KubeCon

North America 2017

Modern Big Data Pipelines Over Kubernetes

Eliran Bivas *iguazio*

@bivas @iguazio

A Tale of 3 Engineers ...

DevOps

Application

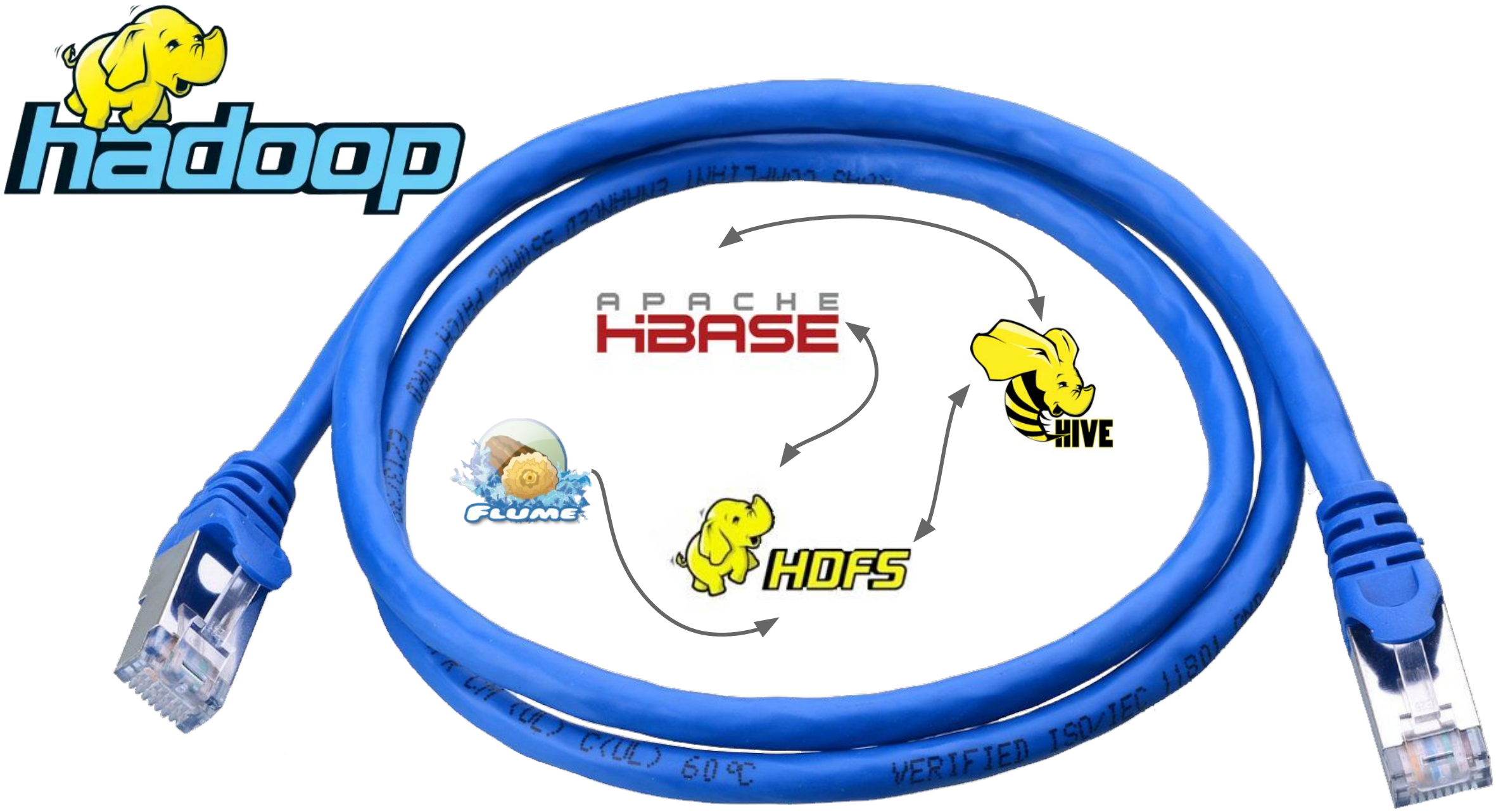
Data

