From batch to streaming to both

Herman Schaaf, Senior Software Engineer

Skyscanner

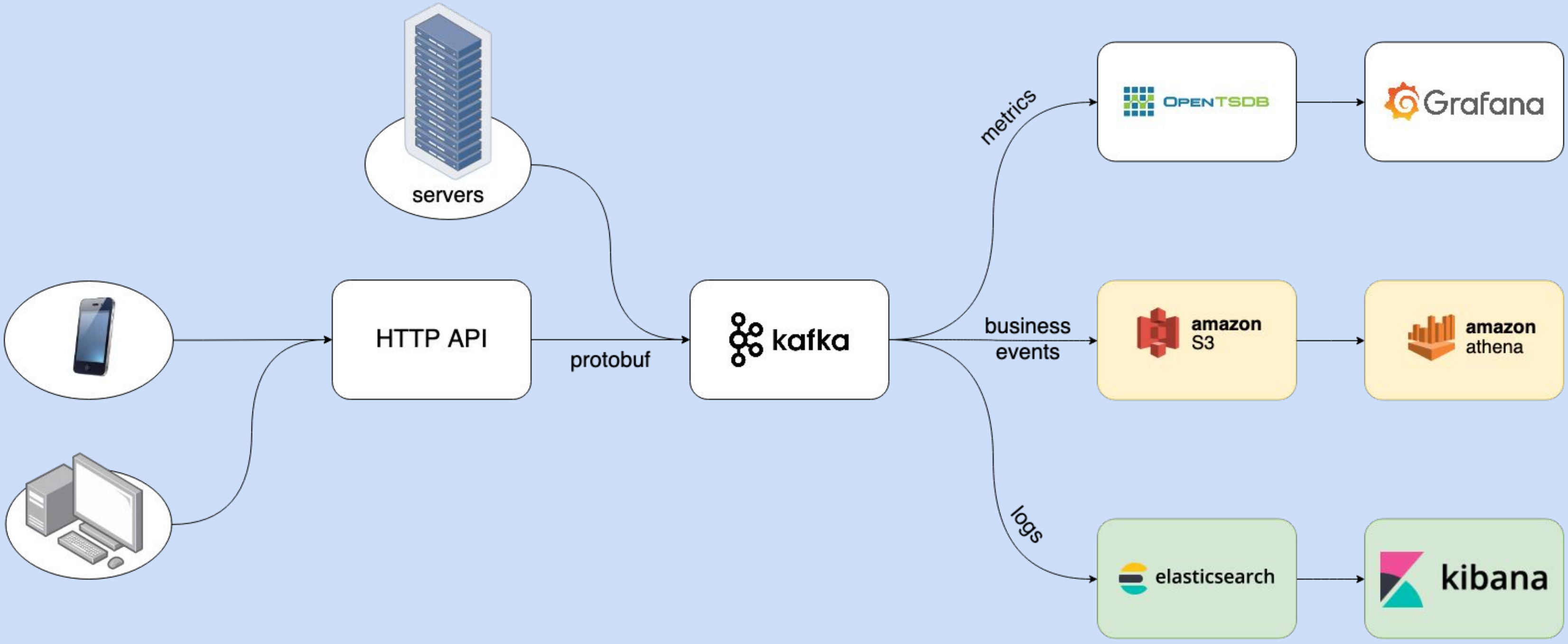
@ironzeb

