Choosing The Right Boat For Your Stream

Kamal Aboul-Hosn
Uber Tech Lead, Google Pub/Sub & Kafka

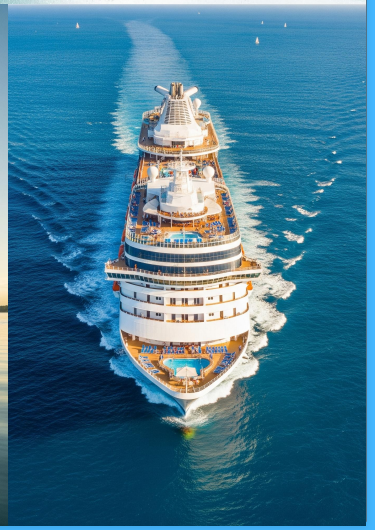
BEAM SUMMIT NYC 2025

What We Will Cover

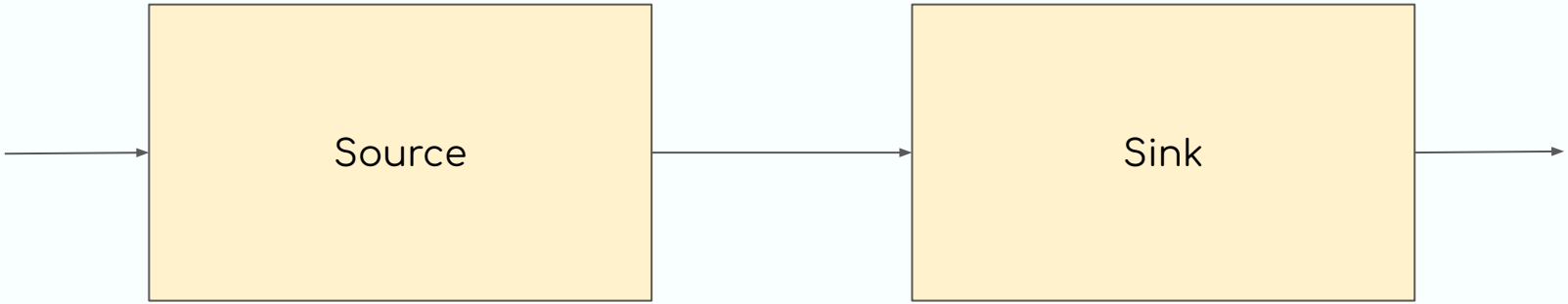
A Holistic View Of Streaming

How To Think About Decisions

Where Streaming Is Going



What Is Streaming Anyway?



There Are A Lot Of Ways To Do This

Dataflow, Flink, Spark, Kafka Connect, Data Transfer Service, Datastream, Continuous Queries, Import Topics & Export Subscriptions, Single-Message Transforms (SMTs), DIY Solutions…

Everyone loves to (re)build ways to move data!

Multiple Options Always Exist

There will never be "one solution to rule them all"