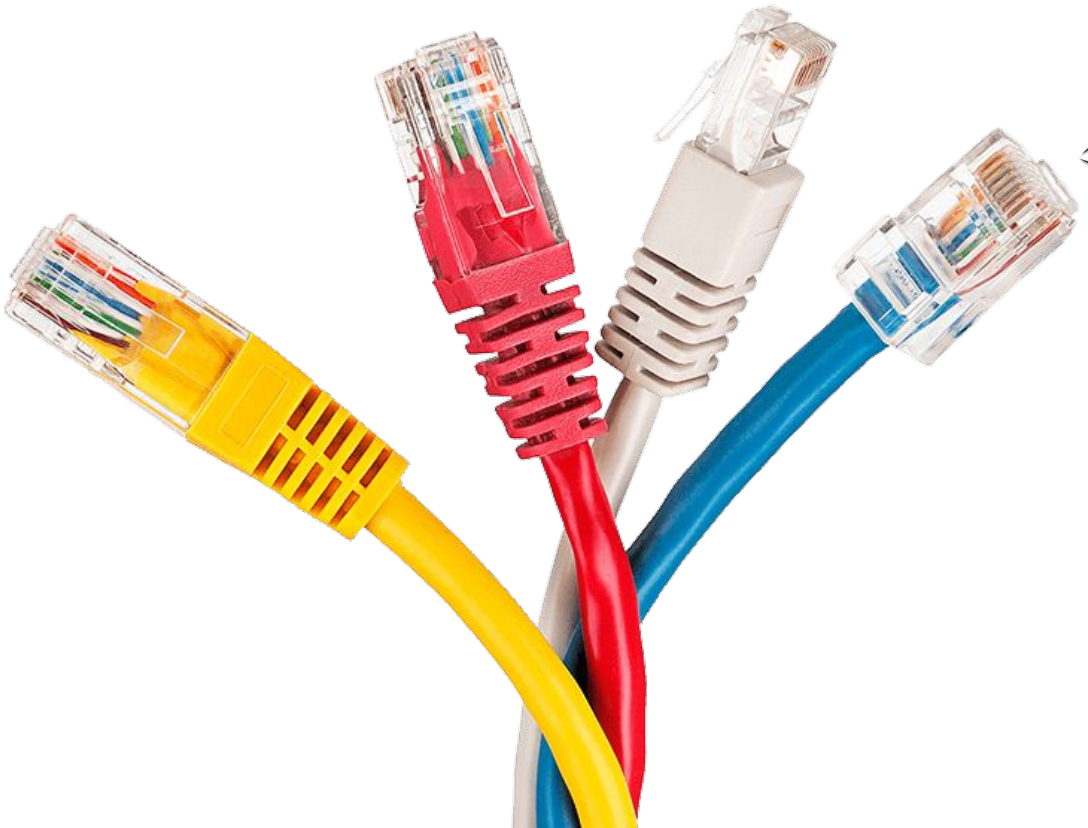
@bivas

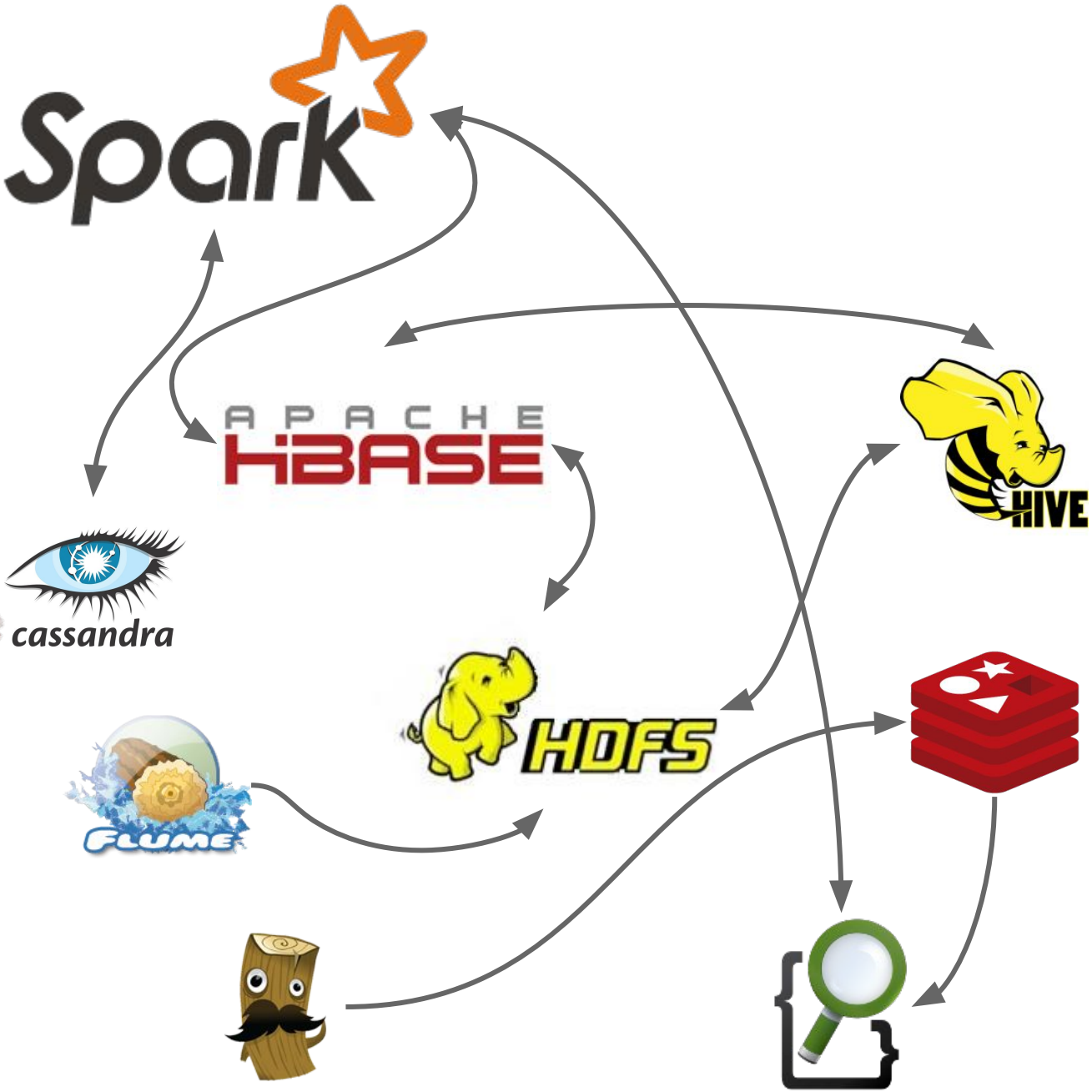
https://github.com/bivas

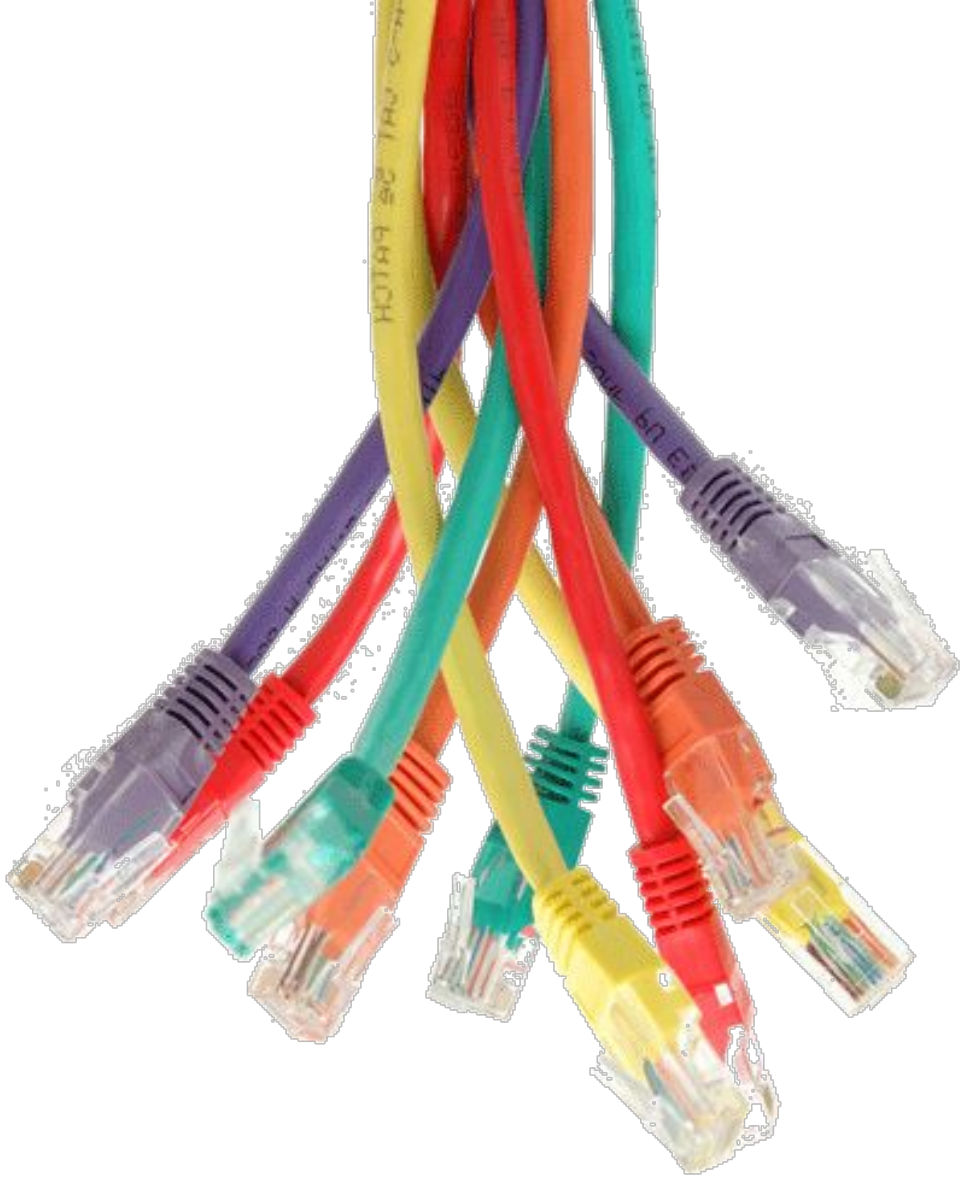
https://linkedin.com/in/bivas

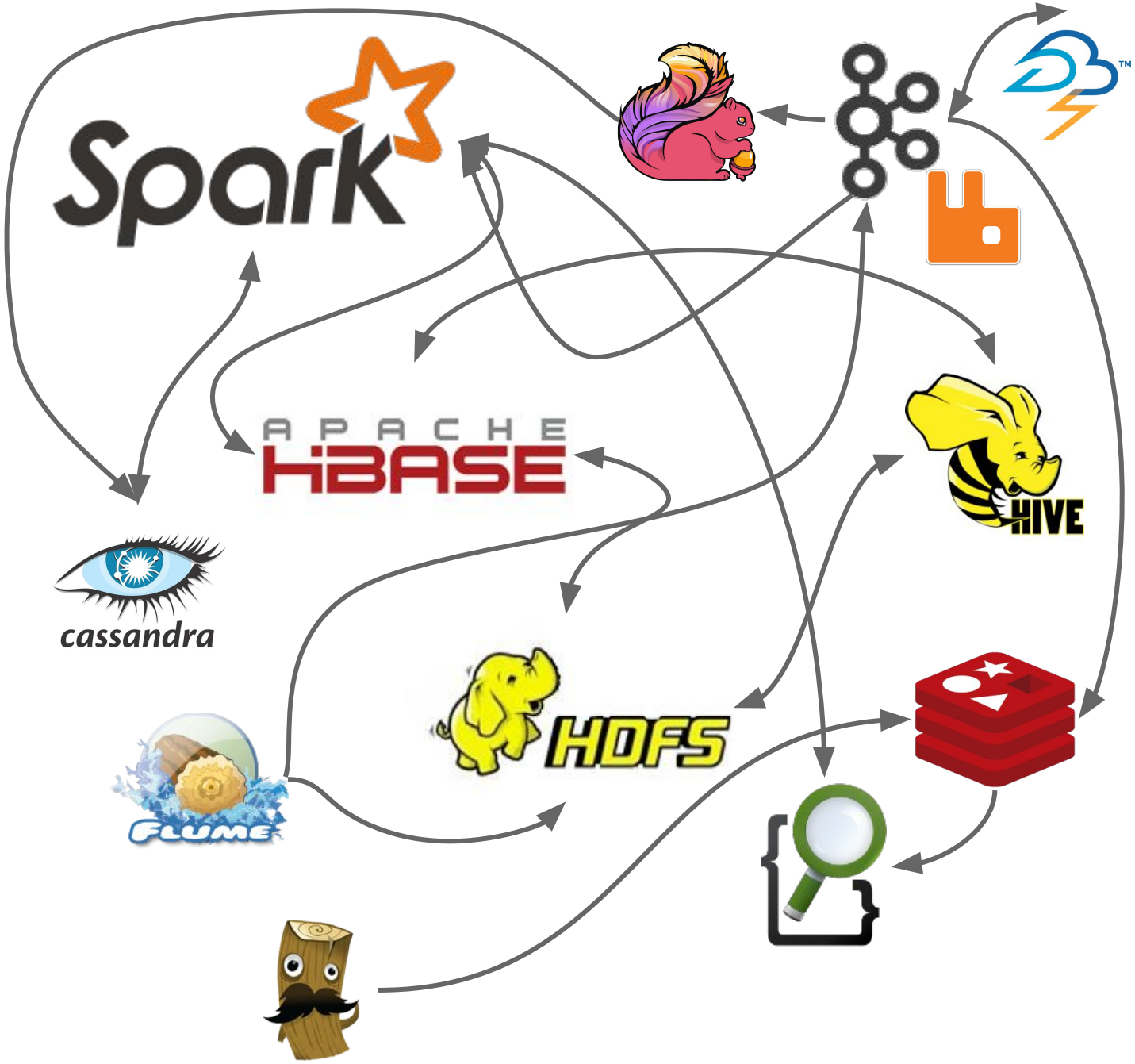
https://www.iguazio.com