A Story

About me

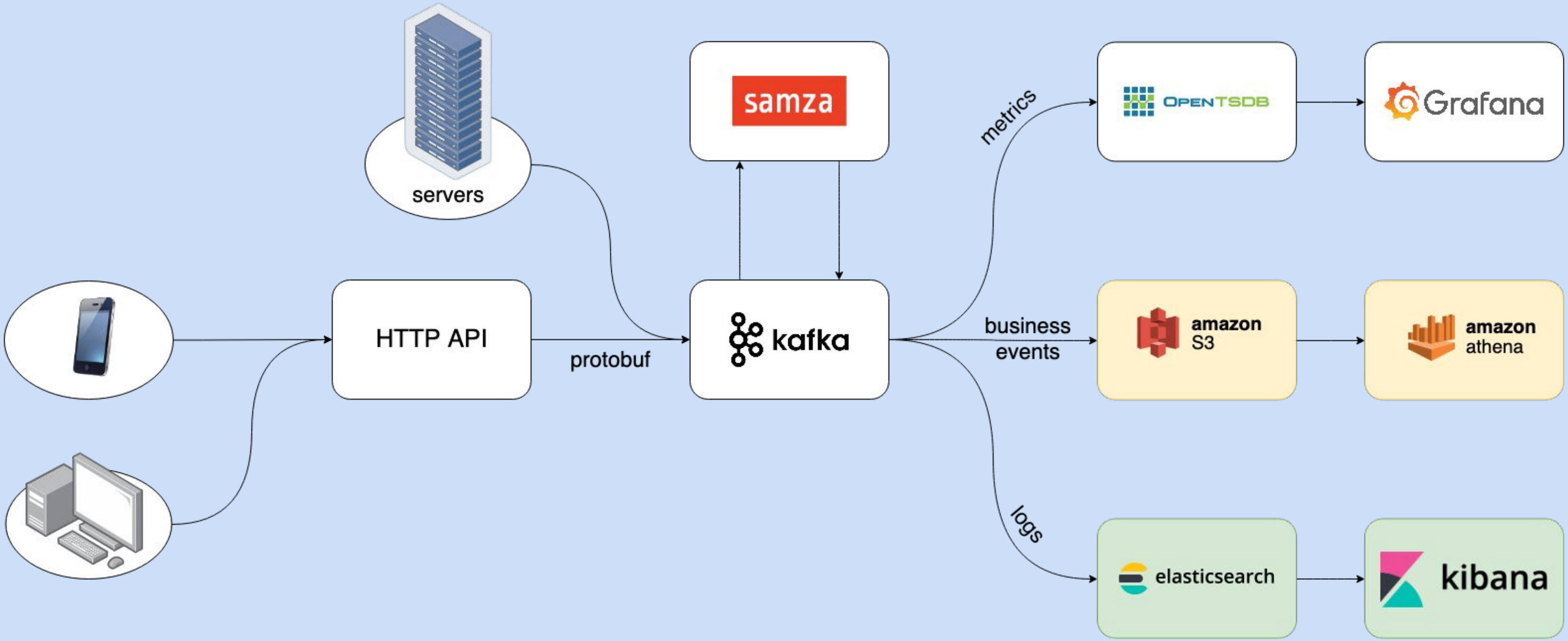
Herman Schaaf, Senior Software Engineer
Data Platform Tribe

“The Cube”

