PYCON 2017

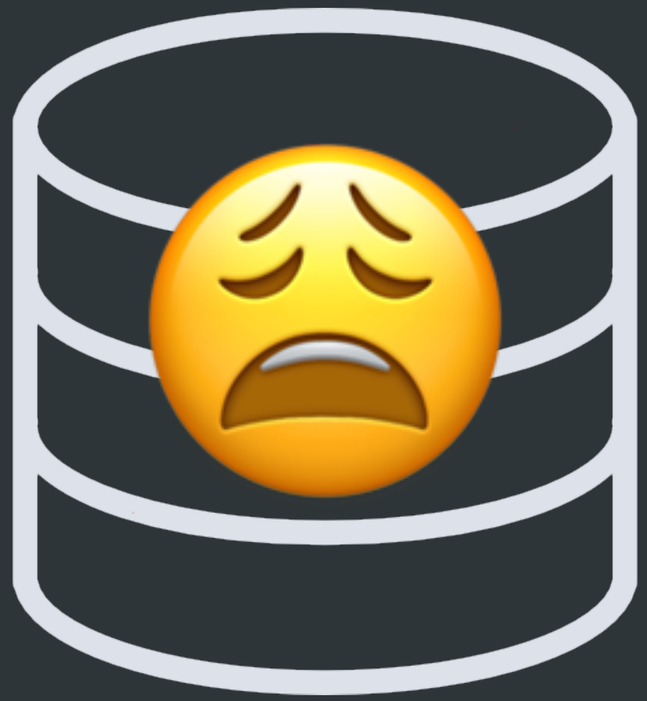
ONE DATA PIPELINE TO RULE THEM ALL

IT'S 11 PM. DO YOU KNOW WHERE YOUR DATA IS?

