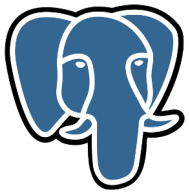
101: Introduction to Change Data Capture (CDC) patterns

Franesco Tisiot

Why do we need Change Data Capture?

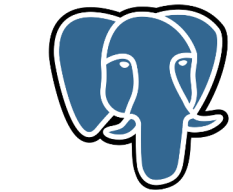
The Data’s Journey

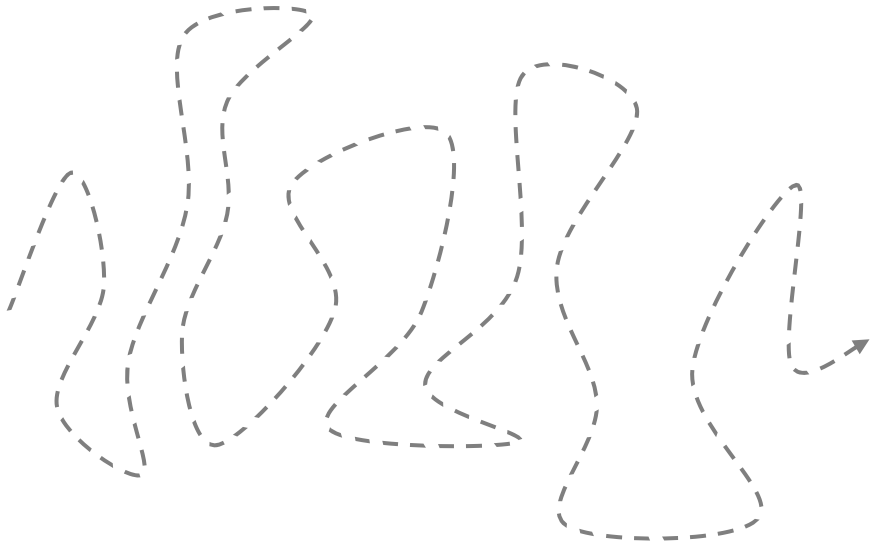
Where data is stored



The Data's Journey

Where data is stored

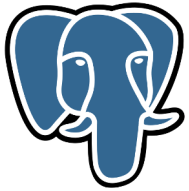




Where data is accessed/analyzed

The Data's Journey

Where data is stored



Where data is accessed/analyzed

Change Data Capture

“The process of identifying and capturing changes made to data in a database and then delivering those changes in real-time to a downstream process or system.”

Use cases

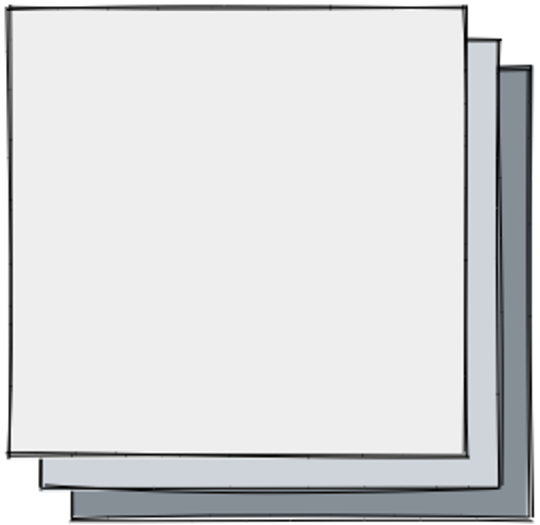
- Notifications
- Caching
- Analytical databases
- Technology Migration
- Event driven architecture

How can we implement Change Data Capture?

