

Big Data Integration Patterns

Michael Häusler

Jun 12, 2017

ResearchGate is built for scientists.

The social network gives scientists new tools to connect, collaborate, and keep up with the research that matters most to them.



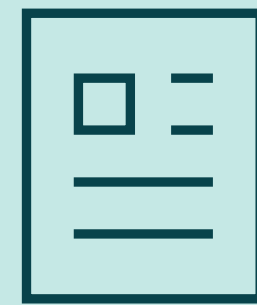


Our mission is to connect the world of science
and *make research open to all.*



12+ million

Members



100+ million

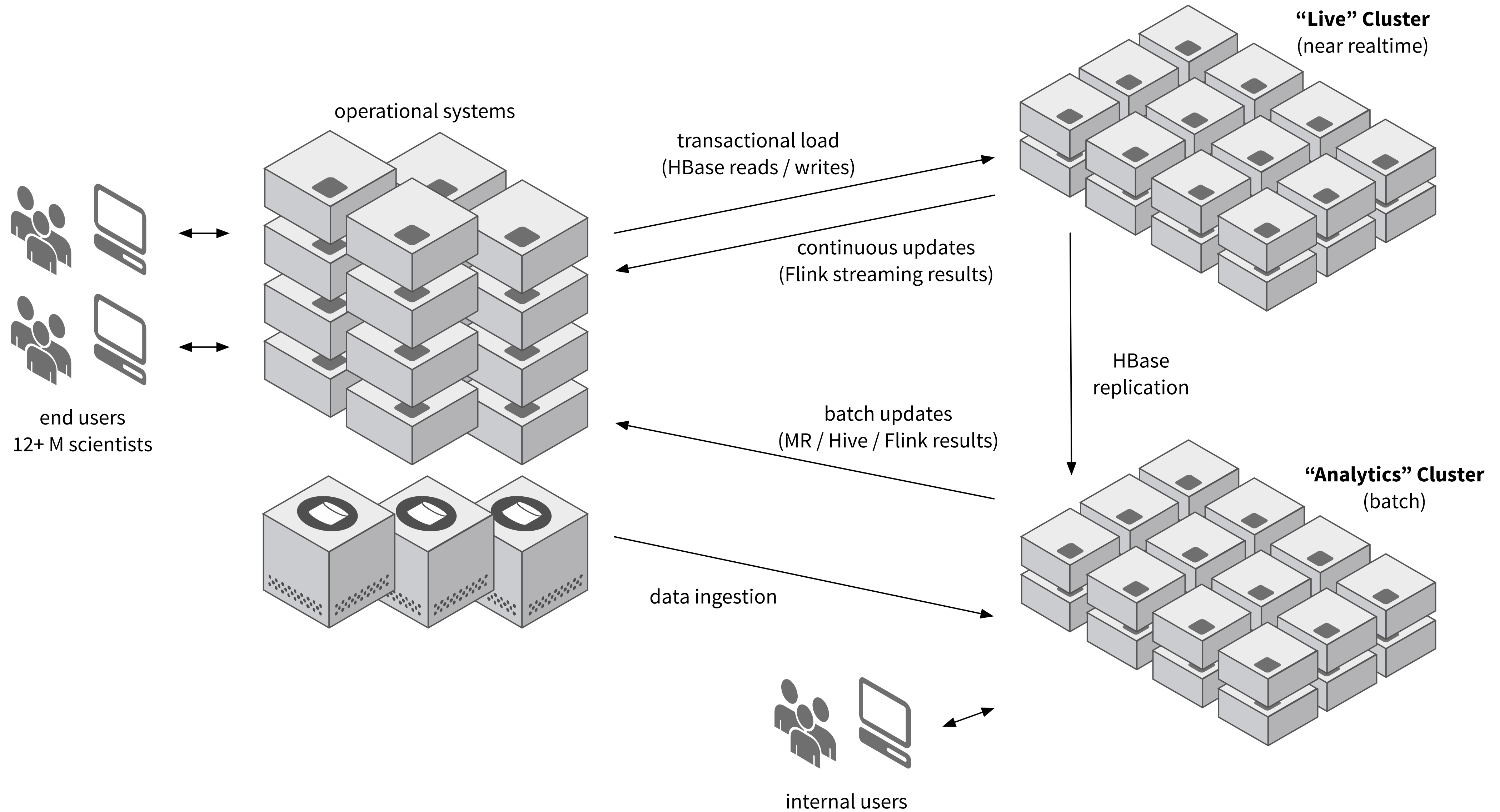
Publications



1,500+ million

Citations

Big Data



Big Data



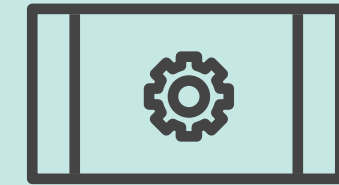
65+

Engineers



370+

Data Ingestion Jobs per Day



3,000+

Yarn Applications per Day



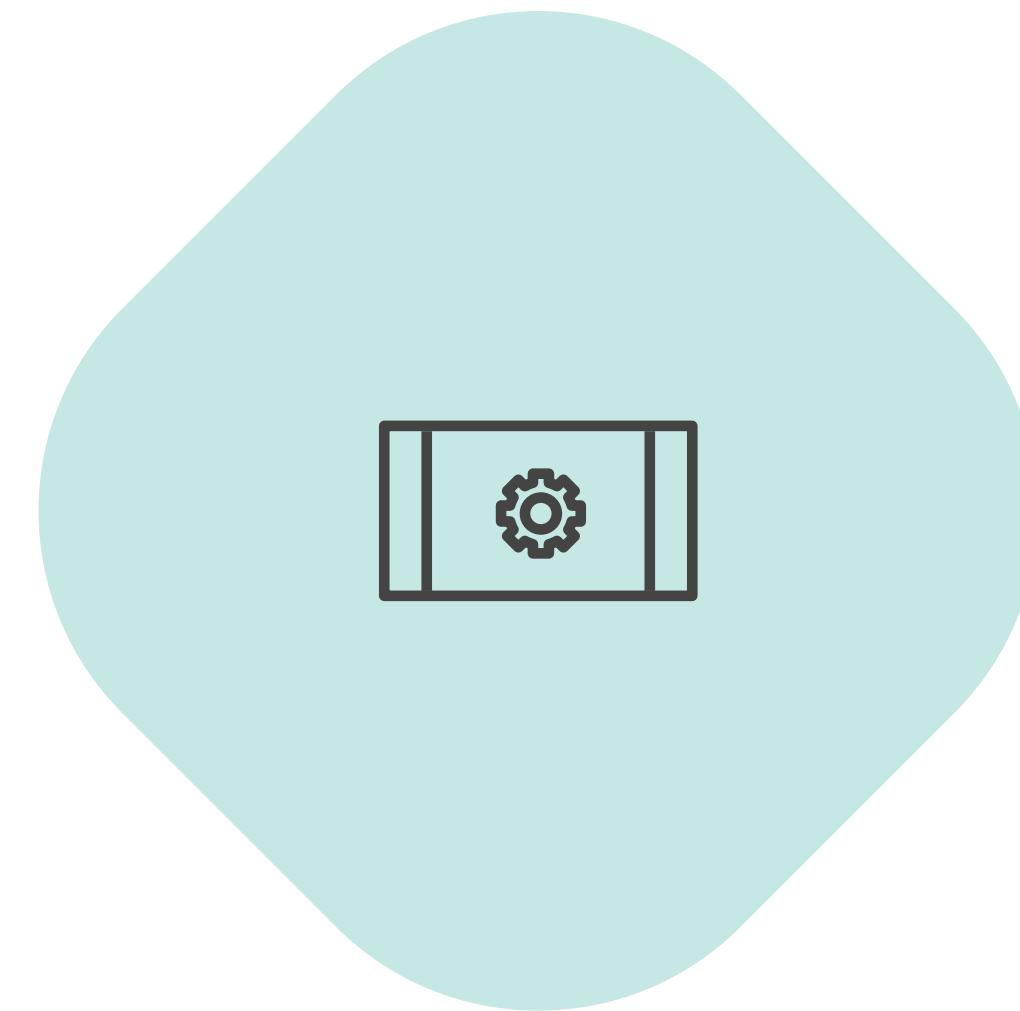
65+

Engineers

Developer Productivity

Ease of Maintenance

Ease of Operations



3,000+

Yarn Applications per Day

Big Data Architecture

Integration Patterns & Principles

Patterns & Principles

Integration patterns should be strategic, but also ...

should be driven by **use cases**

should tackle real world **pain points**

should **not** be dictated by a **single technology**

Patterns & Principles

Big data is still a fast moving space

Big data **batch processing** today is quite different compared to 5 years ago



Big data **stream processing** is evolving heavily right now

Big data architecture

must evolve over time

First Big Data Use Case
Early 2011, Author Analysis

Author Analysis – Clustering and Disambiguation

	Article: Evidence of Molecular Evolution Driven by Recombination Events Influencing Tropism in a Novel Human Adenovirus That Causes Epidemic Keratoconjunctivitis
Source	Michael P Walsh · Ashish Chintakuntlawar · Christopher M Robinson · Ijad Madisch · Balázs Harrach · Nolan R Hudson · David Schnur · Albert Heim · James Chodosh · Donald Seto · Morris S Jones
1k Reads	95 Citations
	Article: Unique sequence features of the Human Adenovirus 31 complete genomic sequence conserved in clinical isolates
Source	Soeren Hofmayer · Ijad Madisch · Sebastian Darr · Fabienne Rehren · Albert Heim
5k Reads	4 Citations

Author Analysis – High Product Impact

The screenshot shows the ResearchGate interface for an article. At the top, there is a navigation bar with the ResearchGate logo (R^G) and links for HOME, PUBLICATIONS, QUESTIONS, and JOBS. A search bar is located to the right of these links. Below the navigation bar, the article's statistics are displayed: 4 CITATIONS, 79 REFERENCES, and 5 FIGURES. The article title is prominently displayed in the center. Below the title, there is a section for the article's details, including a 'FULL TEXT' button, the journal name (BMC GENOMICS), issue information (10:557), date (NOVEMBER 2009), and the number of reads (4947). The authors are listed in a grid format, each with a profile picture, name, and ResearchGate score.

R^G HOME PUBLICATIONS QUESTIONS JOBS

4 CITATIONS **79 REFERENCES** **5 FIGURES**

Unique sequence features of the Human Adenovirus 31 complete genomic sequence conserved in clinical isolates

ARTICLE in BMC GENOMICS 10:557 · NOVEMBER 2009 *with* 4947 READS
Institut für Virologie, Medizinische Hochschule Hannover, Hannover, Germany.
Impact Factor: 3.99 · DOI: 10.1186/1471-2164-10-557 · Source: PubMed

FULL TEXT

1st Sören Hofmayer
| 14.31 · ResearchGate

2nd Ijad Madisch
| 41.88 · Massachusetts General Hospital

3rd Sebastian Darr
| 27.29 · Hannover Medical School

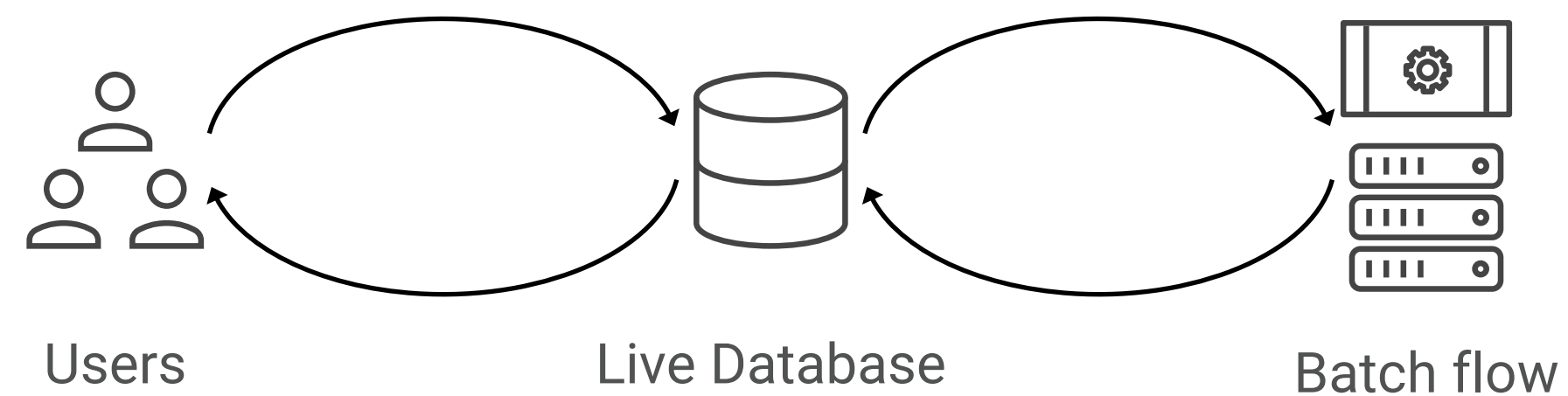
4th Fabienne Rehren
Institut für Virologie, Medizinische Hochsch...