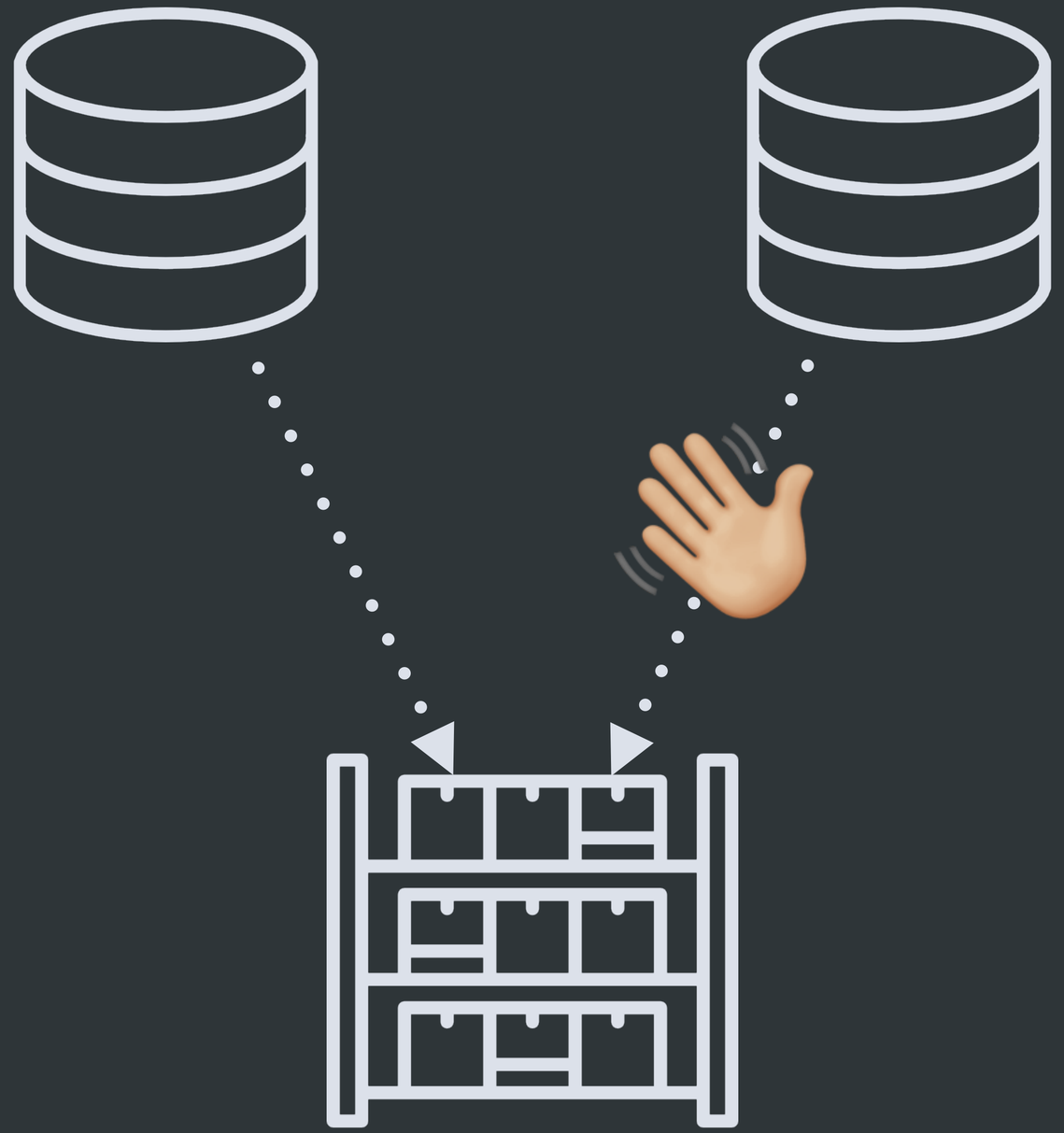
@SKIMBREL

TWILIO // DATA PLATFORM

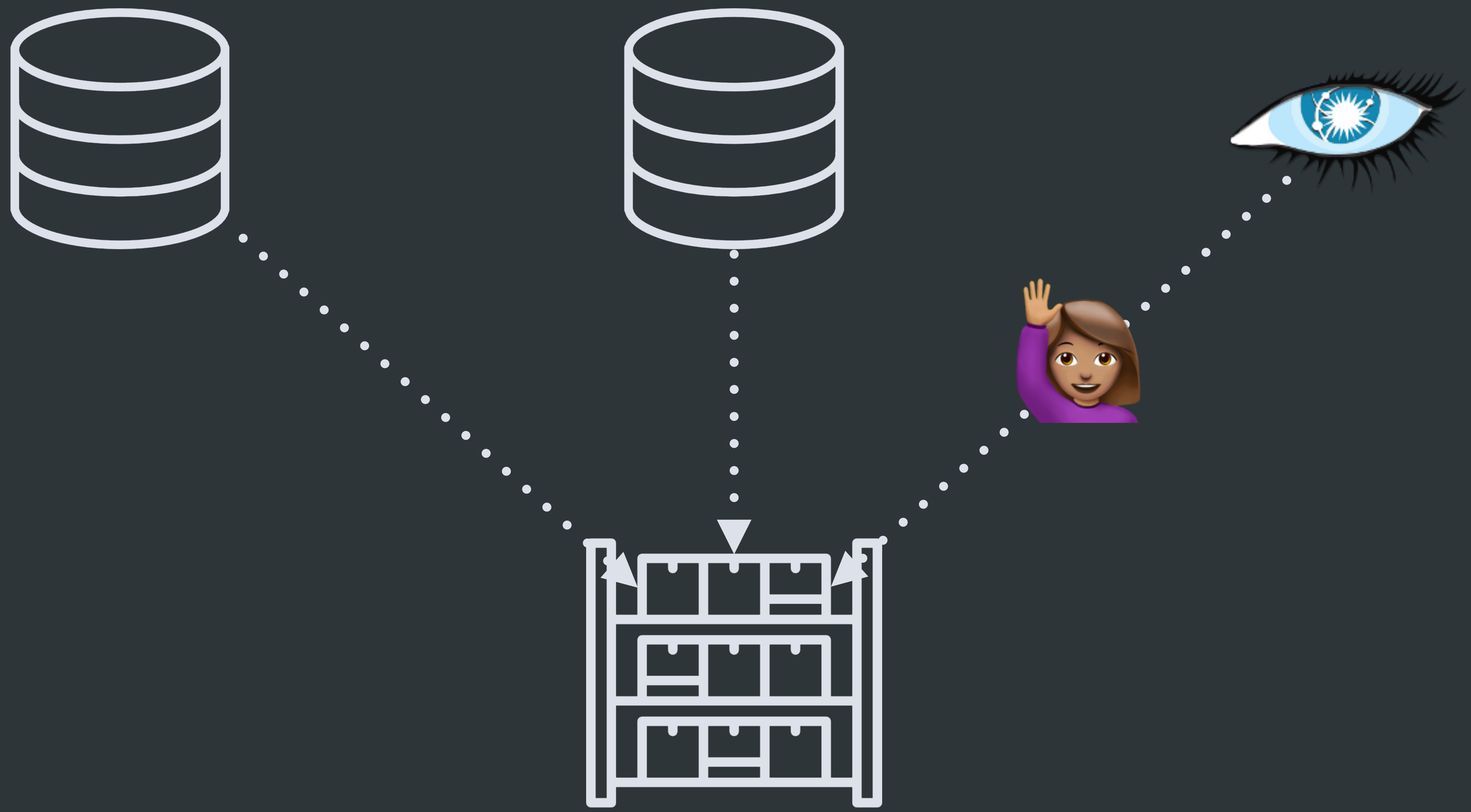
HEY, I'M SAM

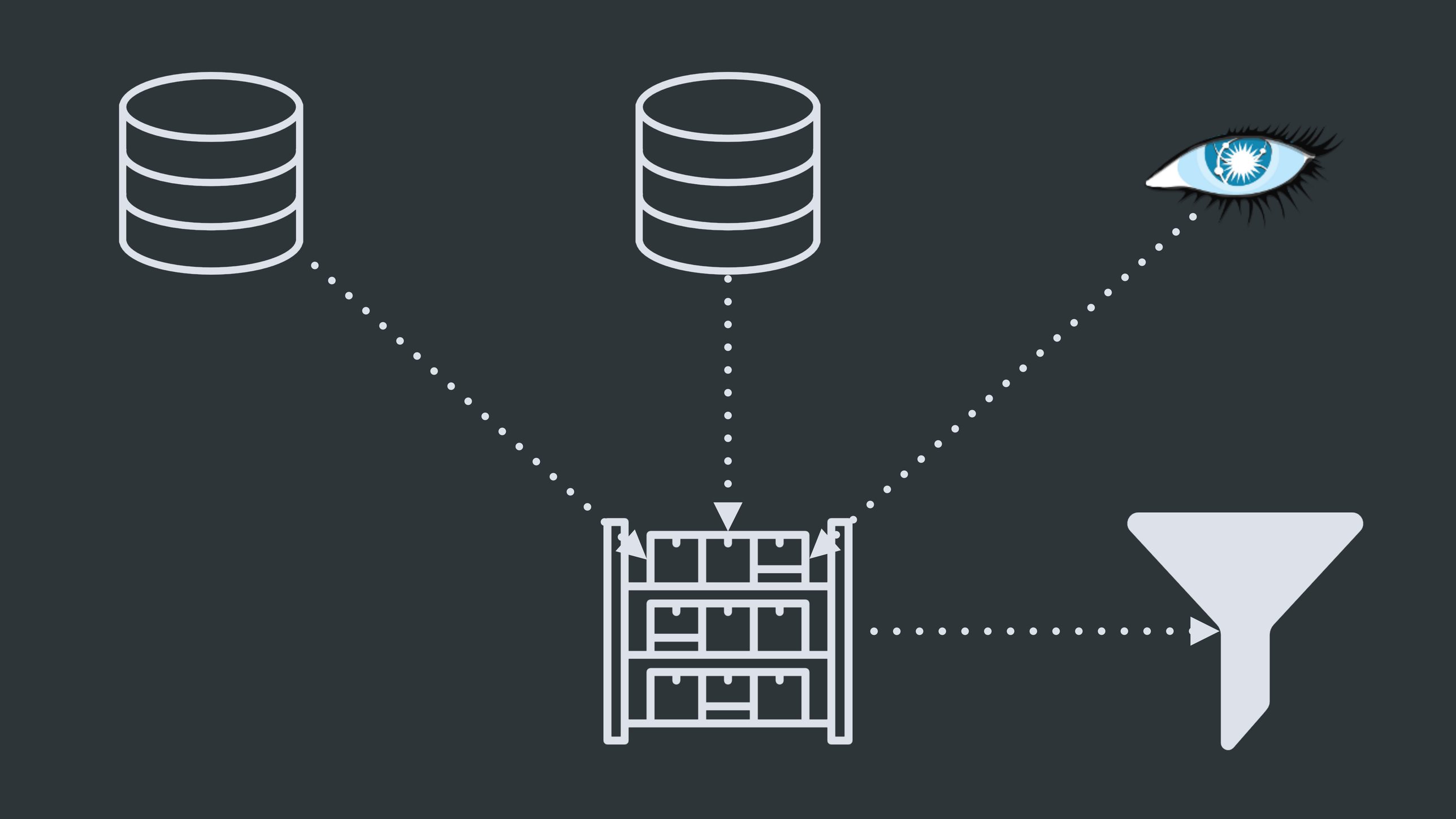


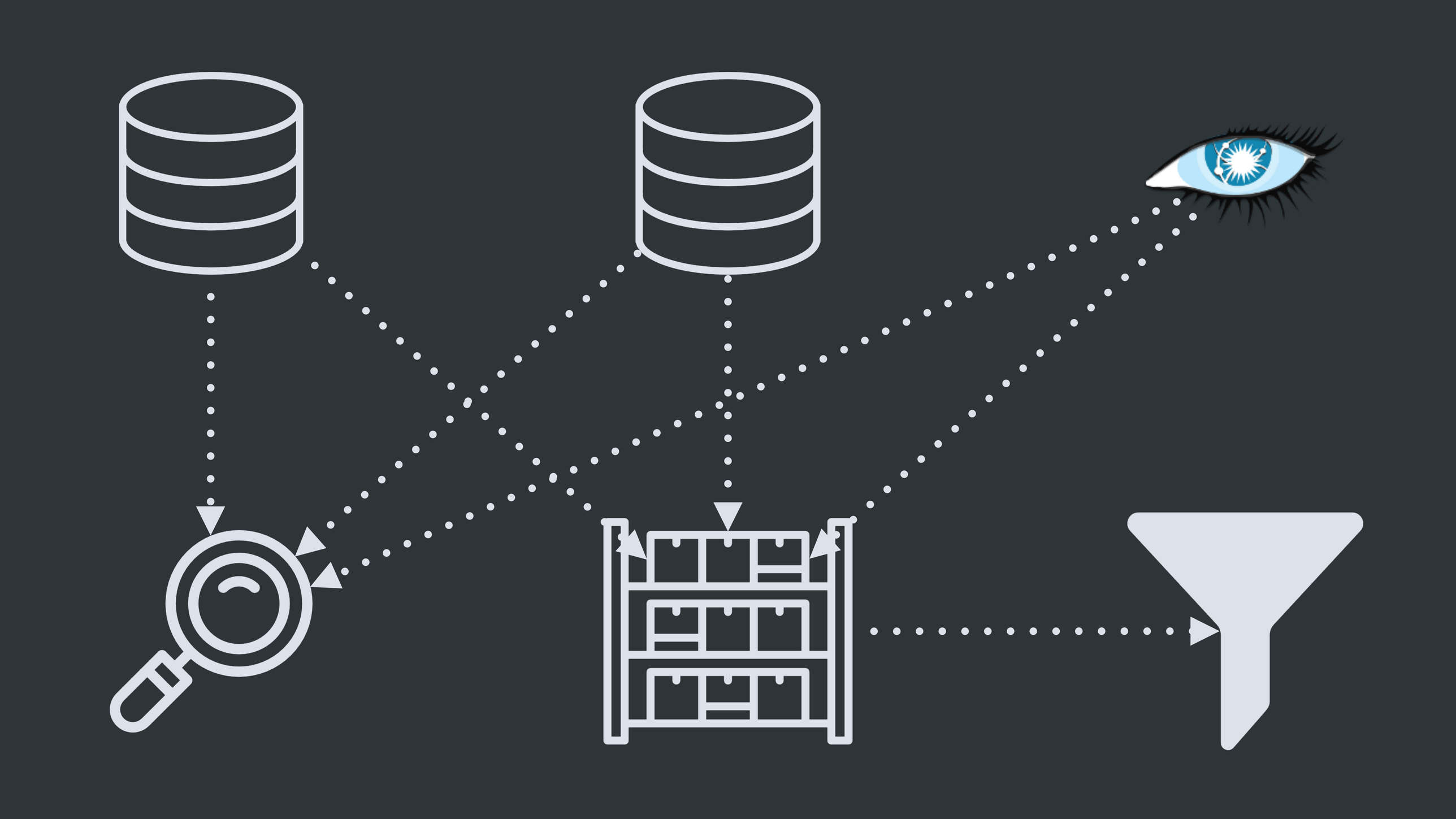

ETL'd!!!

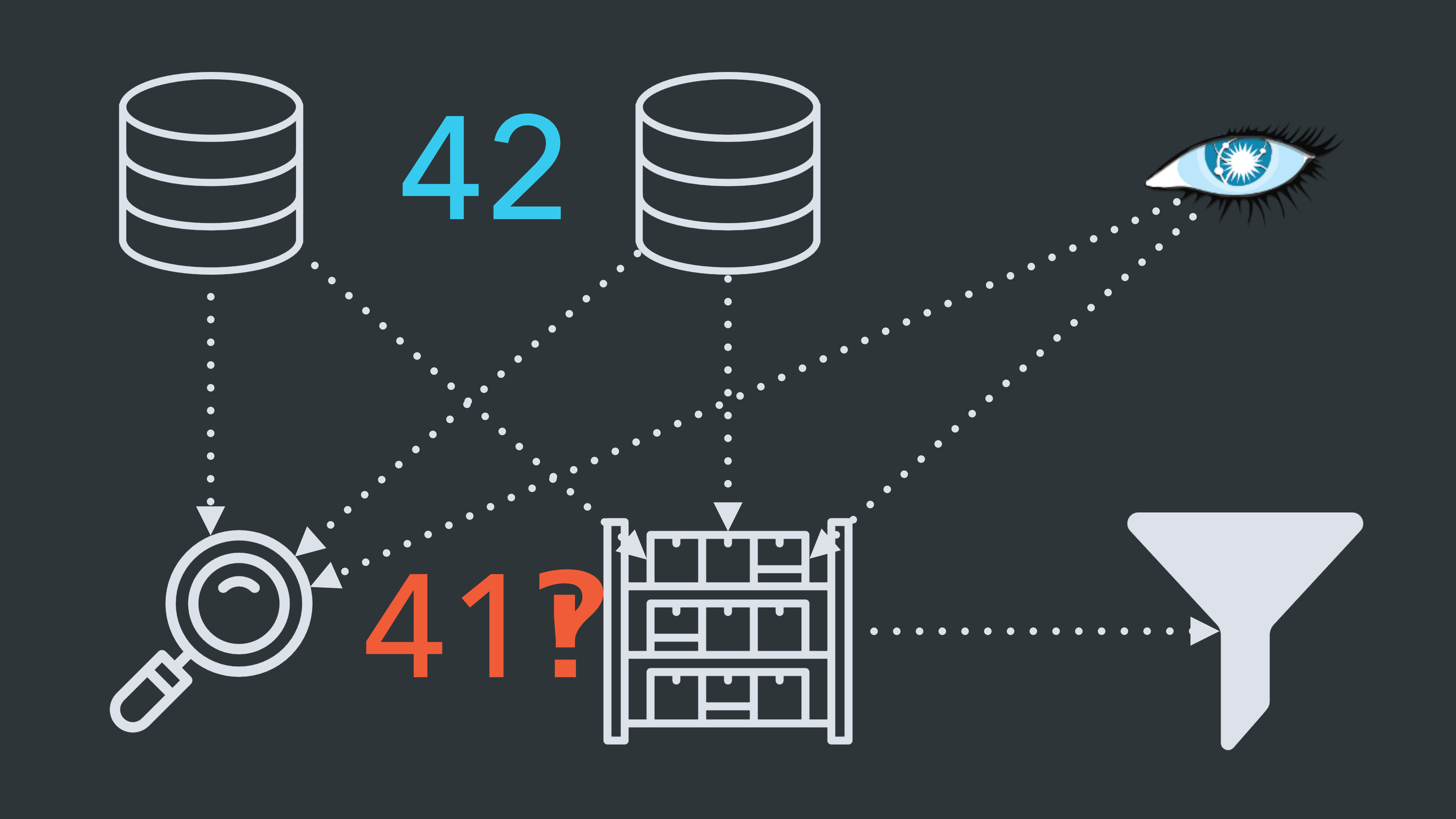



NEW
?








42
41?

YIKES.

WHAT HAPPENED?

