From batch to streaming to both

Herman Schaaf, Principal Software Engineer

Skyscanner

A Story

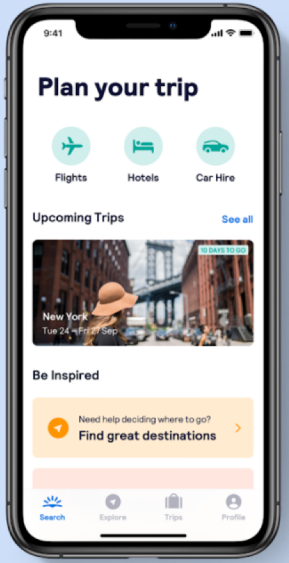
From batch to streaming to both *to streaming...again?*

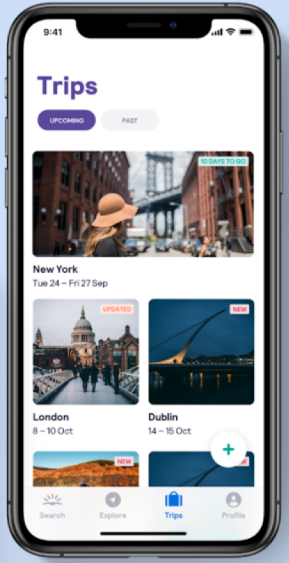
Herman Schaaf, Principal Software Engineer

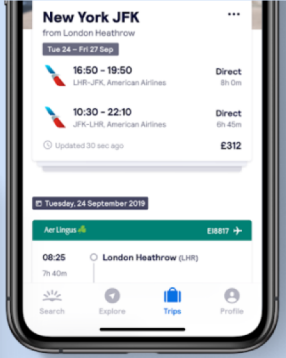
Data Platform Tribe

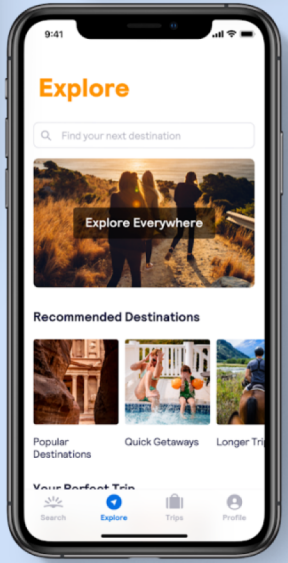
Skyscanner

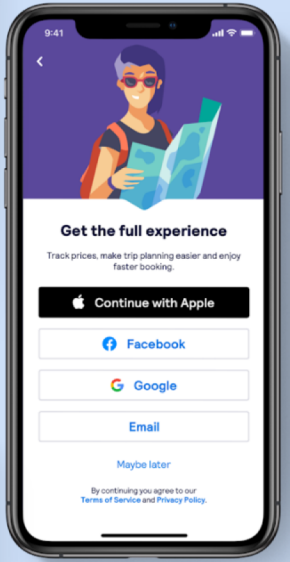
@ironzeb

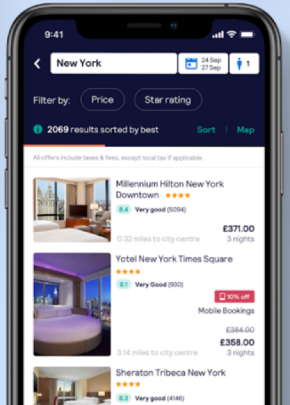

9:41
Plan your trip
Flights
Hotels
Car Hire
Upcoming Trips
See all
New York
Tue 24 – Fri 27 Sep
Be Inspired
Need help deciding where to go?
Find great destinations
Search
Explore
Trips
Profile

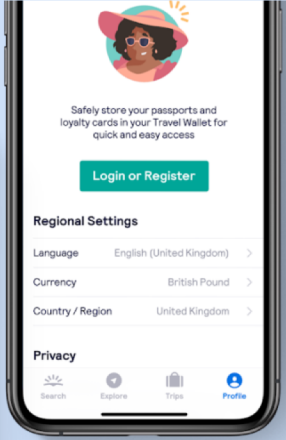

9:41
Trips
UPCOMING
PAST
New York
Tue 24 – Fri 27 Sep
London
8 – 10 Oct
Dublin
14 – 15 Oct
Search
Explore
Trips
Profile

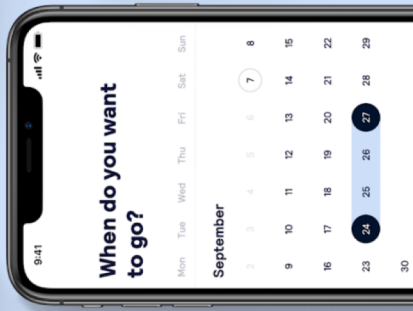

New York JFK
from London Heathrow
Tue 24 – Fri 27 Sep
16:50 – 19:50
LHR–JFK, American Airlines
Direct
8h 0m
10:30 – 22:10
JFK–LHR, American Airlines
Direct
6h 45m
£312
Tuesday, 24 September 2019
Aer Lingus
08:25
London Heathrow (LHR)
7h 40m
Search
Explore
Trips
Profile

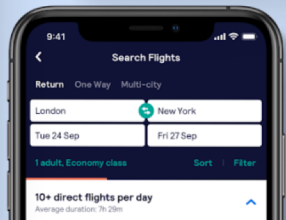

9:41
Explore
Find your next destination
Explore Everywhere
Recommended Destinations
Popular Destinations
Quick Getaways
Search
Explore
Trips
Profile


9:41
Get the full experience
Track prices, make trip planning easier and enjoy faster booking.
Continue with Apple
Facebook
Google
Email
Maybe later
By continuing you agree to our
Terms of Service and Privacy Policy.


9:41
New York
Filter by:
Price
Star rating
2069 results sorted by best
Sort
Map
Millennium Hilton New York Downtown
£371.00
3 nights
Yotel New York Times Square
Mobile Bookings
£358.00
3 nights
Sheraton Tribeca New York


Safely store your passports and loyalty cards in your Travel Wallet for quick and easy access
Login or Register
Regional Settings
Language
English (United Kingdom)
Currency
British Pound
Country / Region
United Kingdom
Privacy
Search
Explore
Trips
Profile