5th Albert Heim
| 40.24 · Hannover Medical School

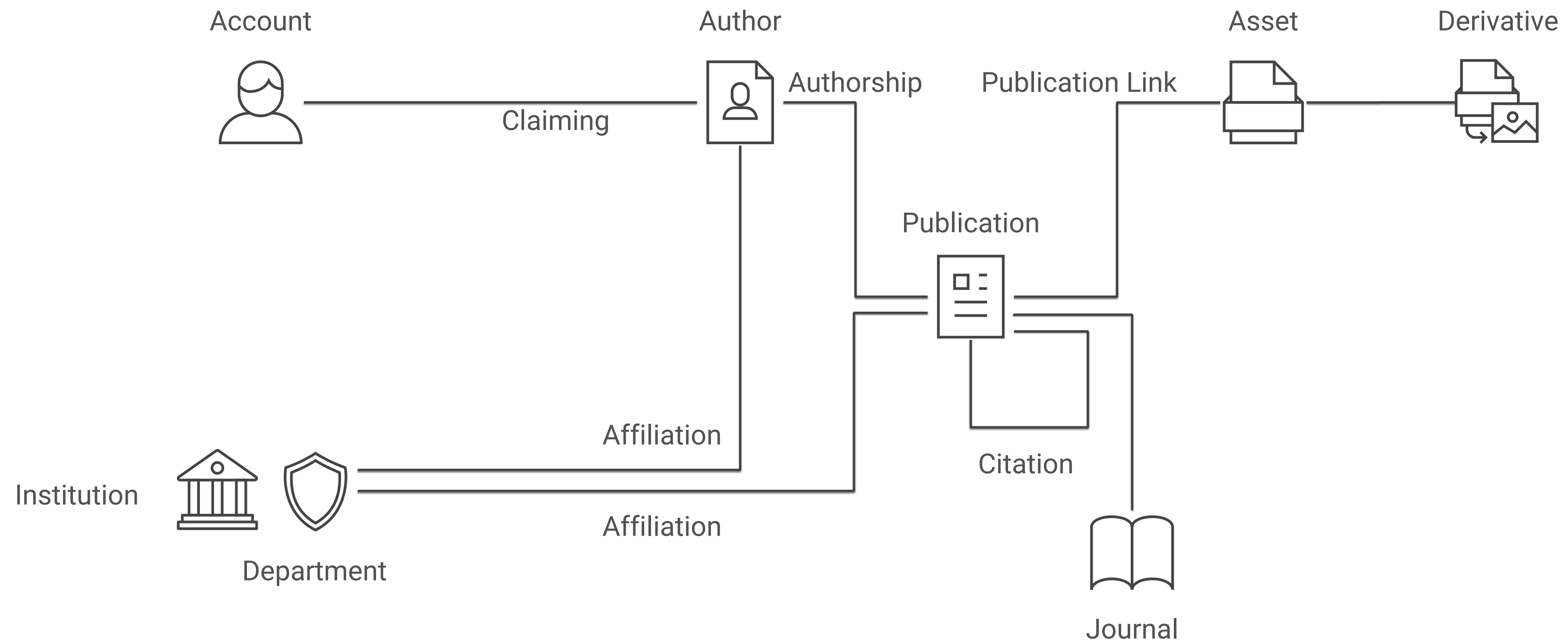
Enriching User Generated Content

Users and **batch flows** continuously enrich an evolving dataset

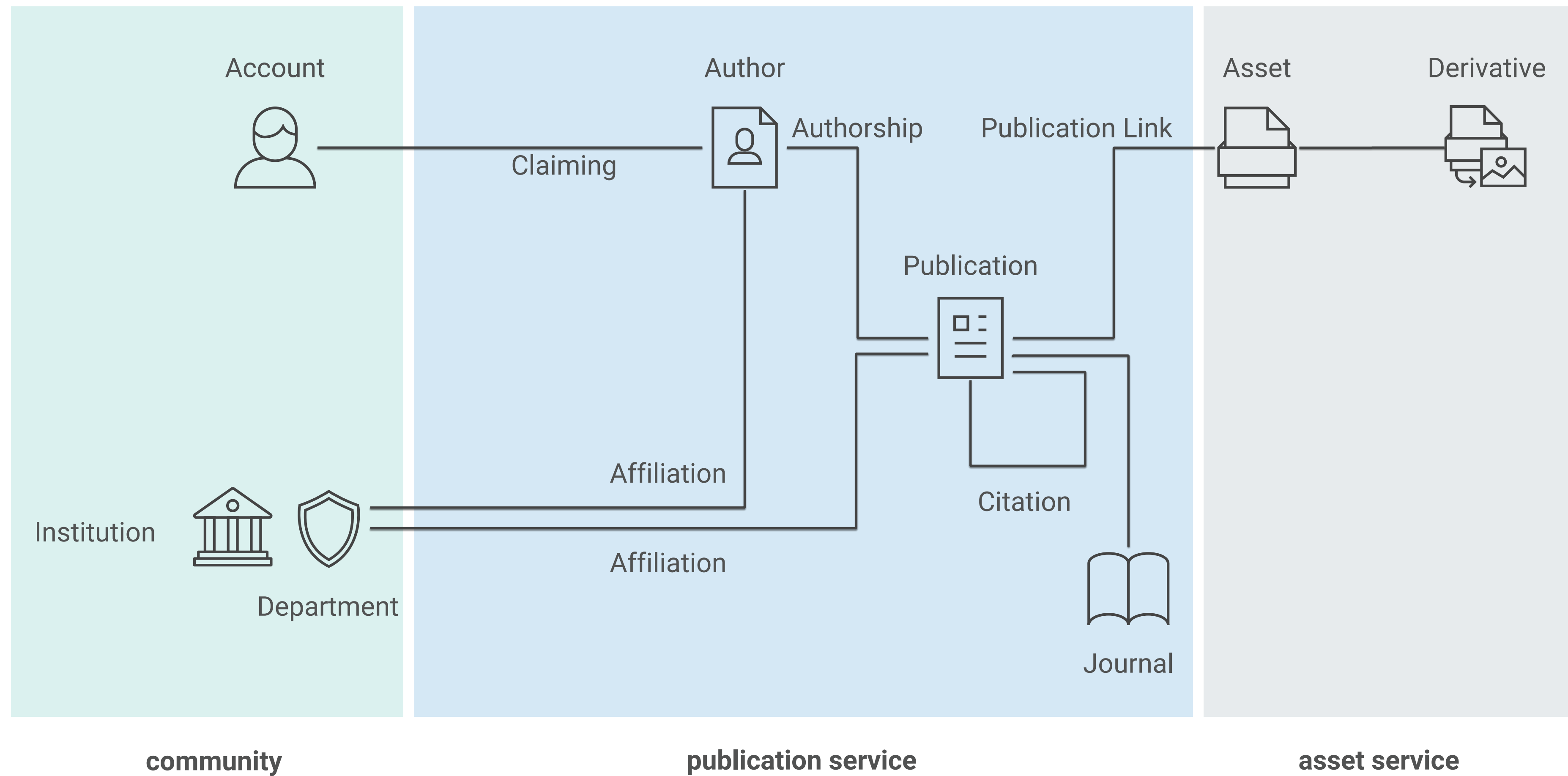
Both user actions and batch flow results **ultimately affect the same live database**