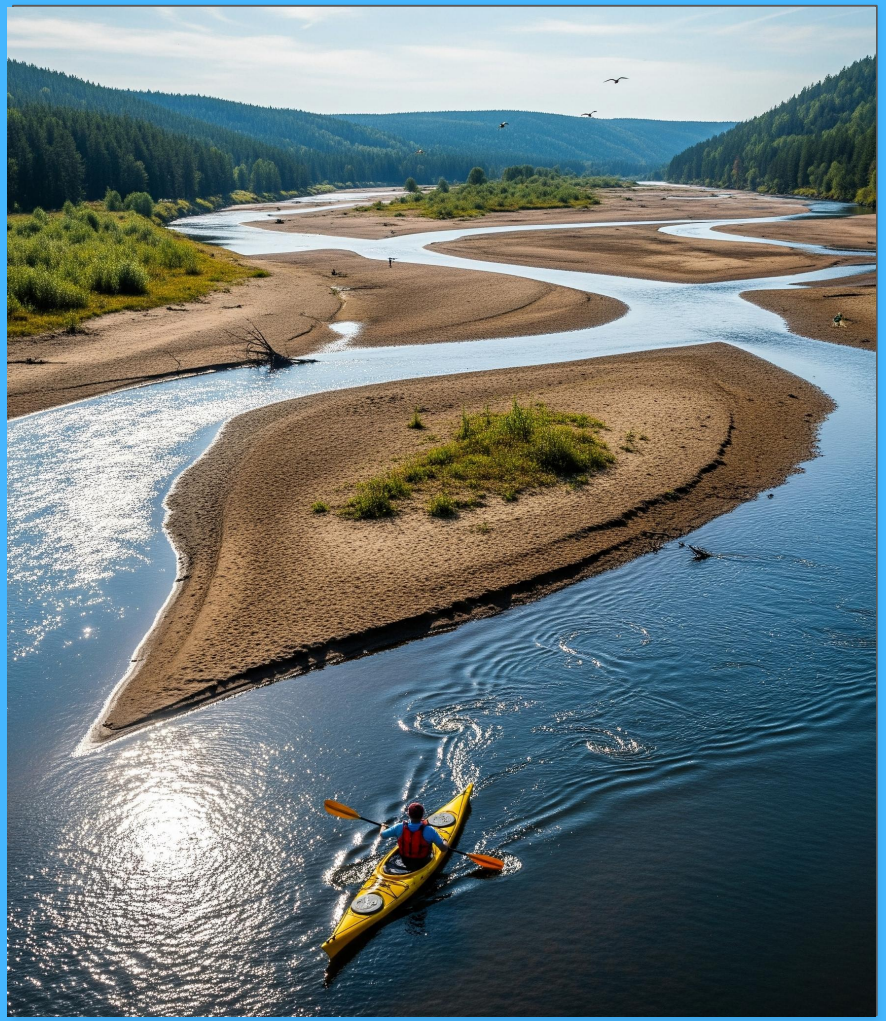
One solution for everything

↓

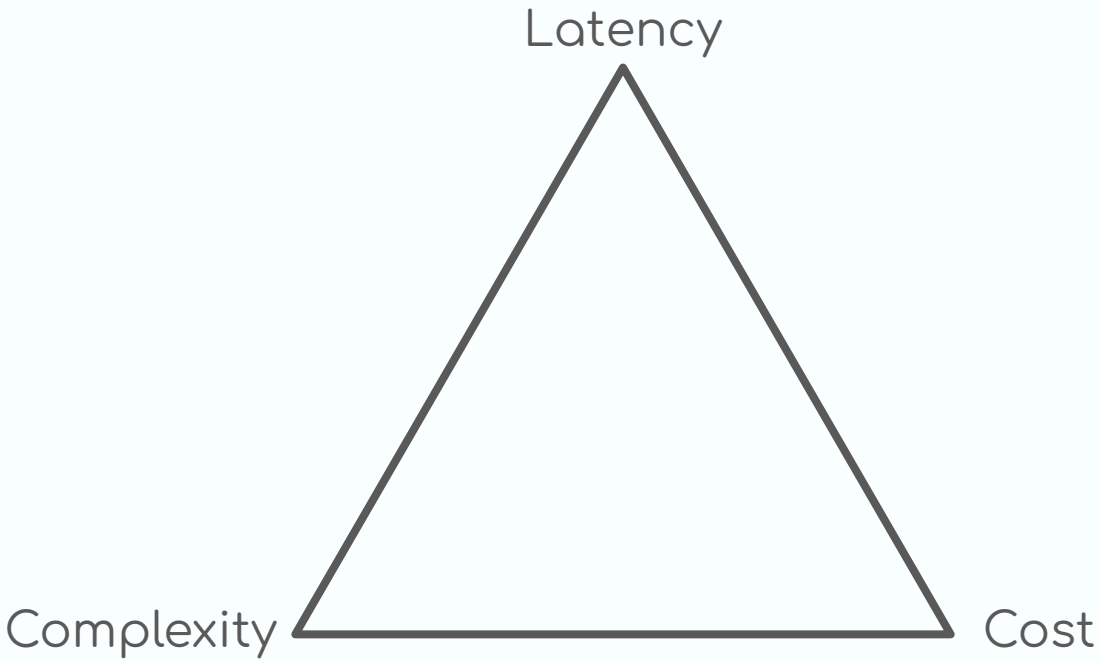
One solution per pipeline

↓

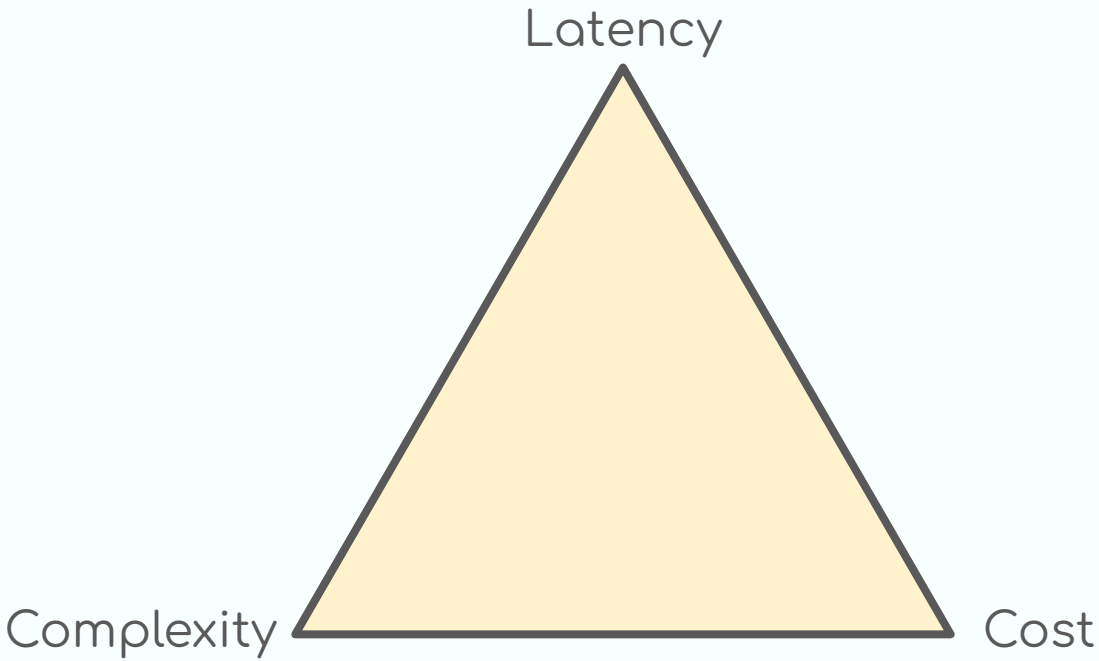
One solution per connection



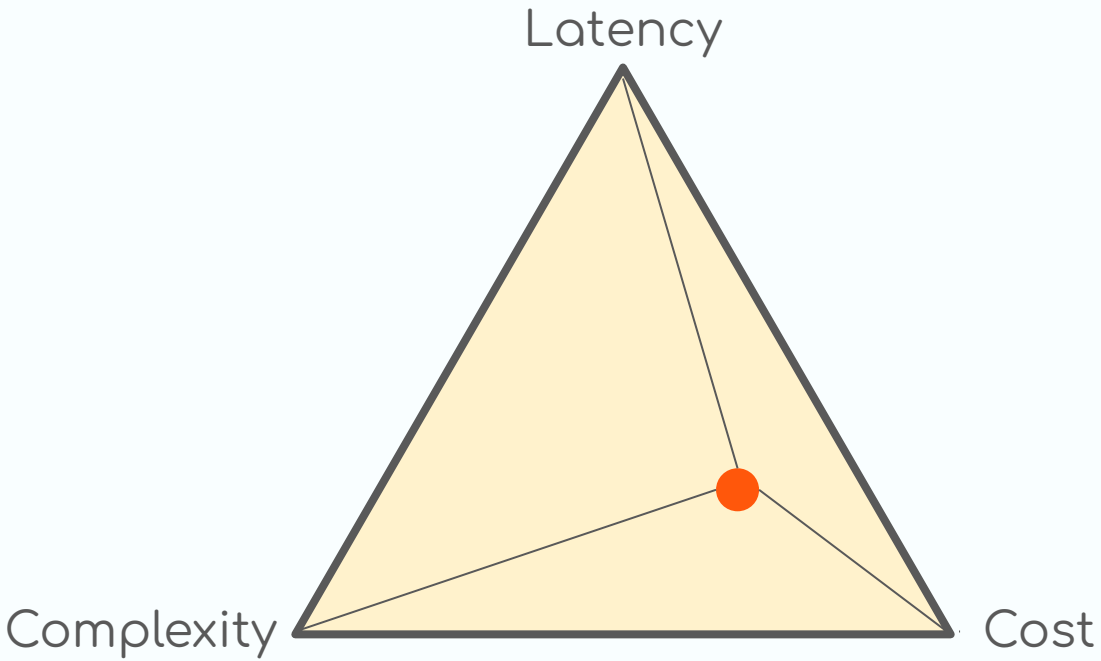
Life Is A Bunch Of Tradeoffs



Life Is A Bunch Of Tradeoffs



Life Is A Bunch Of Tradeoffs



Ask The Right Questions

Where is the data coming from and going?

How much data and **how** fast do I need it to move?

Why might I prefer one tool over another?

What do I need to do to the data?

When should I do it?

Who is the owner of the data?

Ask The Right Questions

Where is the data coming from and going?

How much data and **how** fast do I need it to move?

Why might I prefer one tool over another?

What do I need to do to the data?

When should I do it?

Who is the owner of the data?

Stateless Transformation

Validate

Filter

Alter (redact, re-encode, project)