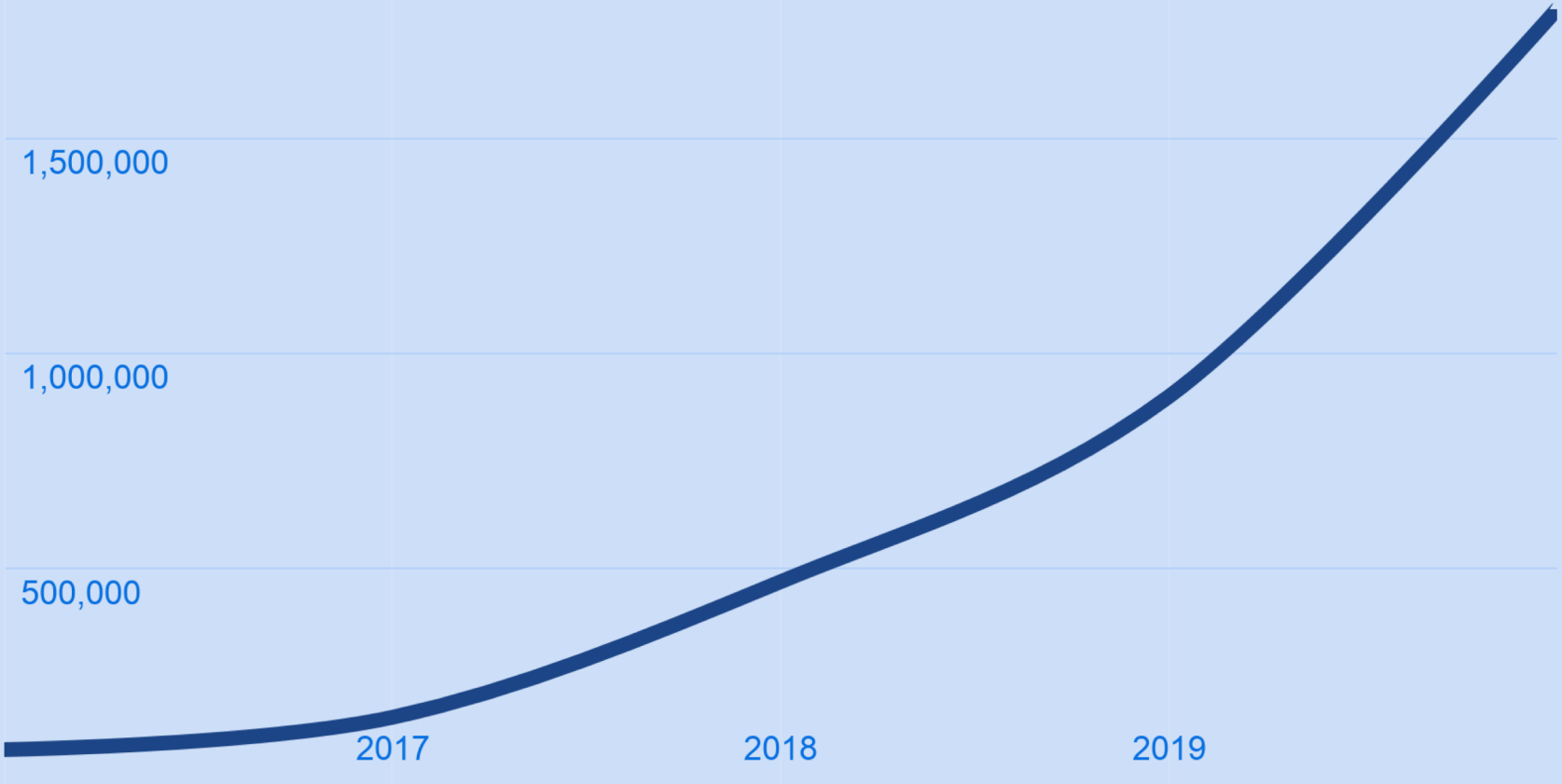

When do you want to go?
September
Sun
Mon
Tue
Wed
Thu
Fri
Sat


9:41
Search Flights
Return
One Way
Multi-city
London
New York
Tue 24 Sep
Fri 27 Sep
1 adult, Economy class
Sort
Filter
10+ direct flights per day

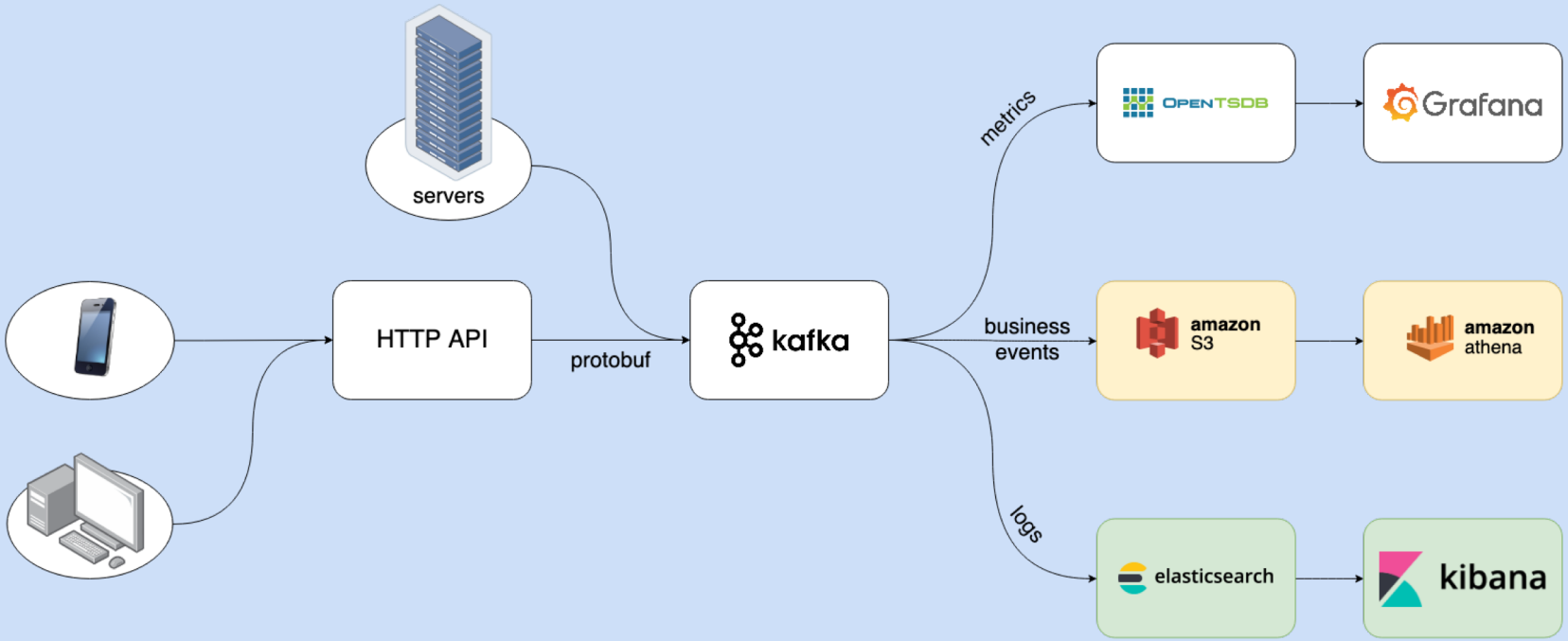
Skyscanner


Global travel company with local expertise
100
million peak monthly active users
30+
languages available
100
million app downloads
1.2k+
partners

