Big Data ecosystem

Worldline

M. Fanilo Andrianasolo

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

Data Analytics Tech Lead & Product Manager

Worldline

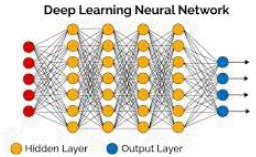
2013 - 2016	2016 - 2017	2017-2019	2019-current
Big Data Engineer	Data Analytics Evangelist	Data Science Tech Lead	Data Analytics Product Manager

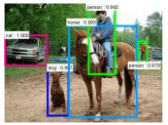
Data is at the center of all IT activities

Data

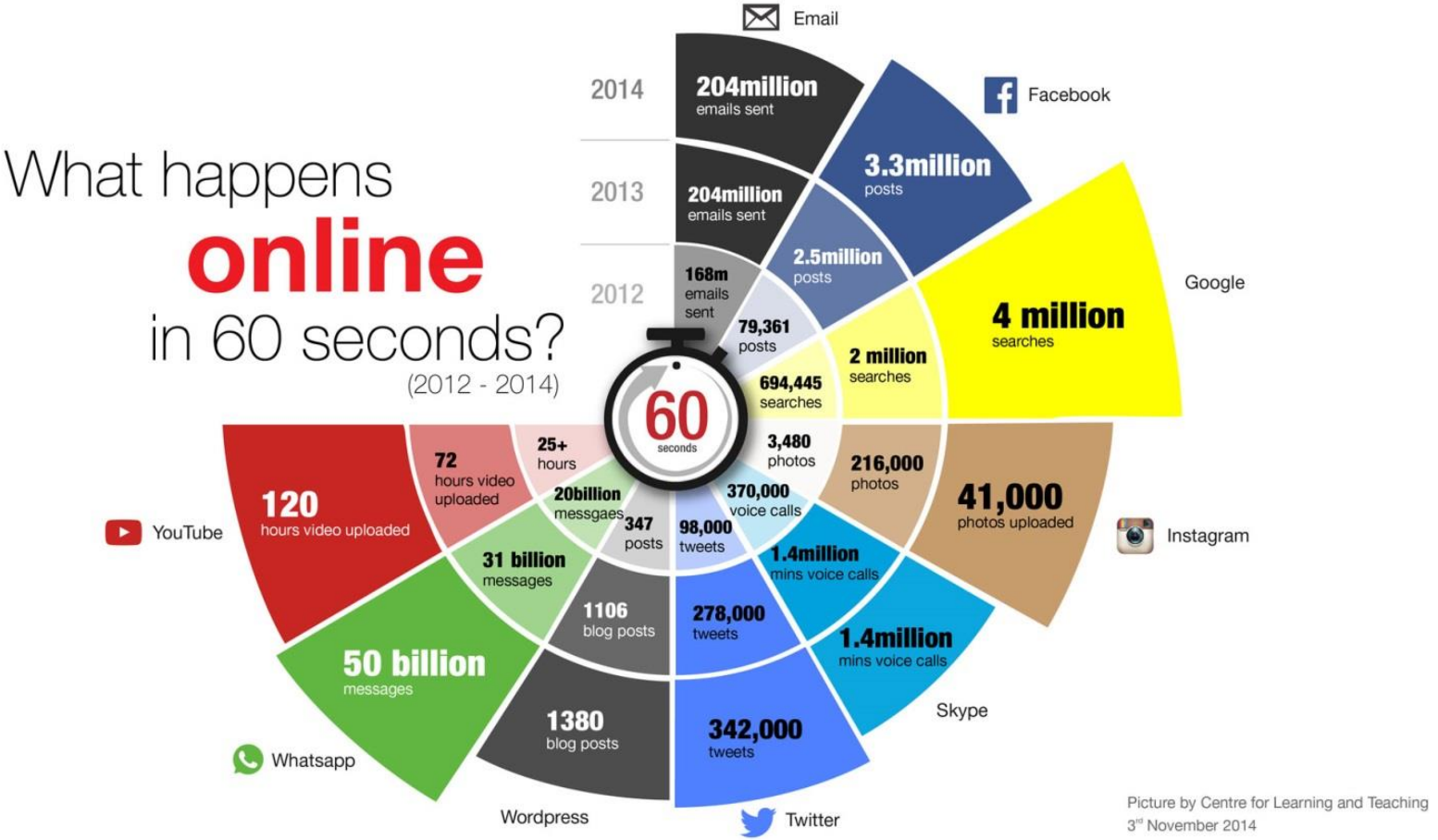
Hardware

Innovations





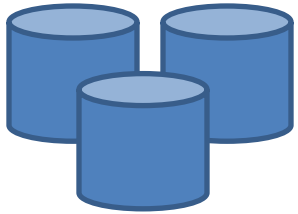
Explosion of data

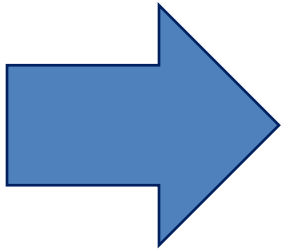


Picture by Centre for Learning and Teaching
3rd November 2014

3Vs of Big Data

Volume





Velocity

Big Data

Variability

Value

Vulnerability

Vwhatever..

Variety

Problem

Daily rate in 2014

How should we store and query such data ?

Scaling ?