Enrich

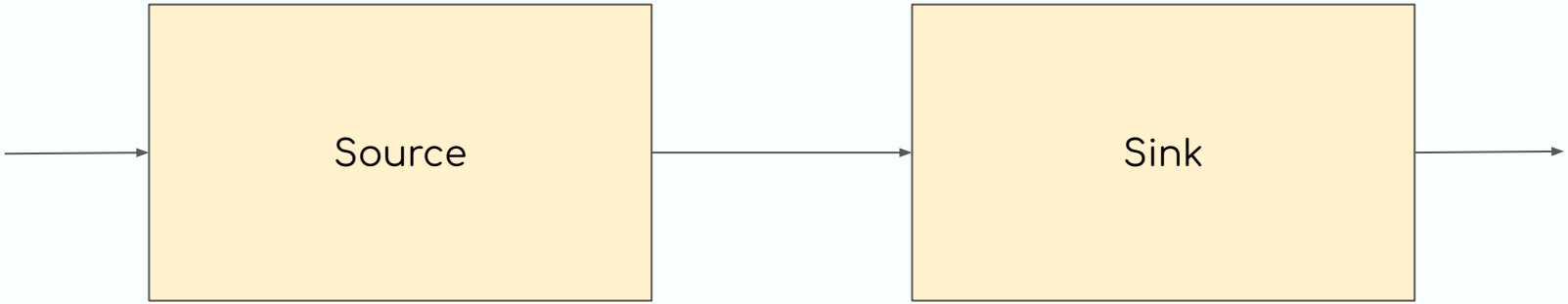
Stateful Transformation

Dedupe

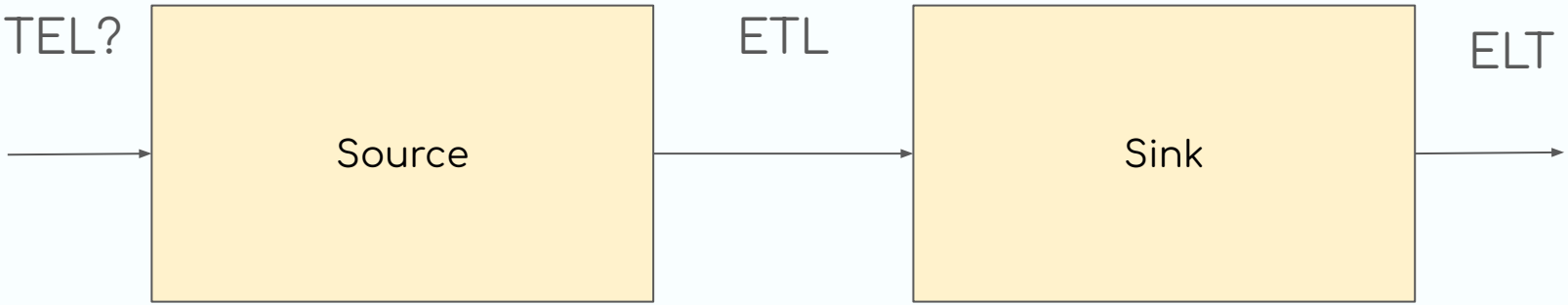