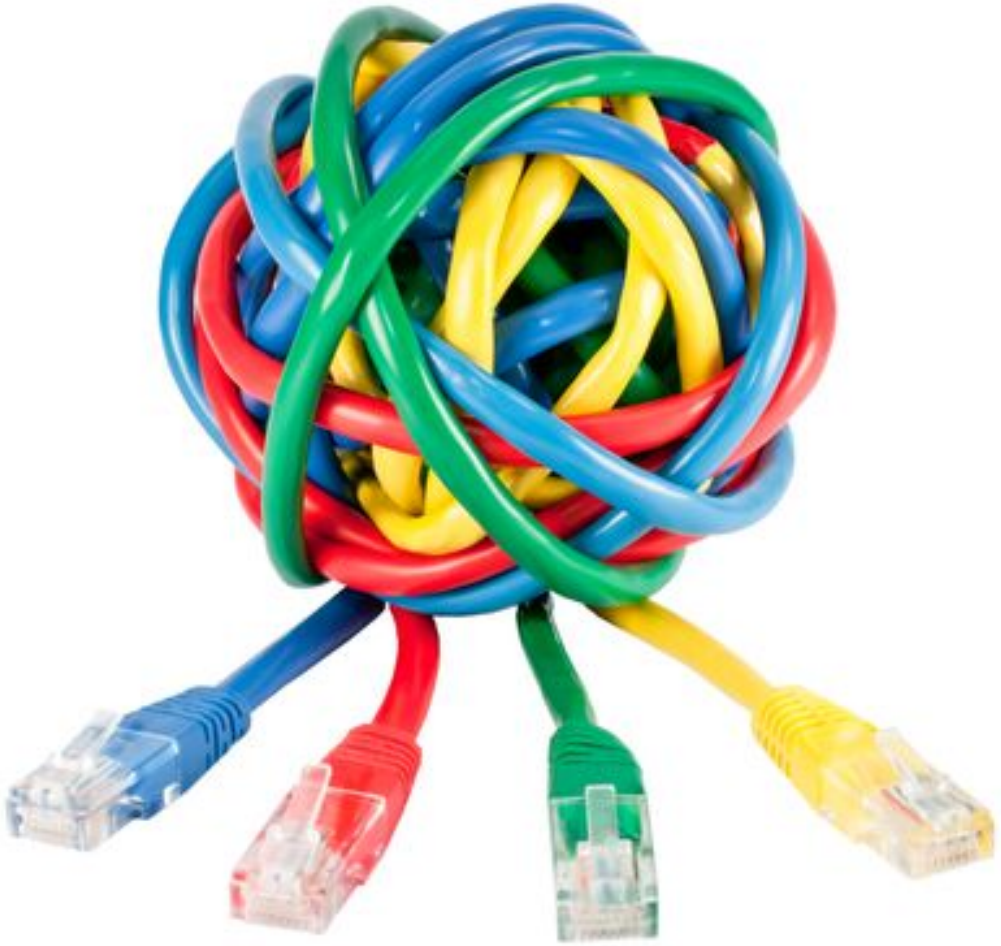

hadoop
APACHE
HBASE
HIVE
FLUME
HDFS
VERIFIED ISO/IEC




Spark
APACHE
HBASE
cassandra
HIVE
HDFS
FLUME

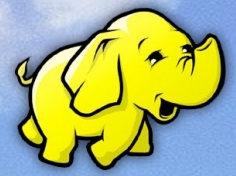



Spark
APACHE
HBASE
cassandra
HIVE
HDFS
FLUME

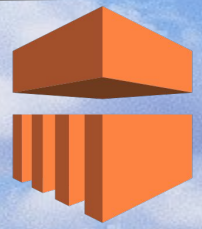


Spark
APACHE
HBASE
HIVE
cassandra
HDFS
FLUME

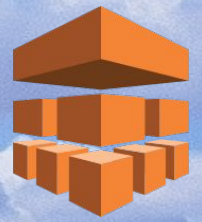
Microsoft
Azure
Google
Cloud Platform
amazon
web services
APACHE
HBASE
HIVE
Spark