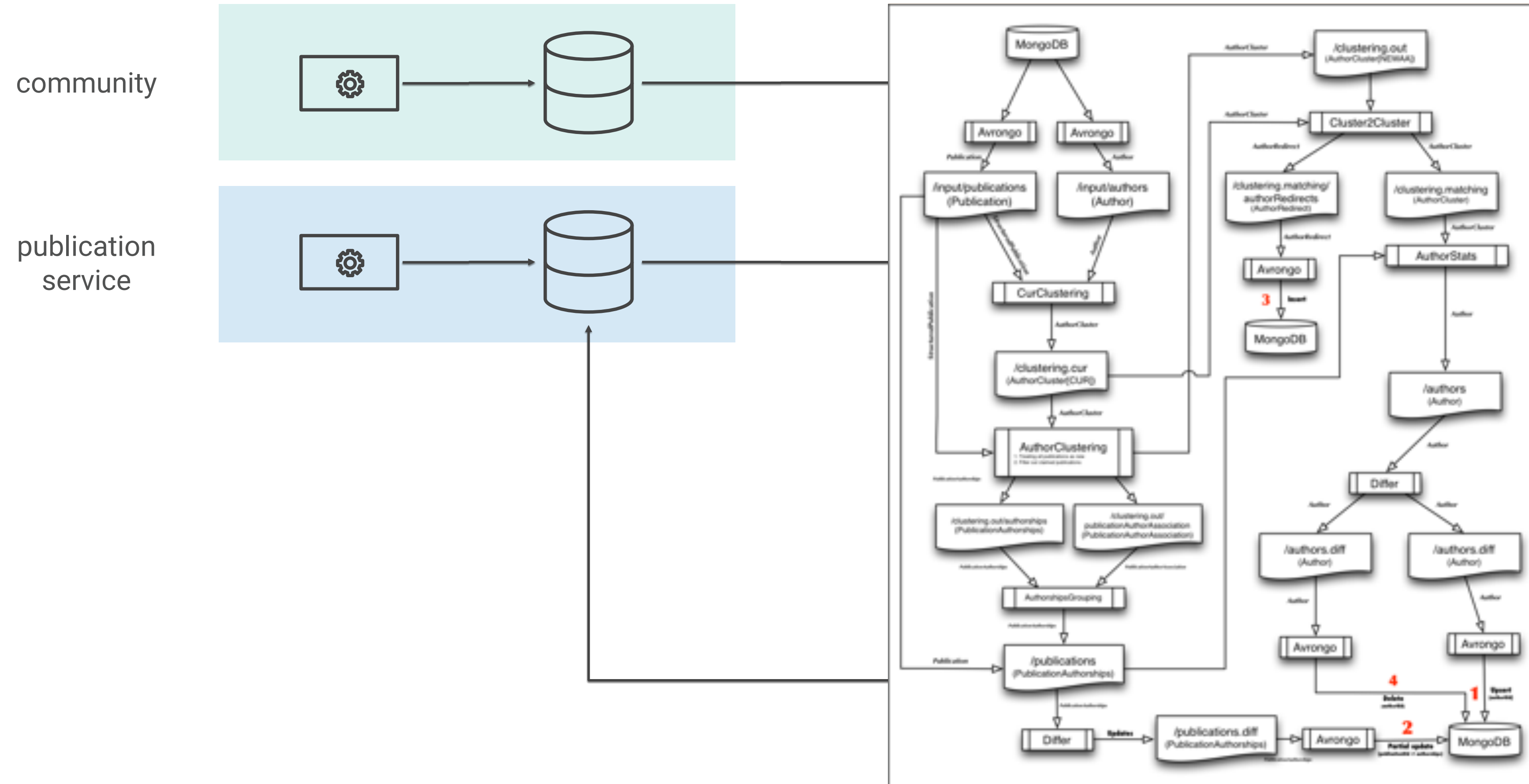
Bibliographic Metadata – Data Model



Bibliographic Metadata – Services

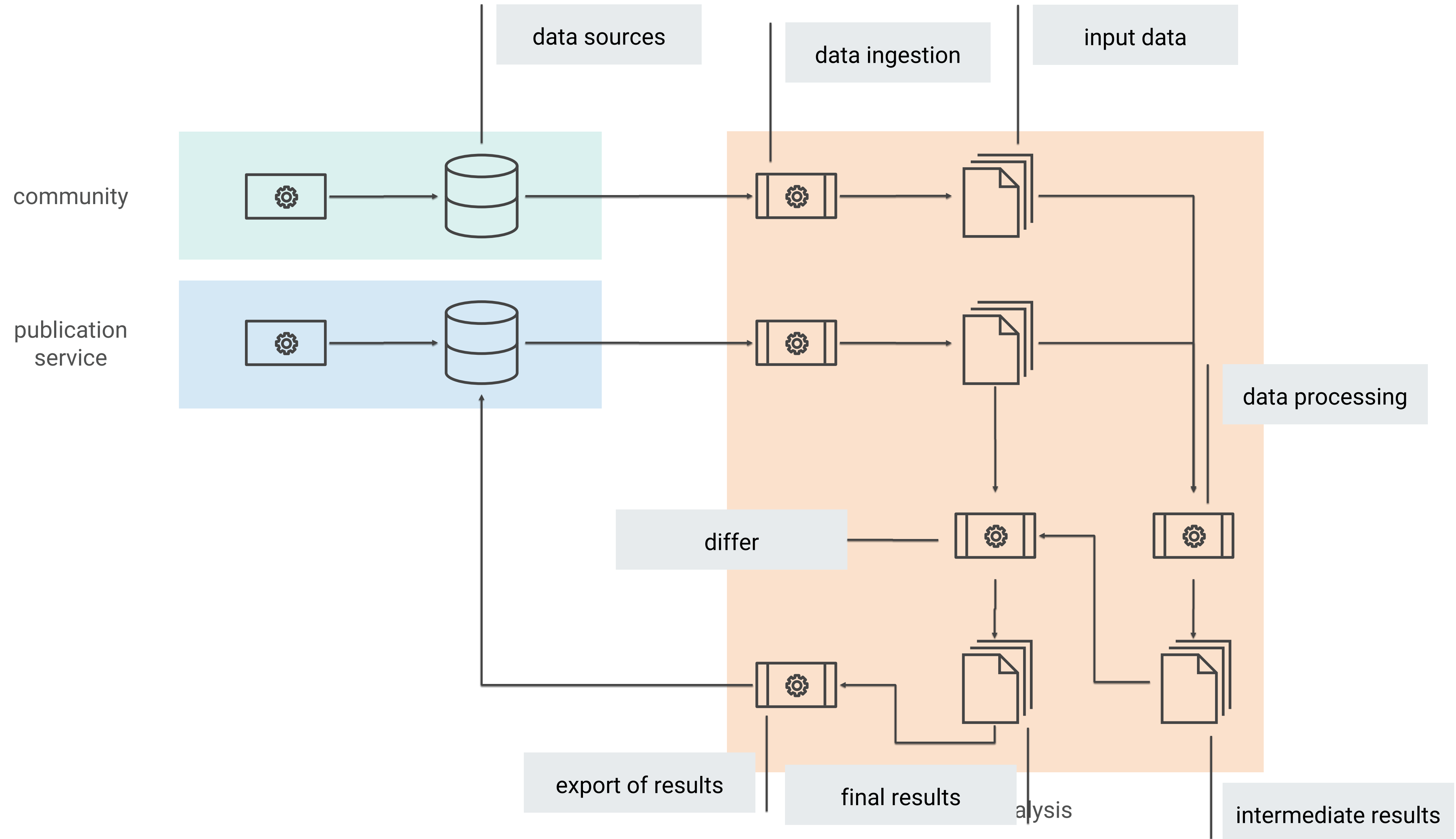


Implementation



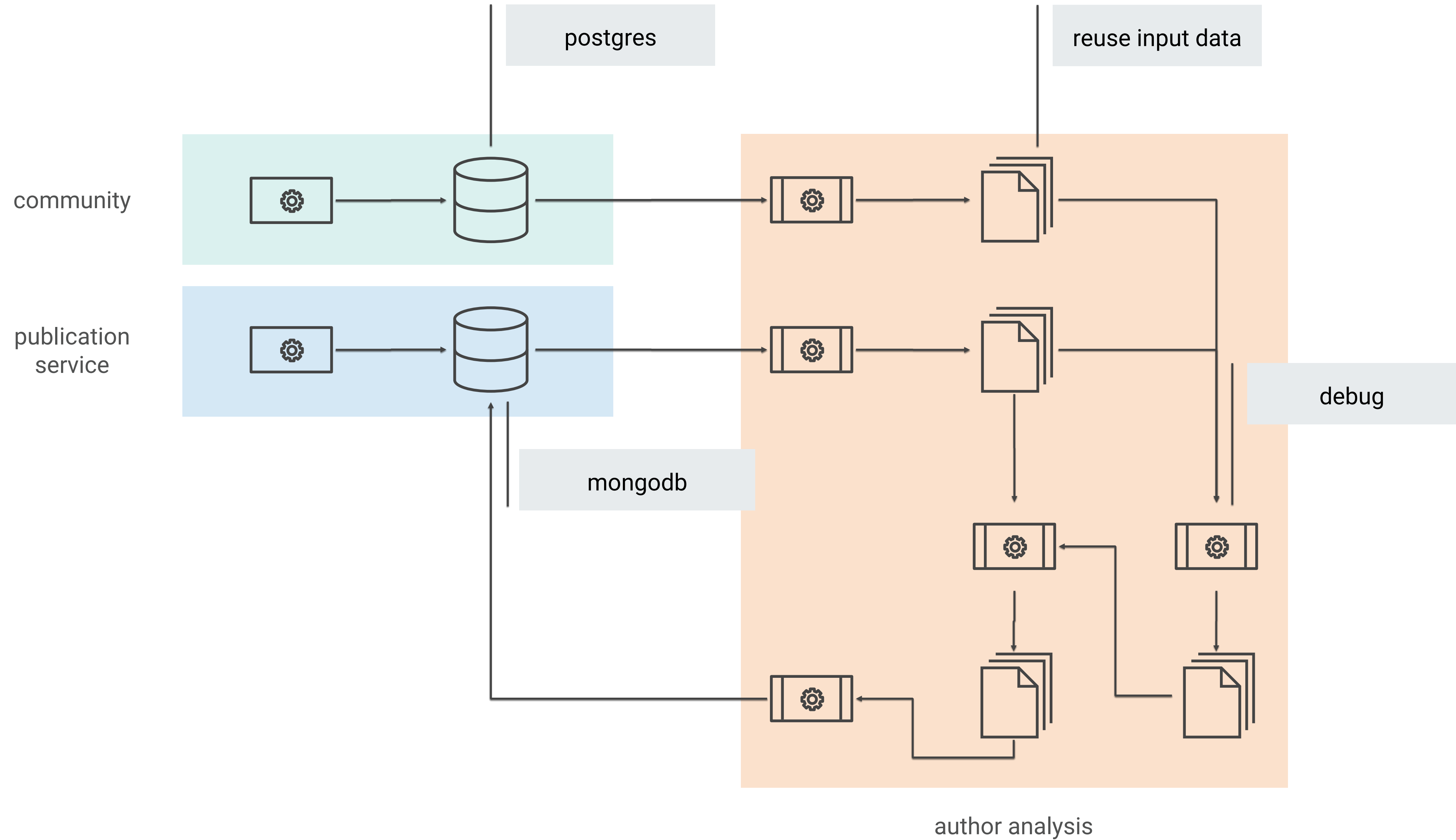
author analysis

Implementation



#1 Decouple Data Ingestion

Implementation



Debugging an Error on Production

Your flow

has unit and integrations tests
but still breaks unexpectedly in production

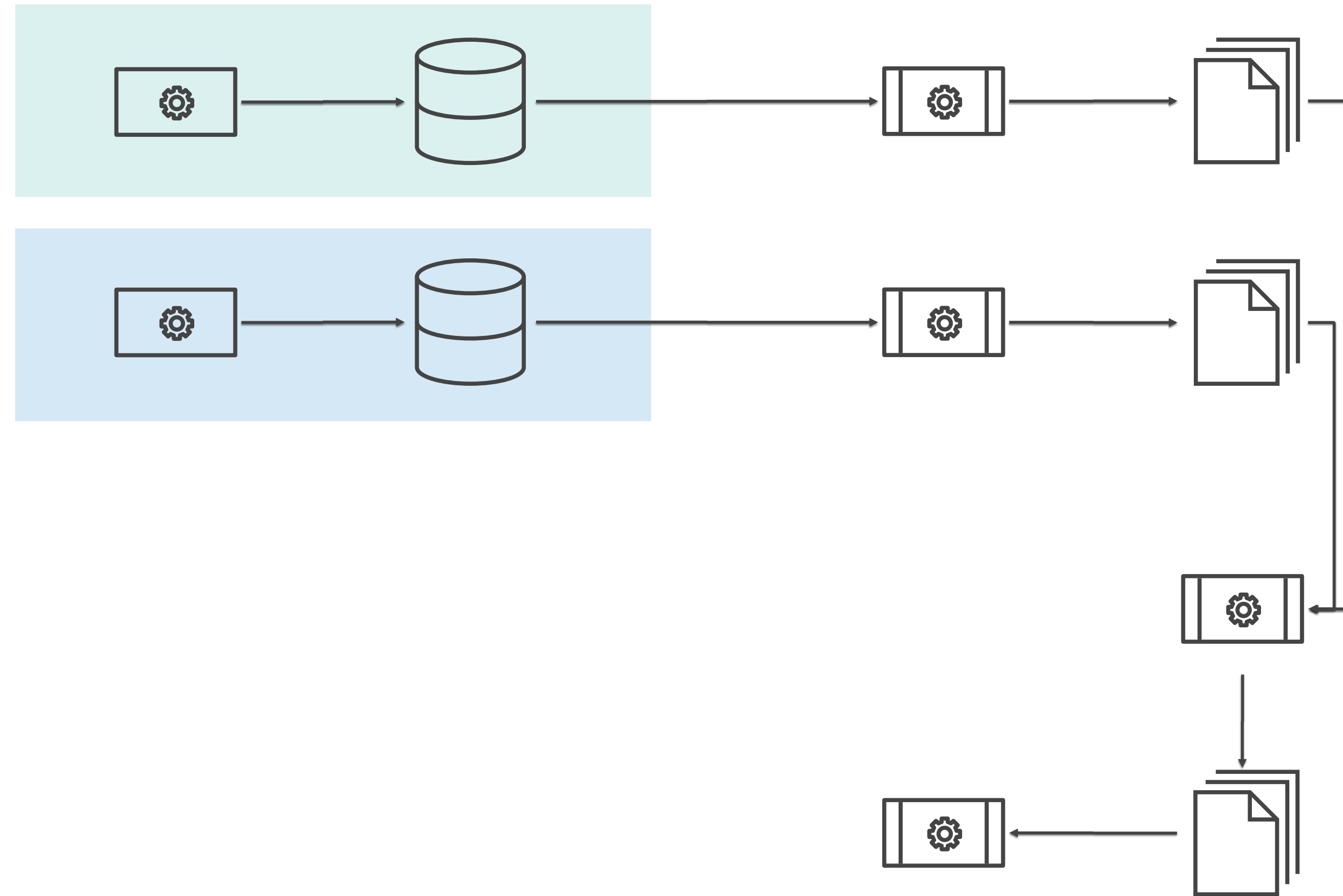
You need to find the root cause

Is it a change in input data?
Is it a change on the cluster?
Is it a race condition?