Detect Changes

Query based CDC

Log based CDC

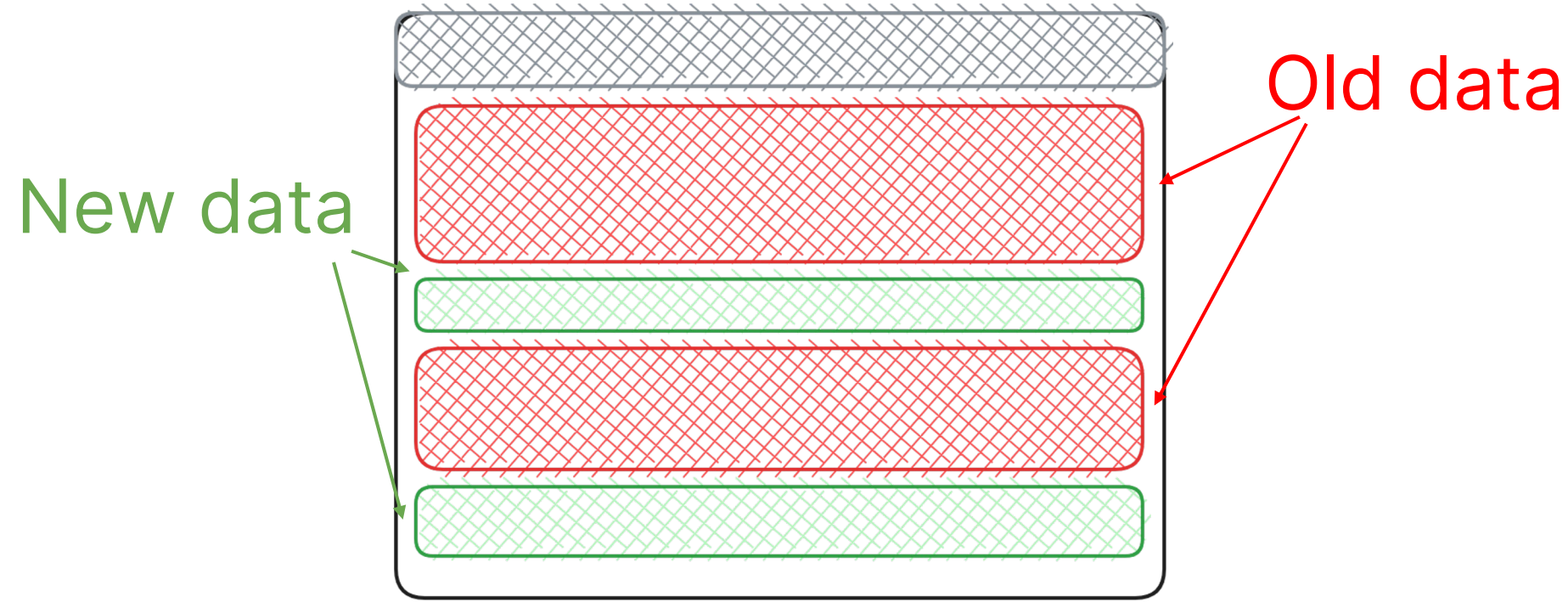


Query based CDC

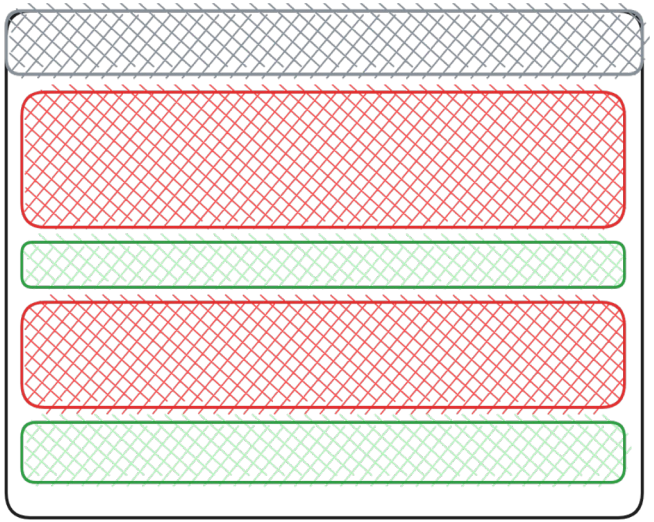
Query based CDC

- Query the Database table(s)
- Extract the results
- Propagate

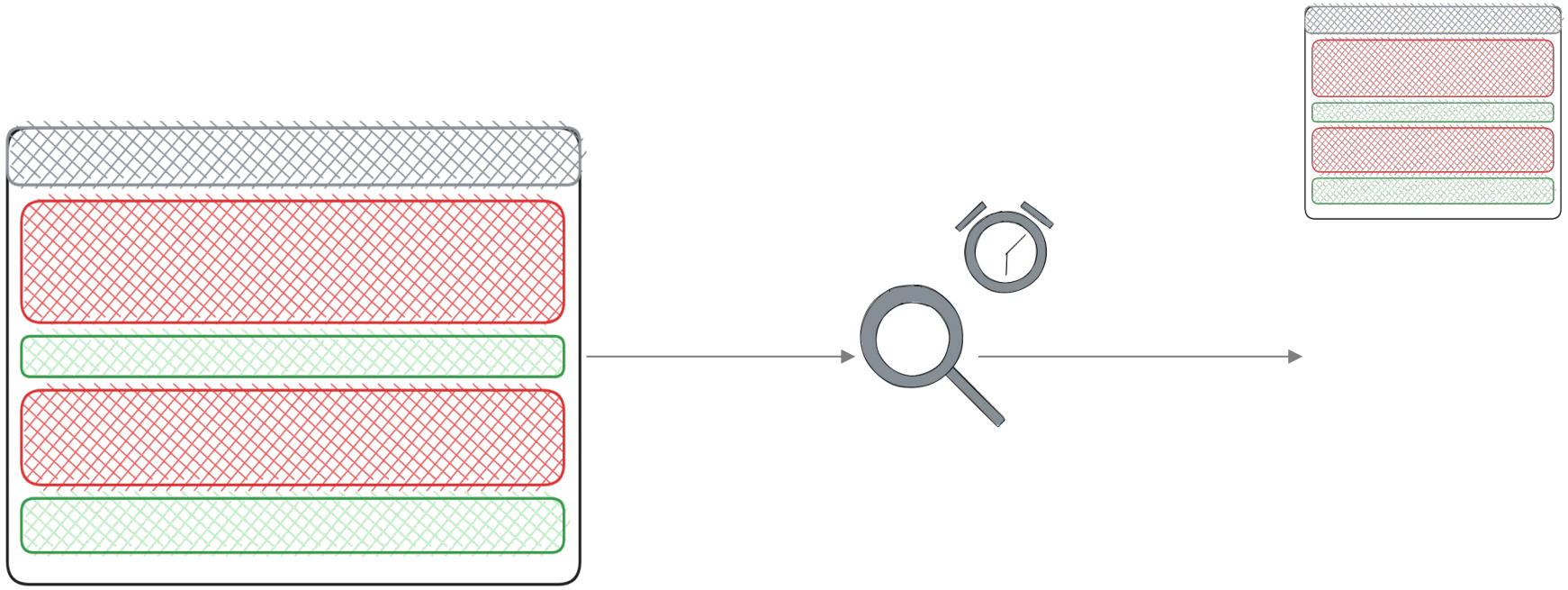
Table structure



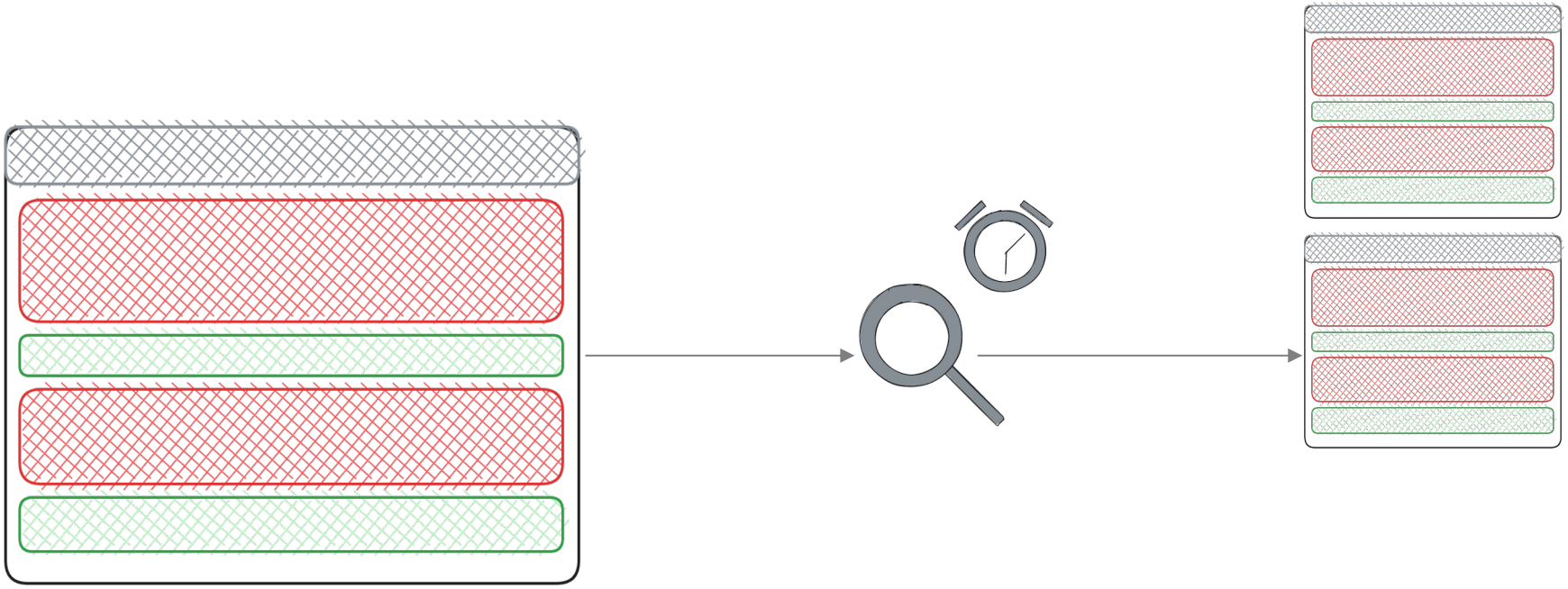
Bulk Approach



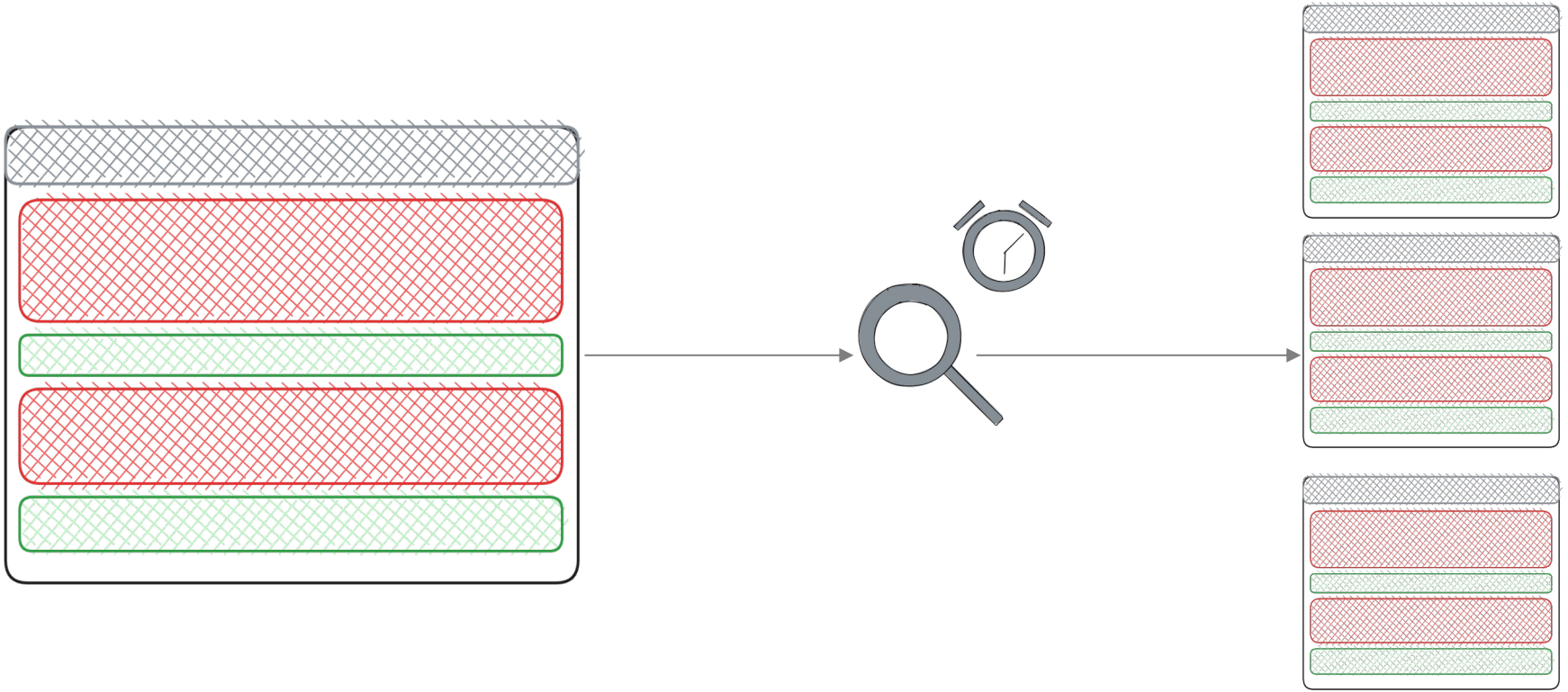