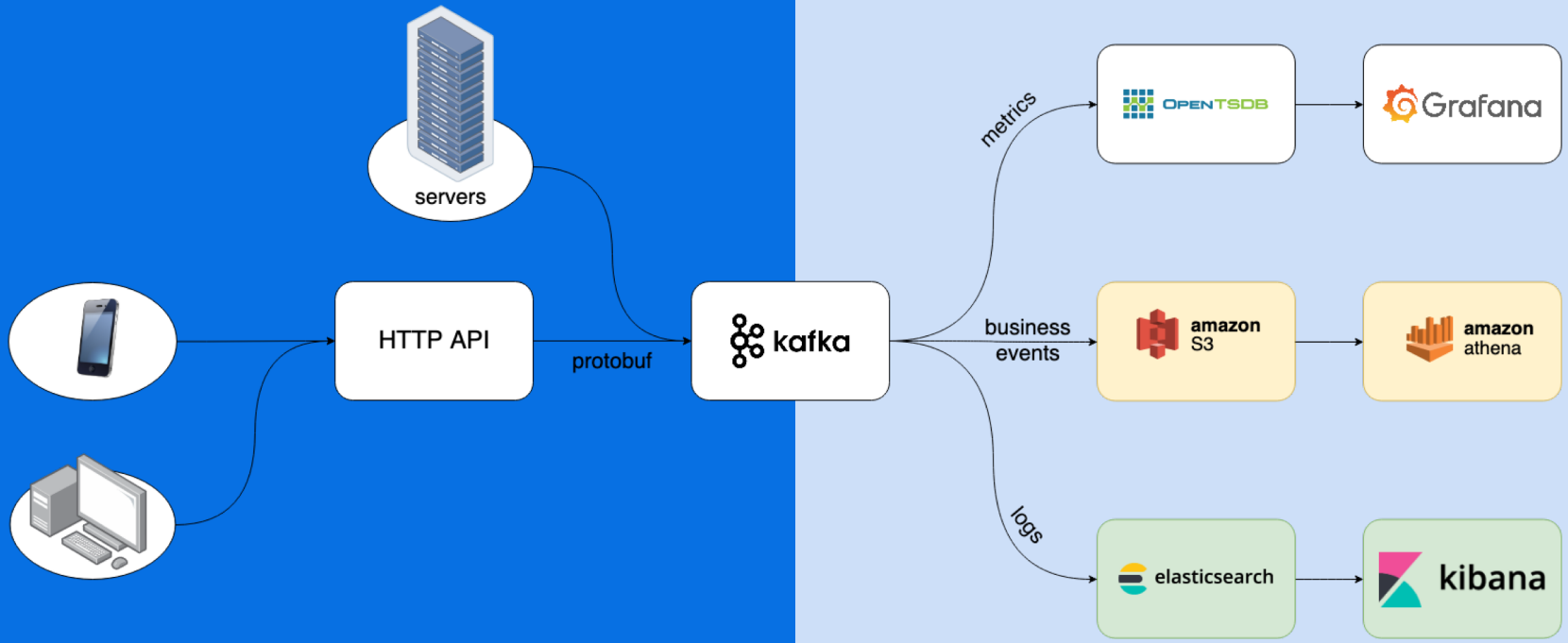
Messages per second through the Data Platform



“The Cube”