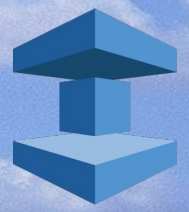
S3



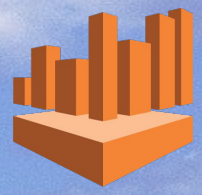
Elastic MapReduce



AI



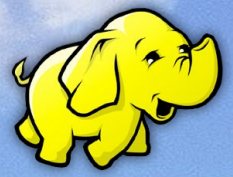
ElastiCache



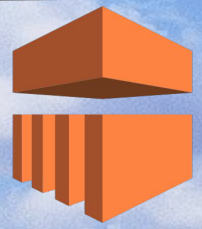
Athena

Kinesis

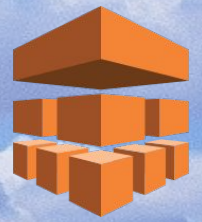
Lambda



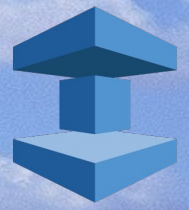
S3



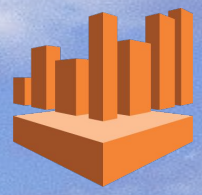
Elastic MapReduce



AI



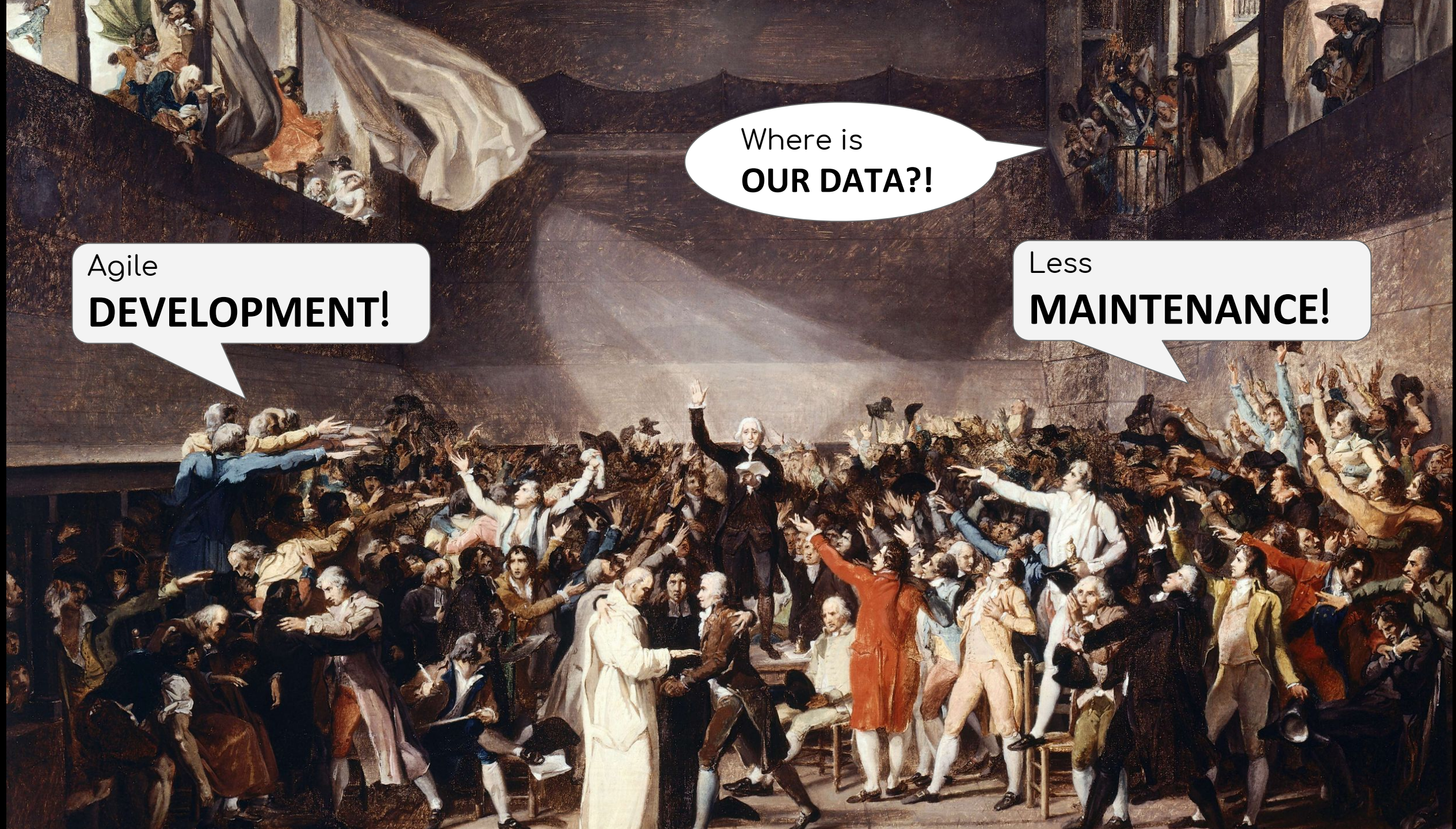
ElastiCache

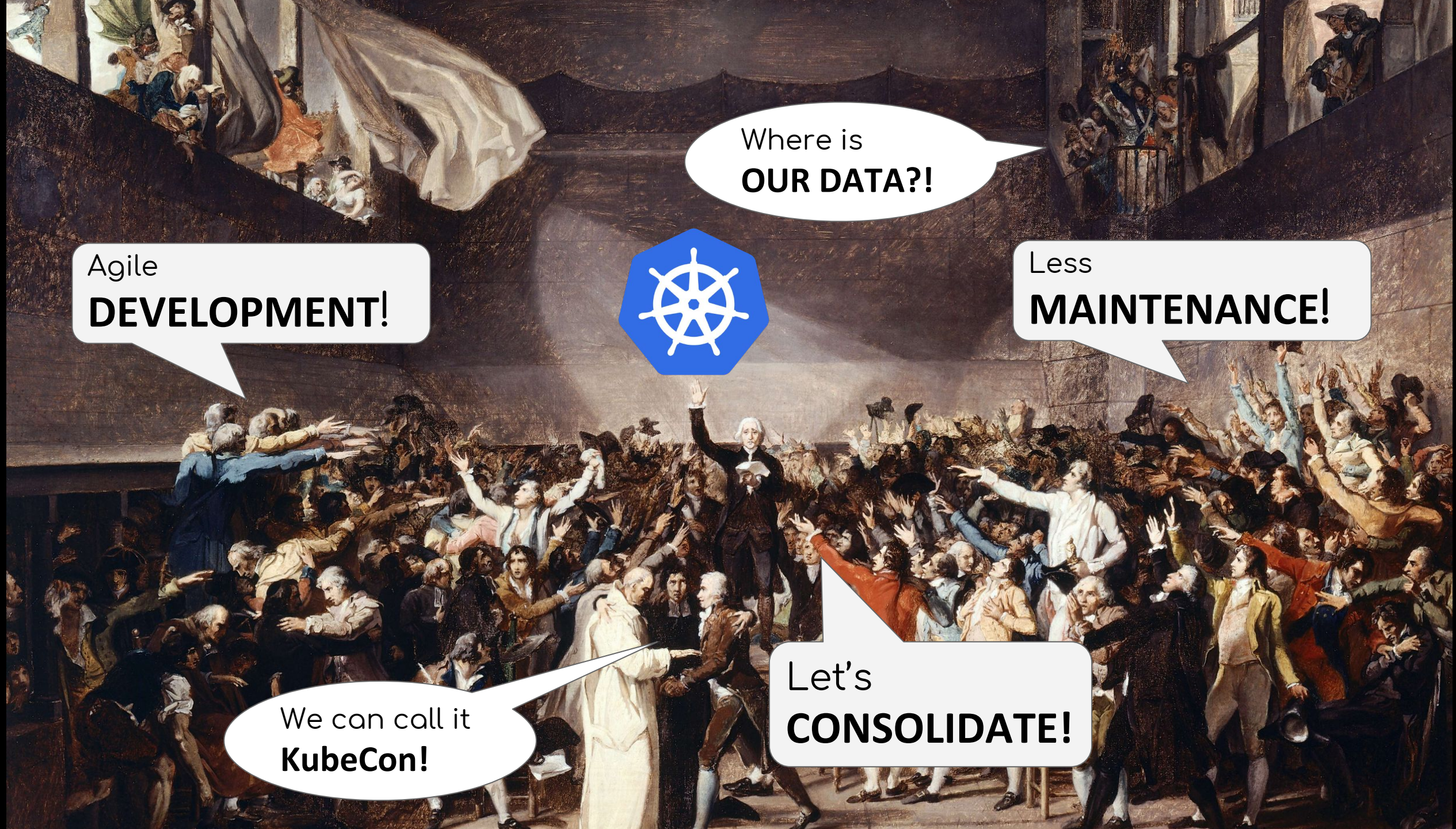


Athena

Kinesis

Pause


Where is
OUR DATA?!
Agile
DEVELOPMENT!
Less
MAINTENANCE!


Where is
OUR DATA?!
Agile
DEVELOPMENT!
Less
MAINTENANCE!
We can call it
KubeCon!
Let's
CONSOLIDATE!