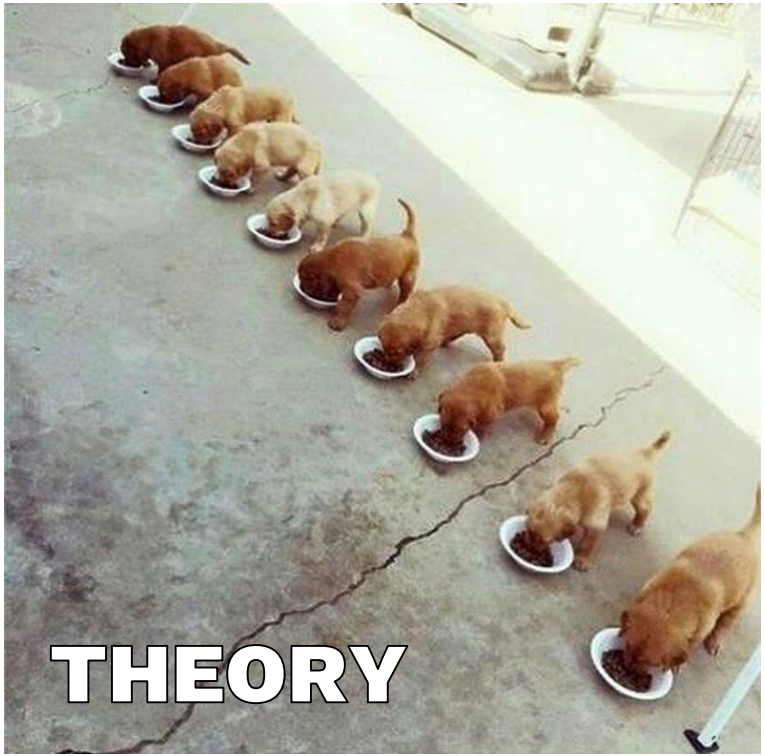
Vertical scaling

Less power consumption, cooling costs Less challenging to implement Less licencing costs (Sometimes) less network hardware	**PRICE** Hardware failure causes bigger outages Vendor lock-in Limited upgradeability

Horizontal scaling

Much cheaper Easier fault-tolerance “Easier” upgrade by adding new machines	Bigger energy footprint Higher utility cost (electricity, cooling) More networking equipment

Scaling is hard



Big Data ecosystem

Apache Hadoop

Open-source software for reliable, scalable, distributed computing

Apache Hadoop ecosystem

More than 30 open source projects for managing and analyzing Big Data

...