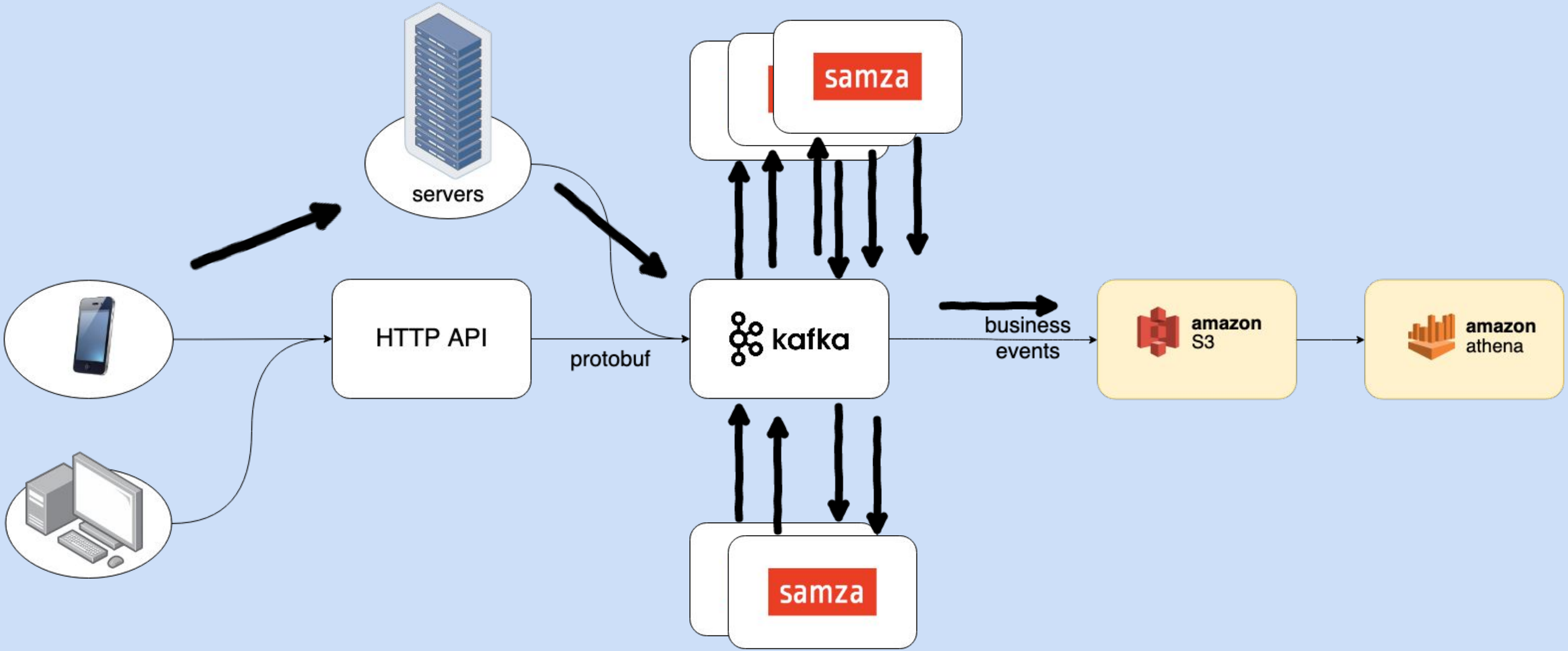


From batch to streaming

@ironzeb



The Single Unified Log



The Single Unified Log

@ironzeb

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.AuditLog.identity.AuditMessage

prod.flyingcircus.applog.applog.Message

prod.raccoon_bandit.experiment.bandit.Metric

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?