Data Services

Unstructured

S3 / GlusterFS / Iguazio

Structured

DynamoDB / Iguazio / Cassandra

Stream

Kinesis / Kafka / Iguazio

Decouple Data Services from Application and Frameworks

Data Services

Unstructured

S3 / GlusterFS / Iguazio

Structured

DynamoDB / Iguazio / Cassandra

Stream

Kinesis / Kafka / Iguazio

Unified Orchestration

Decouple Data Services from Application and Frameworks

Application

Serverless

nuclio

fission

Data Services

Unstructured

S3 / GlusterFS / Iguazio

Structured

DynamoDB / Iguazio / Cassandra

Stream

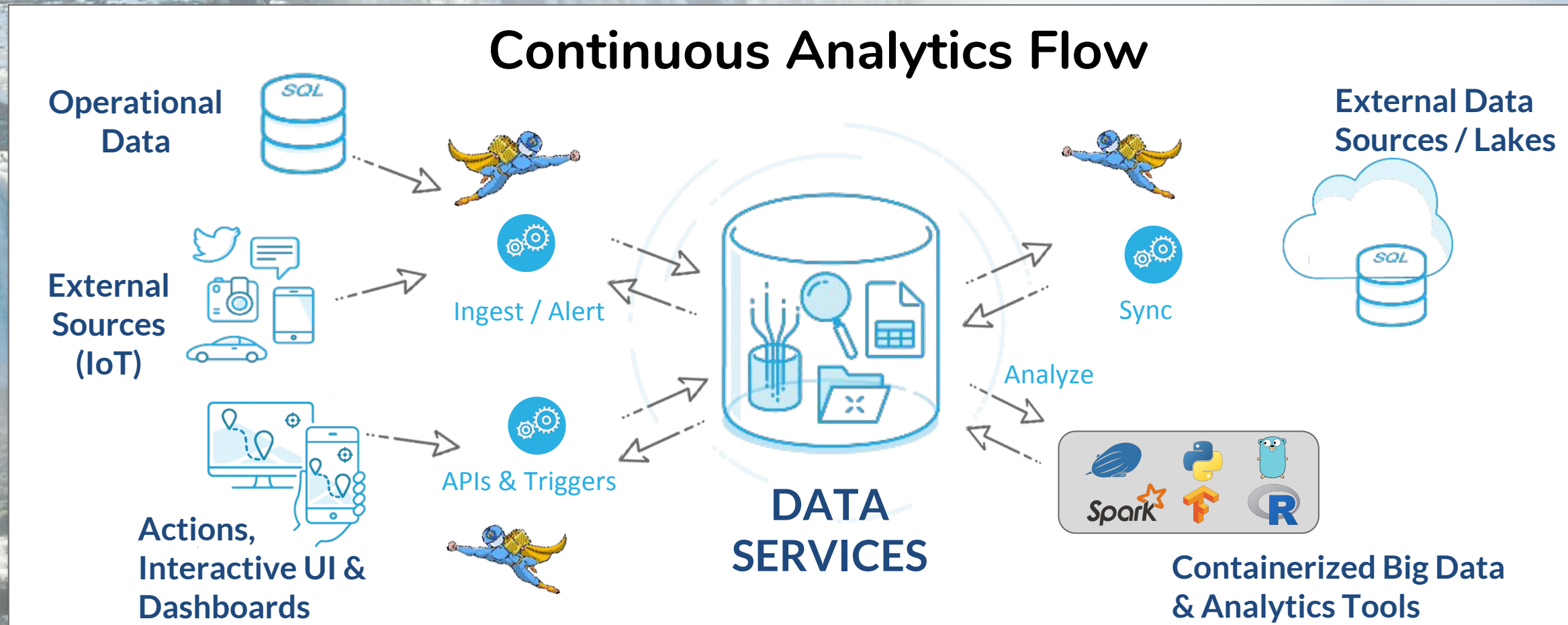
Kinesis / Kafka / Iguazio

Cloud Native Application and Frameworks

Unified Orchestration

Decouple Data Services from Application and Frameworks

Continuous Analytics Flow



Serverless

- ✓ Easy to deploy functions, no YAMLs, Dockerfile, build, etc.
- ✓ Auto-scaling and event triggers are built-in
- ✓ Use any language

BUT

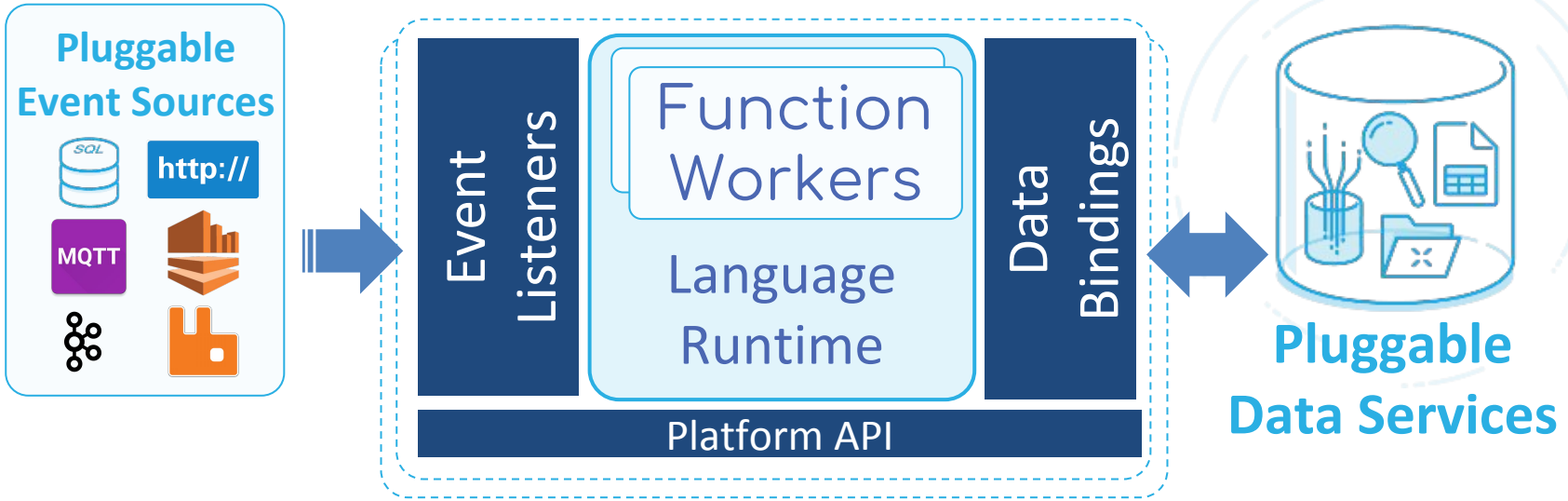
- ✗ Slow performance, lack of concurrency, no stream partitioning
- ✗ Limited number of event sources (mostly HTTP, some kafka)
- ✗ Hard to debug and diagnose
- ✗ Hard to manage complex functions and dependencies