Hadoop distributions

Hadoop distributions vs Cloud providers

aws

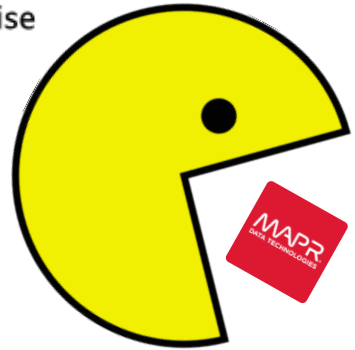
Google Cloud Platform

Microsoft Azure

CLOUDERA

HORTONWORKS®

Hewlett Packard Enterprise



Hadoop ecosystem use cases

Web indexing from web crawlers

Playlist generation from every listens

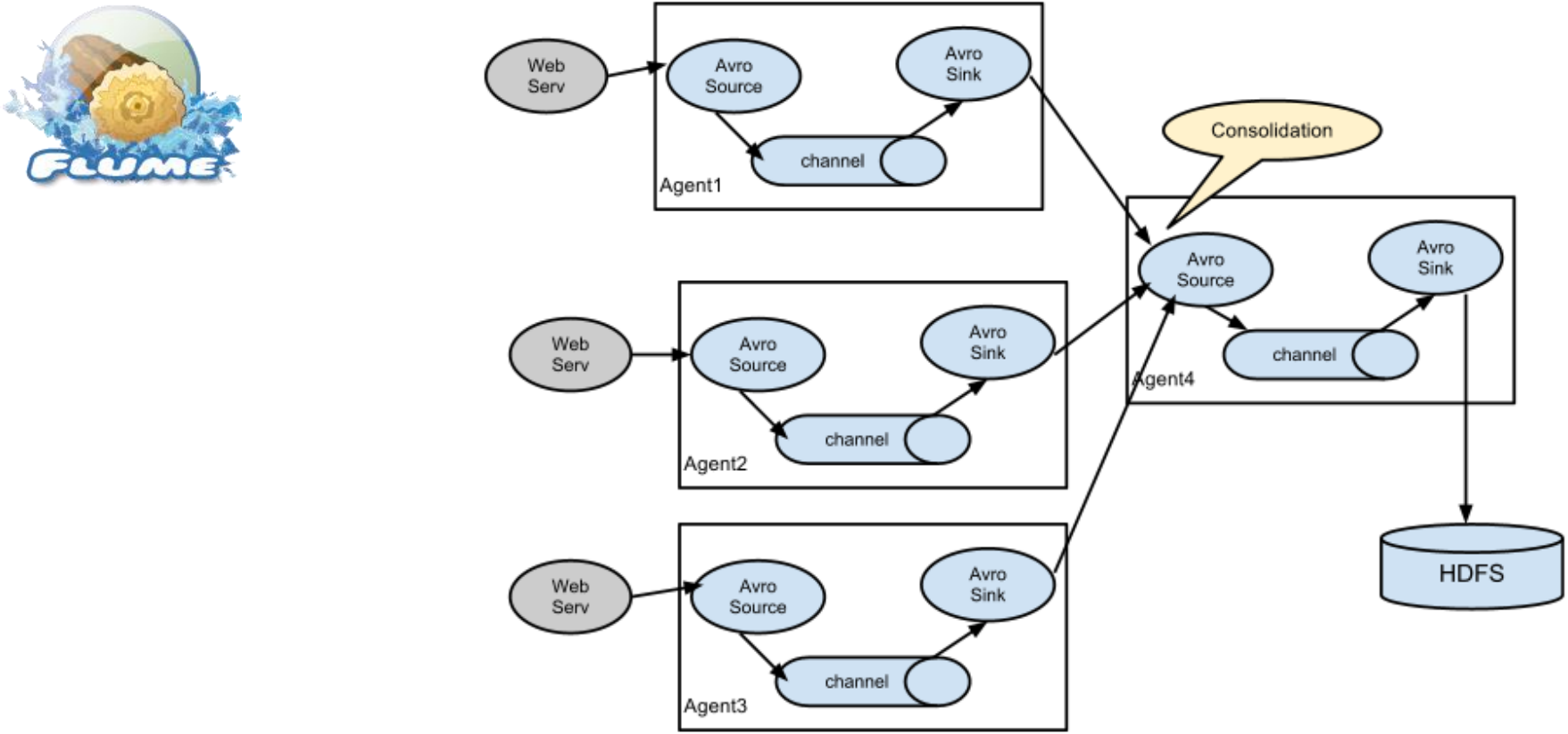
Log analysis

Product recommendation from purchases

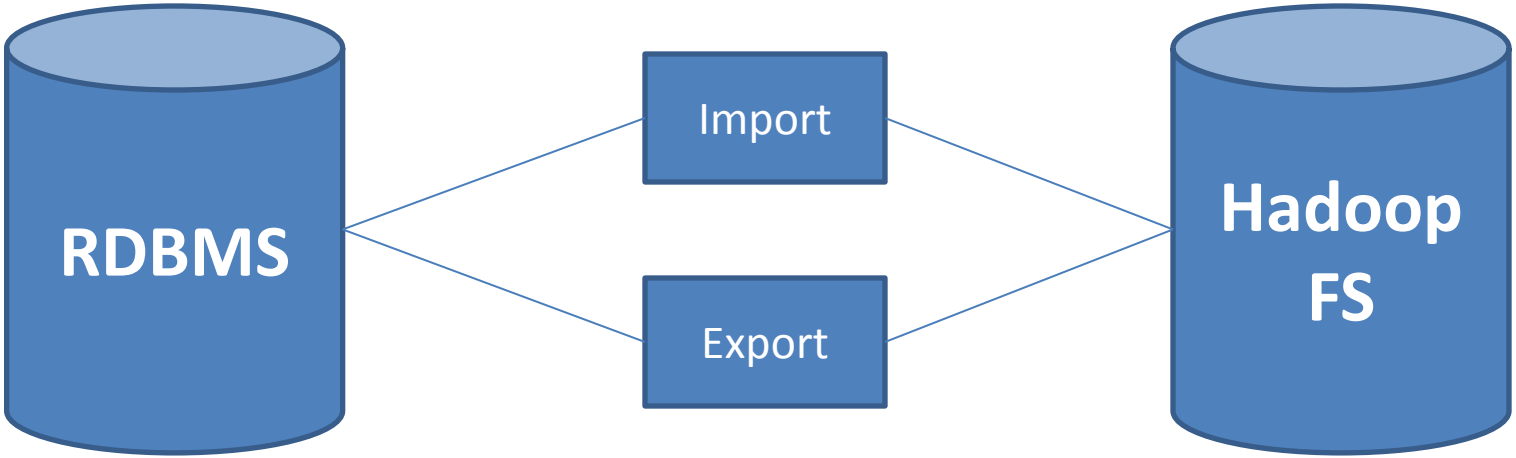
A data platform canvas

A data platform canvas

Acquisition



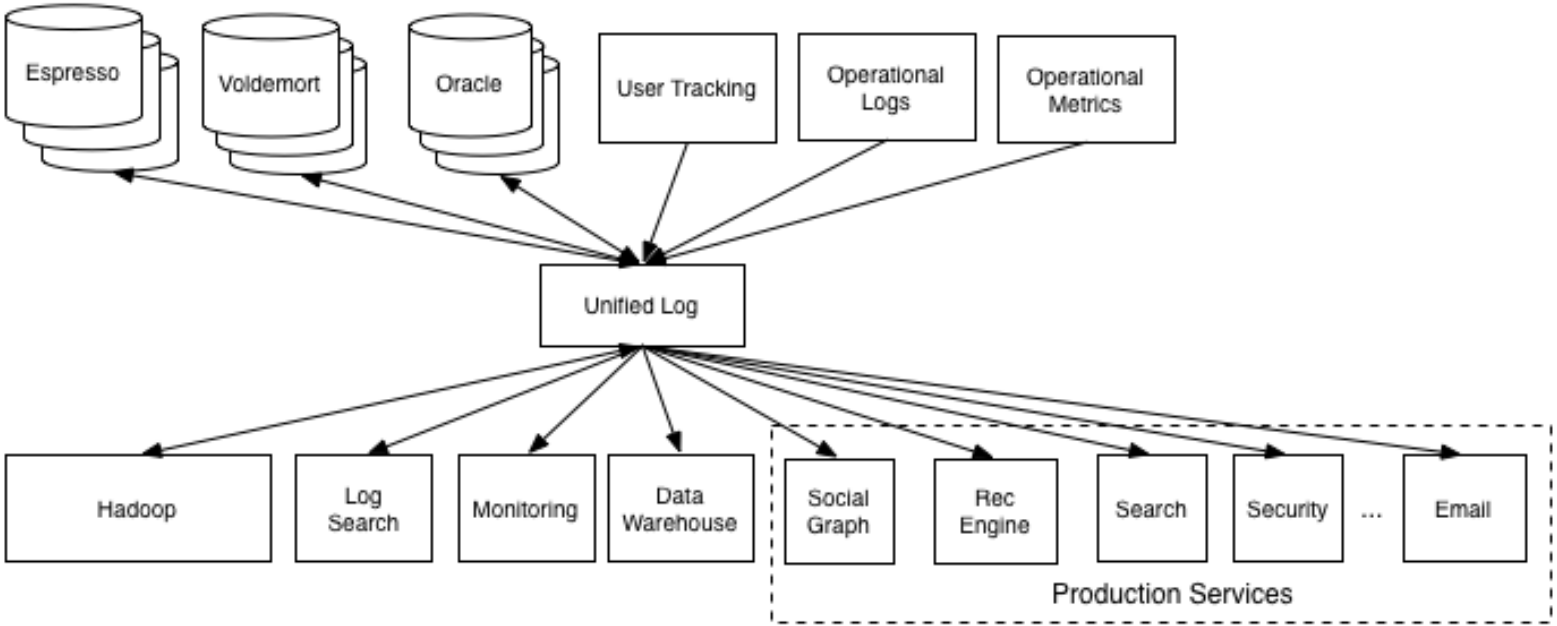
Acquisition



A data platform canvas

Transport

APACHE kafka
A distributed streaming platform



Transport


```
{
   "type" : "record",
   "namespace" : "test",
   "name" : "Employee",
   "fields" : [
       { "name" : "Name" , "type" : "string" },
       { "name" : "Age" , "type" : "int" }
    ]
}
```