Order

Aggregate

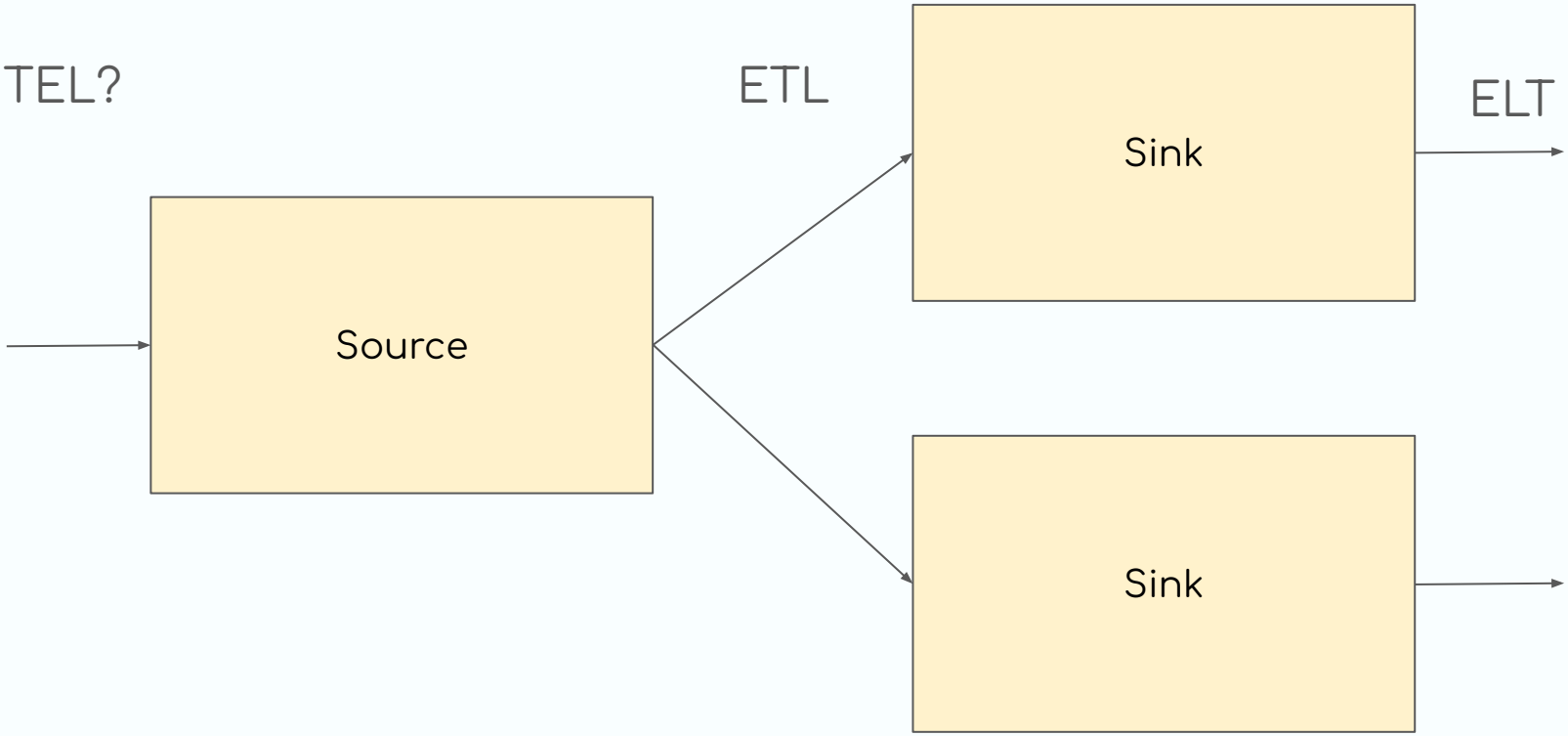
When Should I Do It?



When Should I Do It?