nuclio A New Real-Time Serverless Platform

Comprehensive, Open, Super Fast and runs ANYWHERE

400K Events/sec per Process

Real-time, Auto-scaling Function Processors

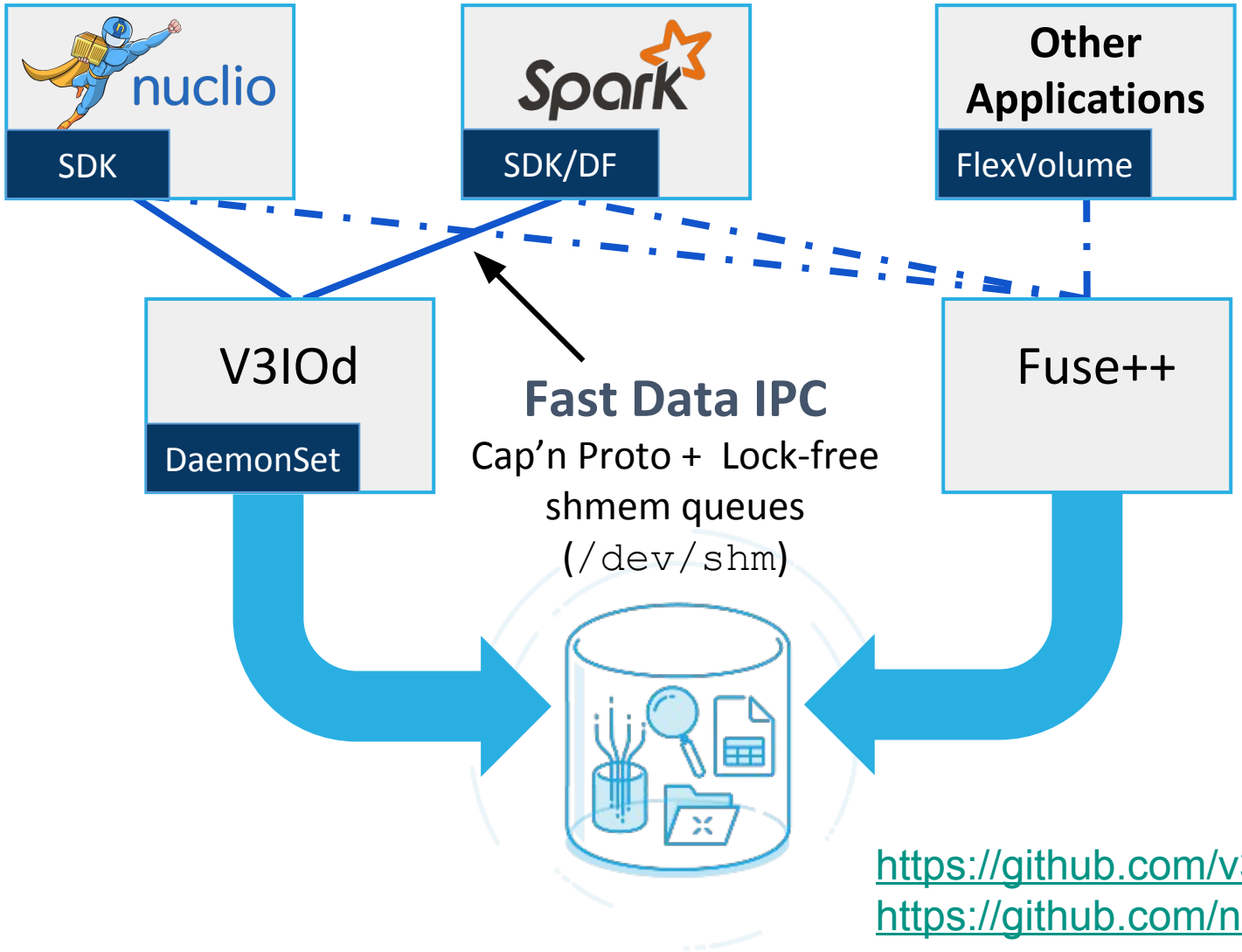


https://nuclio.io

- Any Source and Any Workload
- Maximum Performance
- Dev, Test, and Run ANYWHERE
- Simple, Fast, Secure data integration

Accelerate Performance Using Shared Memory

- Fast Data Access
- Fast Microservice Initialization
- Share TCP/RDMA Connections
- NO Kernel Changes!