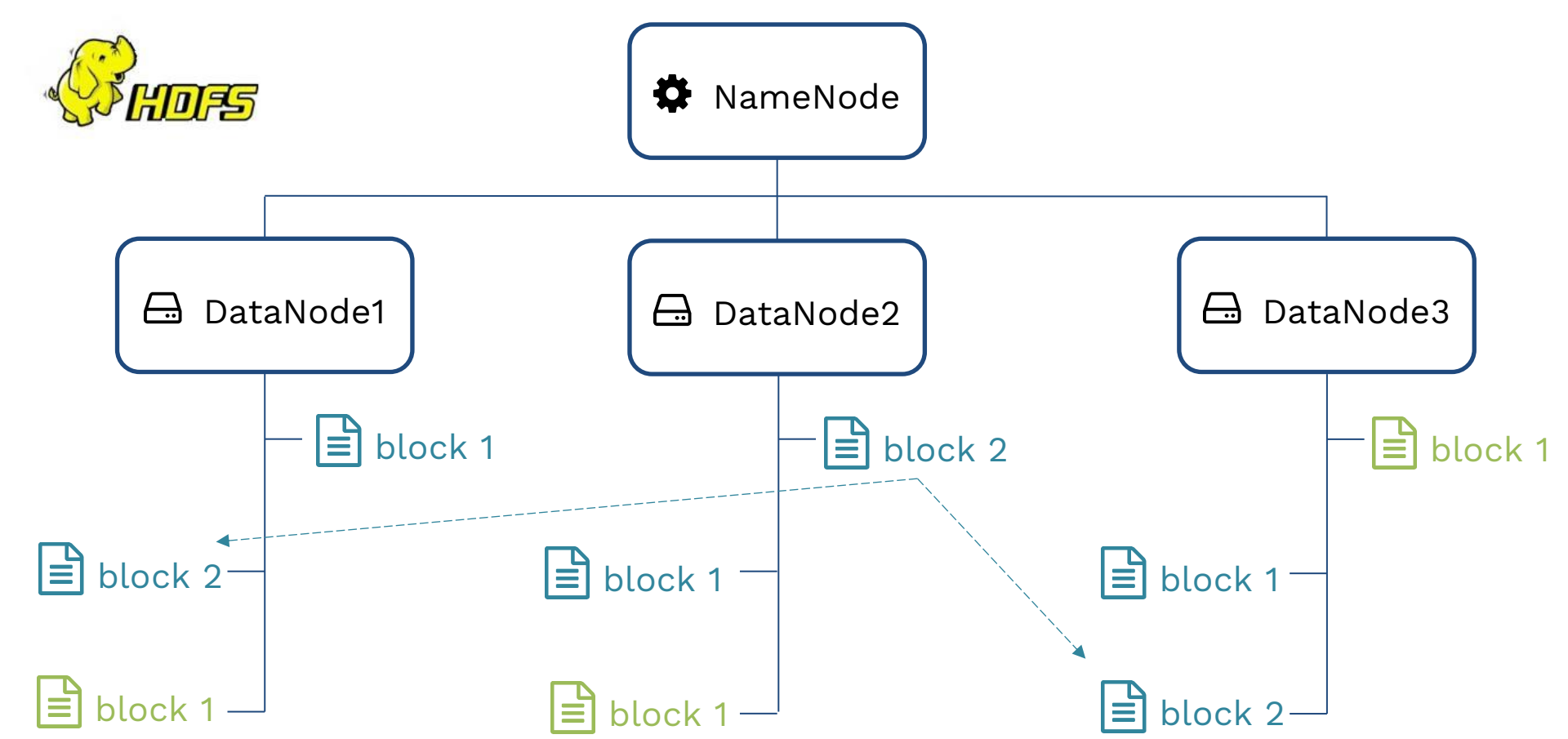
.ascv

```
emp e1=new emp( );
e1.setName("omar");
e1.setAge(21);
```