ARCHITECTURAL PROBLEMS

Multiple data sources

Multiple data sinks

N^2 custom connection paths

ARCHITECTURAL PROBLEMS

Multiple data sources

Multiple data sinks

N^2 custom connection paths

No source of truth for schemas

ARCHITECTURAL PROBLEMS

Multiple data sources

Multiple data sinks

N^2 custom connection paths

No source of truth for schemas

No correctness guarantees

THE NEW SHINY

ONE (AND ONLY ONE) WAY TO PUBLISH DATA

COMMON STORAGE TOOLING

COMMON SCHEMAS

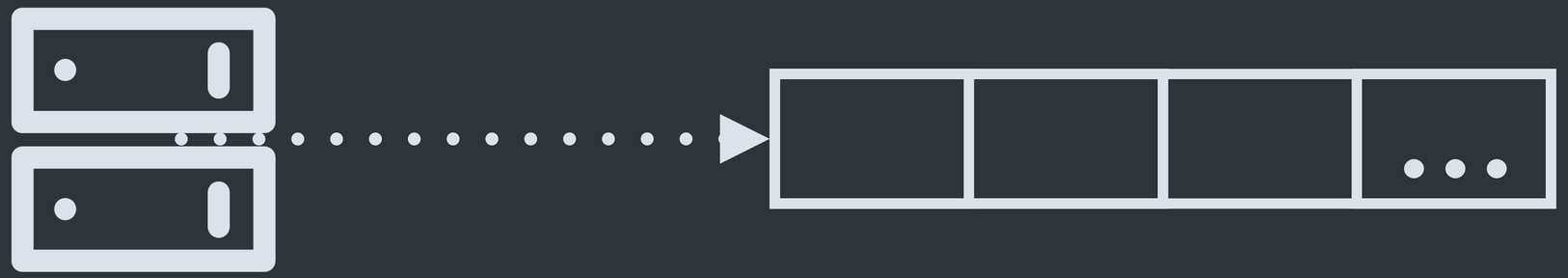
VERIFIABLE DELIVERY AND CORRECTNESS

EVENT PIPELINE

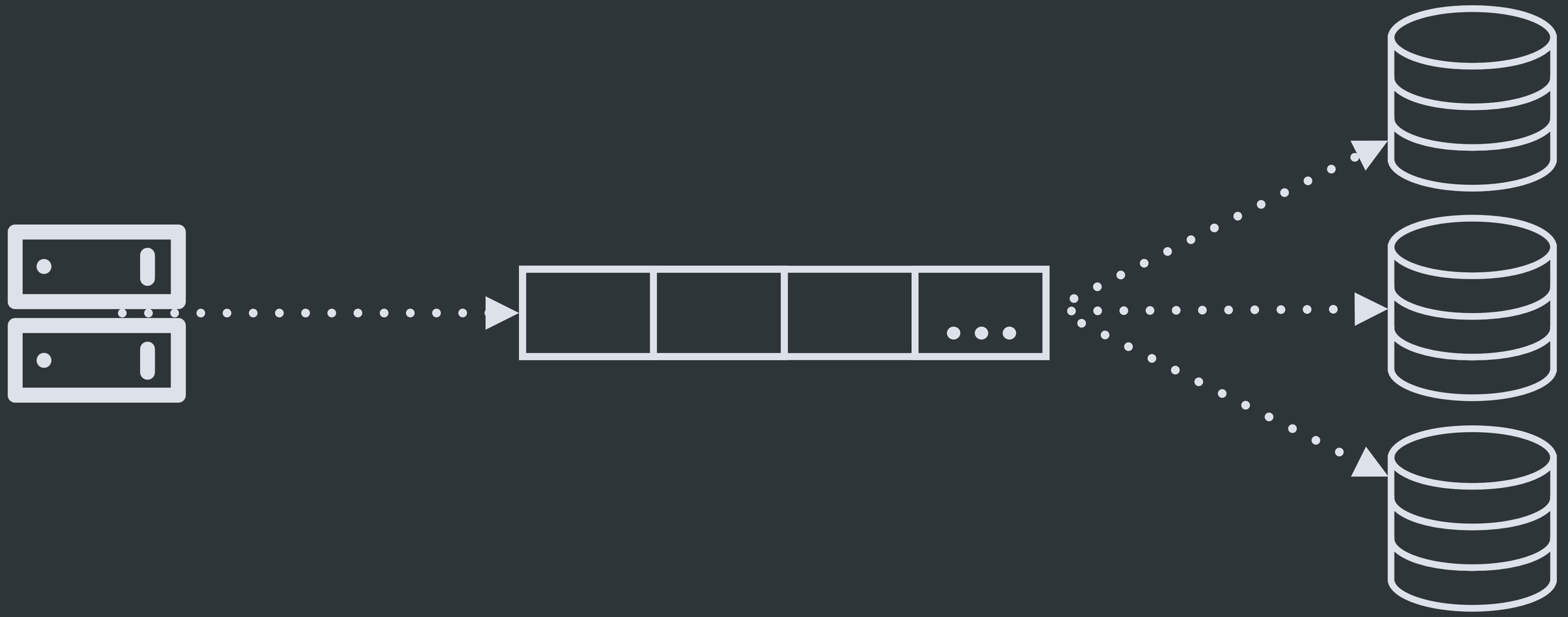
EVENT SOURCING ARCHITECTURE

EVENT SOURCING ARCHITECTURE

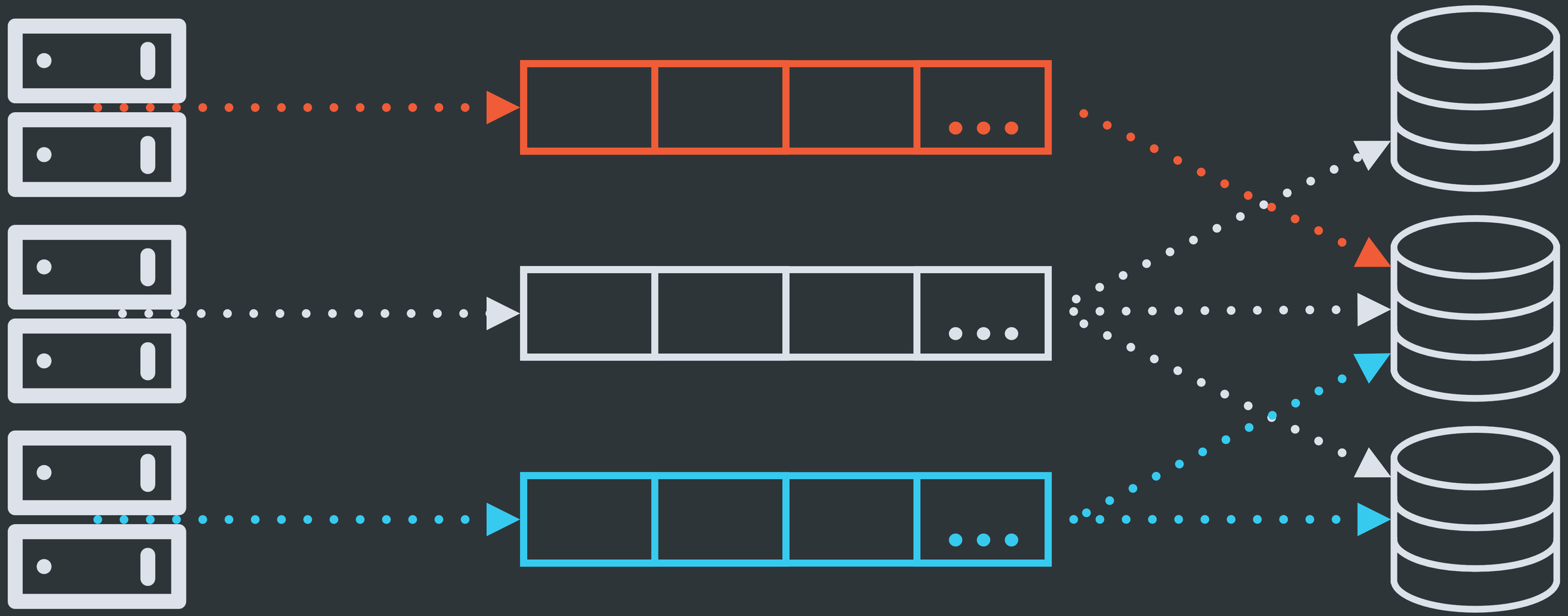
EVENT SOURCING ARCHITECTURE



EVENT SOURCING ARCHITECTURE



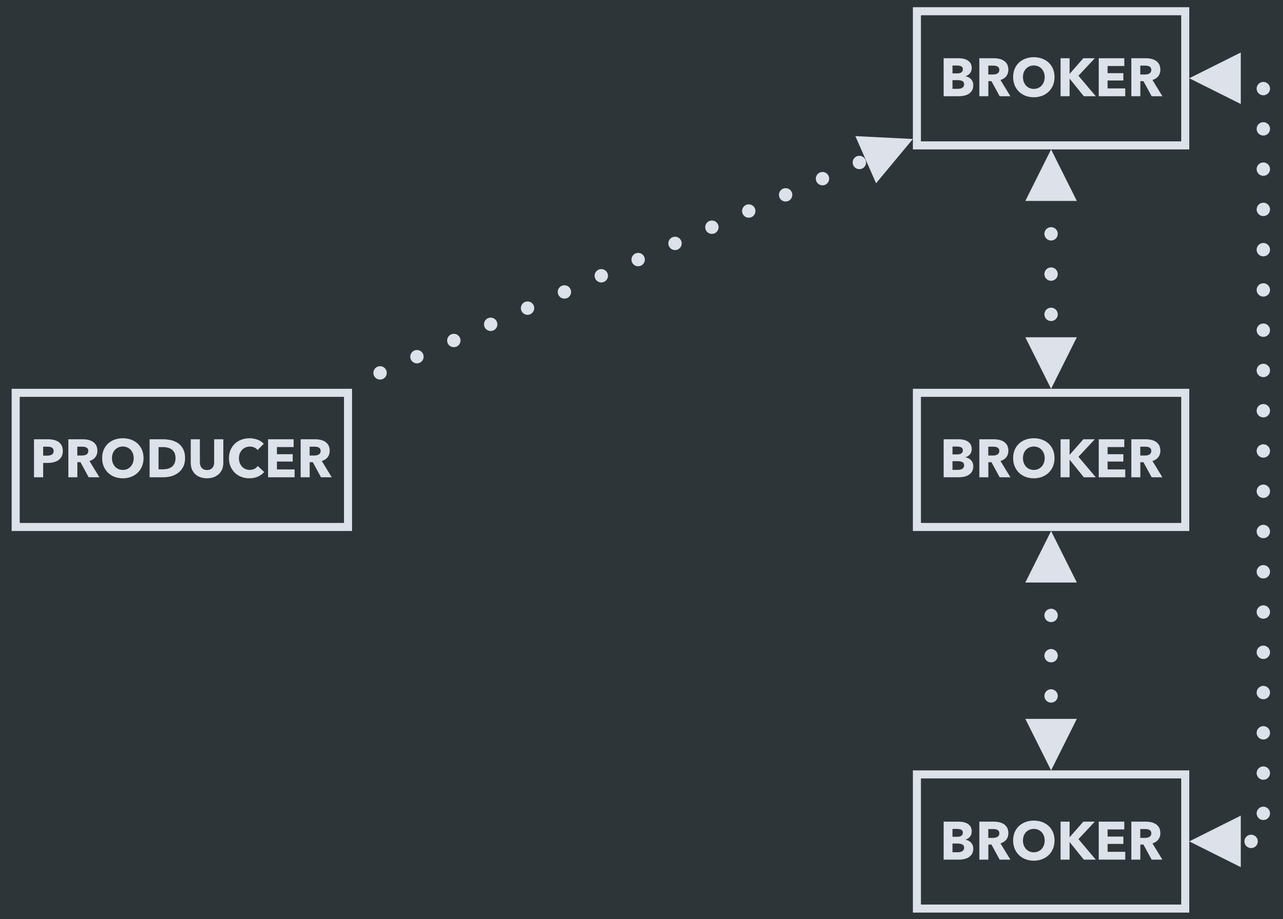
EVENT SOURCING ARCHITECTURE