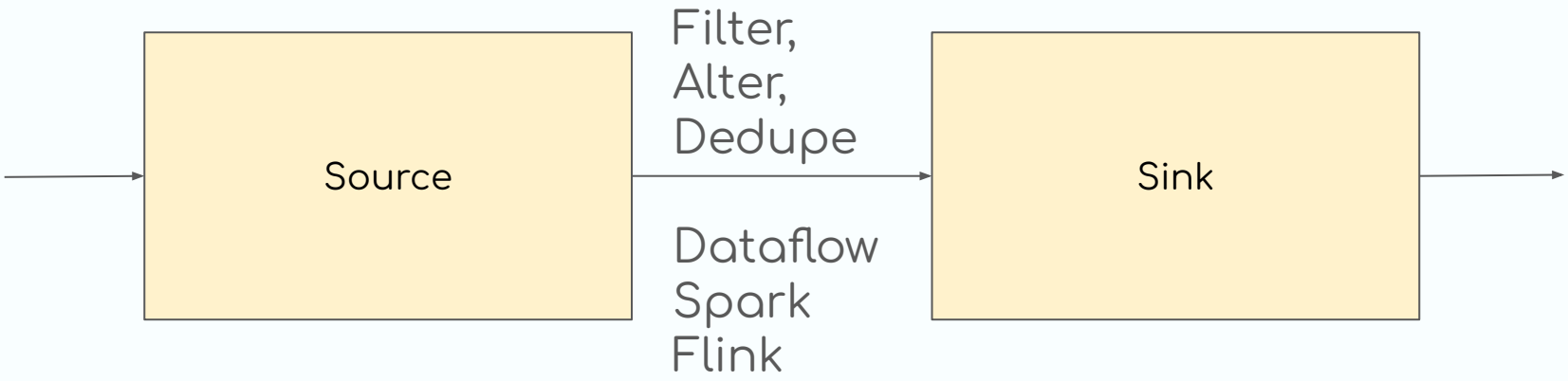
When Should I Do It?



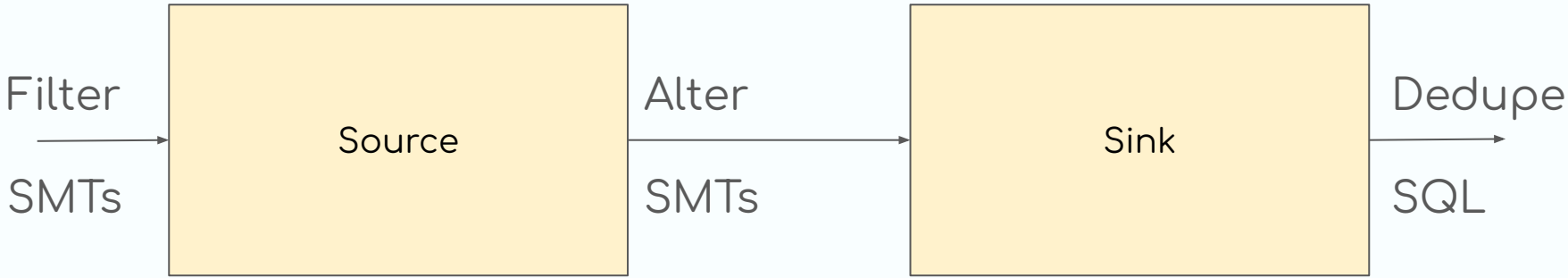
When Should I Do It?

Operation	TEL	ETL	ELT
Filter	Cheaper		Change retroactively
Validate	Report failure to producer	What to do with invalid data?	What to do with invalid data?
Alter/Enrich	Edit once for all consumers	Process at rate of enrichments	Need to perform repeatedly
Dedupe /Order	Can't control what is done downstream		Operate on a subset
Aggregate	Slows down producer	Process at rate of enrichments	Need to perform repeatedly

When Should I Do It?

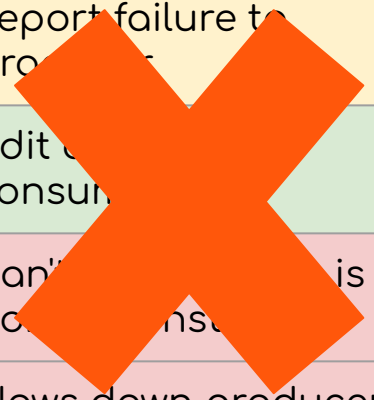


When Should I Do It?