Bulk Approach



Bulk Approach



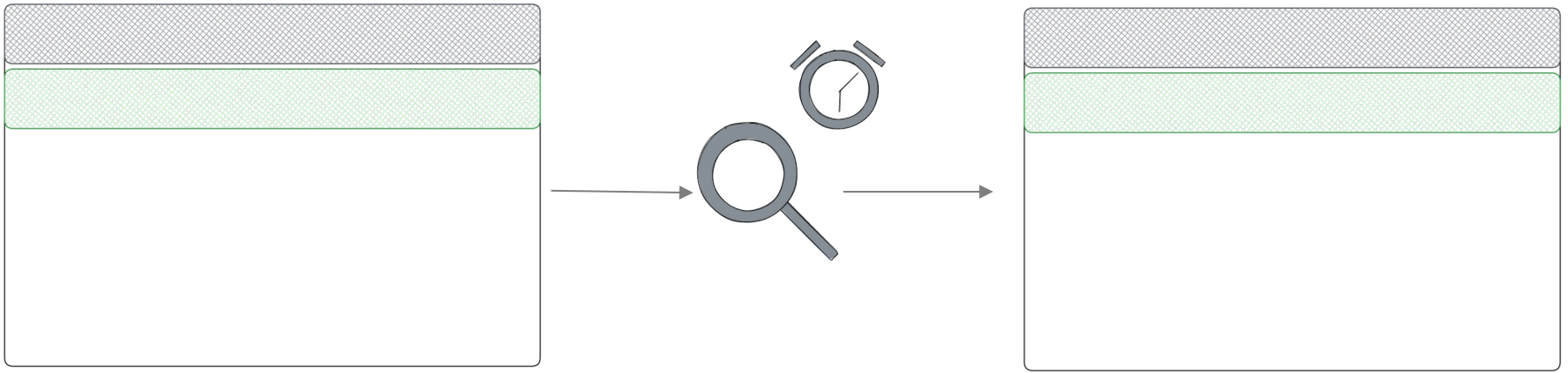
Bulk Approach



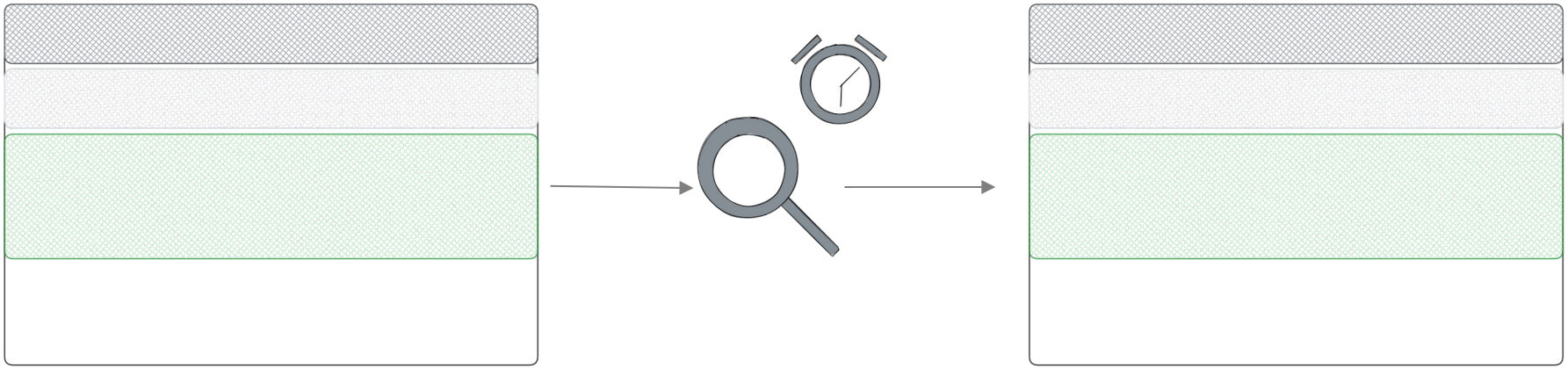
Bulk approach

- Retrieve **all data** every polling period

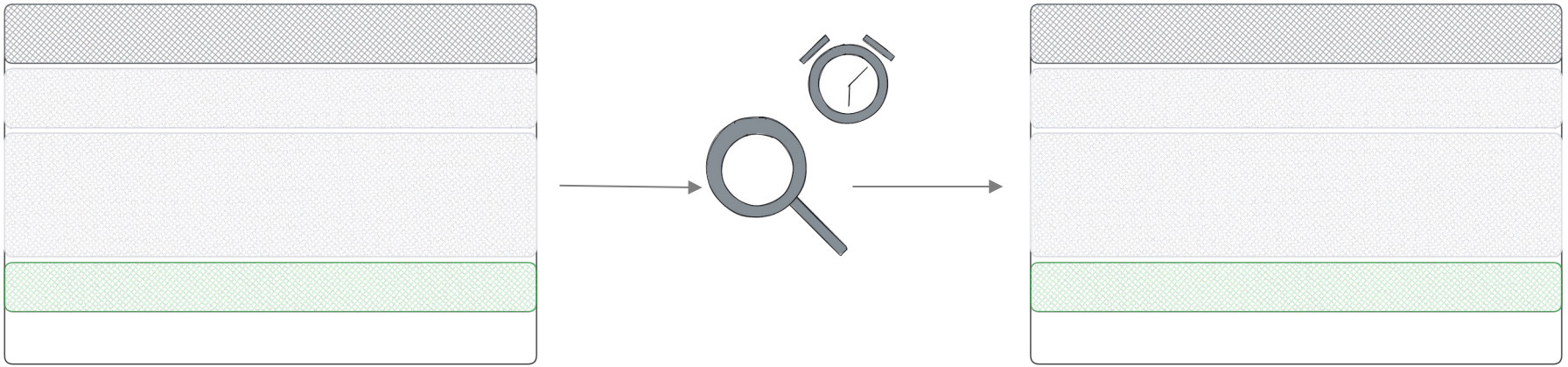
Incremental/Timestamp Approach



Incremental/Timestamp Approach



Incremental/Timestamp Approach



Incremental/Timestamp approach