.java

A data platform canvas

Hadoop Distributed File System

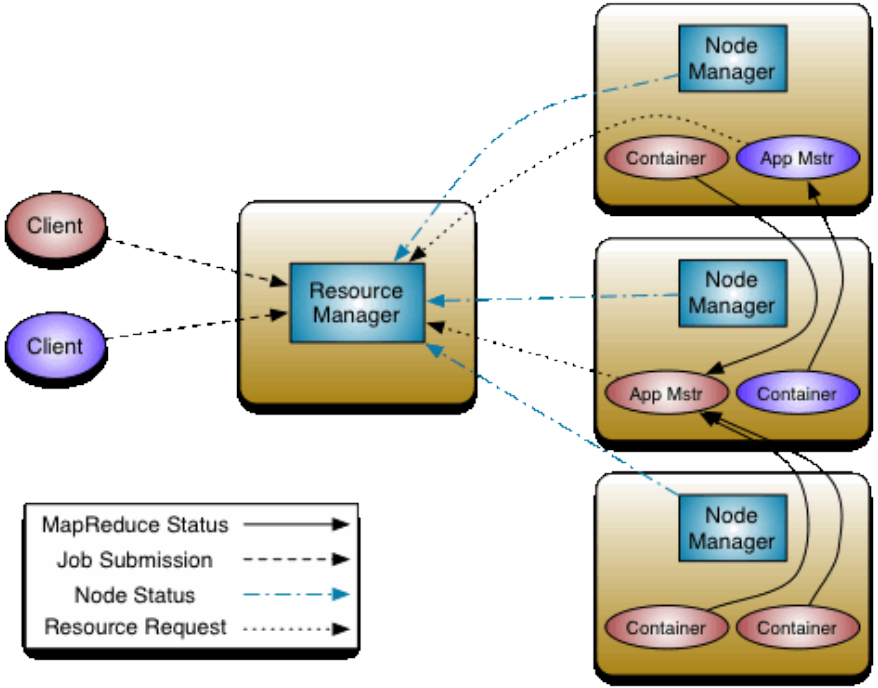


HBase

Key	U:cookie	U:is_auth	U:has_t	P:Product1	P:Product2	P:Product3
1960:Fanilo	c13e	1				3
2001:Fanilo	c13e		1			
1990:Omar	d45				1	

A data platform canvas

YARN – Yet Another Resource Negotiator



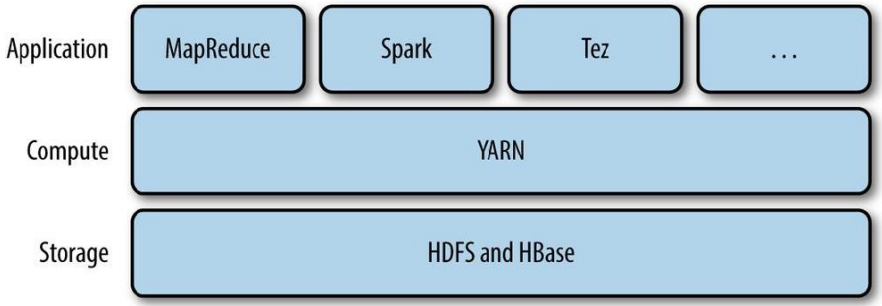
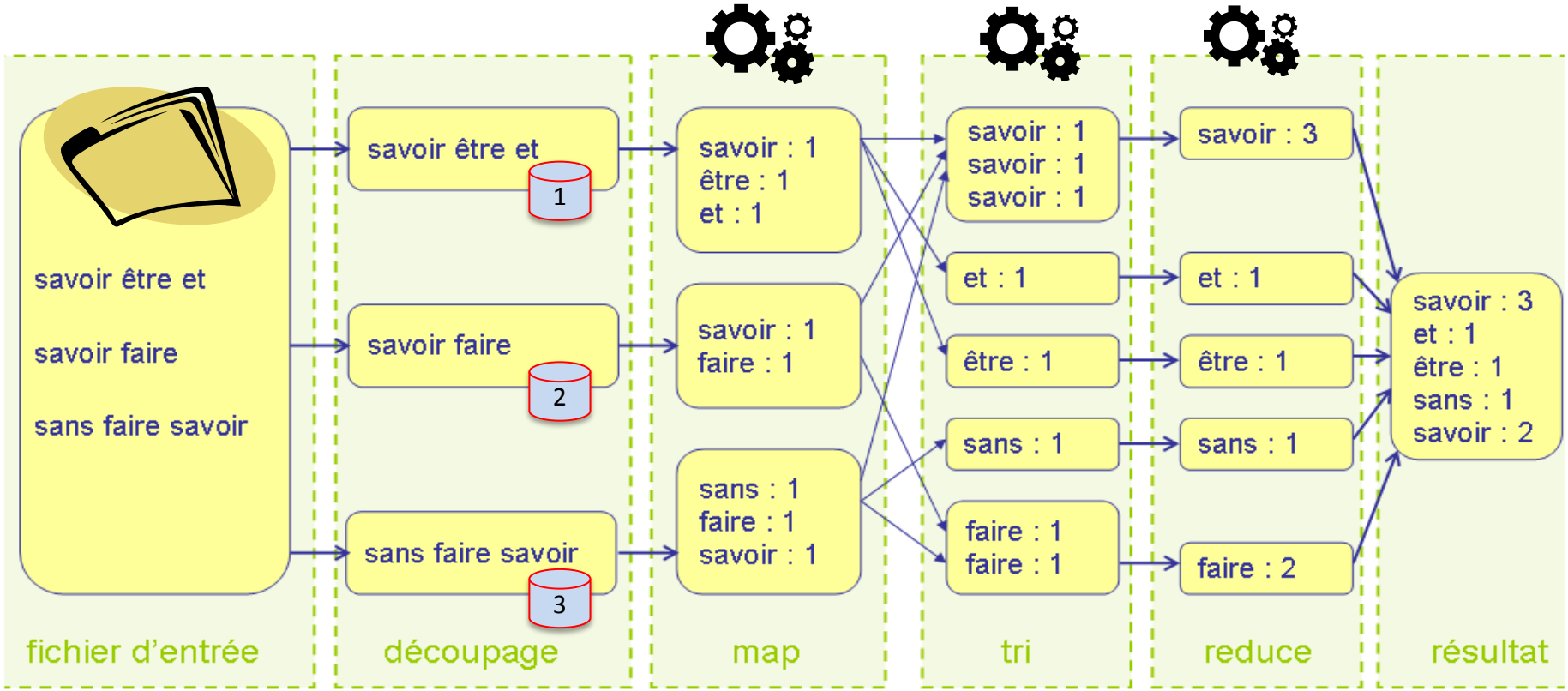


Figure 4-1. YARN applications

YARN hides the resource management details from the user to facilitate the management of parallel applications.