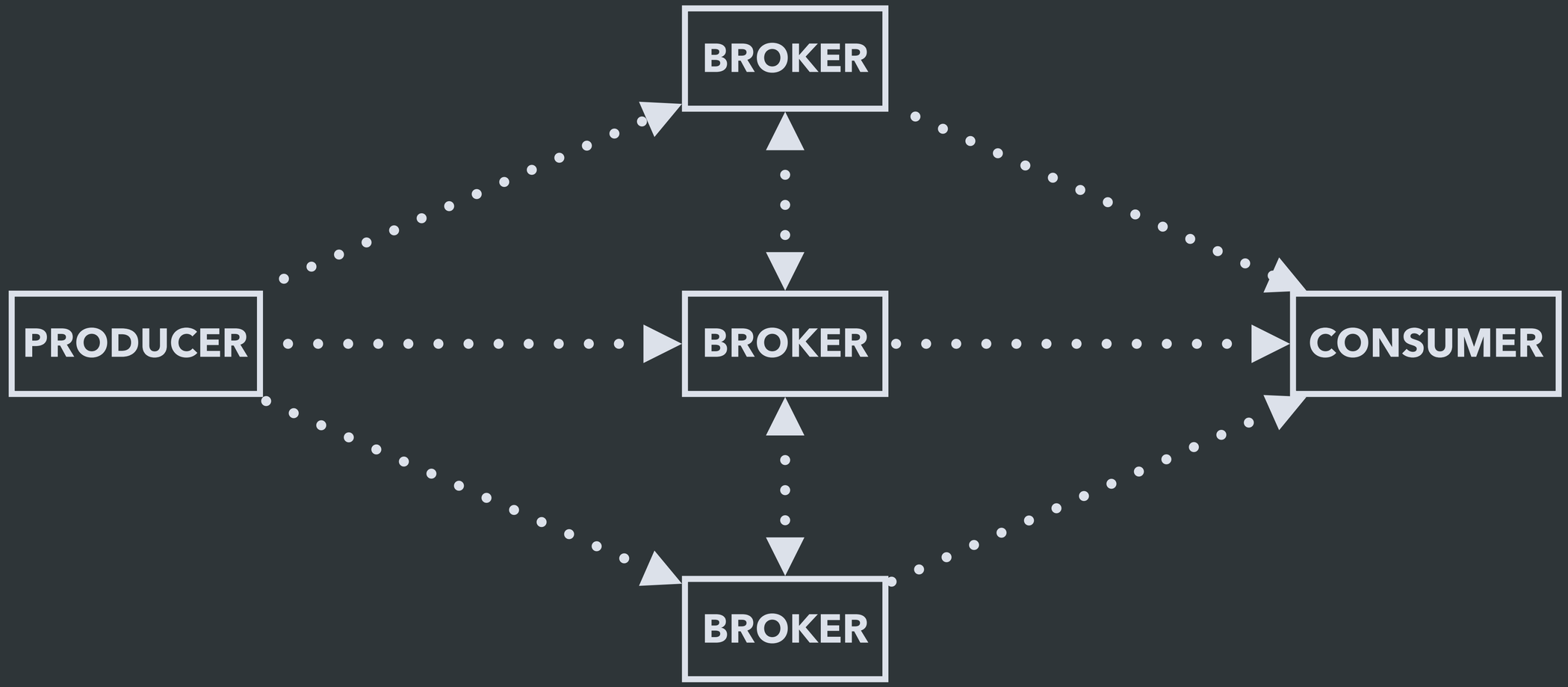

kafka

KAFKA ARCHITECTURE

KAFKA ARCHITECTURE



KAFKA ARCHITECTURE

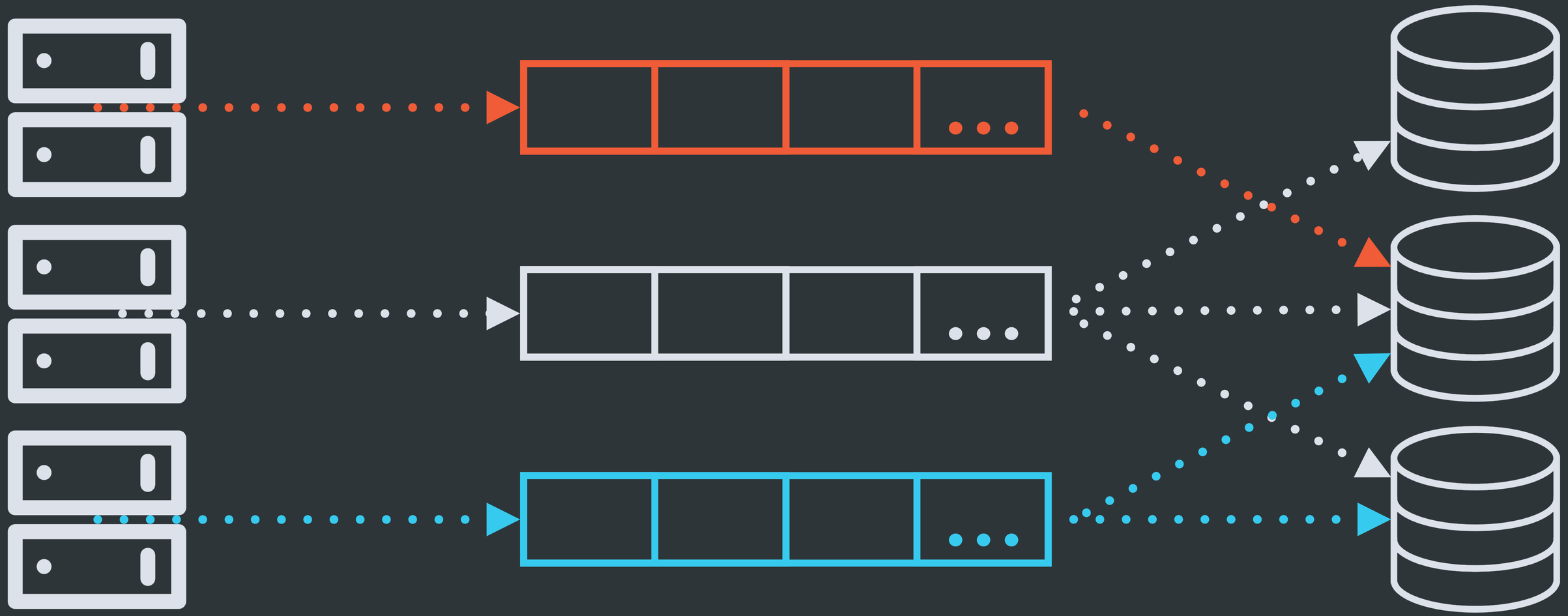


MANAGED KAFKA(-LIKE) SERVICES

Heroku

AWS Kinesis

KAFKA PIPELINE ARCHITECTURE



EVENT SOURCING: WHAT THE BLOG POSTS DON'T TELL YOU

SCHEMAS ARE IMPORTANT. LIKE, REALLY REALLY IMPORTANT.

SCHEMA LIBRARIES

Avro

Protocol Buffers

Thrift

msgpack

…

AVRO

Multiple language platforms

Dynamic and static bindings

Automatic cross-grading

Compact binary serialization

ENFORCE SCHEMAS AT PRODUCE TIME

TOPIC METADATA

What schema is in this topic?

TOPIC METADATA

How do we resolve duplicates?

TOPIC METADATA

What fields uniquely ID a record?

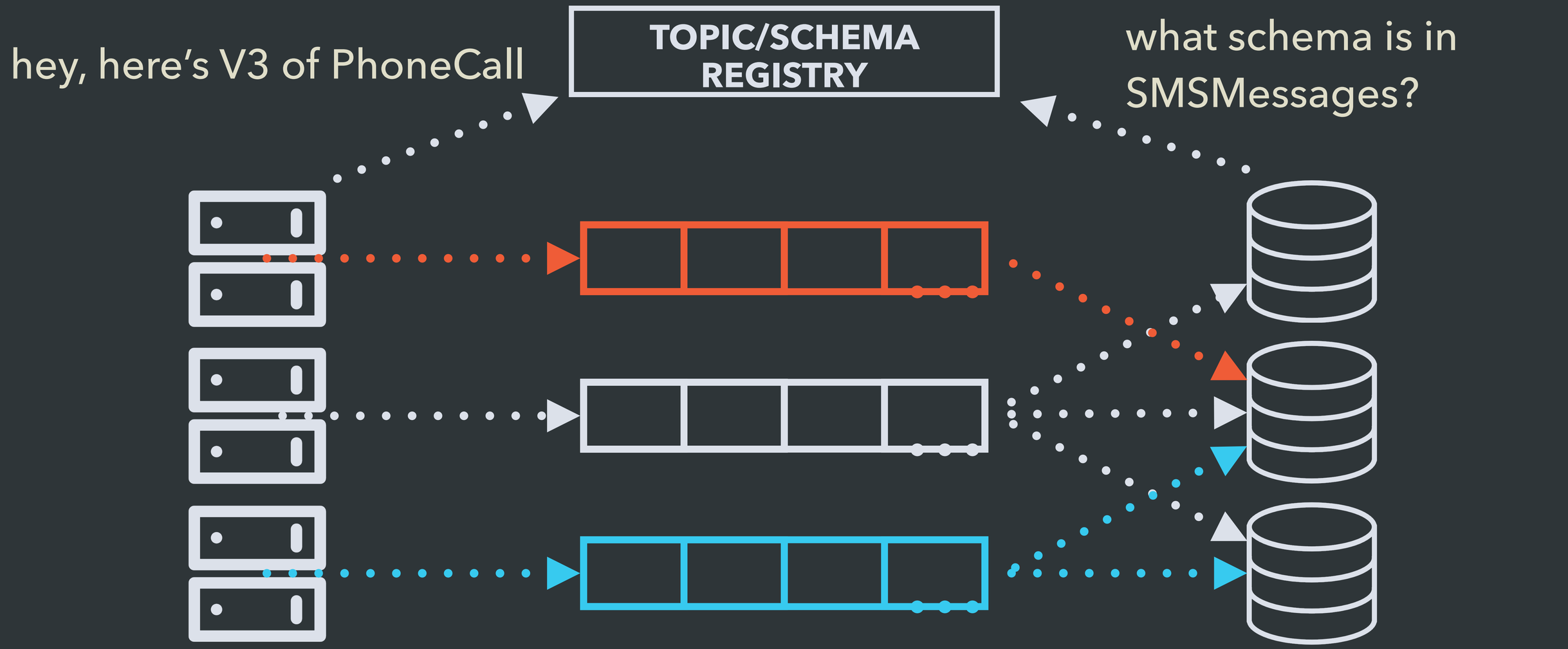
What fields and logic let us choose among or merge multiple versions?

