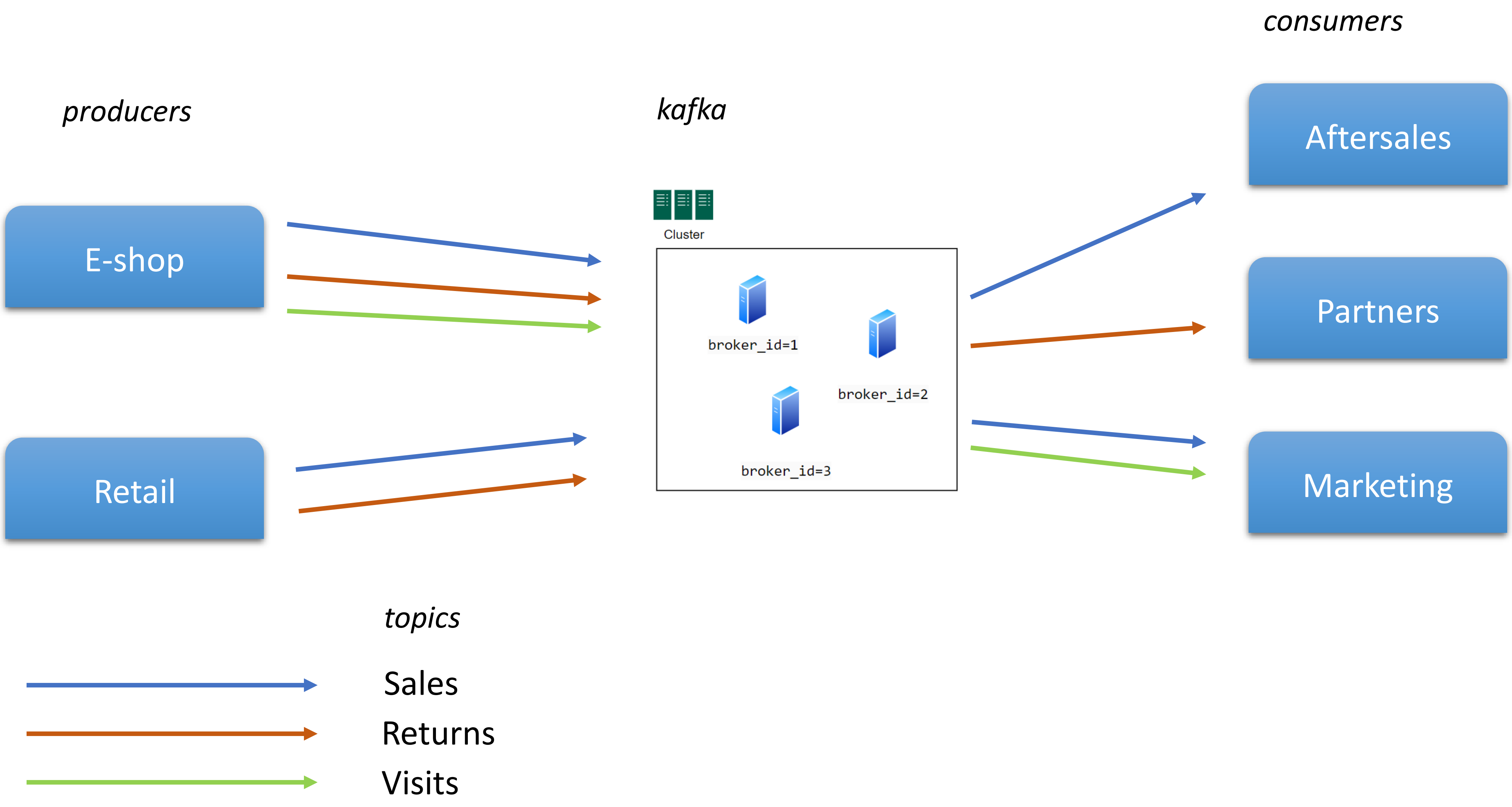
topics

Sales

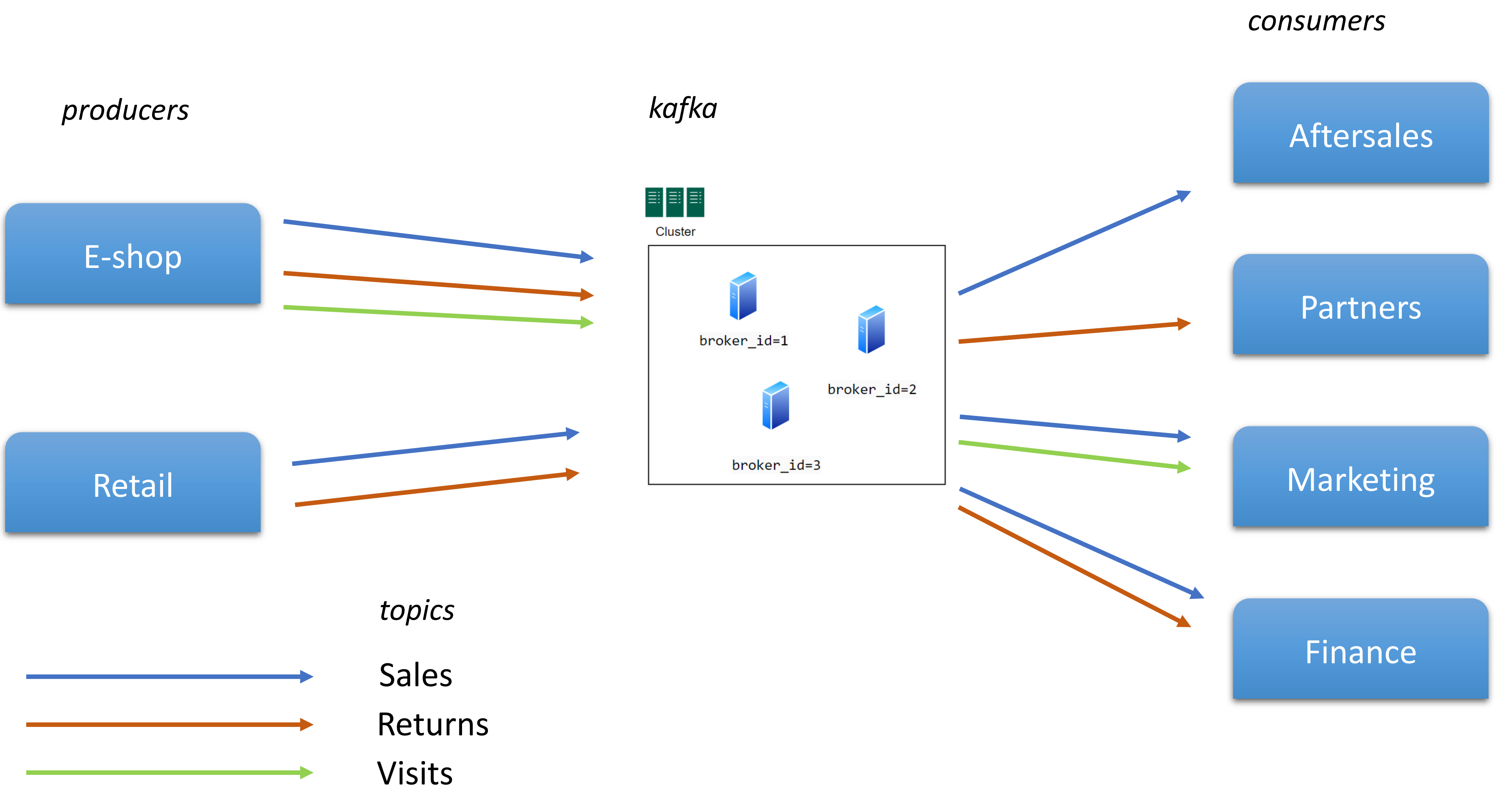
Returns

Visits

Kafka topic

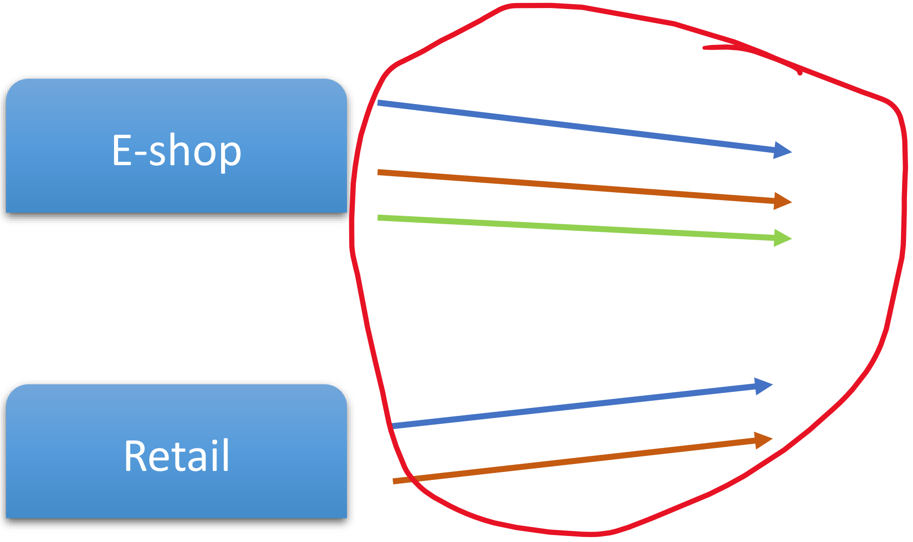

consumers
producers
kafka
Aftersales