Batch processing - Map Reduce



Data locality : Moving Computation is Cheaper than Moving Data

Batch processing

HiveQL → MapReduce

Metastore

music_sales.csv

1, « Let it go », 4.99€, 5

2, « Snow », 7.99€, 1

3, « Lion King », 0.99€, 1

4, « SISE », 1.99€, 2

5, « Lyon is great », 2.99€, 3

Batch processing

```
recordings = LOAD '$file' USING PigStorage(',') AS
                        (id, price, artist, title,
duration, year);
limit = LIMIT recordings $size;
DUMP limit;
```

MapReduce

music_sales.csv

1, « Let it go », 4.99€, 5

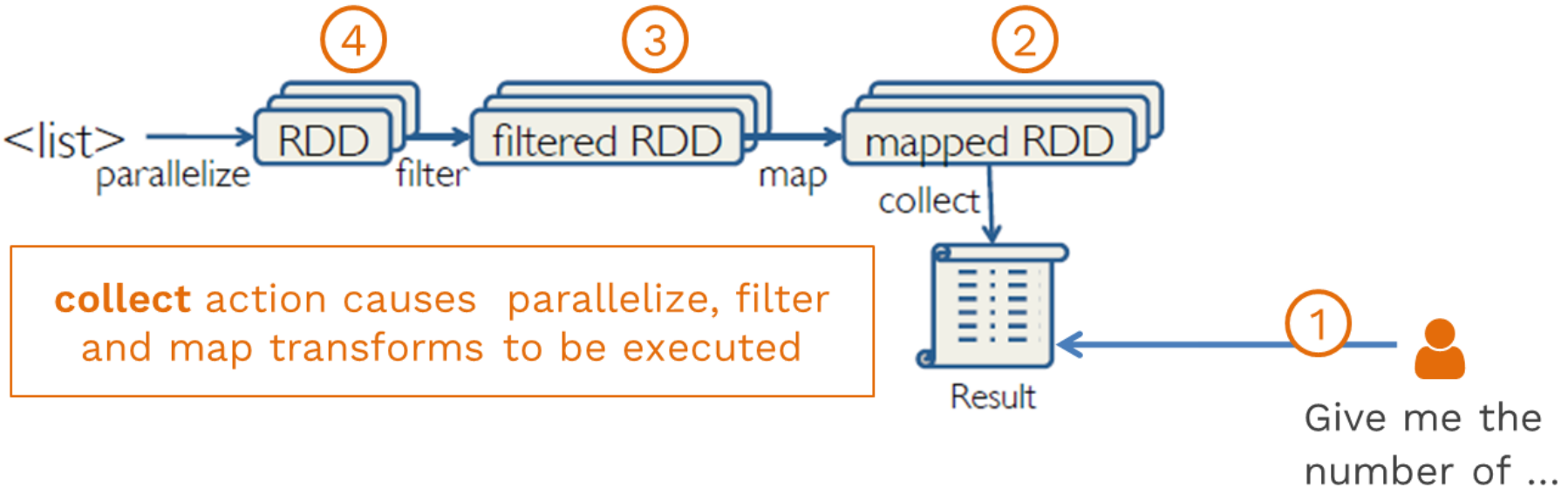
2, « Snow », 7.99€, 1

3, « Lion King », 0.99€, 1

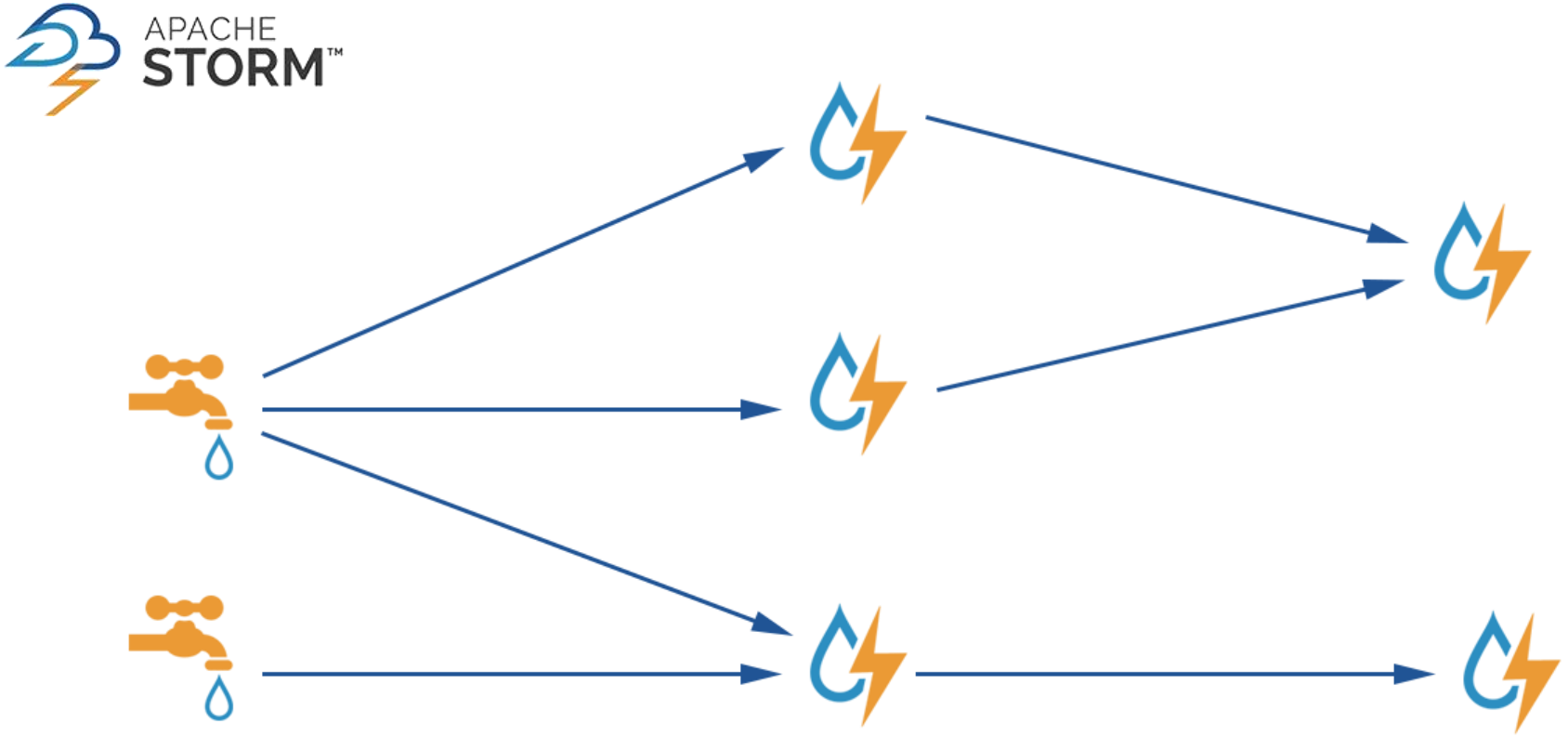
4, « SISE », 1.99€, 2

5, « Lyon is great », 2.99€, 3

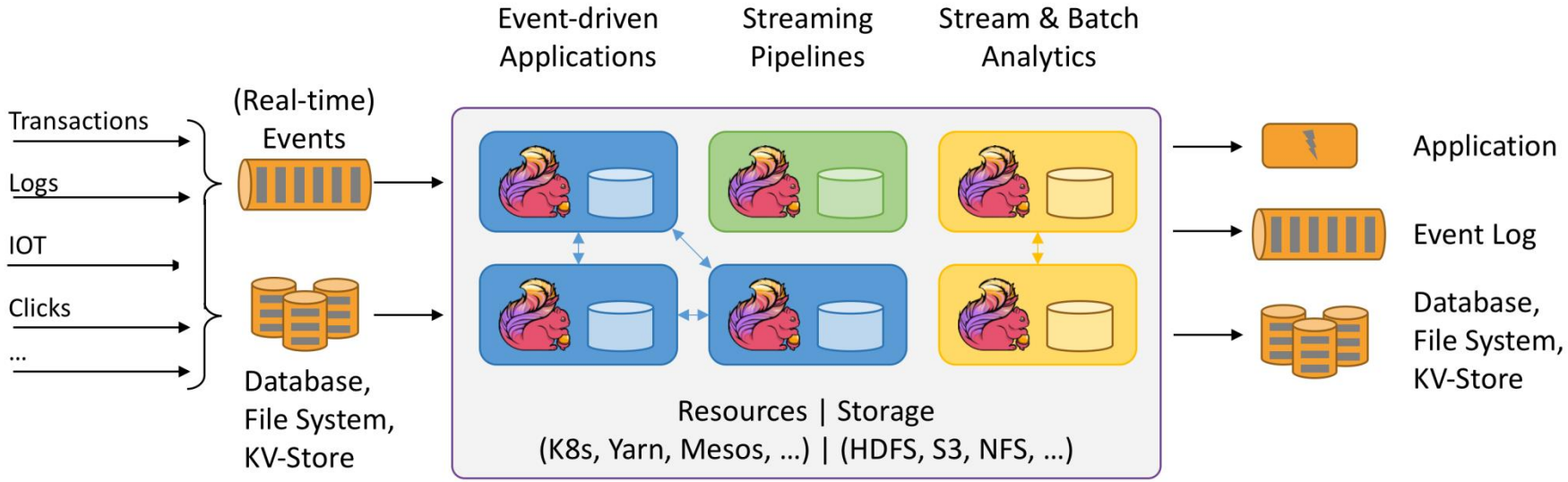
Batch processing