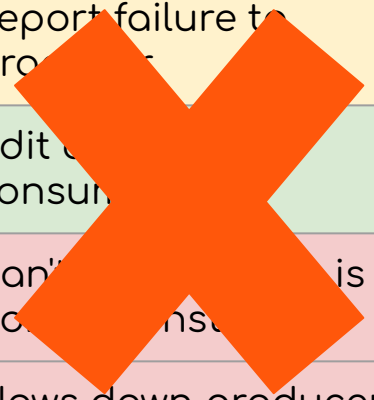
What If I Don't Own The Producer?

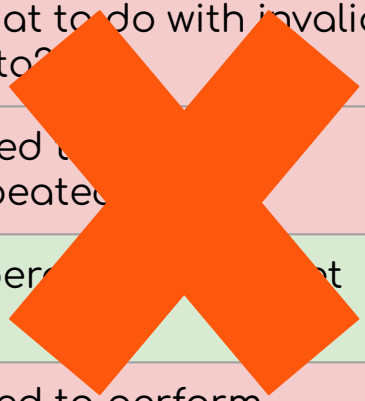
Operation	TEL	ETL	ELT
Filter	Cheaper		Change retroactively
Validate	Report failure to pro	What to do with invalid data?	What to do with invalid data?
Alter/Enrich	Edit consu	Process at rate of enrichments	Need to perform repeatedly
Exactly once /Order	Can' is do nsu		Operate on a subset
Aggregate	Slows down producer	Process at rate of enrichments	Need to perform repeatedly



What If I Don't Own The Consumer?

Operation	TEL	ETL	ELT
Filter	Cheaper		Change retroactively
Validate	Report failure to pro	What to do with invalid data?	What to do with invalid data?
Alter/Enrich	Edit consu	Process at rate of enrichments	Need repeate
Exactly once /Order	Can' is do nsu		Oper t
Aggregate	Slows down producer	Process at rate of enrichments	Need to perform repeatedly





Who Owns The Data?

Is this **operational** data being consumed by an **analyst** or
analytical data being provided by an operational workload?

How many copies of the data should exist?

Is it okay for there to be transient copies of the data?

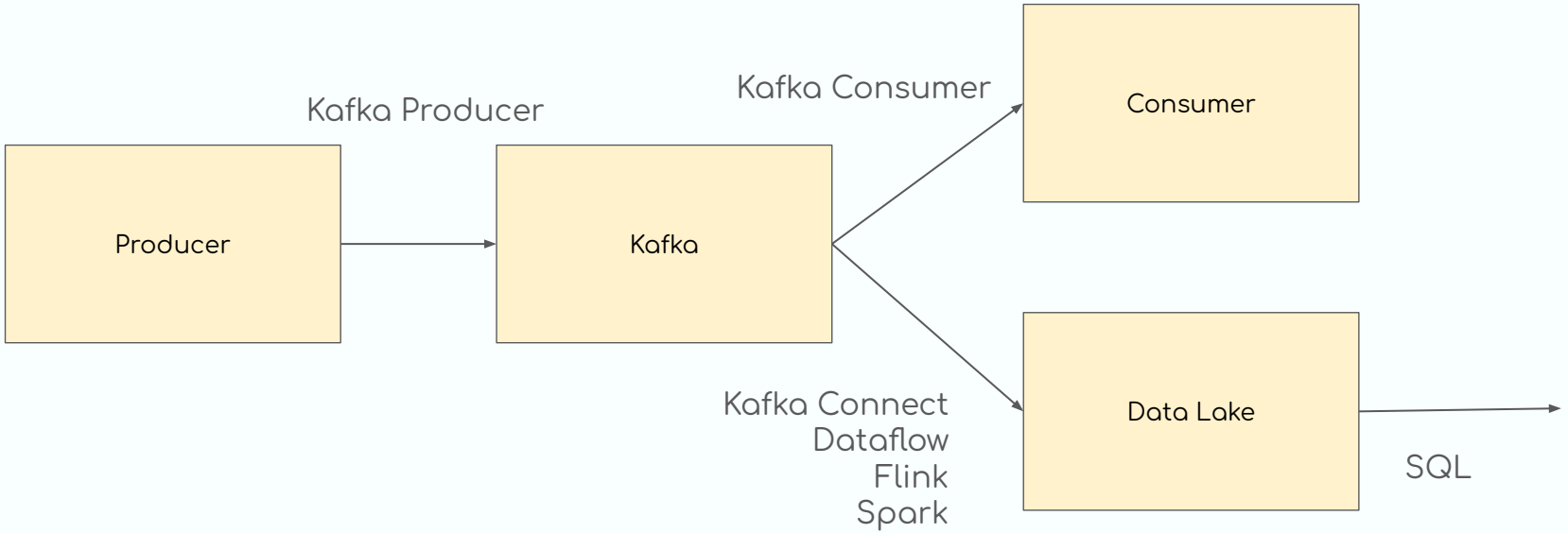
Do Pub/Sub systems count as storage?