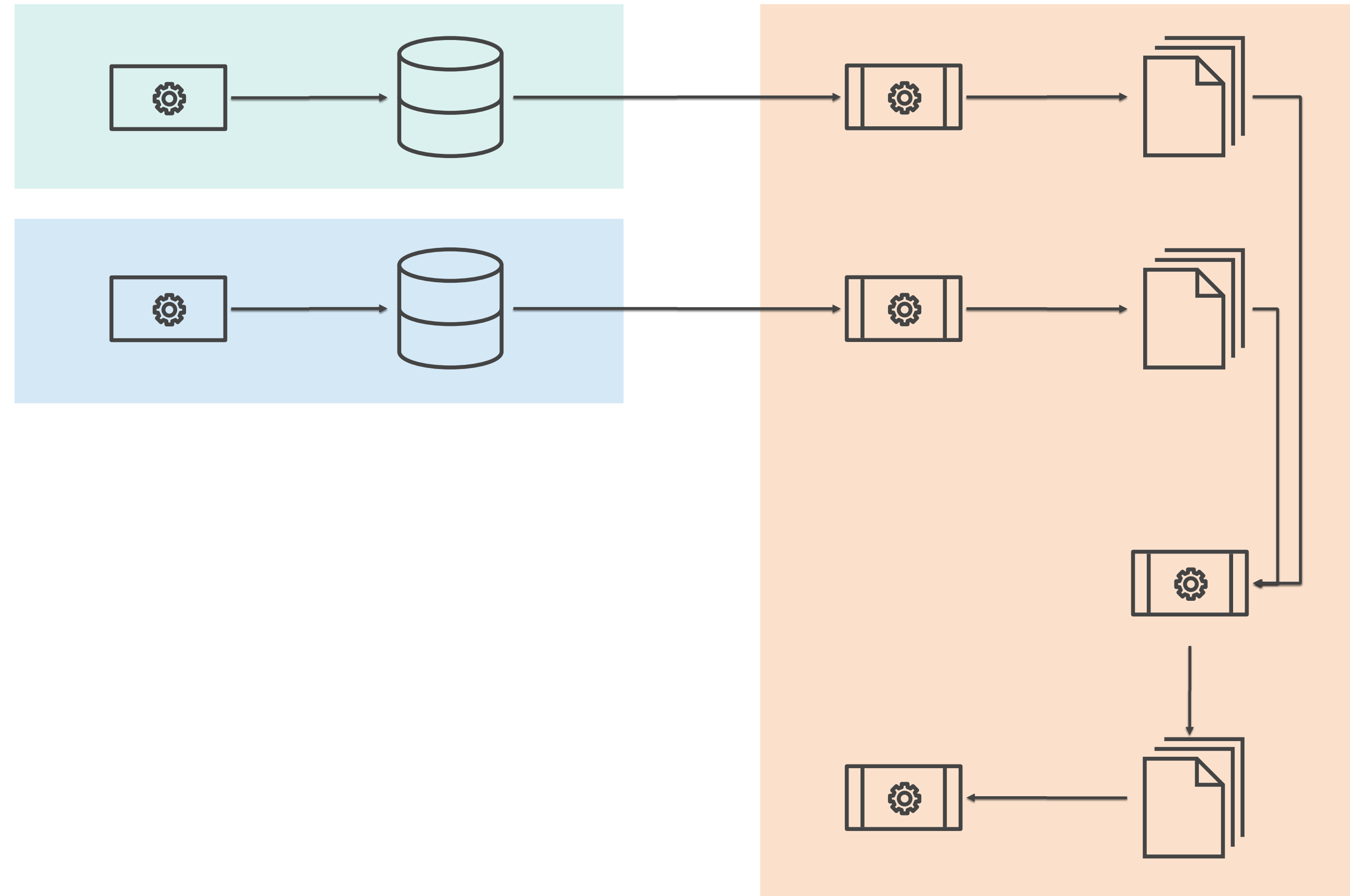
Crucial capabilities

Easy adhoc analysis of all involved data (input, intermediate, result)
Rerun current flow with current cluster configuration on yesterday's data
Confirm hotfix by re-running on today's data (exactly the same data that triggered the bug)

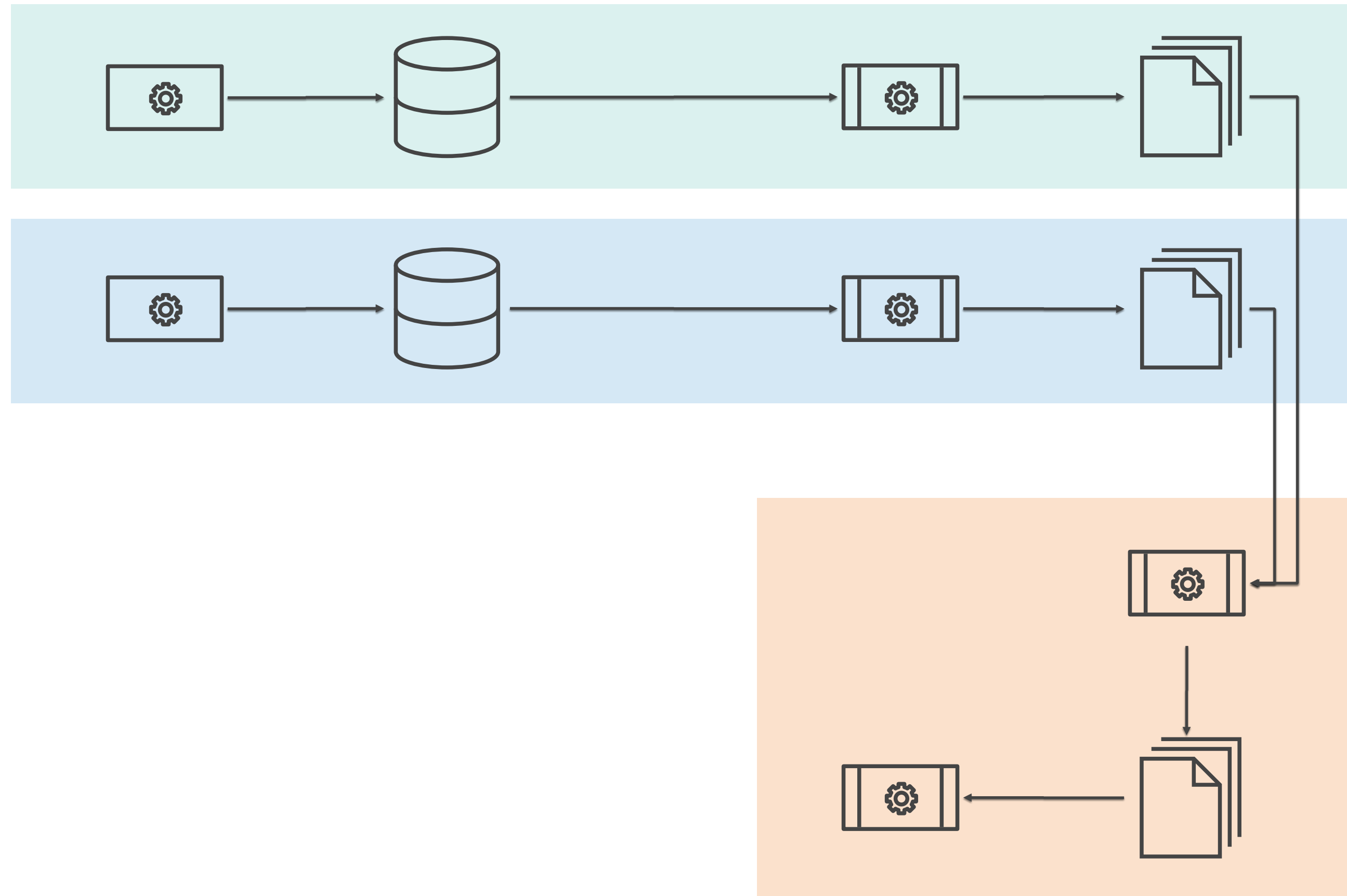
How to decouple?



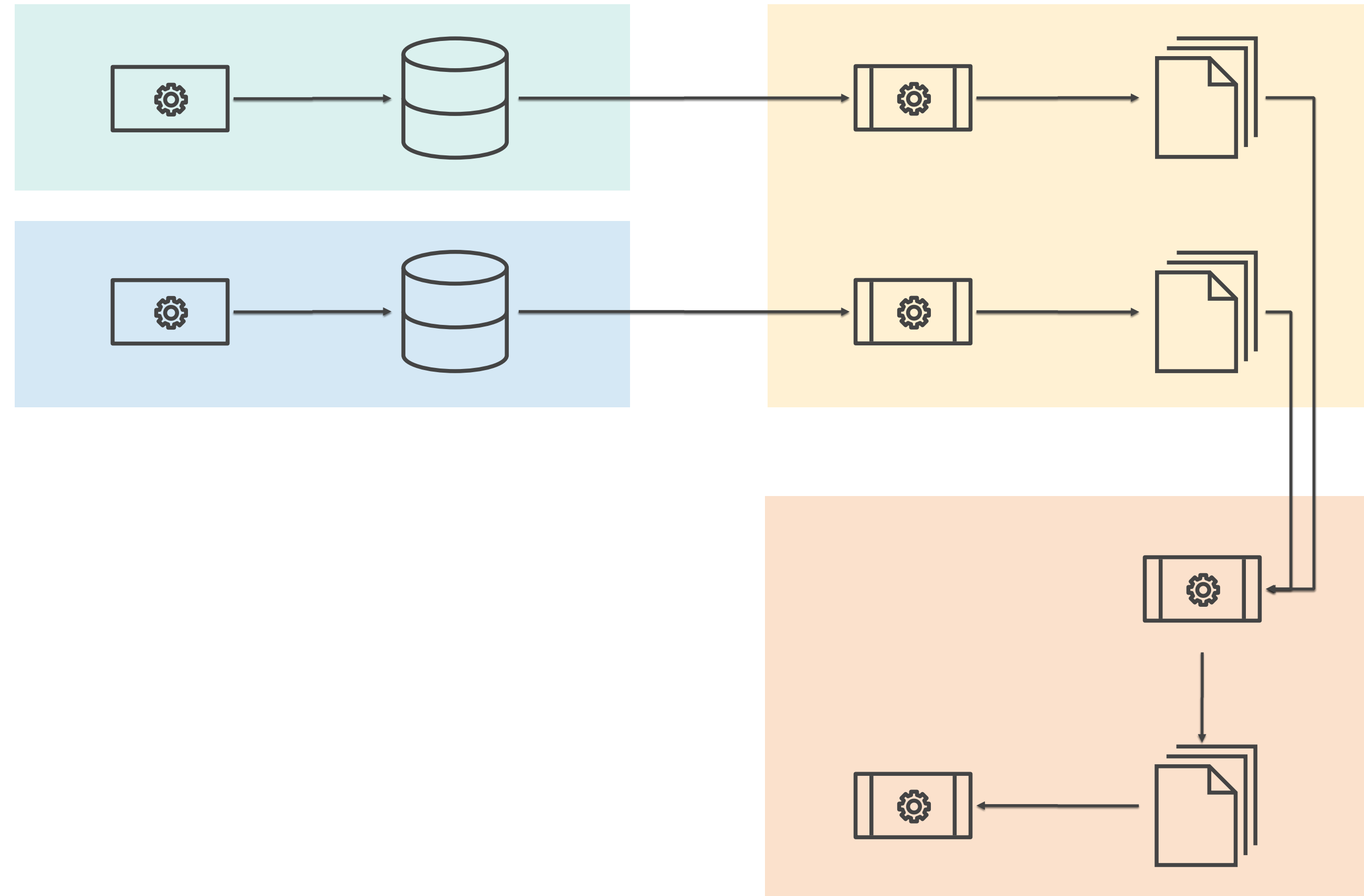
Ingesting Data as Needed?