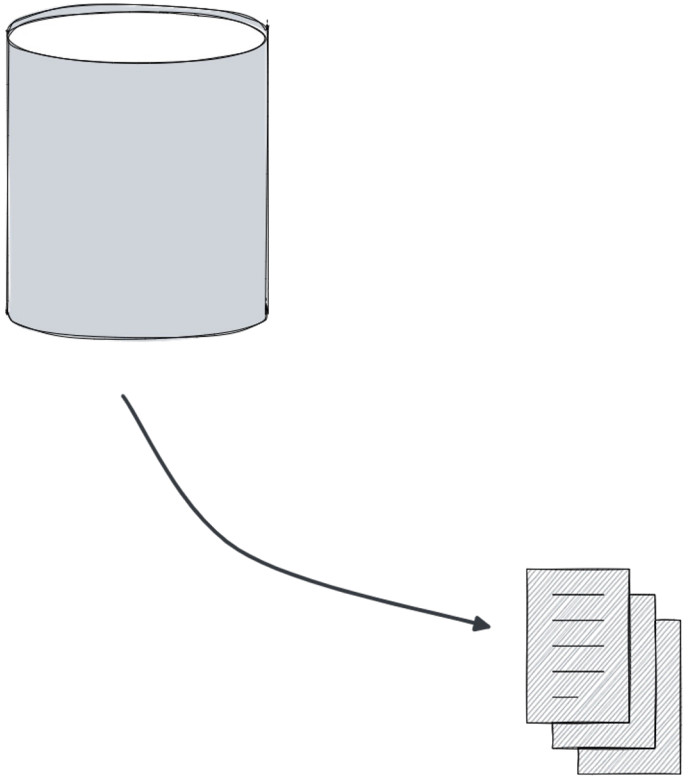
- Replicates only new/updated data
- Requires incremental/timestamp column

Incremental/Timestamp approach

- Batch approach - no data streaming
- Problems with:
  - Fast updates
  - Deletes
- Requires changes in application logic

Log based CDC

Log based CDC



- WAL log - PostgreSQL
- Binlog - MySQL
- Replica set - MongoDB
- Redo Log - Oracle

Debezium

debezium

open source
distributed platform
for change data capture

Debezium Format

id	name	cooking minutes
1	spaghetti	7

Debezium Format

id	name	cooking minutes
1	spaghetti	8

Debezium Format

id	name	cooking minutes
1	spaghetti	8


```
{
  "before": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": "int": 7}},
  "after": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": 8}},
  "source": {...},
  "op": "u",
  "ts_ms": {"long": 1639385467890}
}
```

Debezium Format

id	name	cooking minutes
1	spaghetti	8


```
{
  "before": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": "int": 7}},
  "after": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": 8}},
  "source": {...},
  "op": "u",
  "ts_ms": {"long": 1639385467890}
}
```