TOPIC METADATA

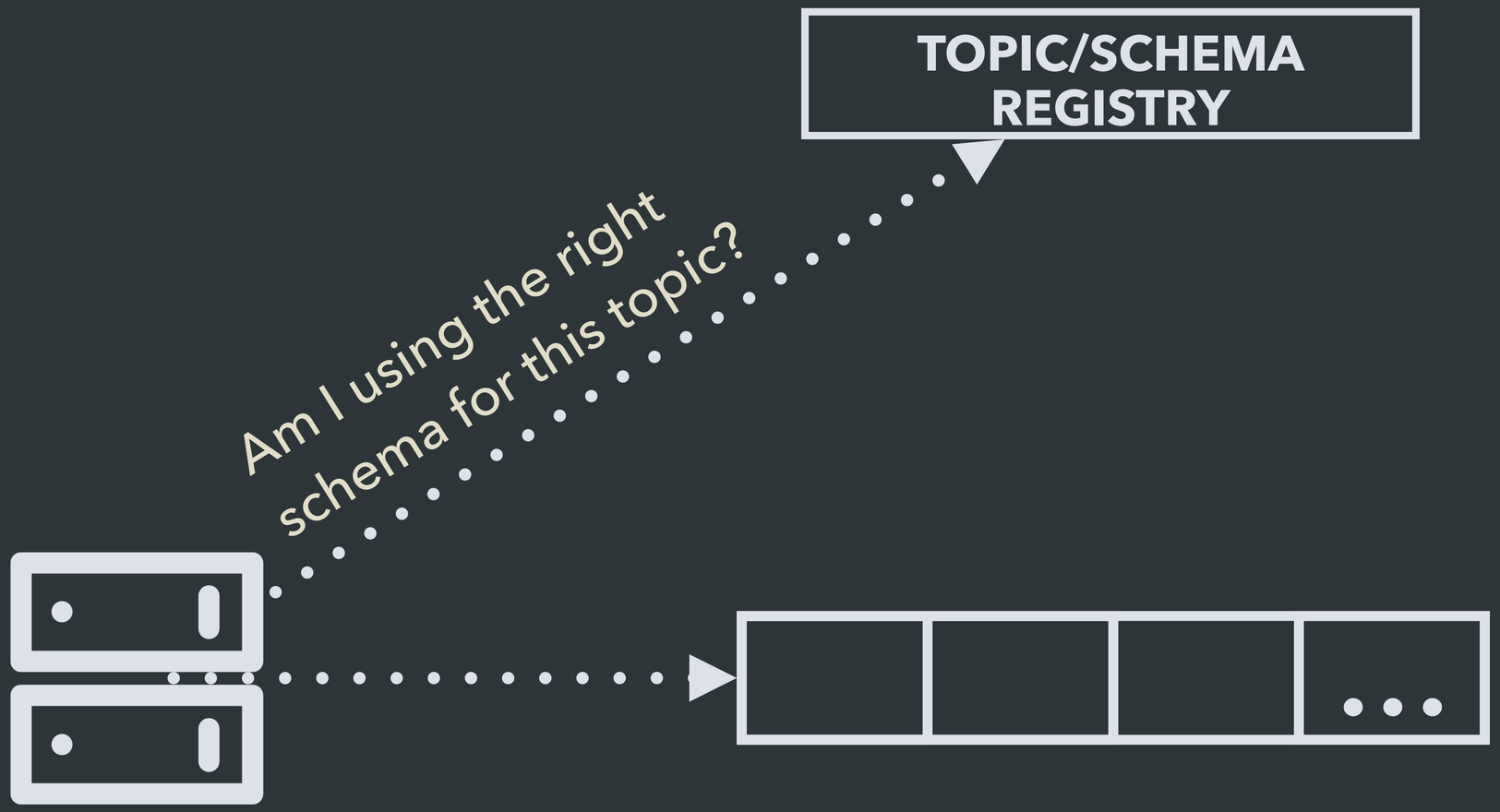
What tells us that the data is correct?