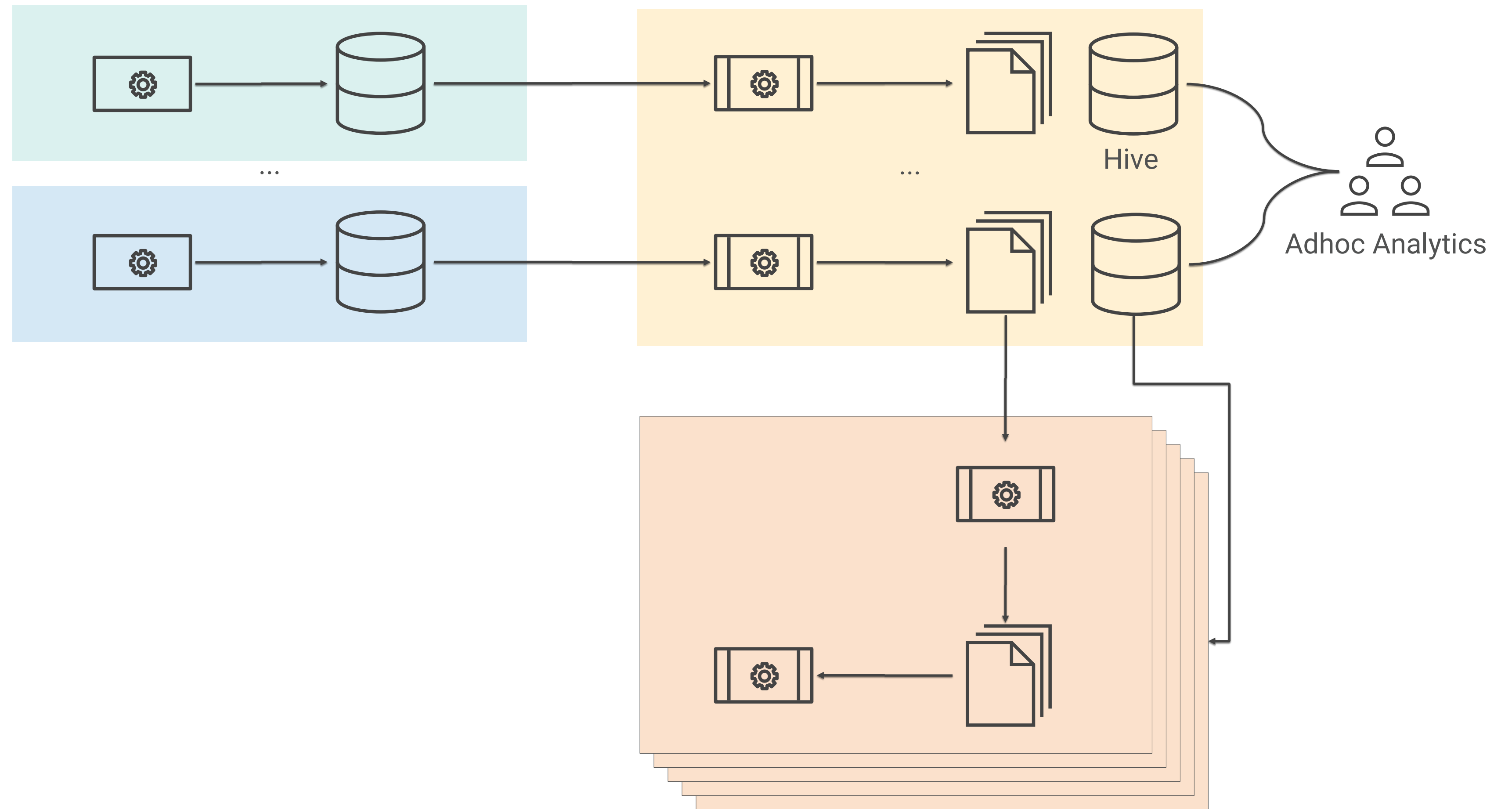
Publishing Data as Needed?



Dedicated Data Ingestion!



Platform Data Import



Platform Data Import

Dedicated component, but **generic**

Every team can onboard new data sources, as required by use cases

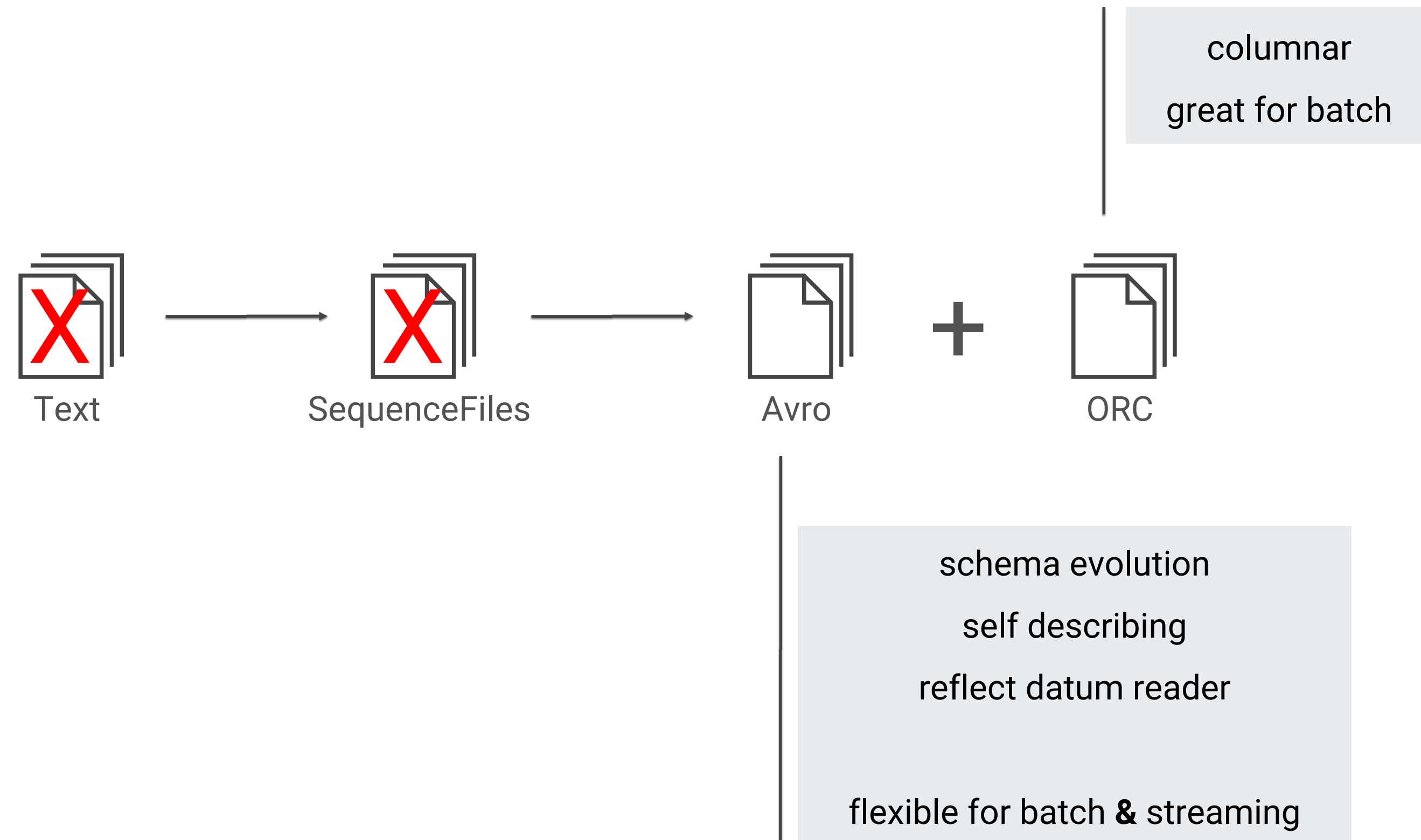
Every ingested source is immediately available **for all consumers** (incl. analytics)

Feature parity for all data sources (e.g., mounting everything in Hive)

#2 Speak a common format*

* have at least one copy of all data in a common format (e.g., avro)

Formats



Speak a common format

Have at least one copy of **all data in a common format**

Your choice of processing framework should not be limited by format of existing data

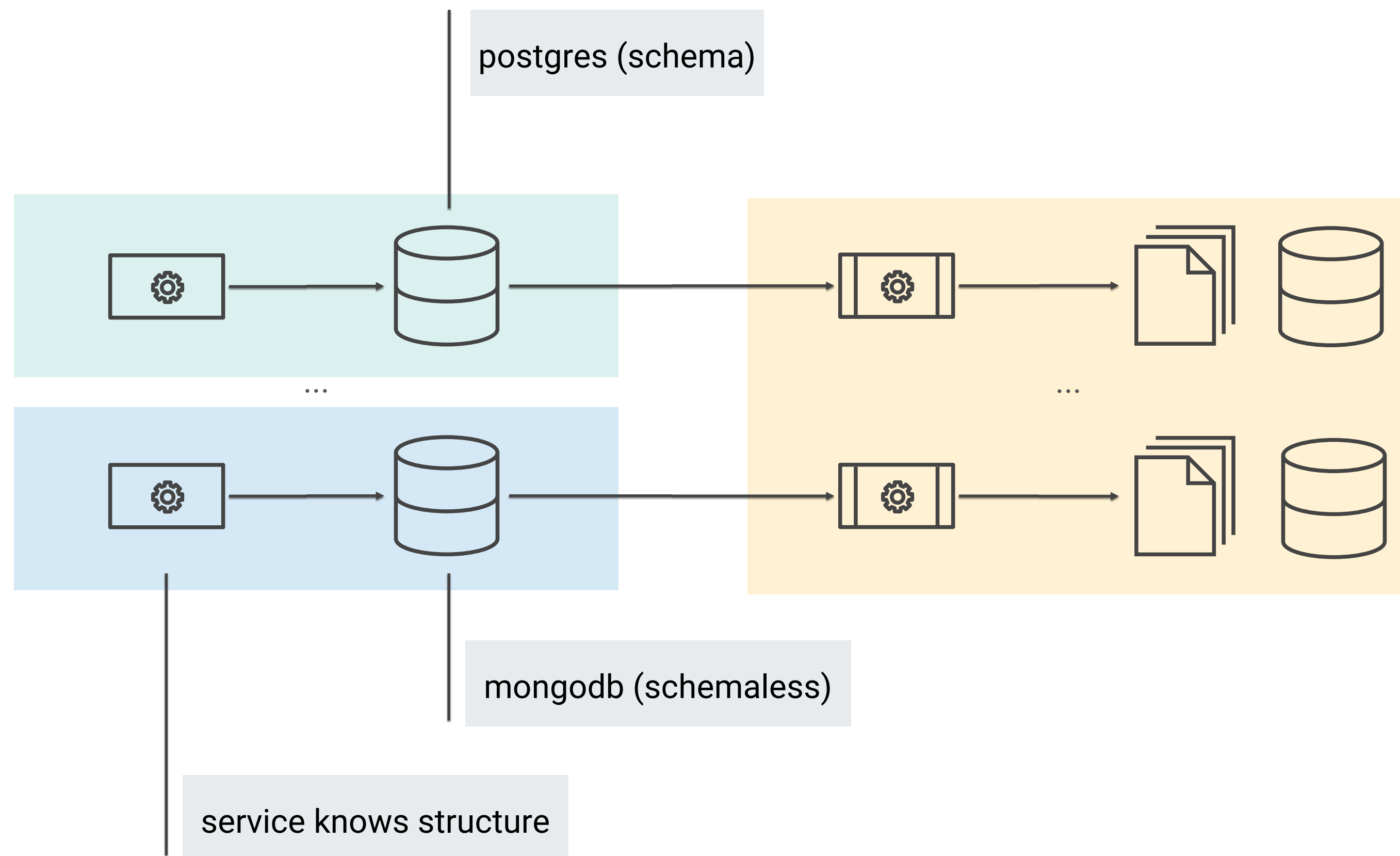
Every ingested source should be available **for all consumers**