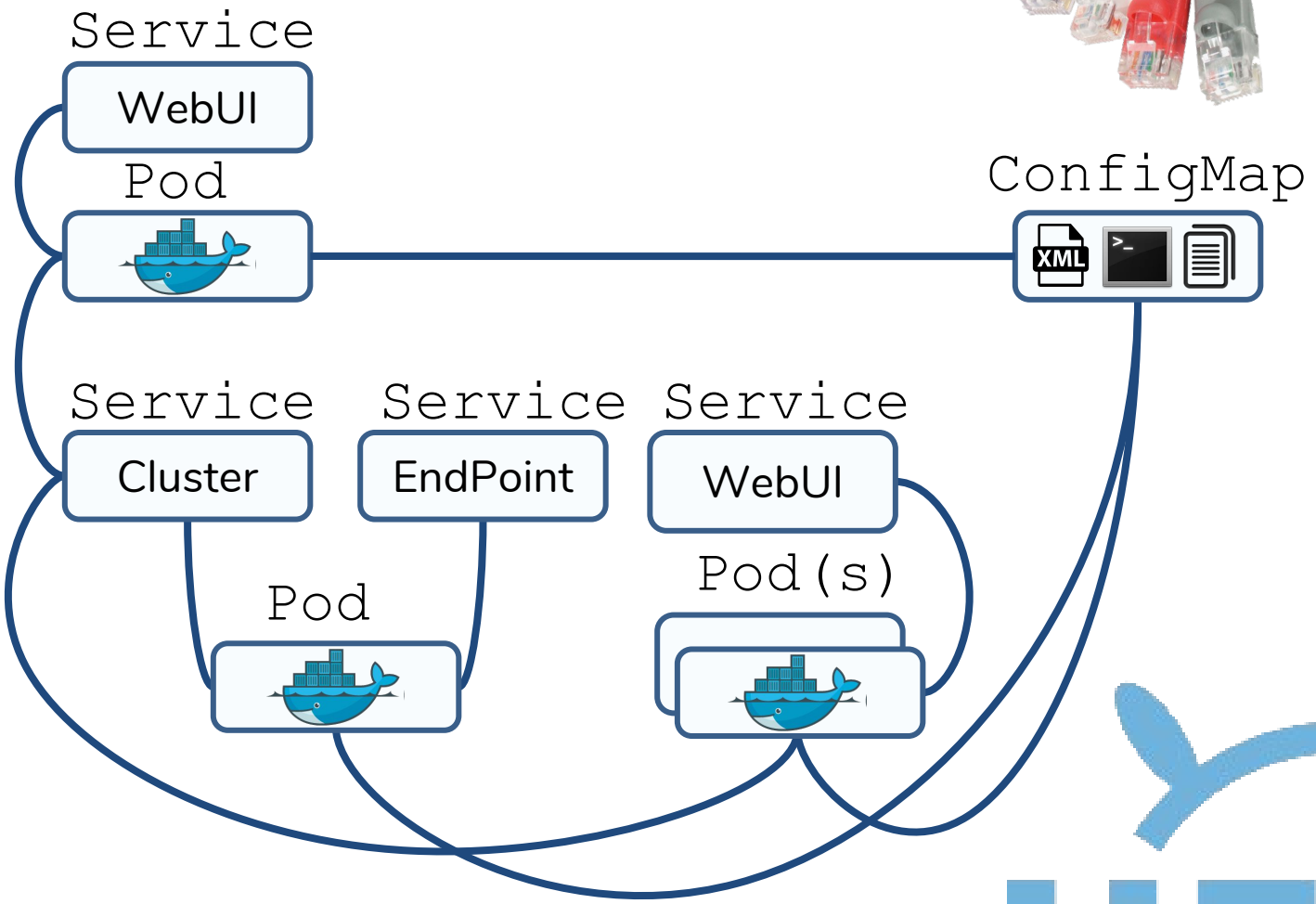


https://github.com/v3io/libfuse
https://github.com/nuclio/nuclio

Spark + HELM USE CASE

Deployment YAML

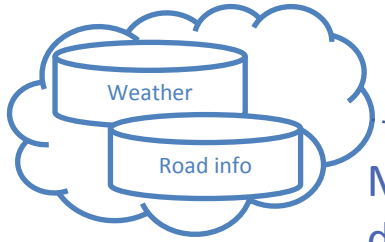
Deployment YAML



CUSTOMER USE CASE

Real time analytics for vehicle maintanance

Streaming vehicles
data via Web APIs



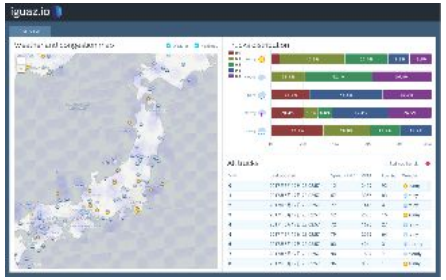
Micro service
data ingestion

Real-time enrichment with weather & road data

Real-time alerts

Find Correlation between weather condition and vehicles components

Real-time
Dashboard

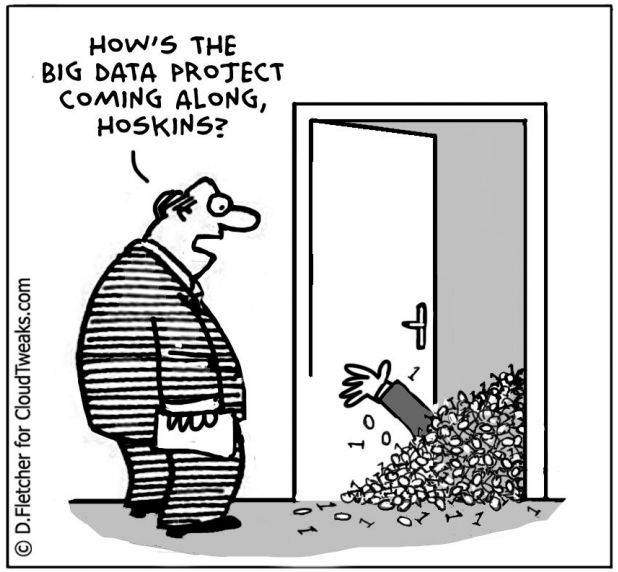


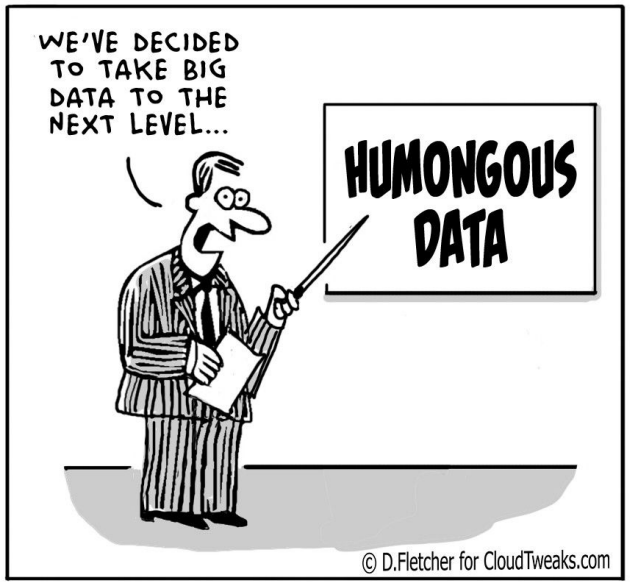
DEMO

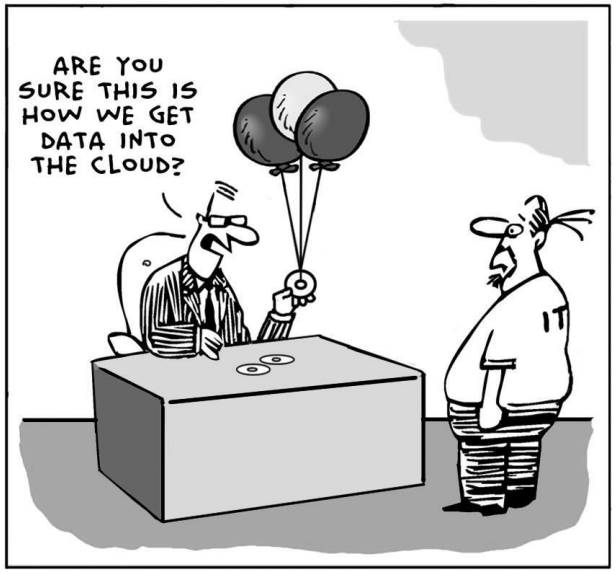
Tips & Lessons Learned

1. RTFM!
2. Community (!/?)
3. Know `kubectl` (`port-forward`...)
4. Navigate with ***Helm***
5. Do NOT overuse `NodePort`
6. Configuration goes to `ConfigMap`
7. Override `command` in YAML is tricky - use init script and keep it in `ConfigMap`
8. Collect operational data and use it!

Thank You







@bivas @iguazio #KubeCon2017

KubeCon

North America 2017