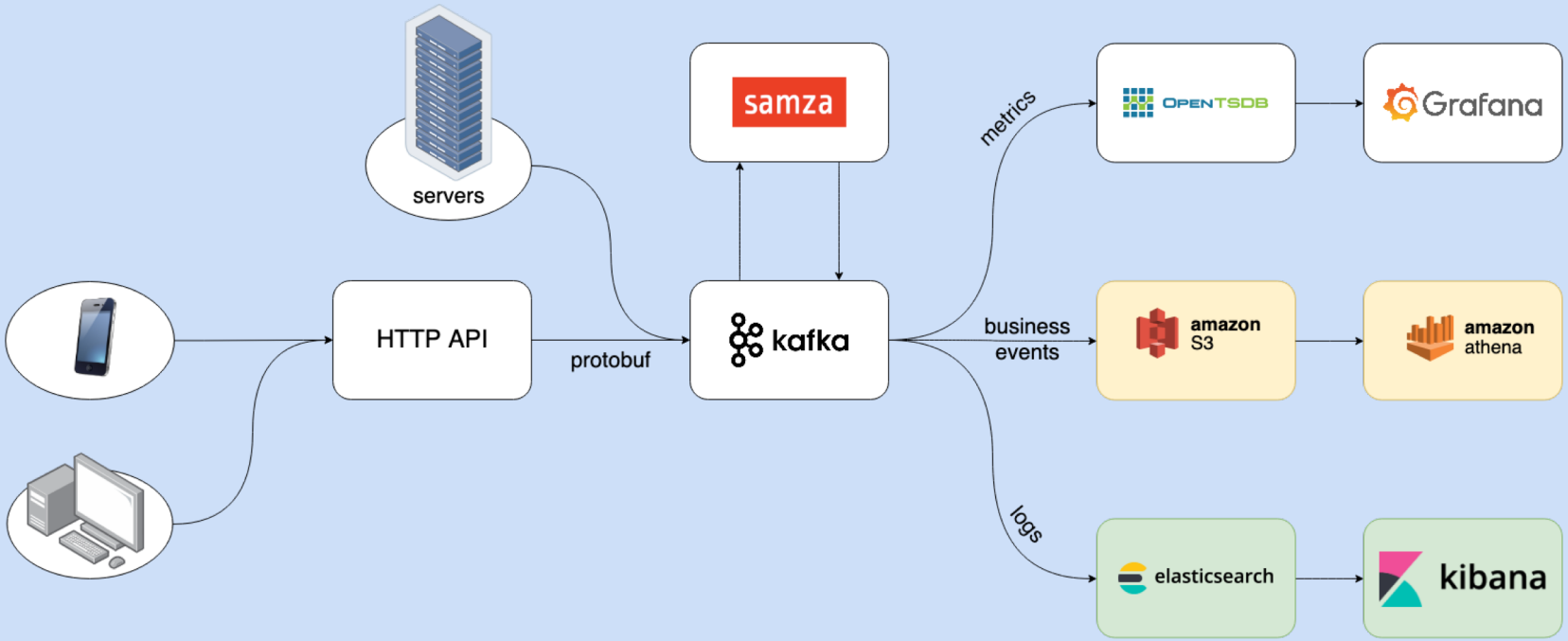


From batch to streaming

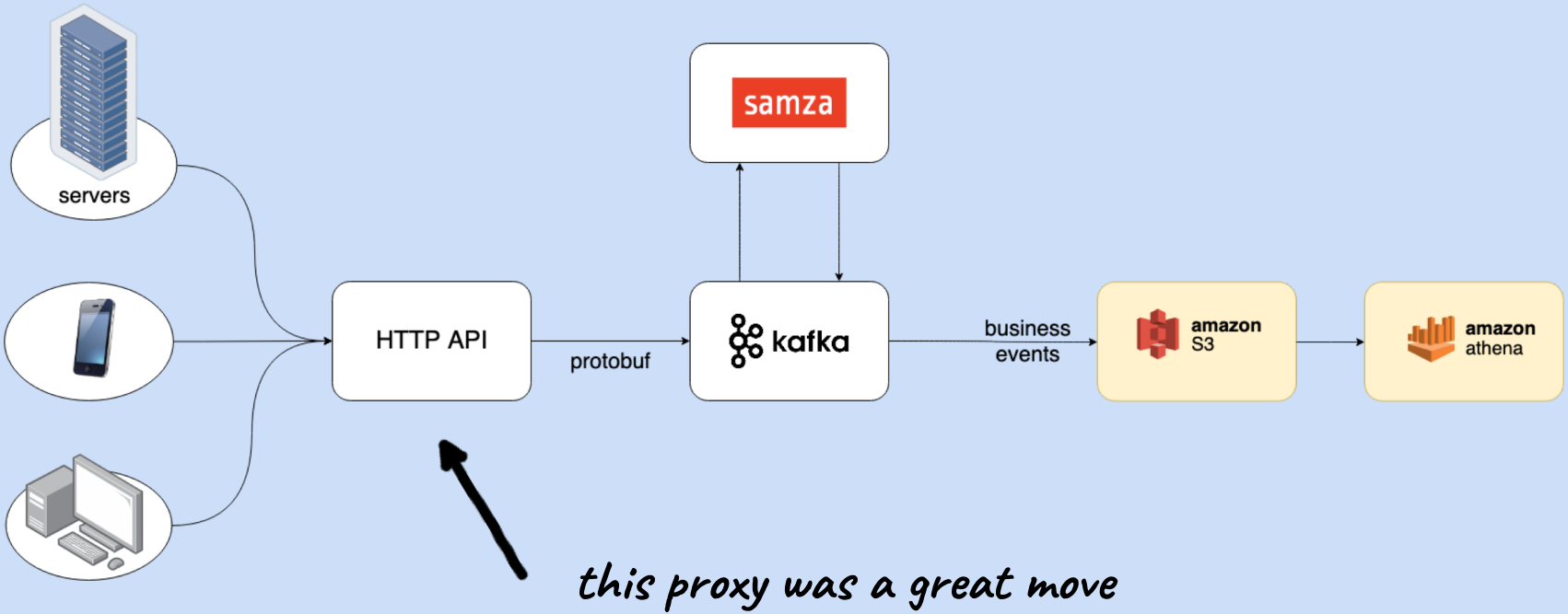


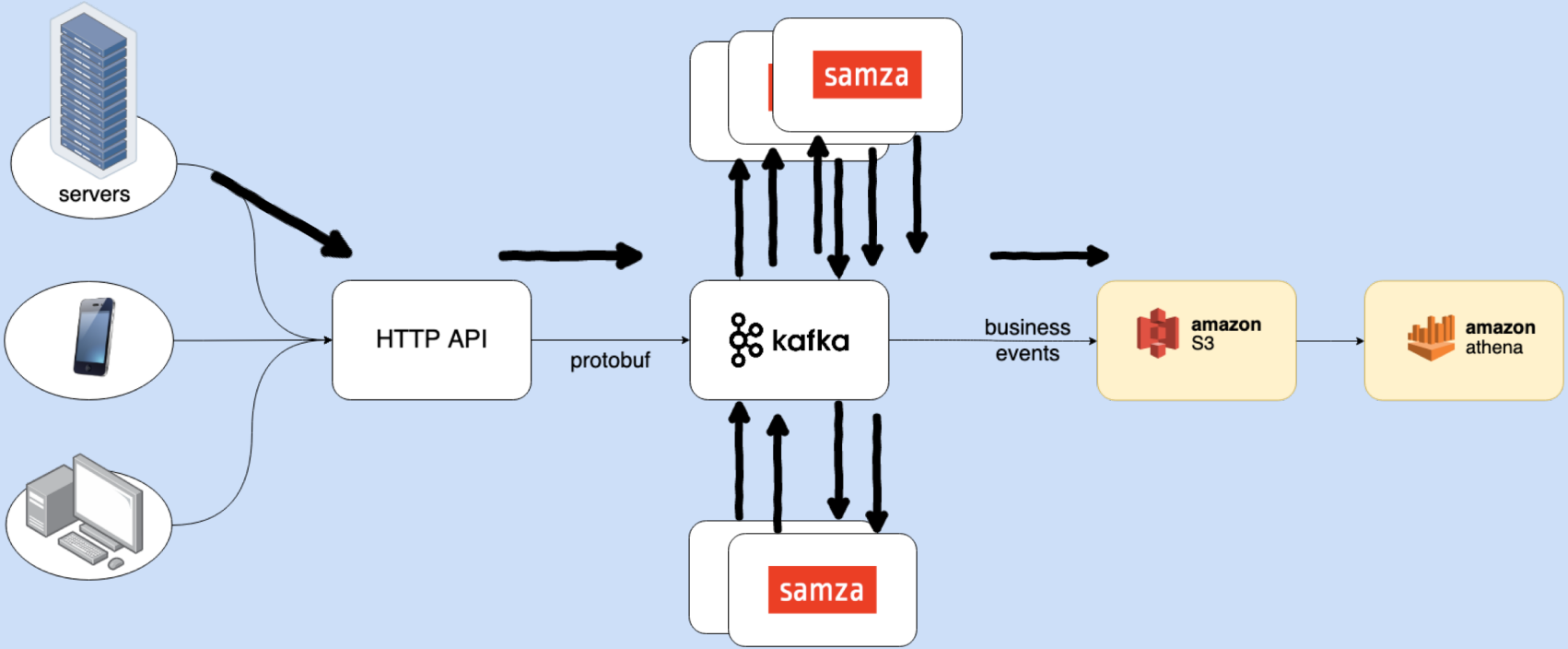
Decoupled

Unified



The "Single Unified Log"


servers
HTTP API
protobuf
samza
kafka
business
events
amazon
S3
amazon
athena
this proxy was a great move



The Single Unified Log

Lesson 1:

Conway's Law is true for data platforms

"Organizations which design **data platforms** are constrained to produce designs which are copies of their communication structures"

Being self-serve is good

...but then metadata is critical

So let’s talk about metadata

prod.identity-service.audit.identity.AuditMessage

prod.flyingcircus.applog.applog.Message

A simple convention

<prod|sandbox|local>.<service-name>.<event-name>.<schema>.<message>

What does it mean?
Who owns it?
What does it contain?
Where does it come from?

Descriptive

we had some of this

How does it relate to other data sets?
How is it organized?
How is it sorted / partitioned?

~~Structural~~

Some, from using protobuf schemas

How far does it date back?
How frequently is it updated?
How large is it?
How complete is it?