When optimizing for a framework (e.g., ORC for Hive) consider a copy

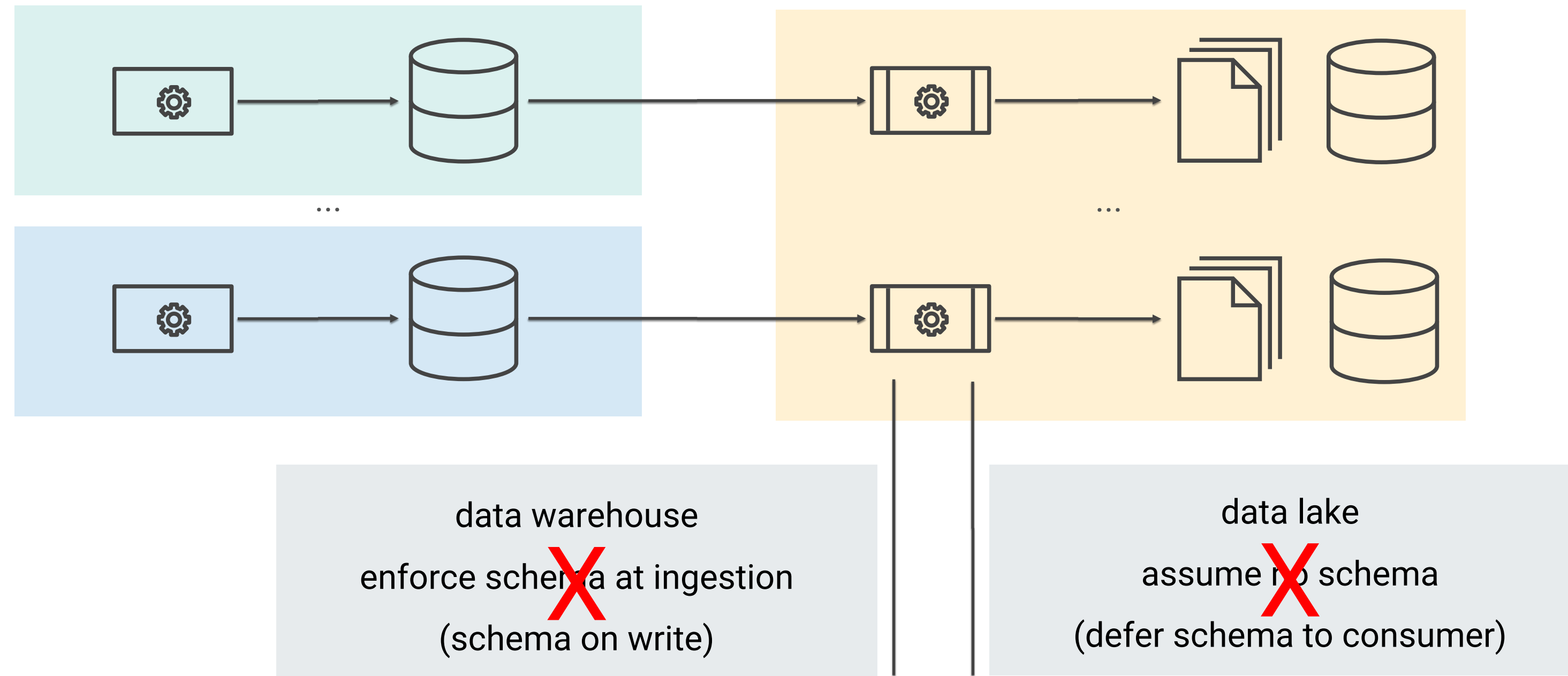
#3 Speak a common language*

* continuously propagate schema changes

Structured or unstructured data?



Data Warehouse vs. Data Lake



Can we have both?

Preserve schema information that is already present

- some times at database level

- many times at application level

Preserve full data – be truthful to our data source

- continuously propagate schema changes

Can we have something like a **Data Lakehouse**?

Entities Define Schema

Code first

entities within owning service define schema

Auto conversion preferred

conversion to other representations via annotations

(JSON, BSON, Avro, ...)

```
@Field("abstract") // Solr
@AvroName("abstract") // Avro
@property("abs") // MongoDB
private String _abstract;
```



```
{
  "type": "record",
  "name": "Publication",
  "namespace": "net.researchgate.refind.domain.entities",
  "fields": [
    {
      "name": "abstract",
      "type": [
        "null",
        "string"
      ],
      "default": null,
      "dbname": "abs"
    }
  ]
}
/* ... */
```

Continuously propagate schema changes

Data ingestion process is generic and driven by avro schema

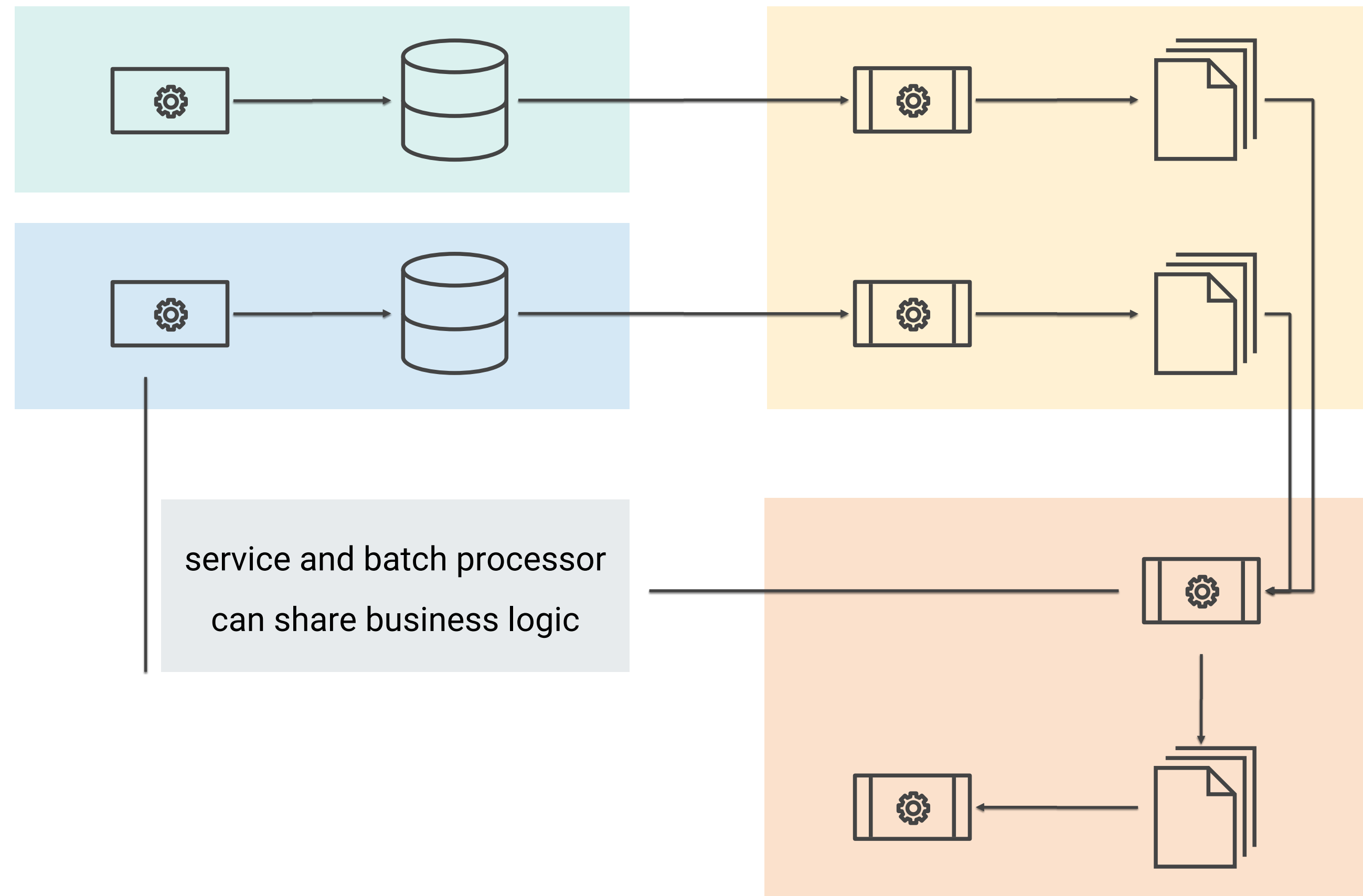
Changes in avro schema are continuously propagated to data ingestion process