Debezium Format

id	name	cooking minutes
1	spaghetti	8


```
{
  "before": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": "int": 7}},
  "after": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": 8}},
  "source": {...},
  "op": "u",
  "ts_ms": {"long": 1639385467890}
}
```

Debezium Format

id	name	cooking minutes
1	spaghetti	8


```
{
  "before": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": "int": 7}},
  "after": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": 8}},
  "source": {...},
  "op": "u",
  "ts_ms": {"long": 1639385467890}
}
```

Debezium Format

id	name	cooking minutes
1	spaghetti	8


```
{
  "before": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": "int": 7}},
  "after": {"Value": {"id": 1,"name": "spaghetti","cooking_minutes": 8}},
  "source": {...},
  "op": "u",
  "ts_ms": {"long": 1639385467890}
}
```

How to create a CDC data pipeline

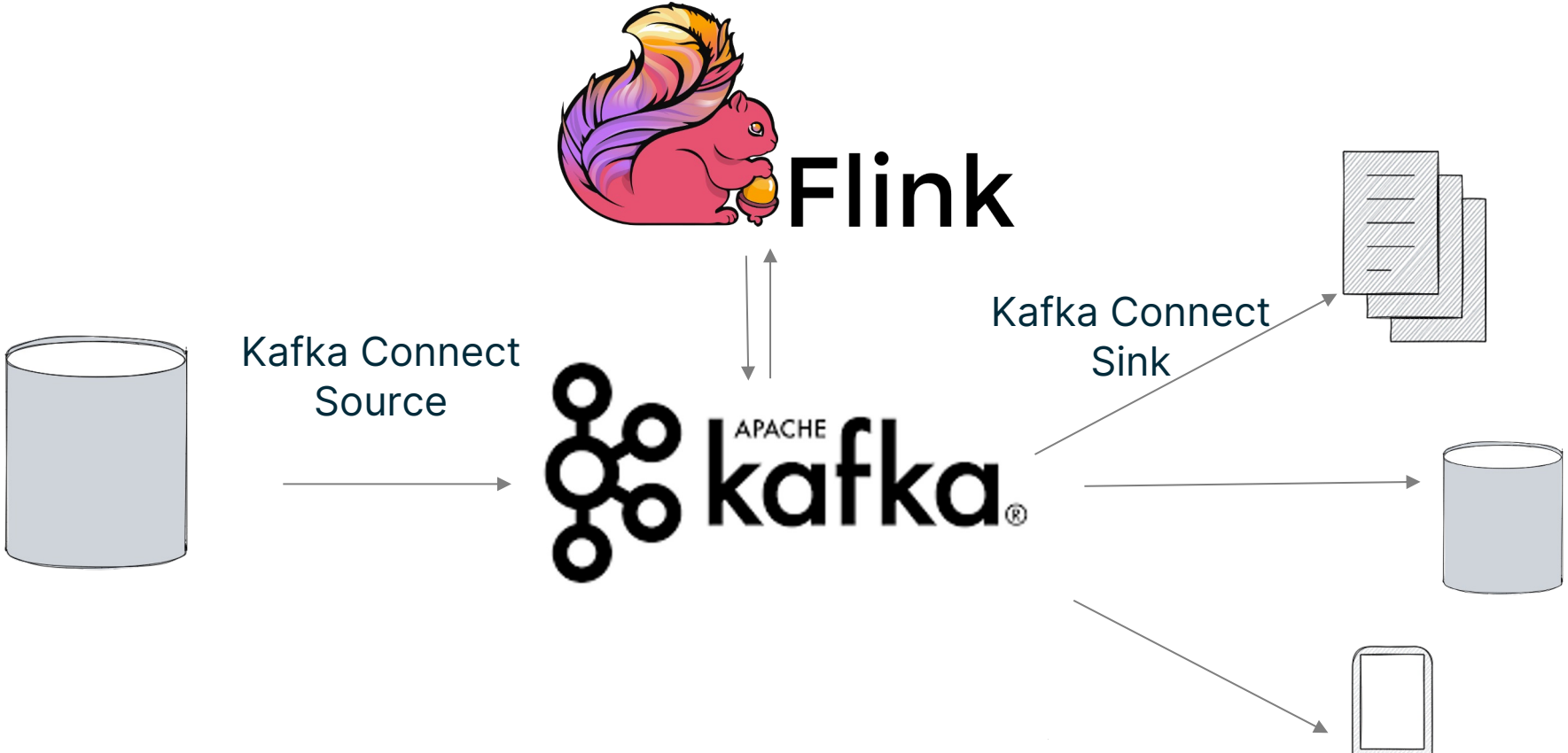
CDC data pipeline

Apache Kafka

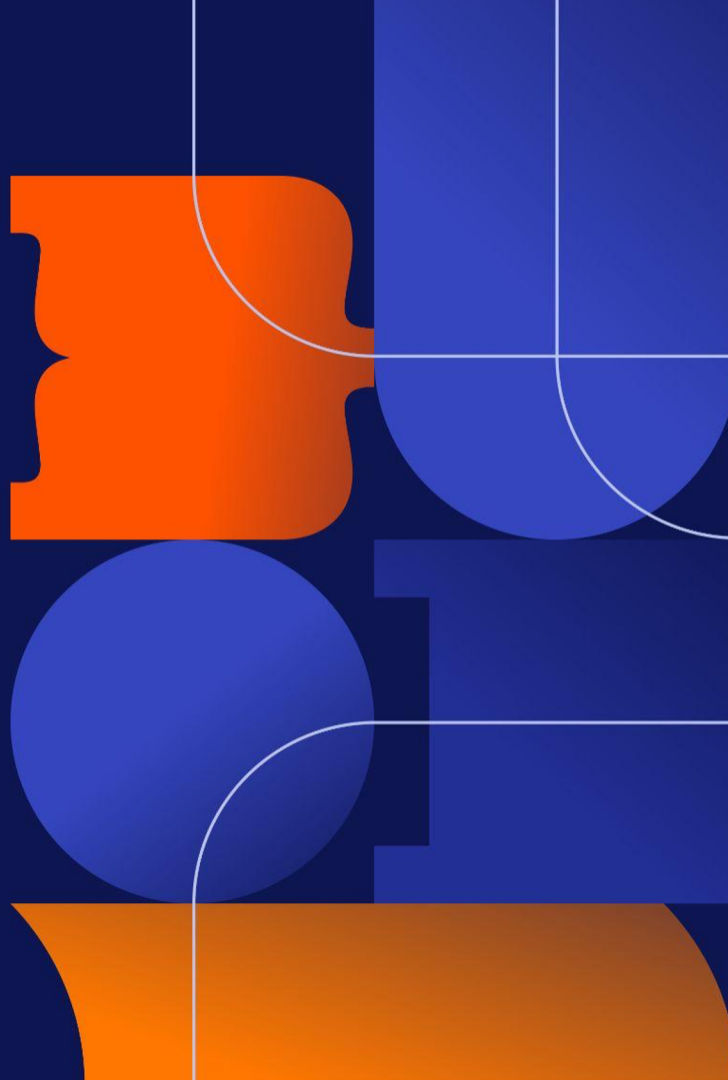
open-source
distributed
event streaming
platform

CDC Data pipeline

Evolved CDC Data pipeline