E-shop
Cluster
broker_id=1
Partners
broker_id=2
broker_id=3
Retail
Marketing
topics
Sales
Returns
Visits

Kafka topic


consumers
producers
kafka
Aftersales
E-shop
Cluster
Partners
broker_id=1
broker_id=2
broker_id=3
Retail
Marketing
Finance
topics
Sales
Returns
Visits

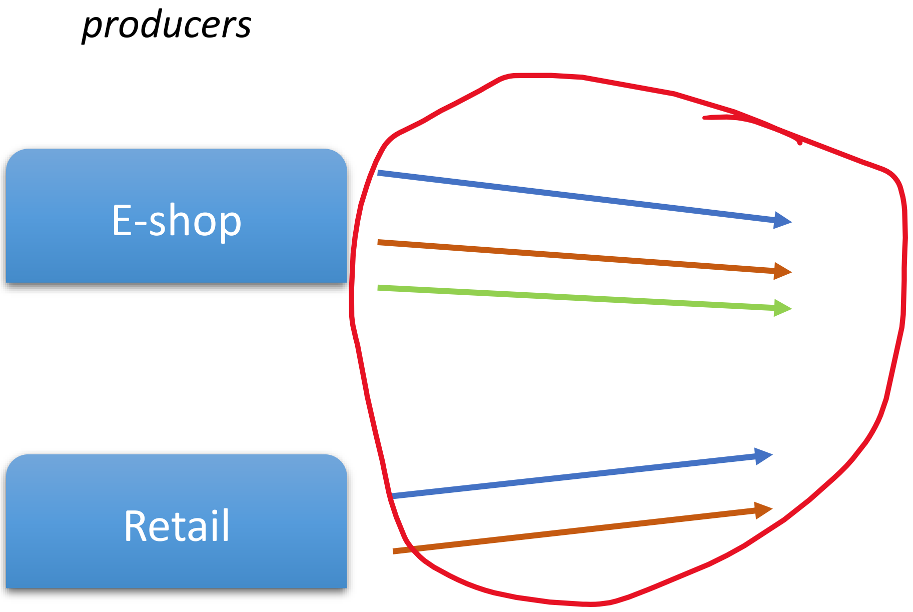
Kafka message?

producers



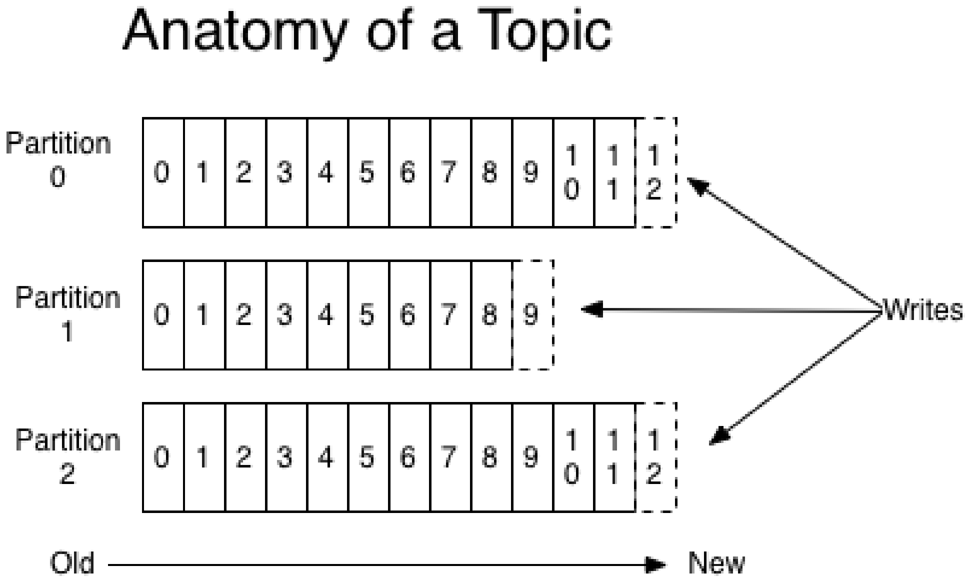
Kafka message?