Consumers with old schema can still read data due to avro schema evolution

Caveat: breaking changes still have to be dealt with by a change process

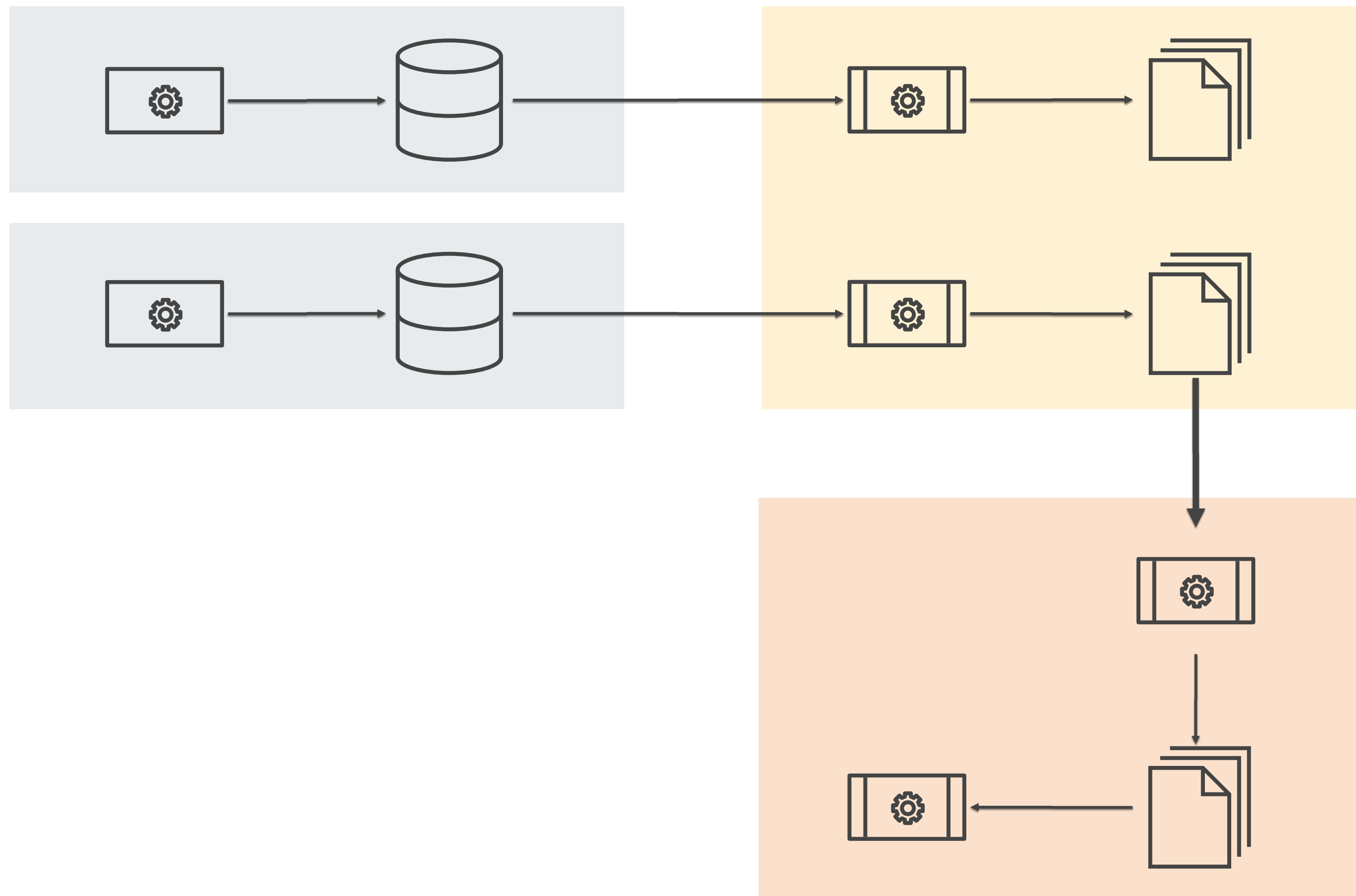
Everyone speaks the same language

Extra Benefit

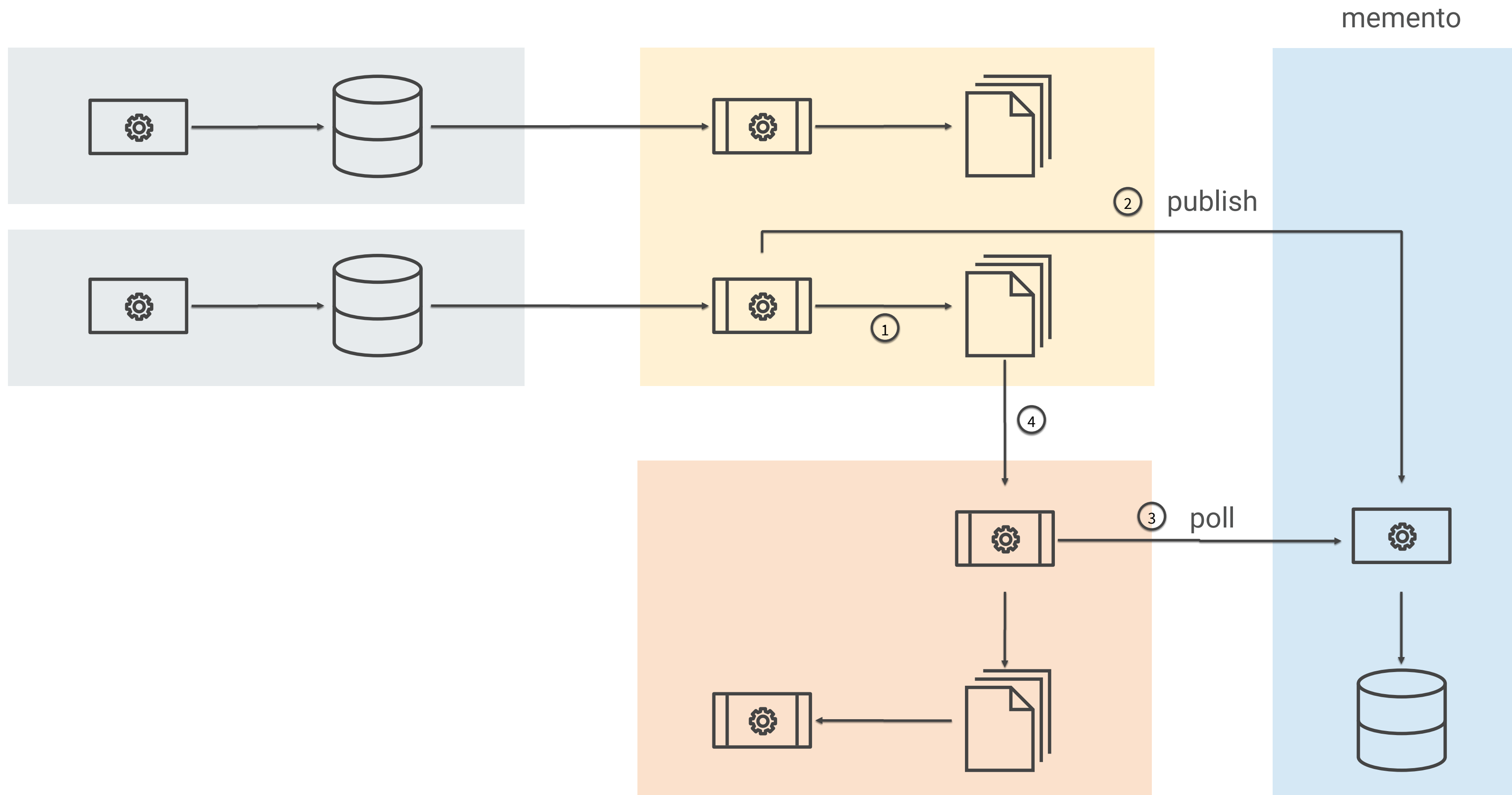


#4 Model Data Dependencies Explicitly

Model Data Dependencies Explicitly



Model Data Dependencies Explicitly



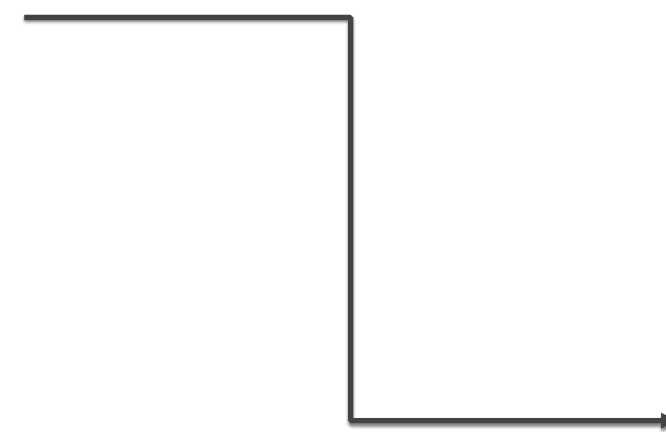
Memento v2

memento publish

```
{  
  "namespace": "platform-data",  
  "name": "refind.authors",  
  "businessDate": "2017-06-11",  
  "type": "hdfs",  
  "path": "hdfs://platform-data/refind/authors/2017-06-12/attempt-0/data",  
  "format": "avro",  
  "attempt": 0,  
  "timestamp": "2015-06-12T00:42:23.125+0000"  
}
```

memento poll <waiting-time>

```
{  
  "namespace": "platform-data",  
  "name": "refind.authors",  
  "businessDate": "2017-06-11",  
  "type": "hdfs"  
}
```



```
{  
  "artifactId": "d114682f02fd",  
  "namespace": "platform-data",  
  "name": "refind.authors",  
  "businessDate": "2017-06-11",  
  "type": "hdfs",  
  "path": "hdfs://platform-data/refind/authors/2017-06-12/attempt-0/data",  
  "format": "avro",  
  "attempt": 0,  
  "timestamp": "2015-06-12T00:42:23.125+0000"  
}
```

unique artifactId

Model Data Dependencies Explicitly

More flexible scheduling – run flows as early as possible

Allows multiple ingestion or processing attempts

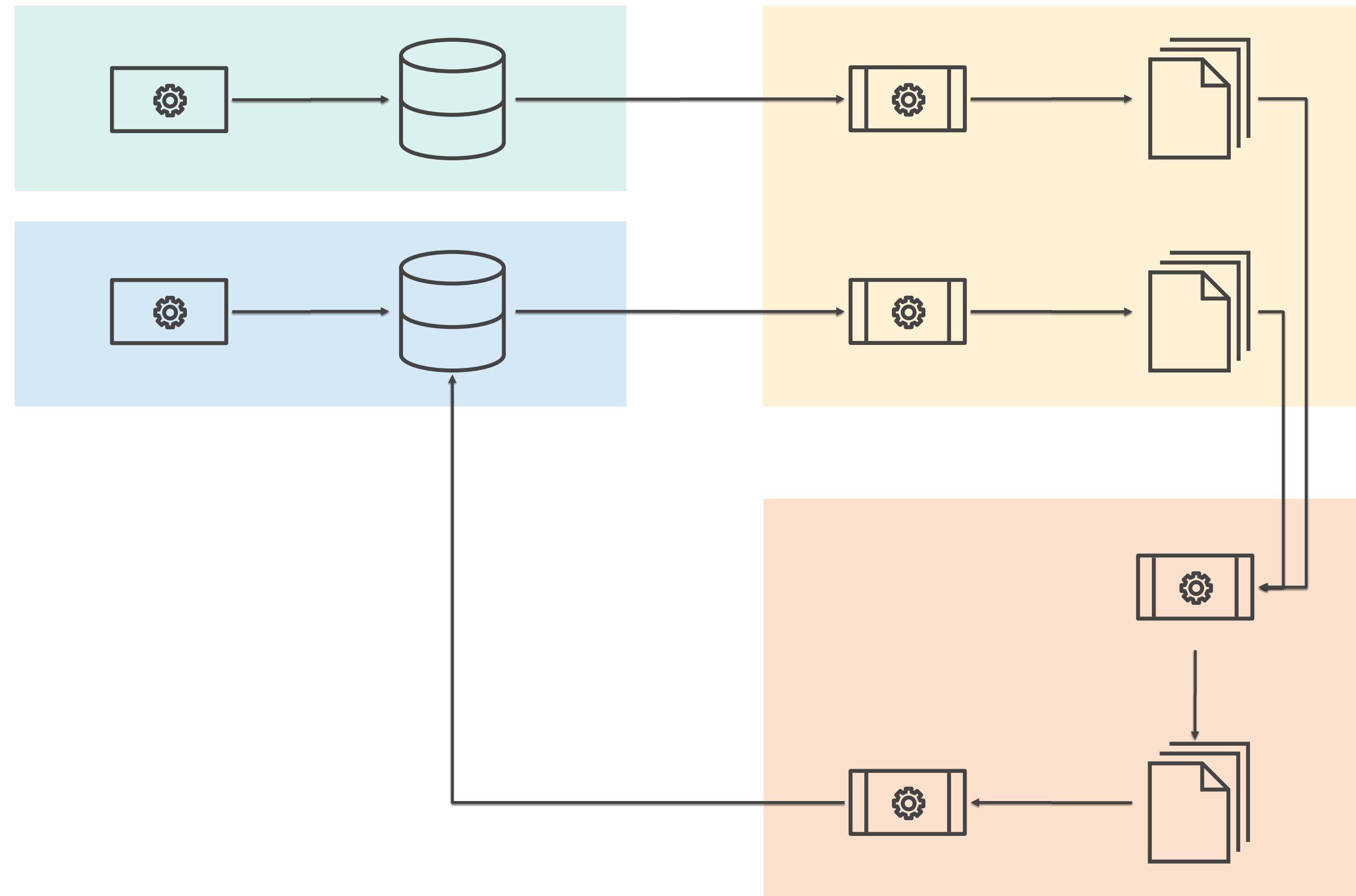
Allows immutable data (repeatable read)

Allows analysis of dependency graph

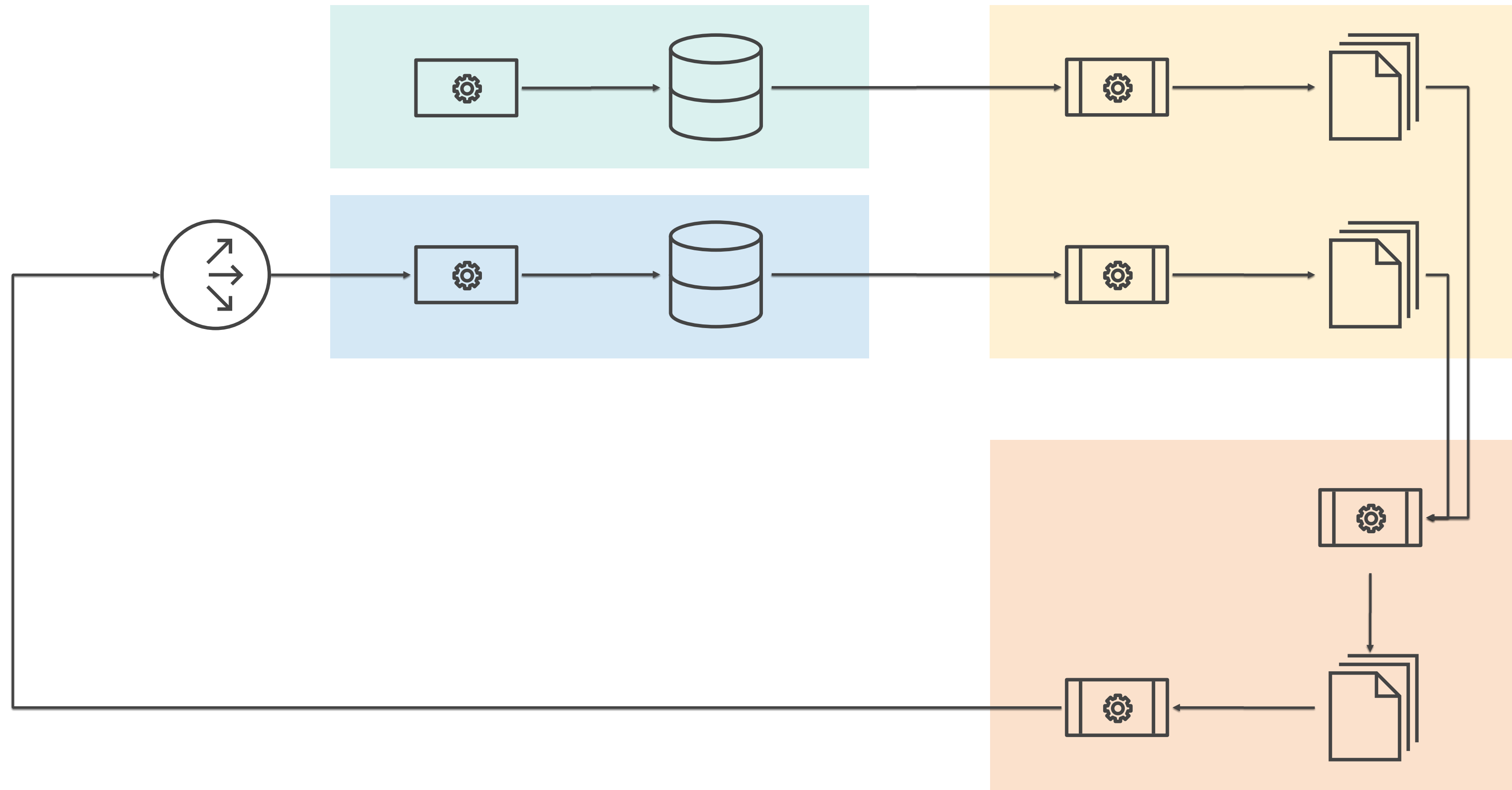
which datasets are used by what flow

#5 Decouple export of results

Decouple export of results



Decouple export of results



Push results via HTTP to service

Export of results just becomes a client of the service

service does not have to be aware of big data technologies

Service can validate results, e.g.,

plausibility checks

optimistic locking

Makes testing much easier

Avro → Http

Part of the flow, but standardized component

Handles tracking of progress

- treats input file as a “queue”