COMPONENTS

METADATA REGISTRY API

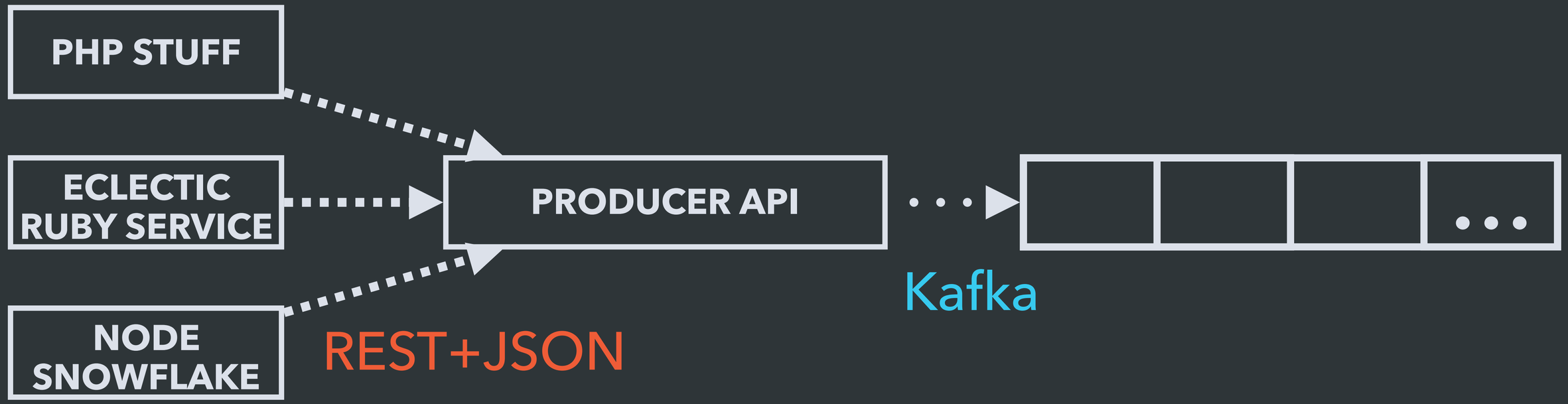


PRODUCER/CONSUMER LIBRARY



PRODUCER/CONSUMER LIBRARY

HTTP PRODUCER SERVICE



CONSUMER SYSTEMS

Archival

Warehousing and structured analytics

Batch processing

Ad-hoc analysis

Stream processing

Online query systems

CONSUMER SYSTEMS

Archival

Warehousing and structured analytics

Batch processing

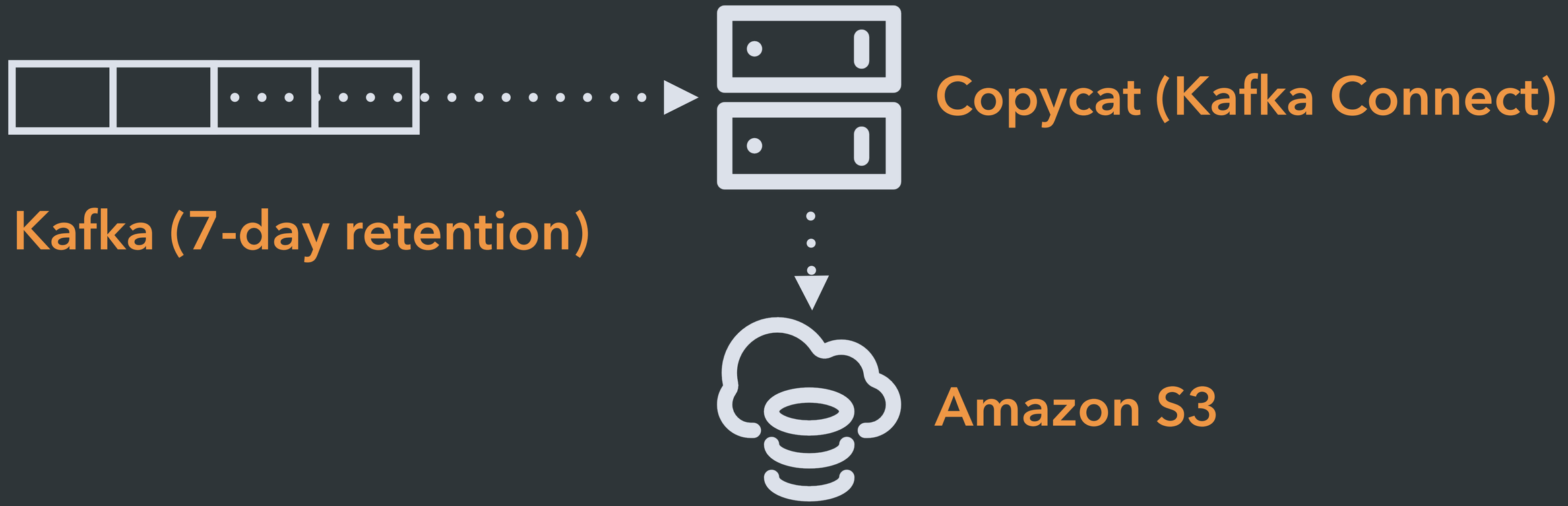
Ad-hoc analysis

Stream processing

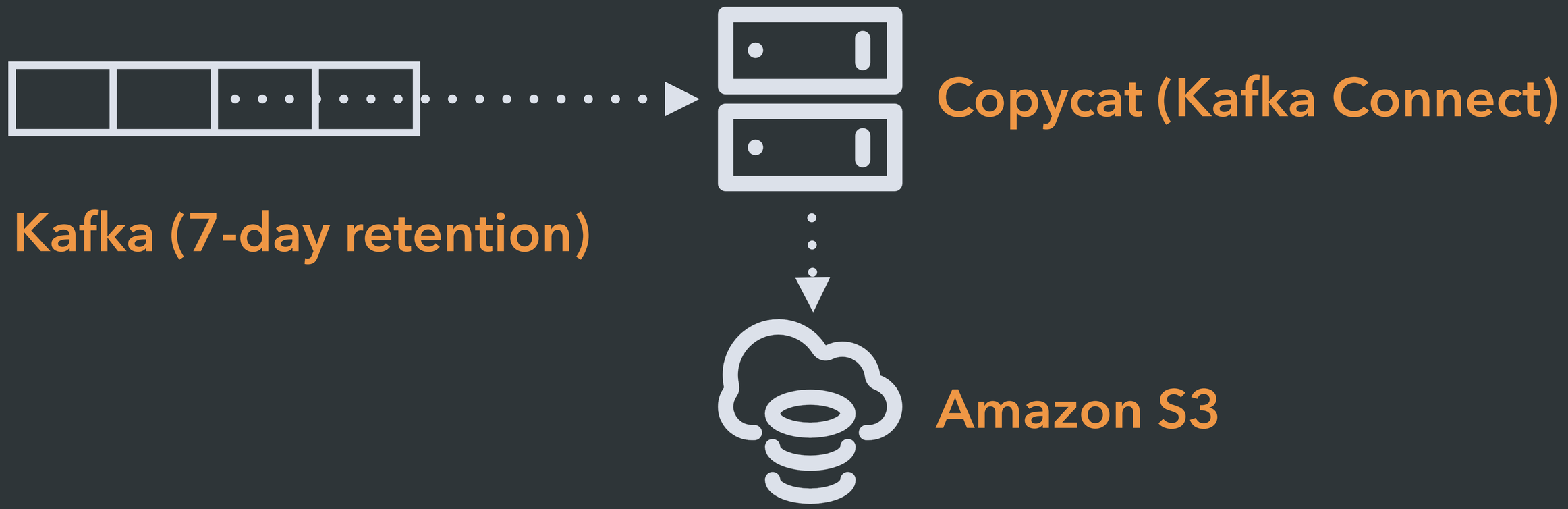
Online query systems

ARCHIVAL: TWILIOFS DATA LAKE

ARCHIVAL: TWILIOFS DATA LAKE

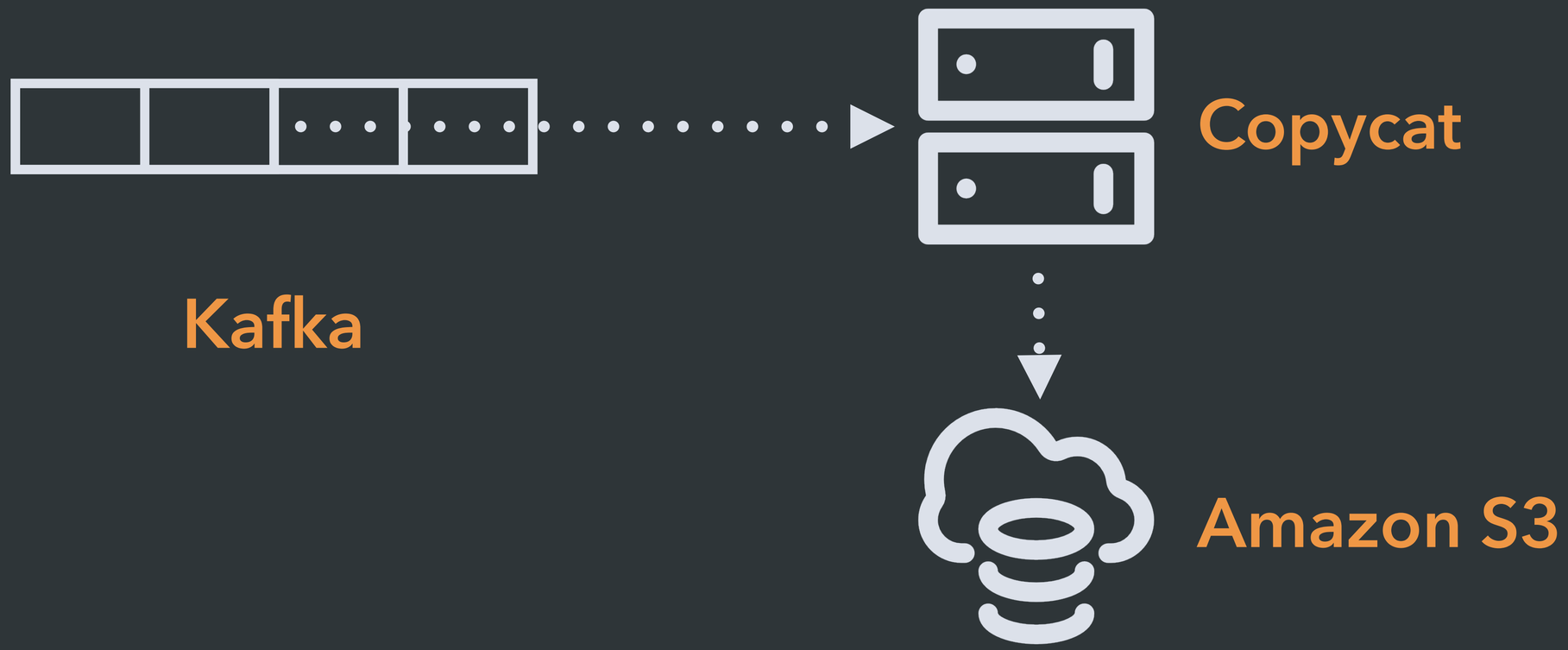


ARCHIVAL: TWILIOFS DATA LAKE

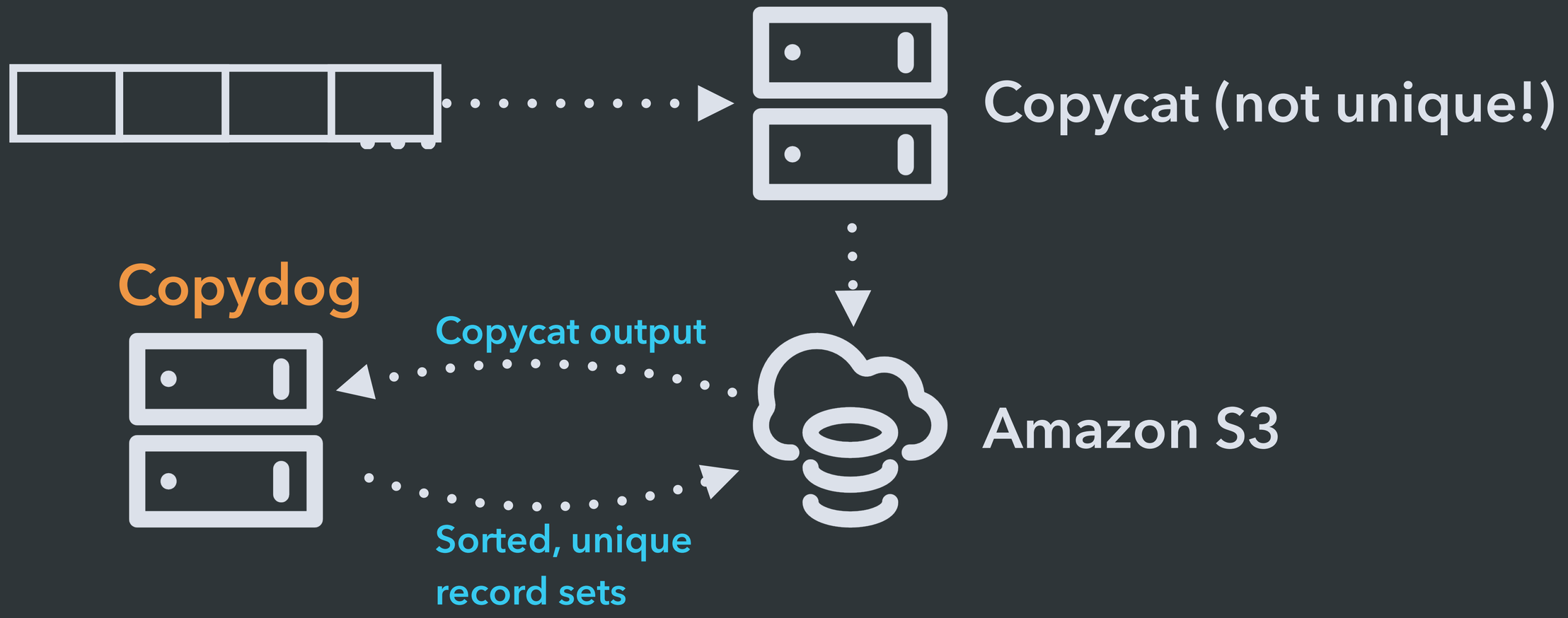



```
Widgets/part3/2017/05/01/01/offsetX.parquet
Widgets/part3/2017/05/01/01/offsetY.parquet
```

ARCHIVAL: TWILIOFS DATA LAKE

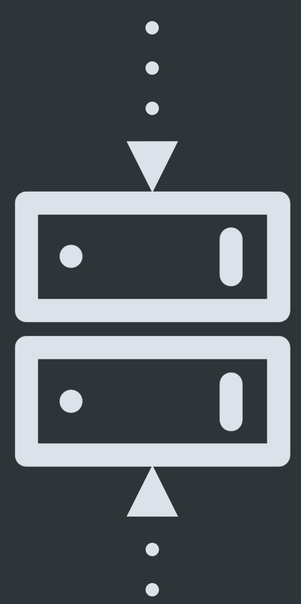


ARCHIVAL: TWILIOFS DATA LAKE



DATA DEDUPLICATION

copycat/Widgets/.../offsetX.parquet

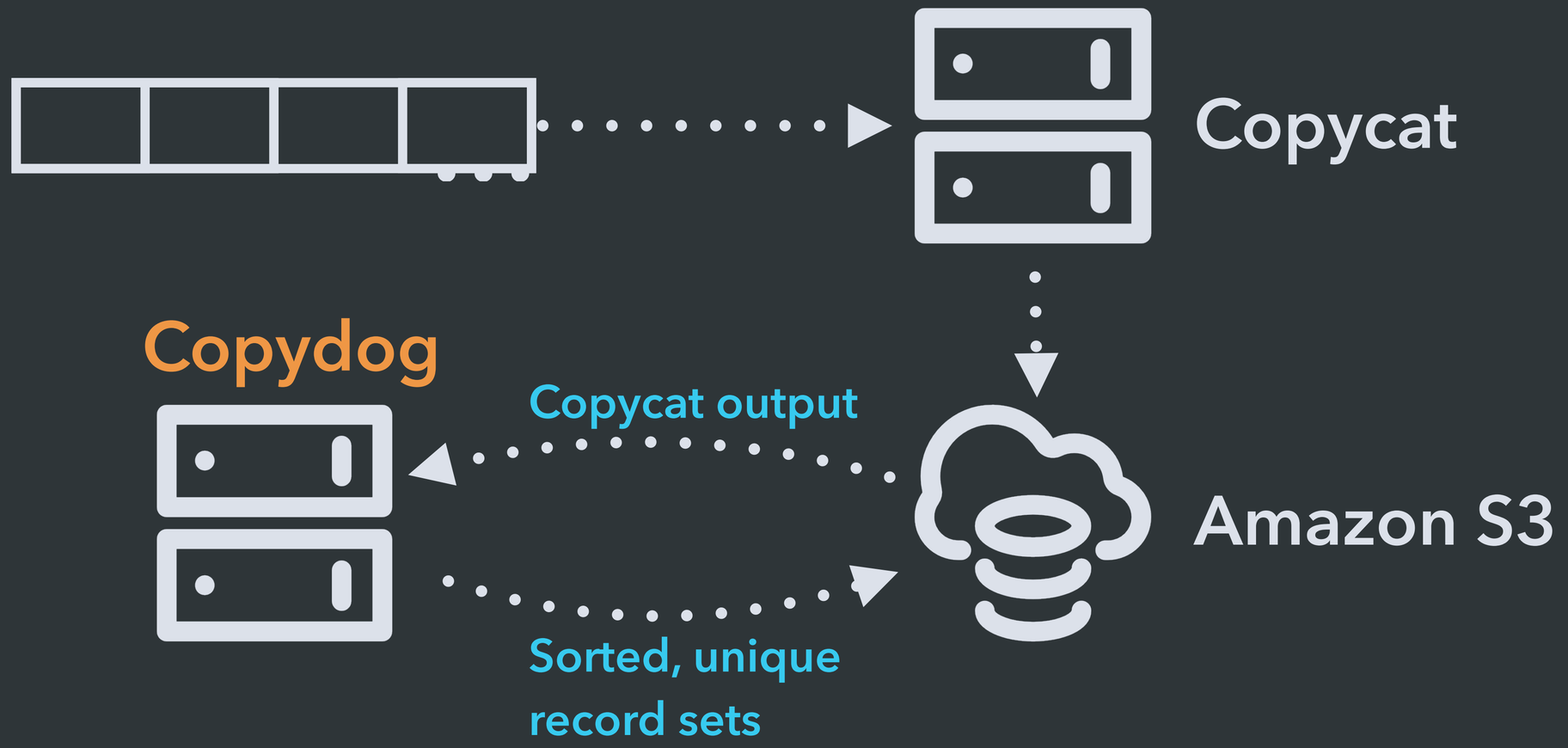


twiliofs/Widgets/2017/05/01/chunkN

twiliofs/Widgets/2017/05/02/chunkM

...