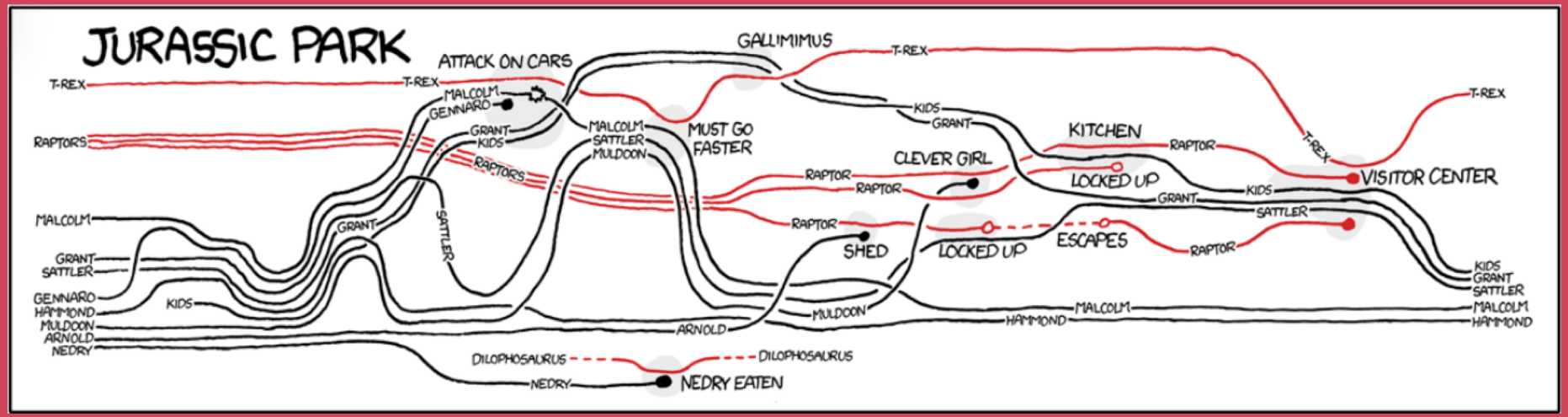
~~Administrative~~

nope.

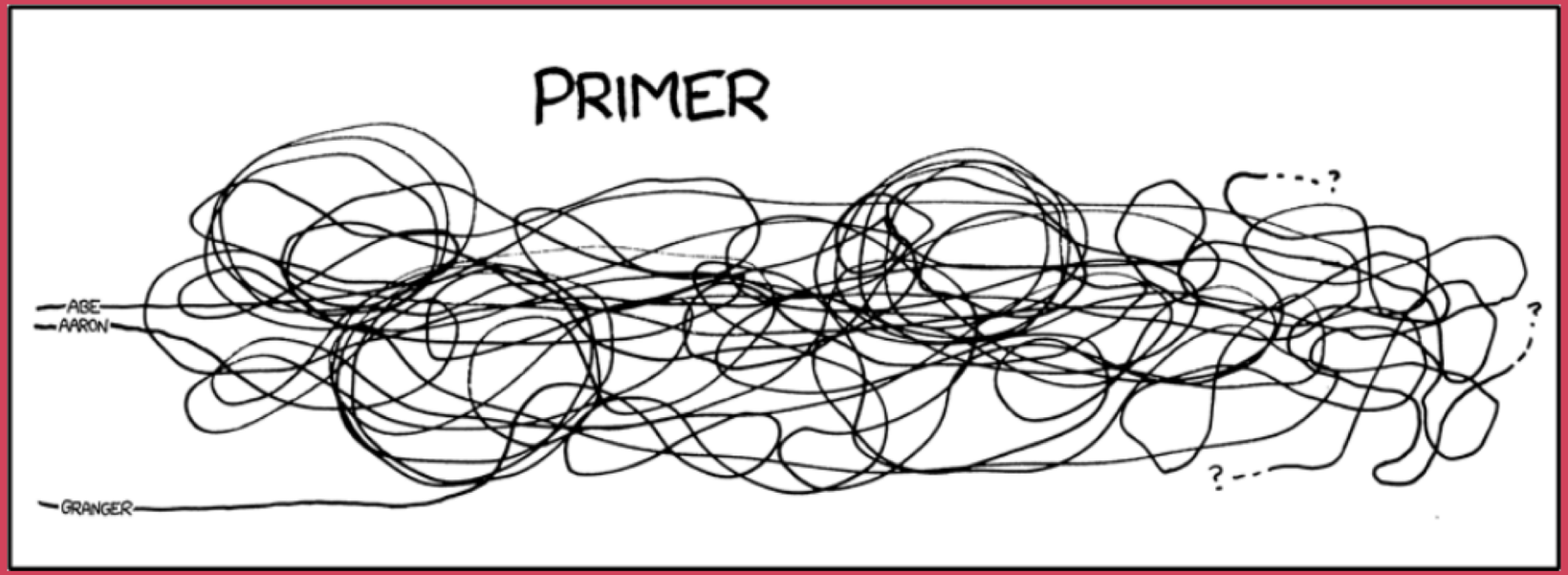
”Creating a Data-driven Enterprise with DataOps”
(Ashish Thusoo.& Joydeep Sen Sarma)

Lesson 2:
Metadata is Critical

- Especially relationships
- Ideally automated
- Ideally from the start
- Tools like Schema Registry are a start, but not the full solution