- converts records to http calls

- can be interrupted and resumed anytime

Sends standardized headers, e.g.,

- `X-rg-client-id: author-analysis`

Handles backpressure signals from services

#6 Model Flow Orchestration Explicitly

Model Flow Orchestration Explicitly

Consider using an execution system like Azkaban, Luigi, or Airflow

Establish coding standards for orchestration, e.g.,

- inject paths from outside – don't construct them in your flow

- inject calculation dates – **never** call `now()`

- inject configuration settings – don't hardcode `-D mapreduce.map.java.opts=-Xmx4096m`

- foresee environment specific settings

Think about

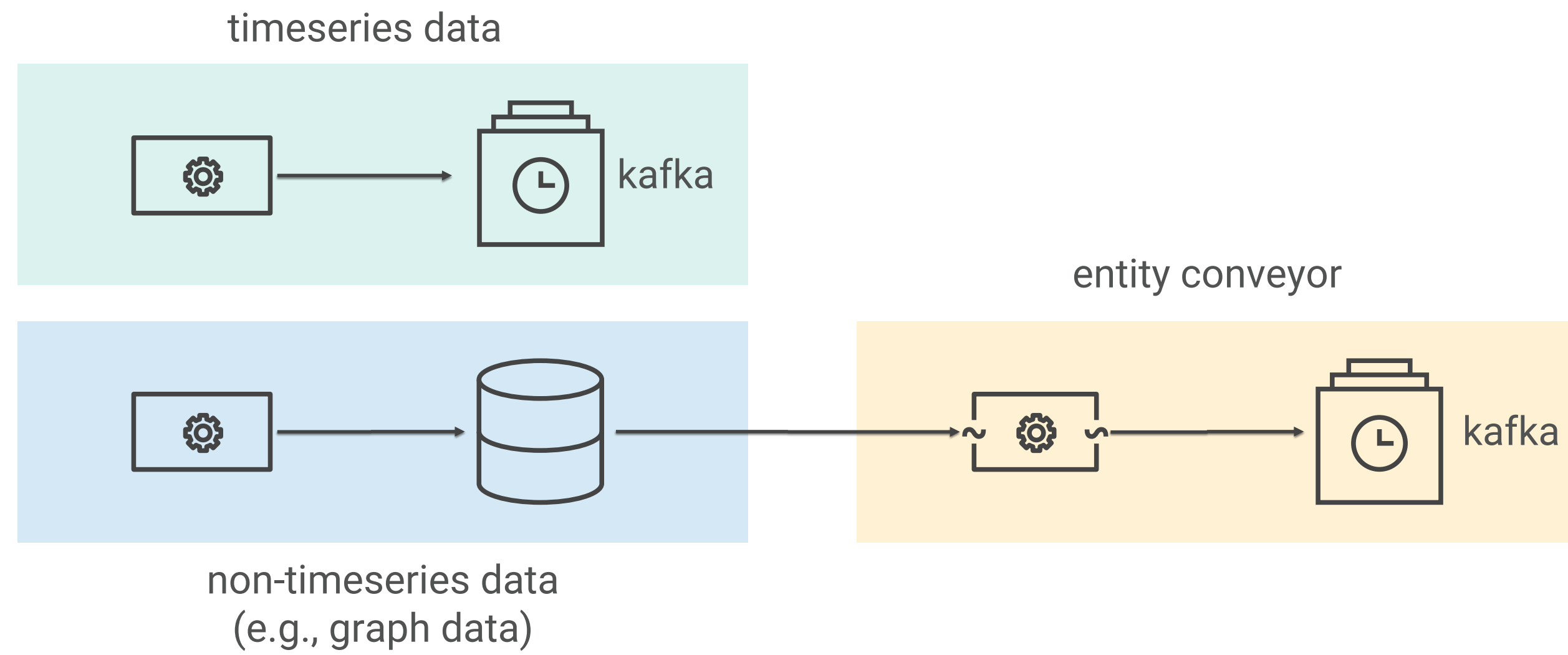
- ease of operations

- tuning of settings

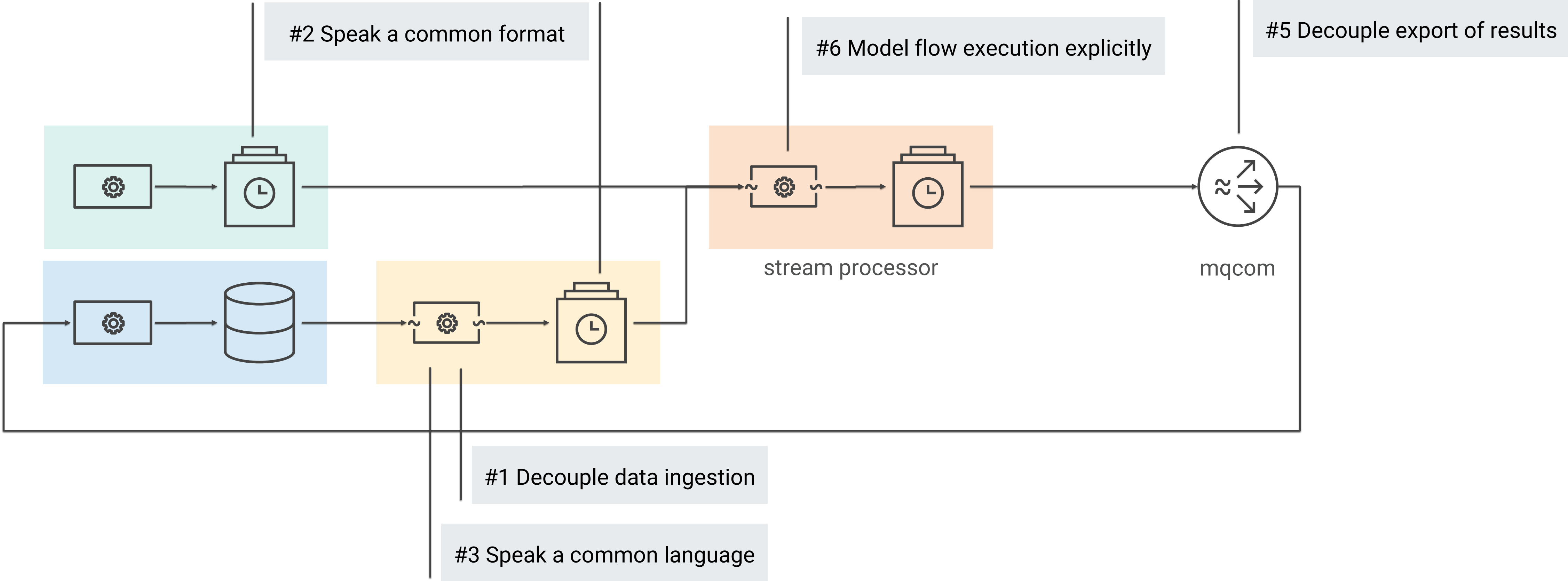
- upgrades

What about Stream Processing?

Sources of Streaming Data



Stream Processing



What about #4 ?

Model Data Dependencies Explicitly

We think about it

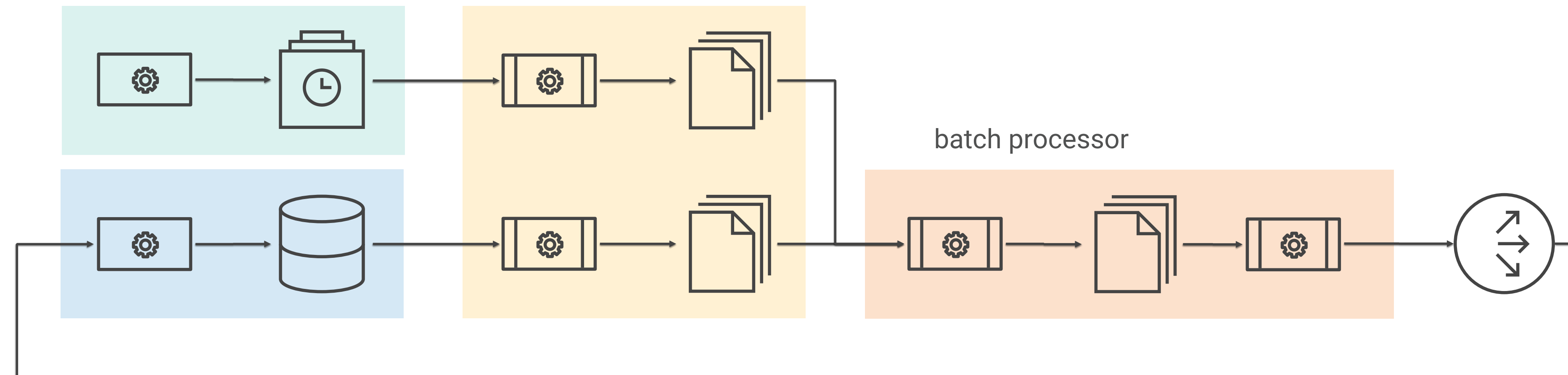
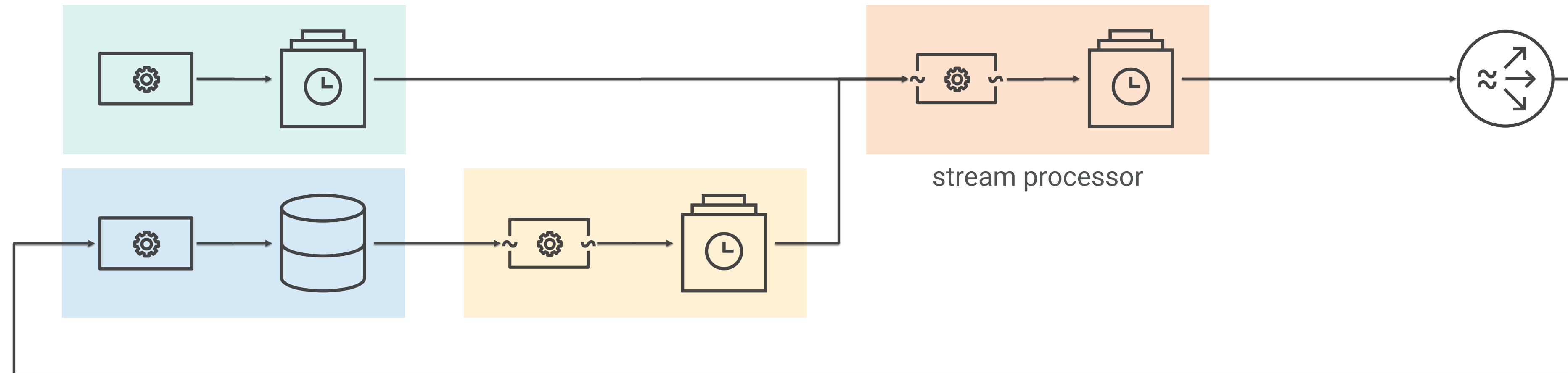
Depends on use cases and pain points

Potentially put Kafka topics into Memento

- storing “offsets of interest” from producers

- facilitate switching between incompatible versions of stream processors

Evolving Big Data Architecture



Thank you!

Michael Häusler, Head of Engineering

https://www.researchgate.net/profile/Michael_Haeusler

<https://www.researchgate.net/careers>