ARCHIVAL: TWILIOFS DATA LAKE



CONSUMER SYSTEMS

Archival

Warehousing and structured analytics

Batch processing

Ad-hoc analysis

Stream processing

Online query systems

DATA MARTS

CONSUMER SYSTEMS

Archival

Warehousing and structured analytics

Batch processing

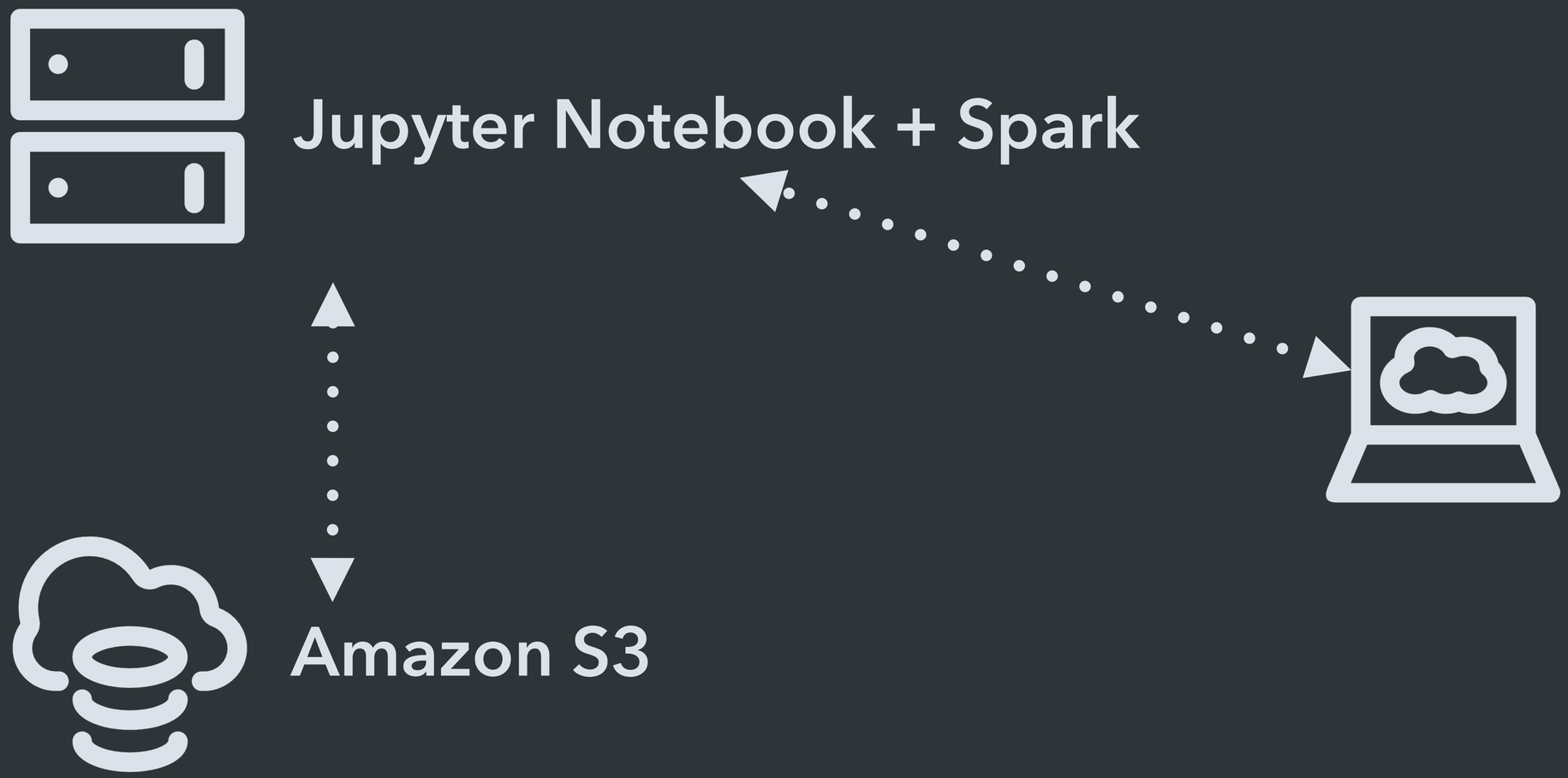
Ad-hoc analysis

Stream processing

Online query systems

BATCH AND AD-HOC PROCESSING

BATCH AND AD-HOC PROCESSING



CONSUMER SYSTEMS

Archival

Warehousing and structured analytics

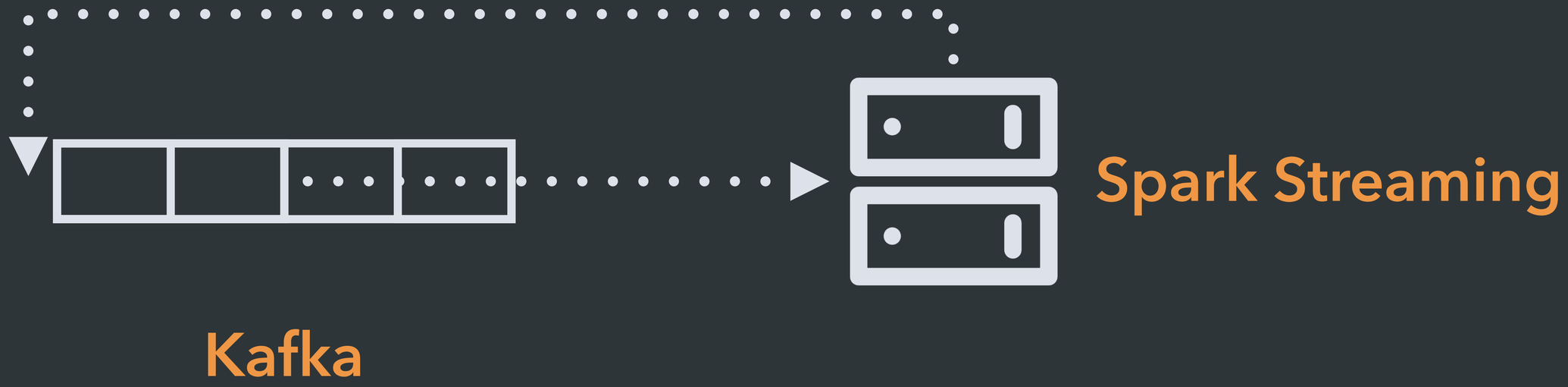
Batch processing

Ad-hoc analysis

Stream processing

Online query systems

STREAM PROCESSING



CONSUMER SYSTEMS

Archival

Warehousing and structured analytics

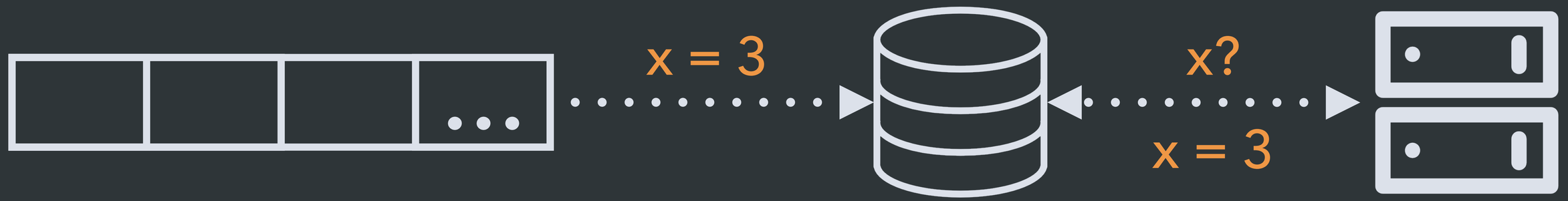
Batch processing

Ad-hoc analysis

Stream processing

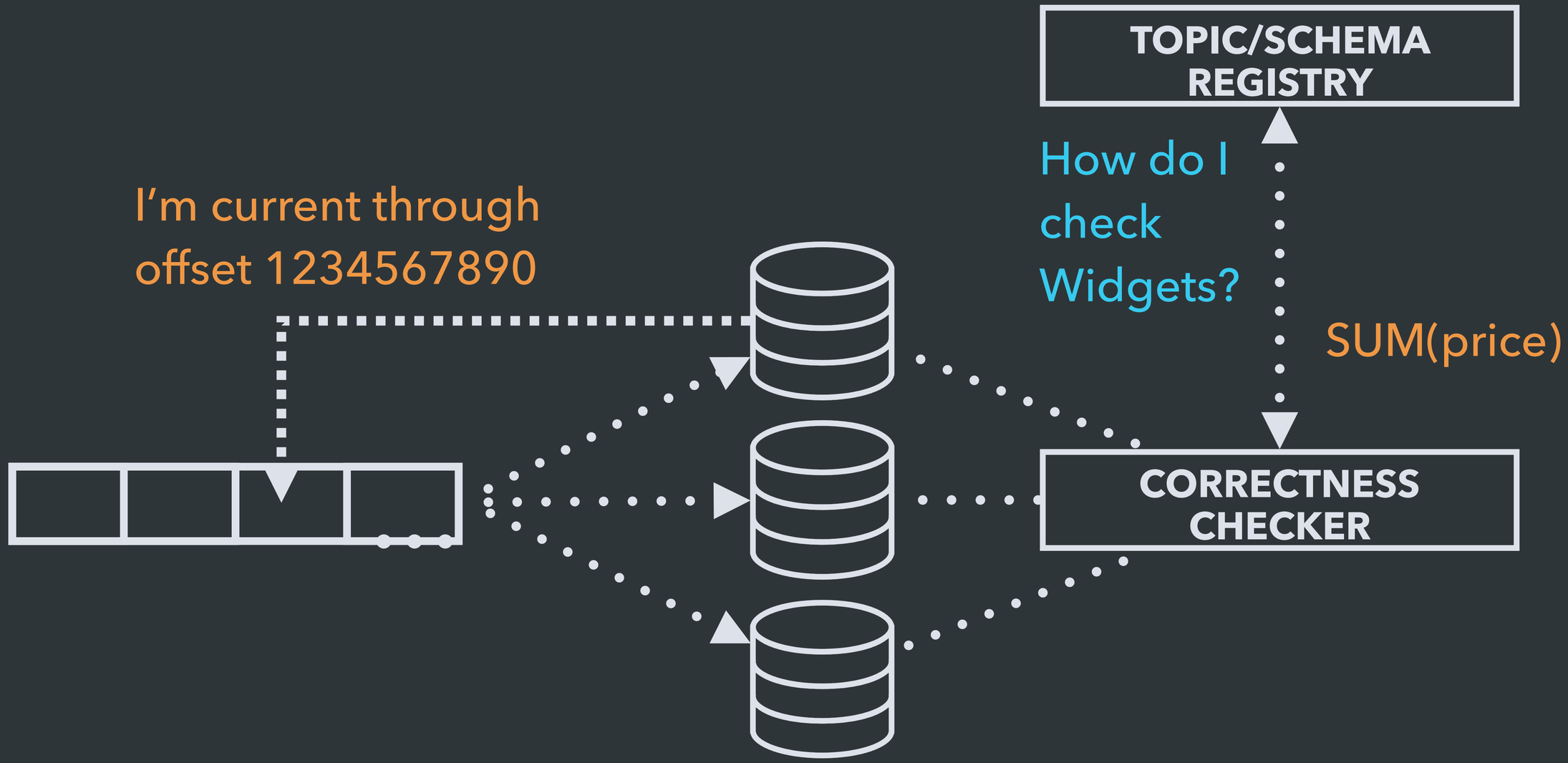
Online query systems

ONLINE STORAGE AND QUERY



VERIFICATION AND CORRECTNESS

MONITORING


TOPIC/SCHEMA REGISTRY
How do I check Widgets?
SUM(price)
I'm current through offset 1234567890
CORRECTNESS CHECKER