- (Key)
- Value



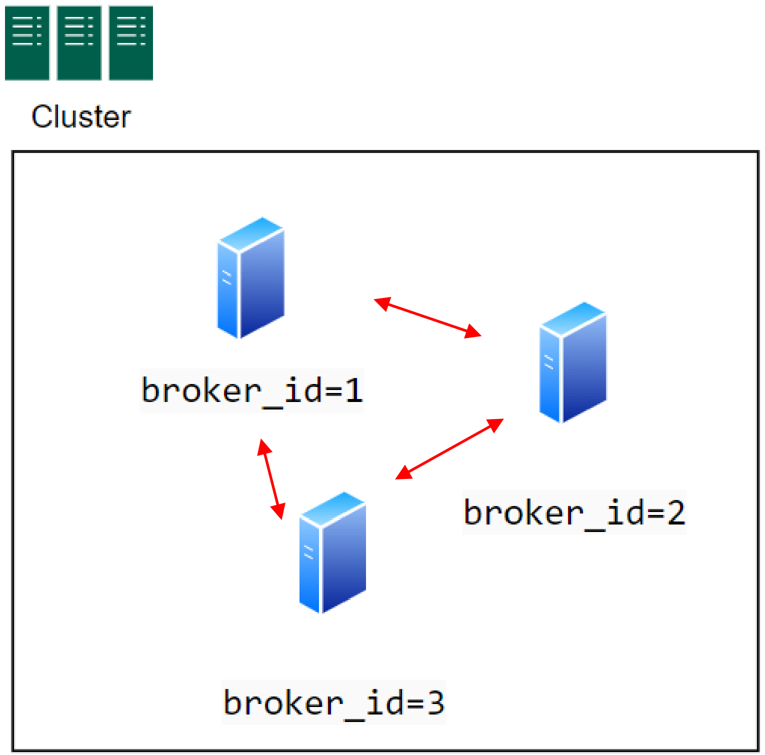
Kafka partitions

- Scalability



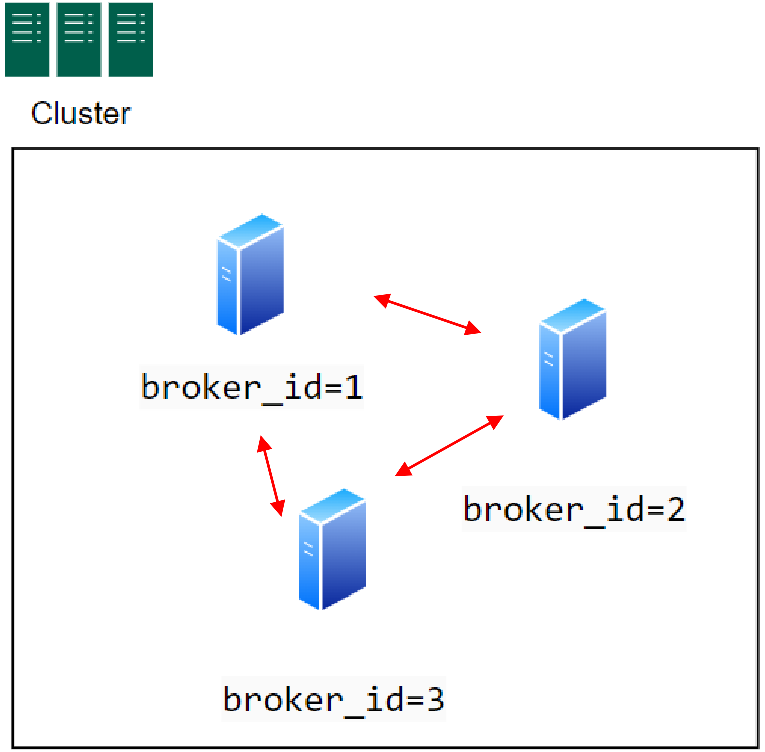
Kafka replication

- Fault-tolerance: if a broker is down, another broker can serve the data



Kafka replication

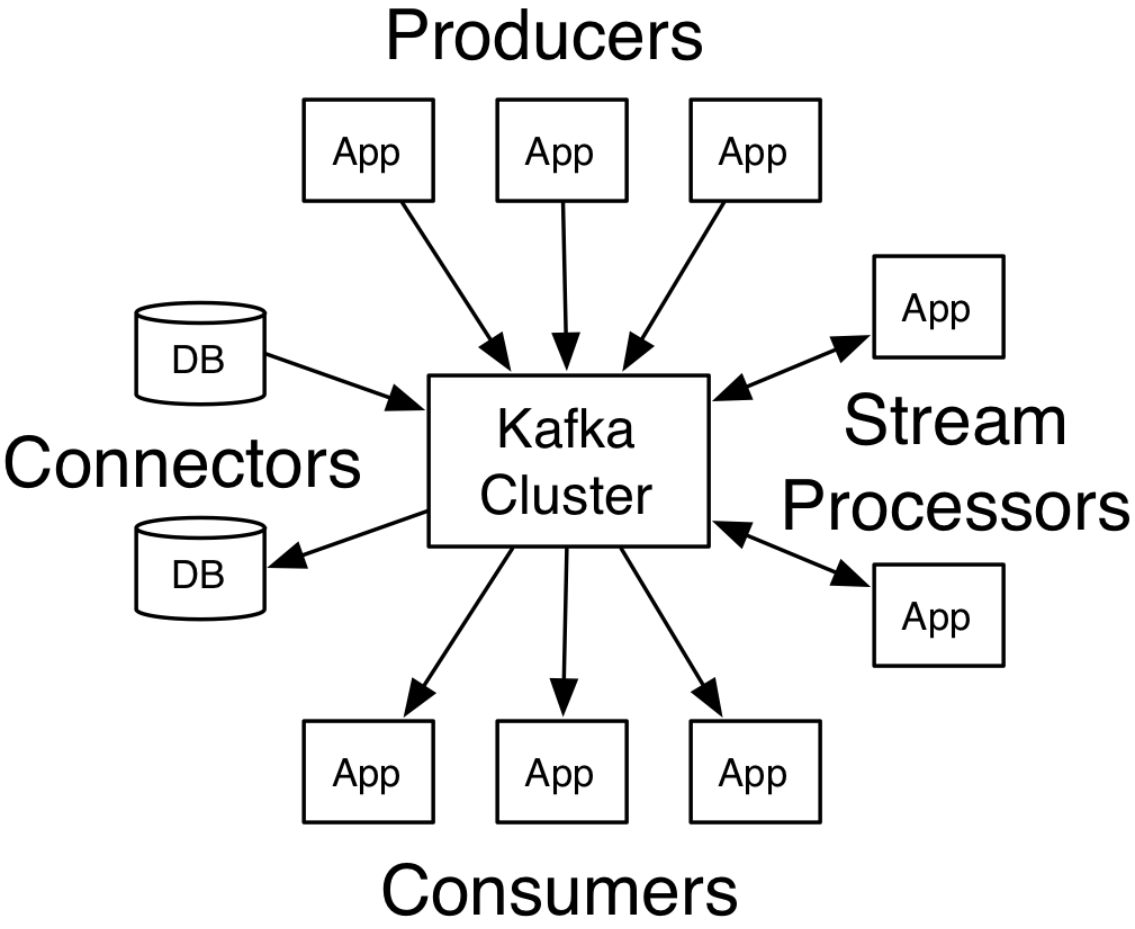
- Fault-tolerance: if a broker is down, another broker can serve the data



ISR = In-Sync Replica

Example

Kafka core


Producers
App
App
App
DB
Connectors
DB
Kafka Cluster
App
Stream Processors
App
App
App
App
Consumers

Kafka extended

- Schema registry
- Kafka Connect
  - CDC (Change Data Capture)
- Zookeeper, Kafka Raft (Kraft)
- ksqlDB

Let's Roll