Realtime processing



Realtime processing



A data platform canvas

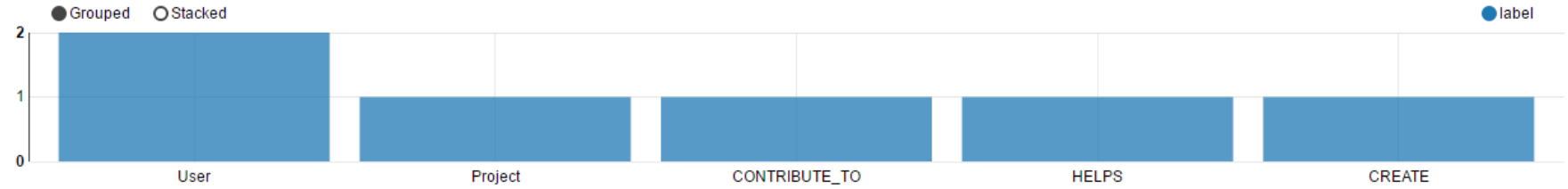
Visualizing

Apache Zeppelin

```
%spark
print(s"""
%network {
    "nodes": [{"id": 1, "label": "User", "data": {"fullName":"Andrea Santurbano"}},{"id": 2, "label": "User", "data": {"fullName":"Lee Moon Soo"}},{"id": 3, "label": "Project", "data": {"name"
        :"Zeppelin"}}],
    "edges": [{"source": 2, "target": 1, "id" : 1, "label": "HELPS"},{"source": 2, "target": 3, "id" : 2, "label": "CREATE"},{"source": 1, "target": 3, "id" : 3, "label": "CONTRIBUTE_TO",
        "data": {"oldPR": "https://github.com/apache/zeppelin/pull/1582"}}],
    "labels": {"User": "#8BC34A", "Project": "#3071A9"},
    "directed": true,
    "types": ["HELPS", "CREATE", "CONTRIBUTE_TO"]
}
""")
```