RECAP

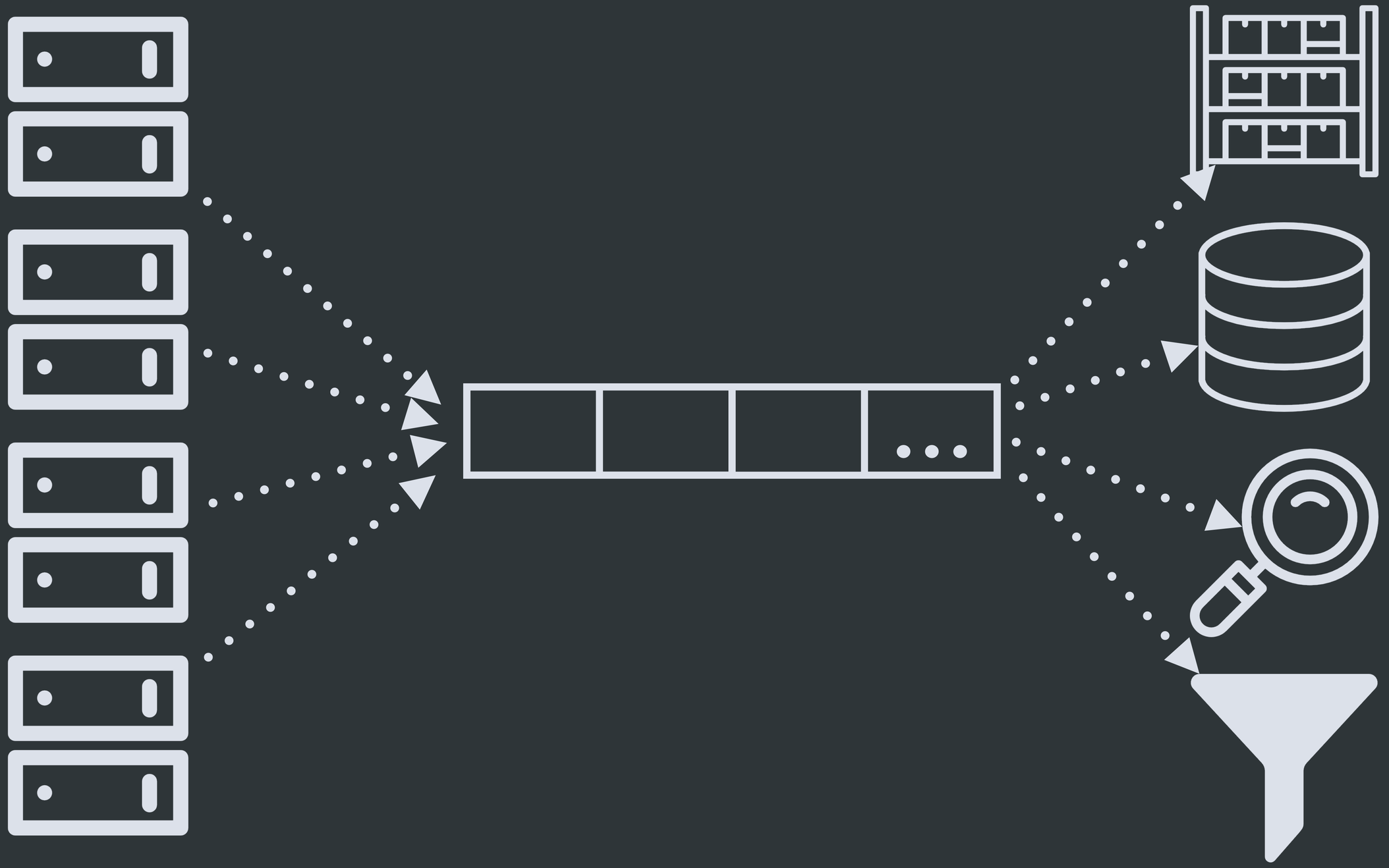
RECAP

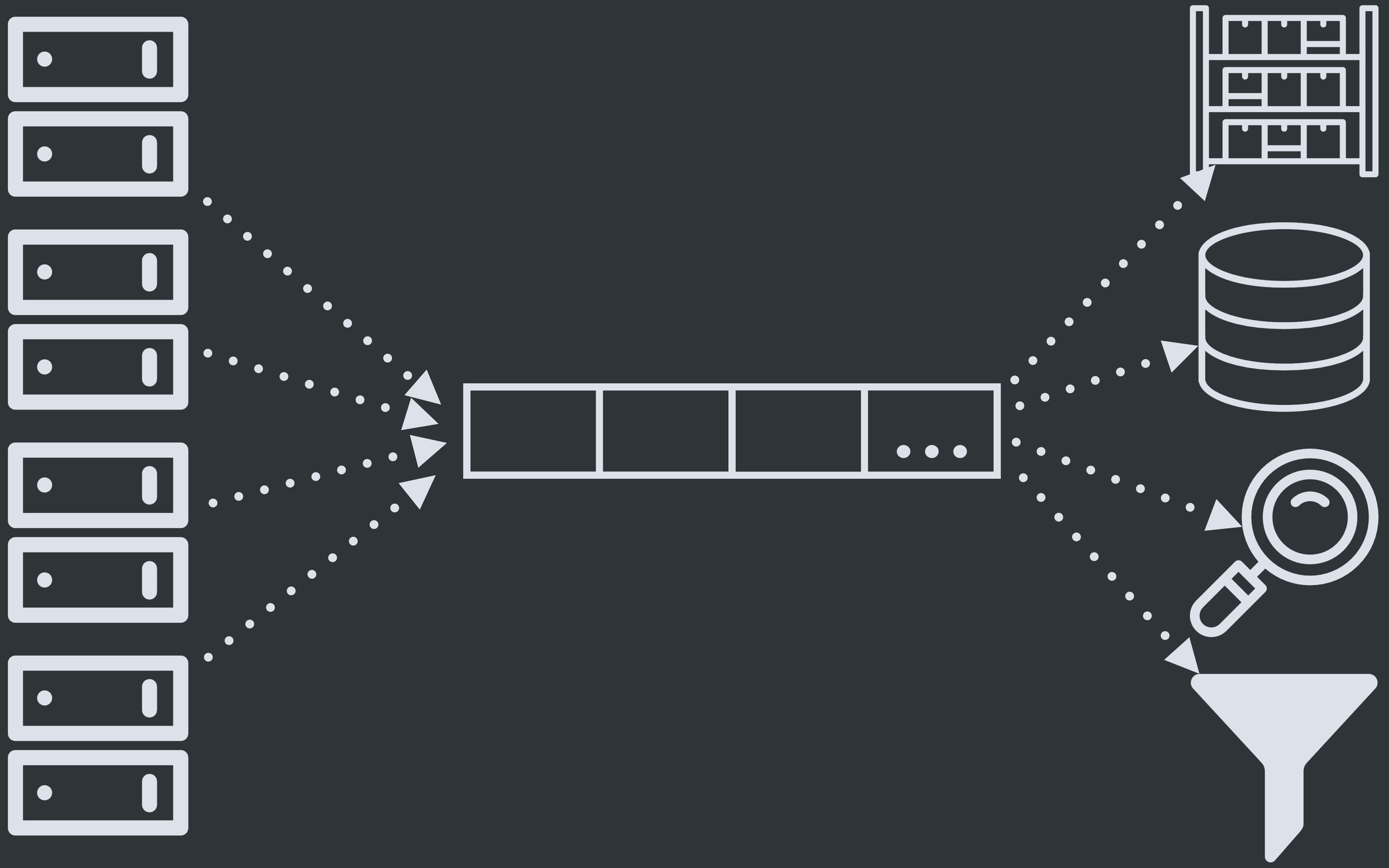
Event-oriented data pipeline

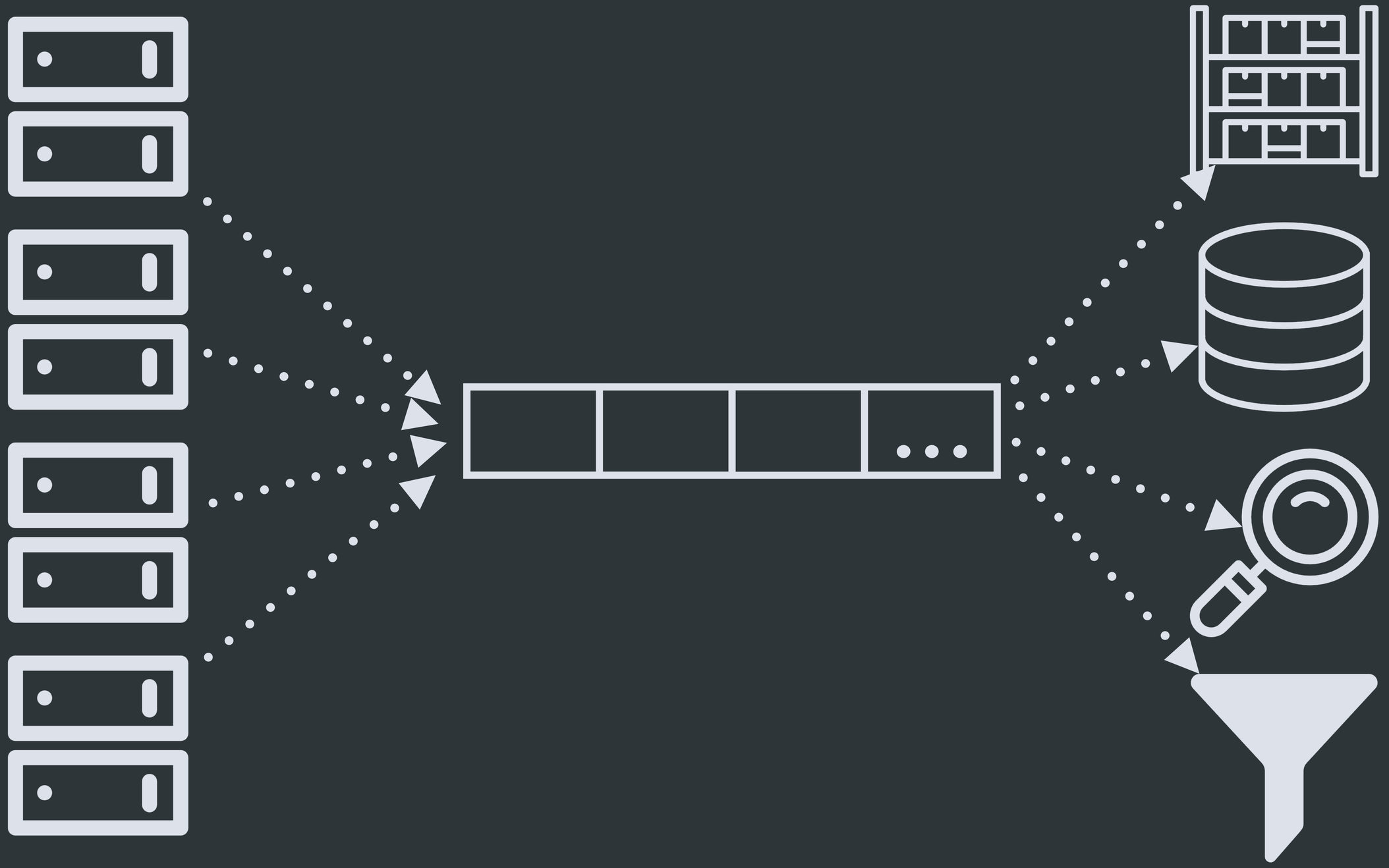
Common producer and consumer libraries

Strong schema validation and planned migrations

Verifiable delivery and correctness







@SKIMBREL // SAM@TWILIO.COM

THANKS!

ATTRIBUTIONS

http://www.flaticon.com/authors/madebyoliver

http://www.flaticon.com/authors/freepik

http://www.flaticon.com/authors/vectors-market

Heat pipes: Bill Ebbesen on Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Heatpipe_tunnel_copenhagen_2009.jpg)