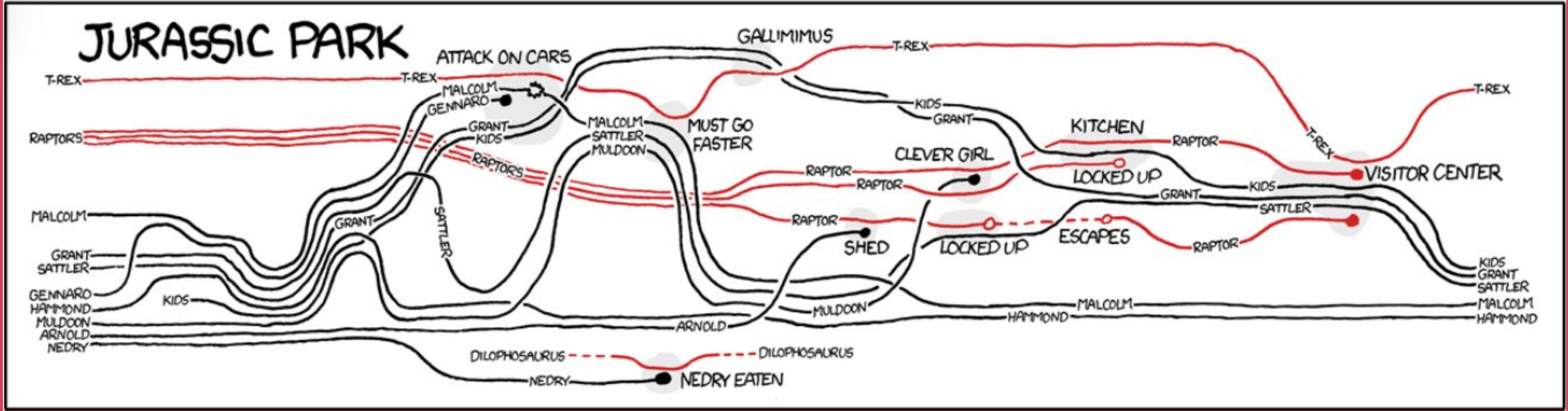
~~Administrative~~

nope.

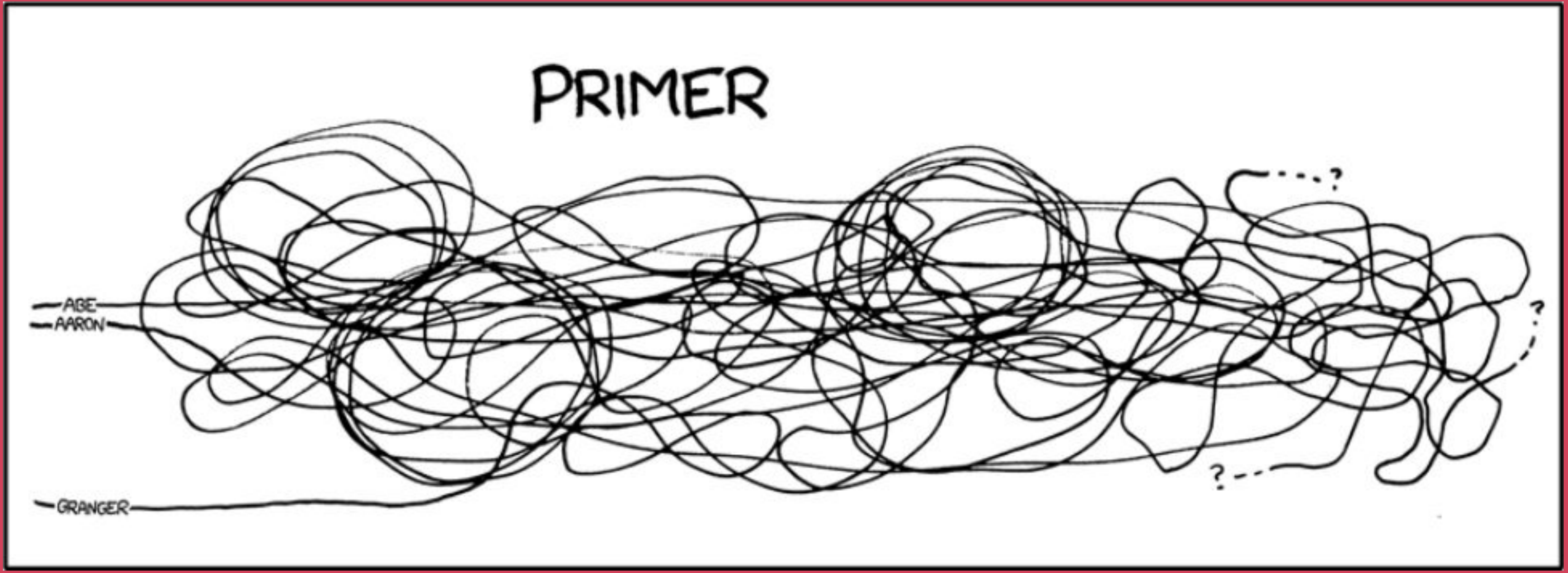
”Creating a Data-driven Enterprise with DataOps”
(Ashish Thusoo.& Joydeep Sen Sarma)