Should all stored data always be valid?

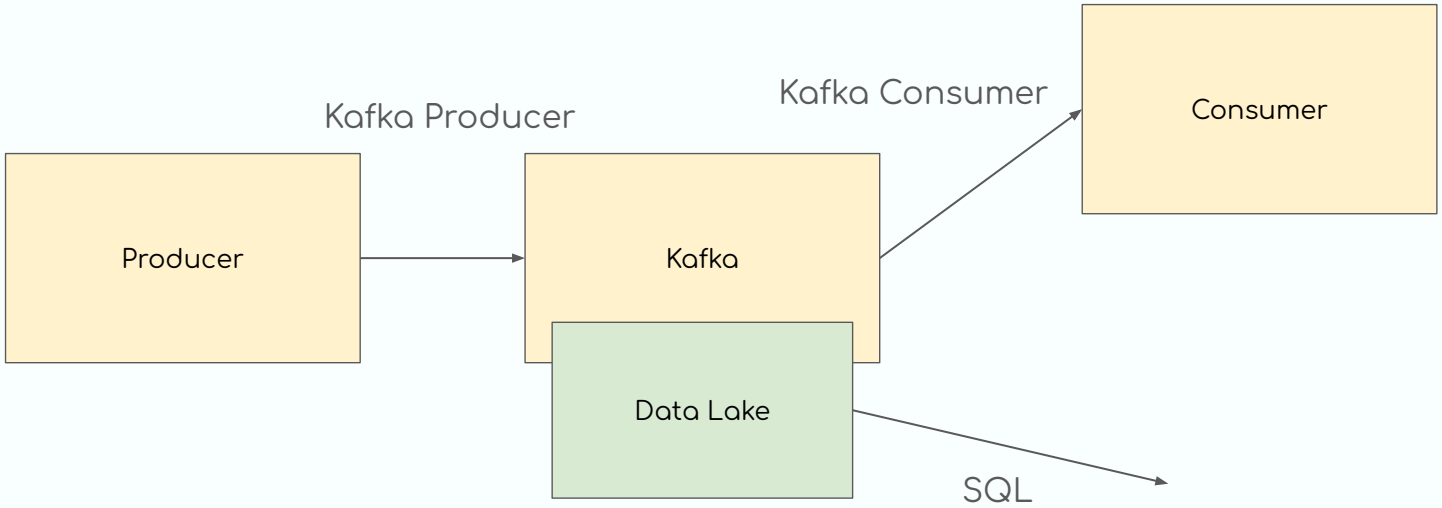
Who defines validity?

Where This Is All Going

Copy/Move Data To Sink



Data Lake As Storage



Data Lake As A Sink

Independent copies of data

Pull-based model

Accept first, clean later

Allows for ETL

Data Lake As Storage

Single copy of data (if it doesn't need to be transformed)

Push-based model

Make producers aware of issues

Relies on ELT (or internal ETL)

More Questions Than Answers

Transforming data is inescapable

Fine-grained decisions are better than coarse-grained decisions

Consider the data owner

Kamal Aboul-Hosn

QUESTIONS?

Linkedin: kamalaboulhosn
Cloud Pub/Sub
Managed Service For Apache Kafka

BEAM SUMMIT NYC 2025