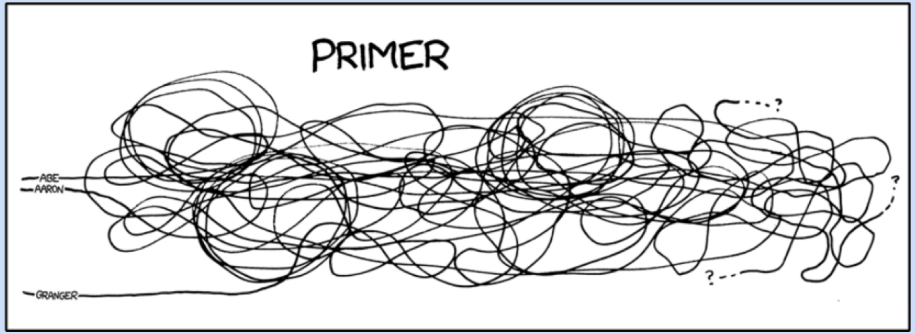
https://xkcd.com/657/

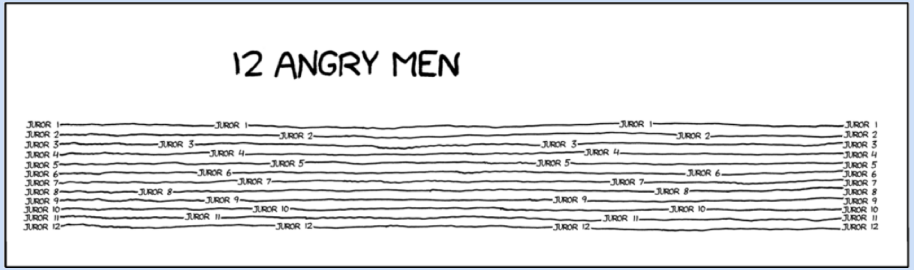


https://xkcd.com/657/

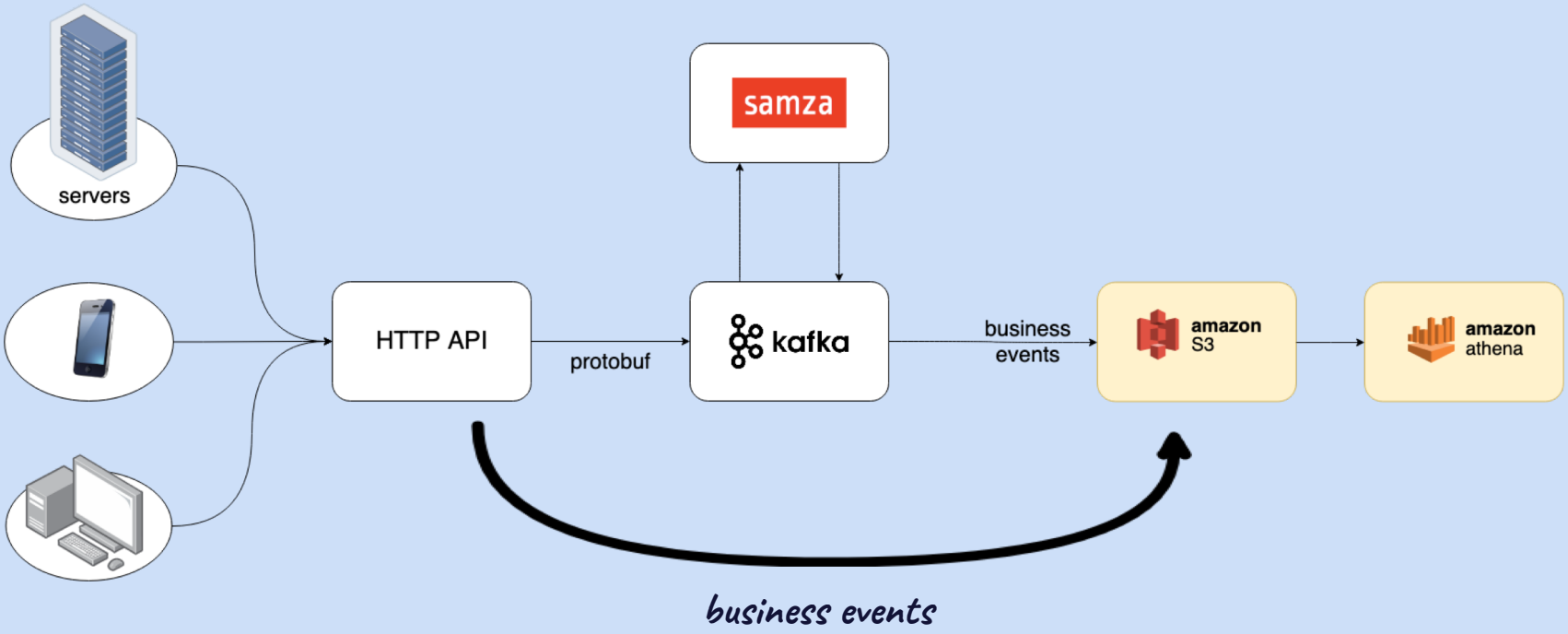
Lesson 3:

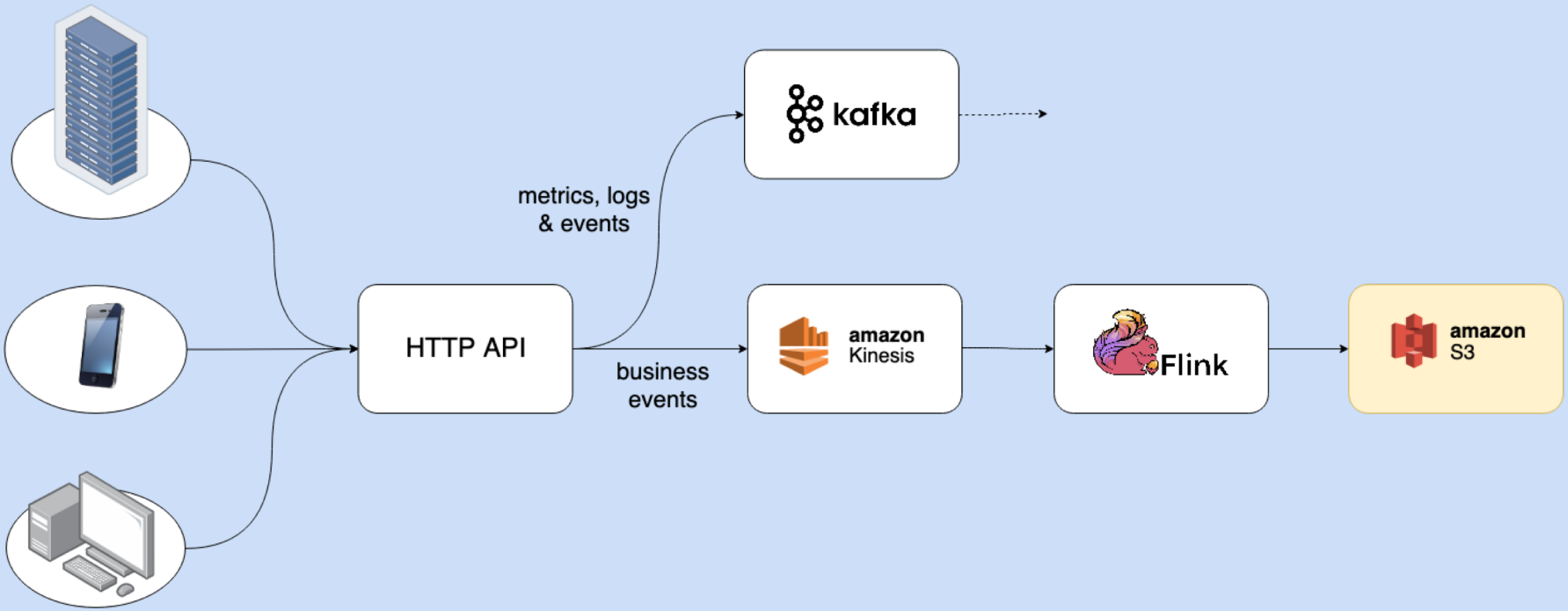
Data Engineers Control the Plot Line



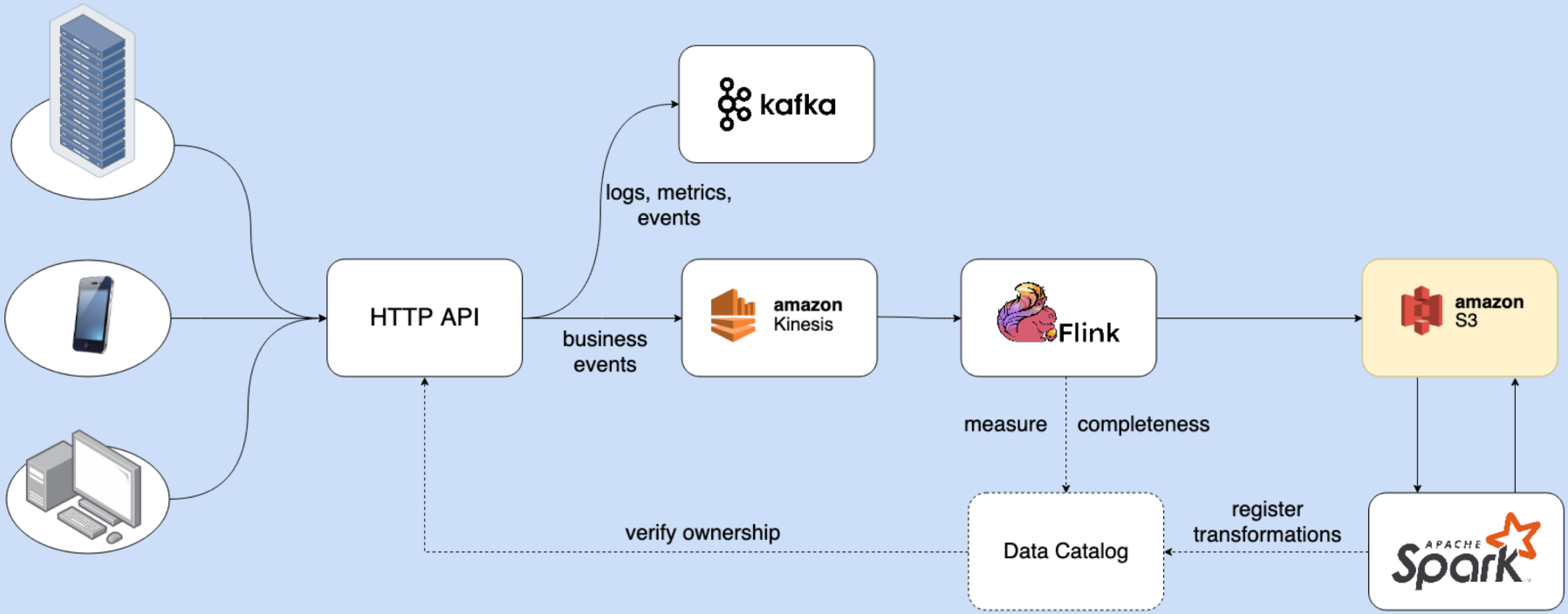


https://xkcd.com/657/


servers
HTTP API
protobuf
samza
kafka
business
events
amazon
S3
amazon
athena
business events