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



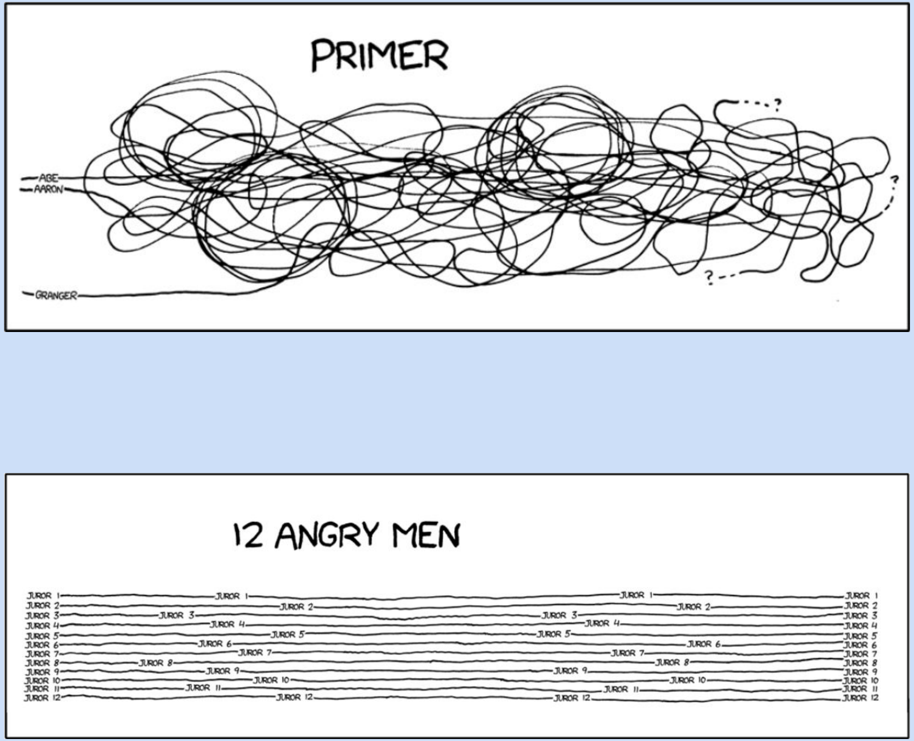
https://xkcd.com/657/



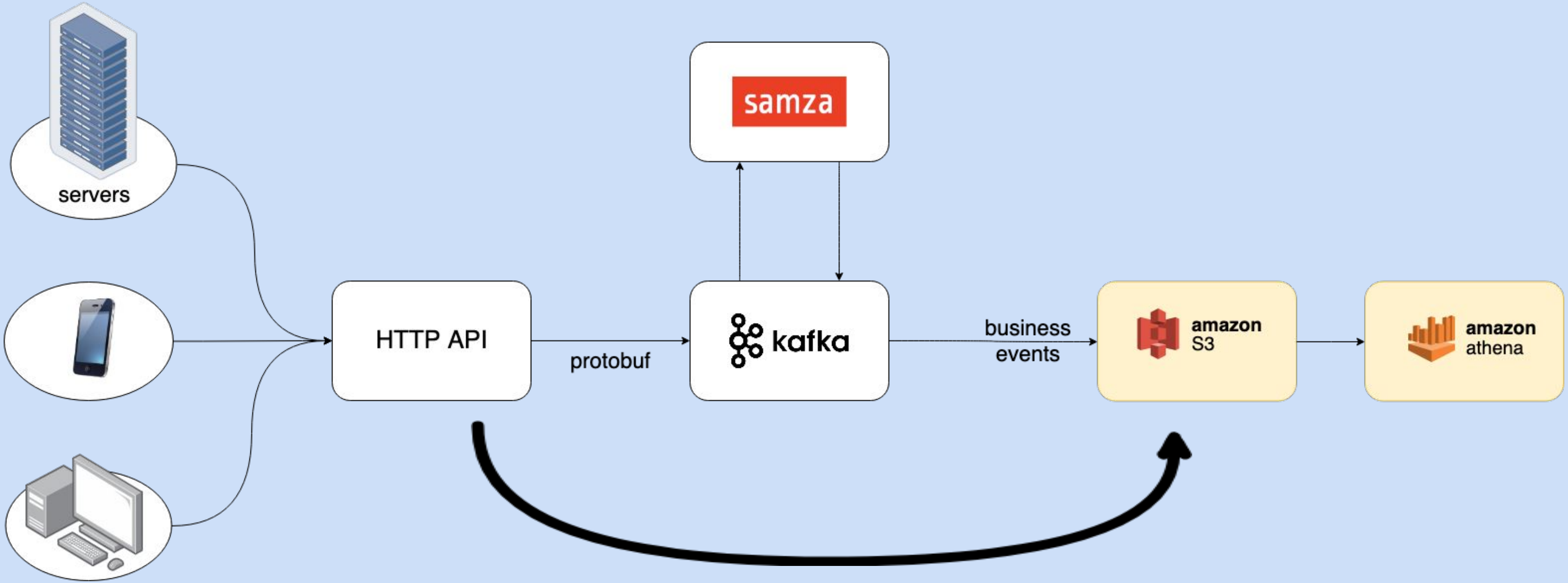
https://xkcd.com/657/

Lesson 3:

Data Engineers Control the Plot Line

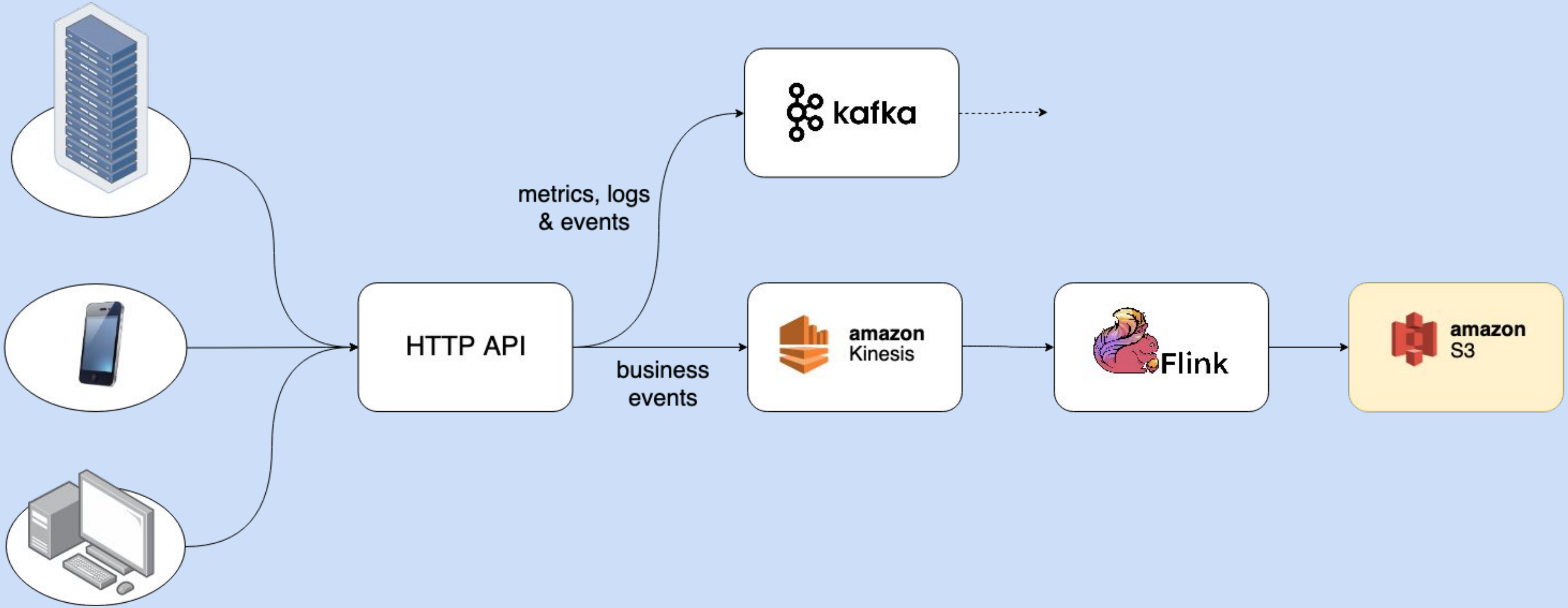


https://xkcd.com/657/

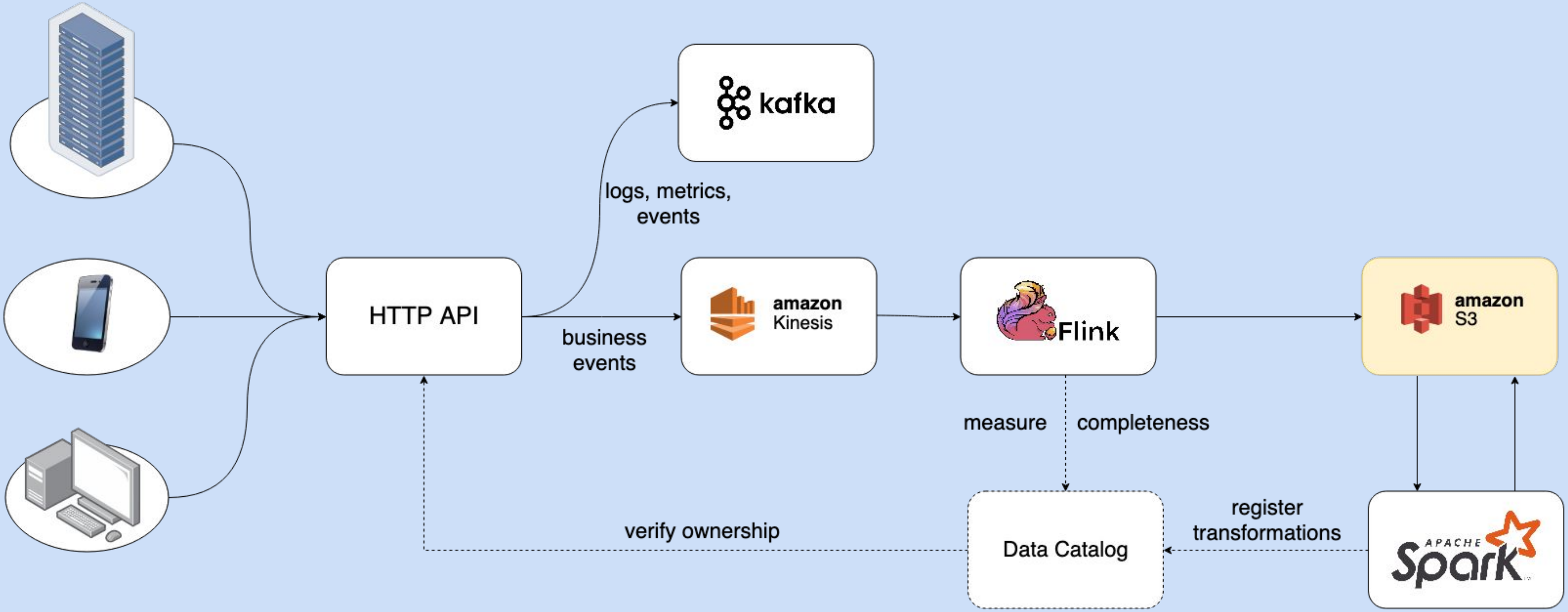


business events

@ironzeb



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

@ironzeb

Lesson 4:
Repeatability is important

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

Key Takeaways

- Conway’s Law is true for data platforms
- Metadata is Critical
- Data Engineers Control the Plot Line
- Repeatability is important

Thanks

Contact
If you have any questions regarding Skyscanner please contact:

Herman Schaaf
herman.schaaf@skyscanner.net

Herman Schaaf
@ironzeb

Skyscanner