The trusted open source data platform for everyone



One data platform for your cloud needs

Event streaming	Event stream processing	Relational databases	Key-value database	Wide column database	Data warehouse	Time series database	Search engine	Data visualization
Aiven for **Apache Kafka®** and **Kafka® Connect**	Aiven for **Apache Flink®**	Aiven for **PostgreSQL®** Aiven for **MySQL**	Aiven for **Redis®**	Aiven for **Apache Cassandra®**	Aiven for **ClickHouse®**	Aiven for **M3**	Aiven for **OpenSearch®**	Aiven for **Grafana®**

STREAM **STORE** **ANALYZE**

Host

aws
Google Cloud
DigitalOcean
Microsoft Azure
Bring your own cloud

Deploy

Terraform
Kubernetes
API REST API
CLI Aiven CLI
Aiven Console

Integrate

Datadog
Prometheus
AWS CloudWatch
GCP Monitoring
MongoDB
AWS S3
GCP BigQuery
Couchbase
Snowflake
Splunk
Sumologic
Debezium
GCP Pub/Sub
GCP Storage

Customers


okta


snyk


DOORDASH


priceline


fiverr.


Norauto


DECATHLON


GTL


ACTIVISION
BLIZZARD


MIRAKL


GOV.UK


goto financial


spare


Schibsted


TOYOTA


paf


CONRAD


adeo


ometria


WÄRTSILÄ

Try it out!

https://go.aiven.io/aiven-live-cdc-signup

300$ - 1 Month!