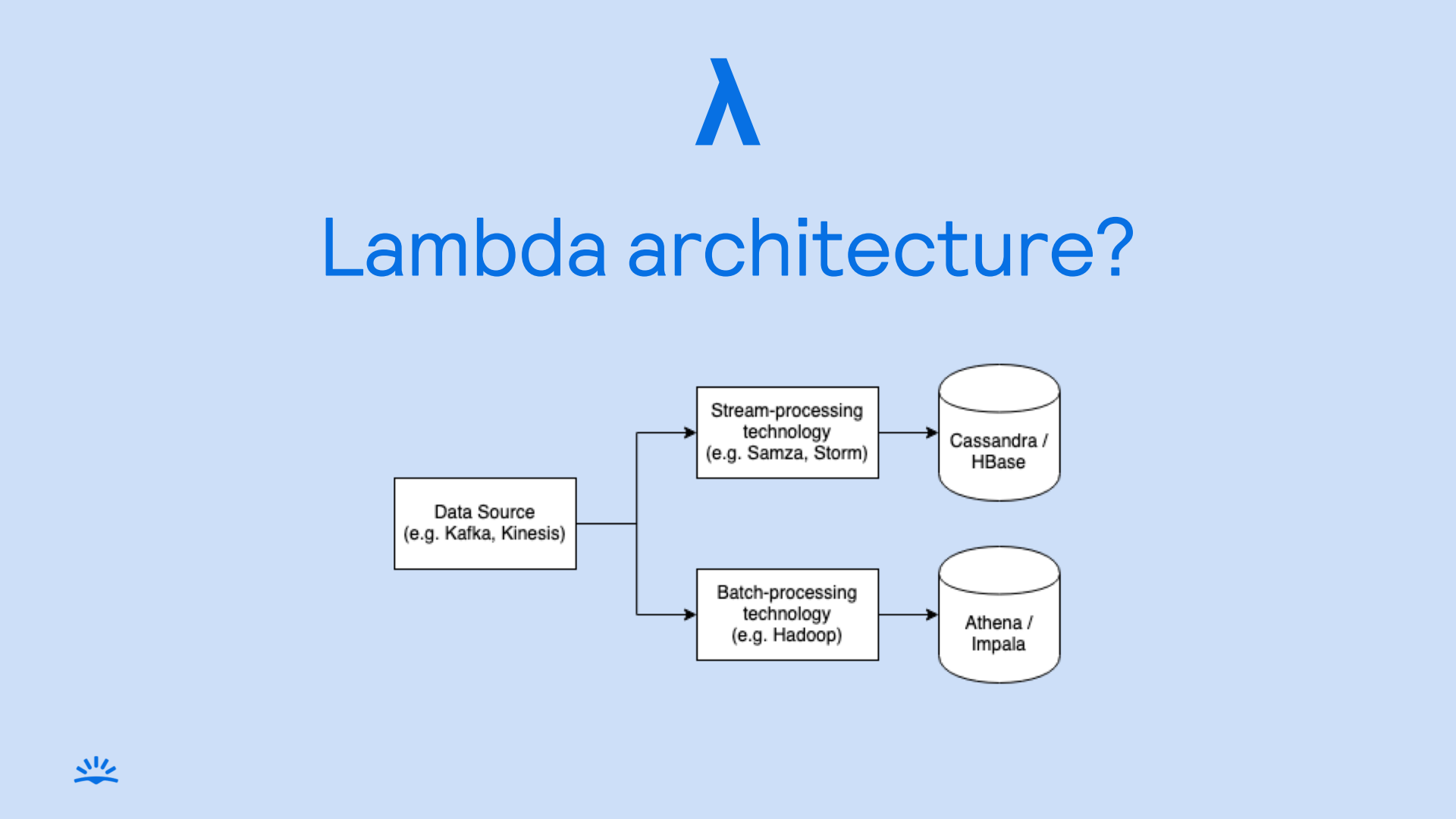
From streaming to both

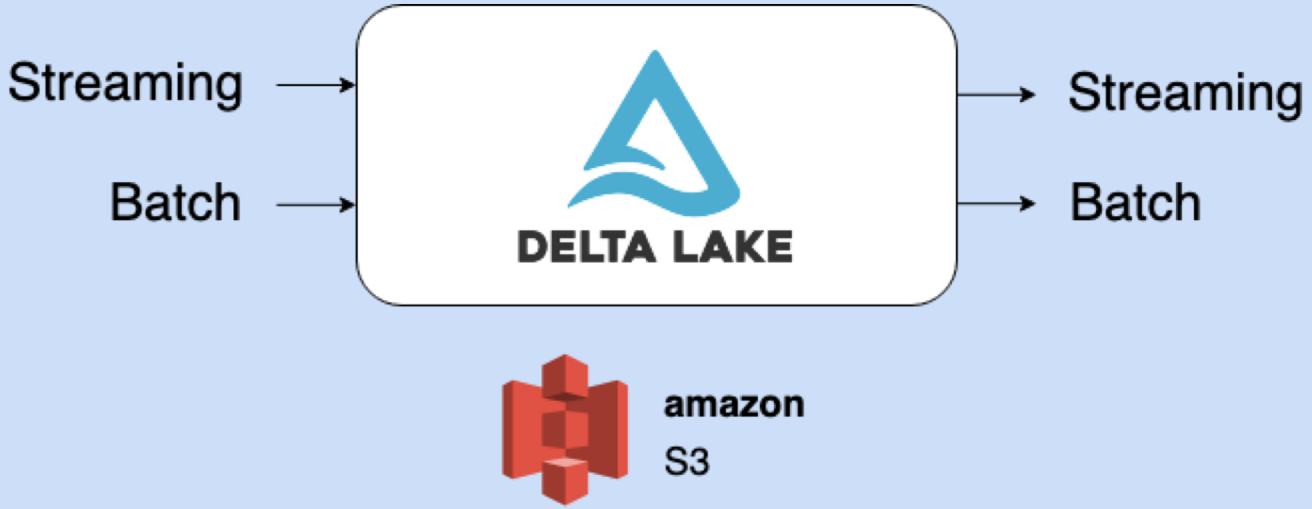

kafka
logs, metrics,
events
HTTP API
business
events
amazon
Kinesis
Flink
amazon
S3
measure
completeness
verify ownership
Data Catalog
register
transformations
APACHE
Spark

Lesson 4:

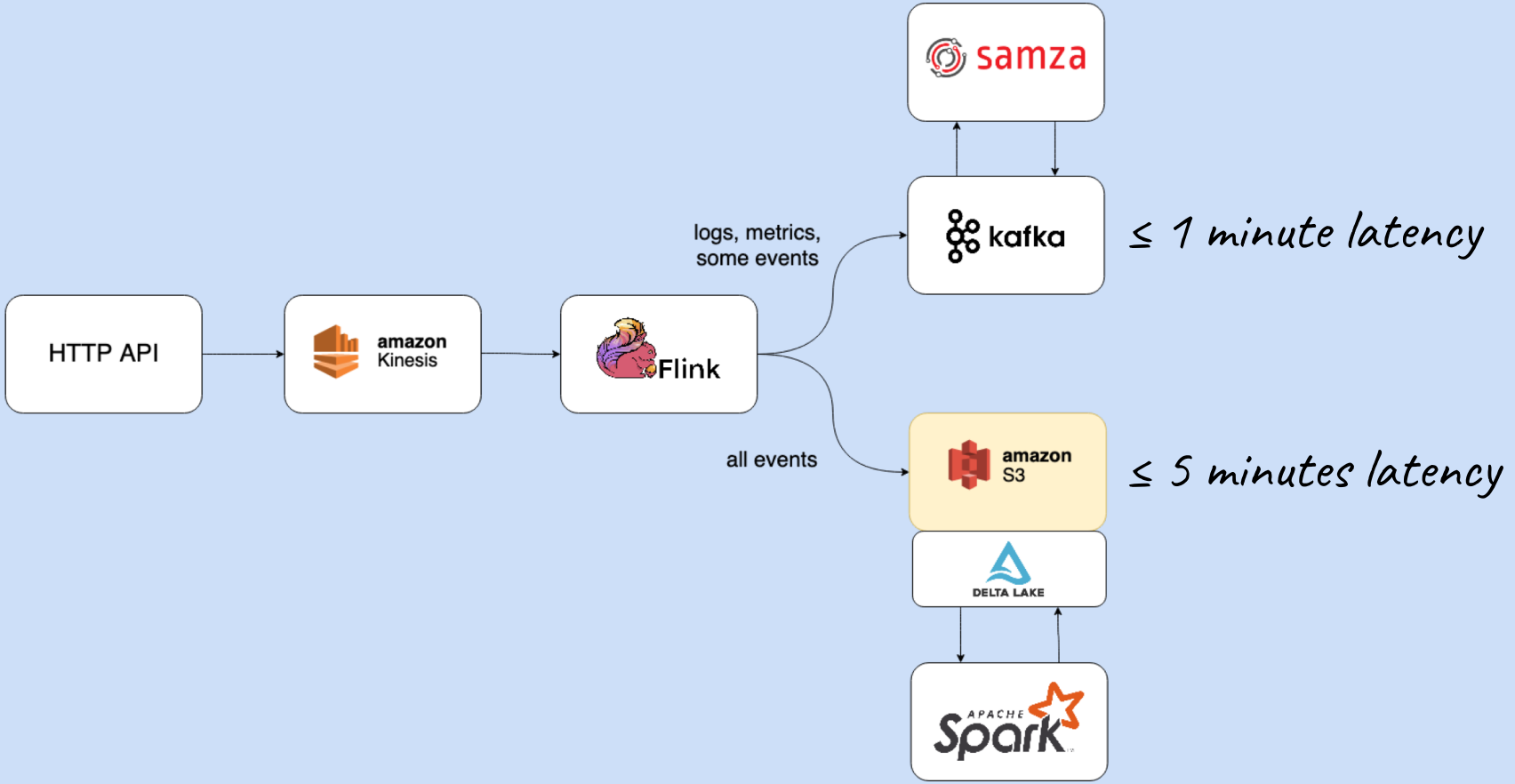
Repeatability is important

- Streams have to choose between replays and accepting errors as permanent
- Replays can be complicated
- Batch processing can be done again any time
- Going straight to the archive in small batches gets benefits of both


λ
Lambda architecture?
Data Source
(e.g. Kafka, Kinesis)
Stream-processing
technology
(e.g. Samza, Storm)
Batch-processing
technology
(e.g. Hadoop)
Cassandra /
HBase
Athena /
Impala



https://delta.io/


samza
kafka
≤ 1 minute latency
logs, metrics,
some events
HTTP API
amazon
Kinesis
Flink
all events
amazon
S3
≤ 5 minutes latency
DELTA LAKE
APACHE
Spark

Key Takeaways

- Conway's Law is true for data platforms
- Metadata is Critical
- Data Engineers Control the Plot Line
- Repeatability is important
- Delta Lake + Structured Streaming can bring the benefits of streaming to batch

Thanks

Contact
If you have any questions regarding Skyscanner please contact:

Herman Schaaf
herman.schaaf@skyscanner.net

Herman Schaaf
@ironzeb

Skyscanner