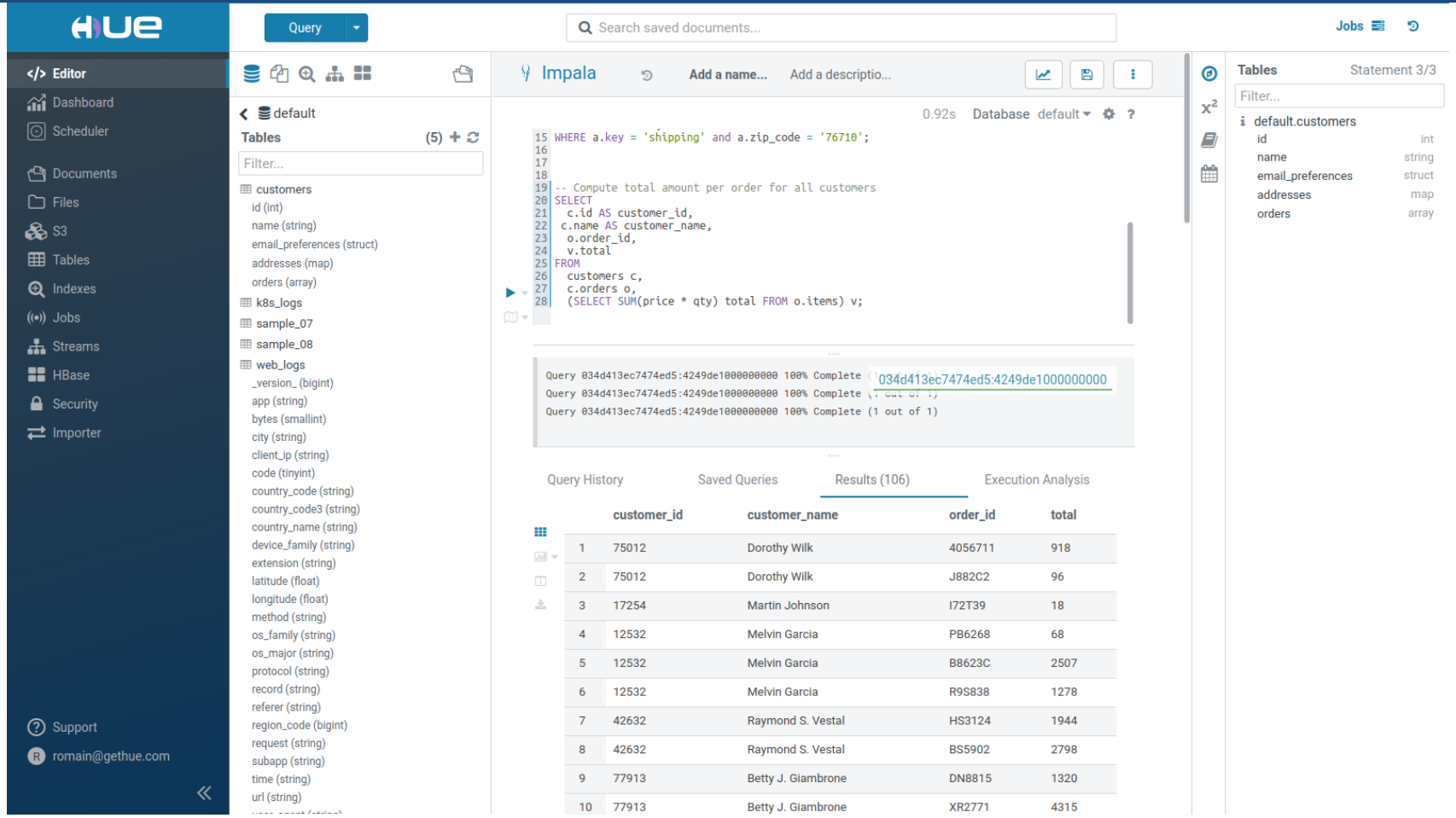
FINISHED

settings



Took 1 sec. Last updated by anonymous at March 06 2017, 10:35:46 PM. (outdated)

Visualizing

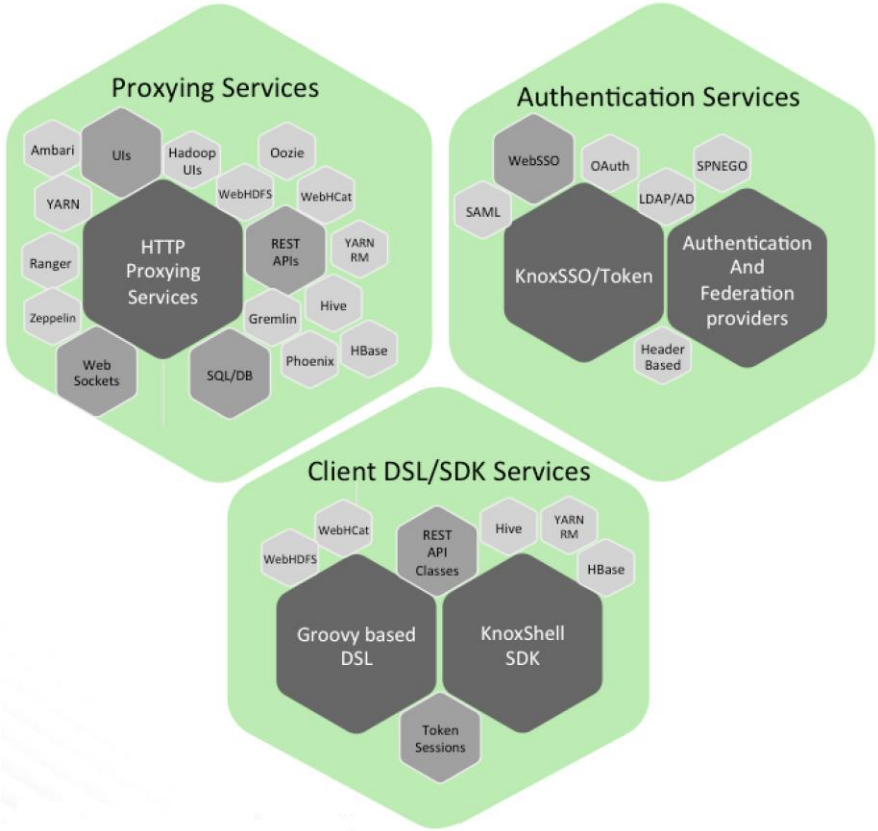


A data platform canvas

Security

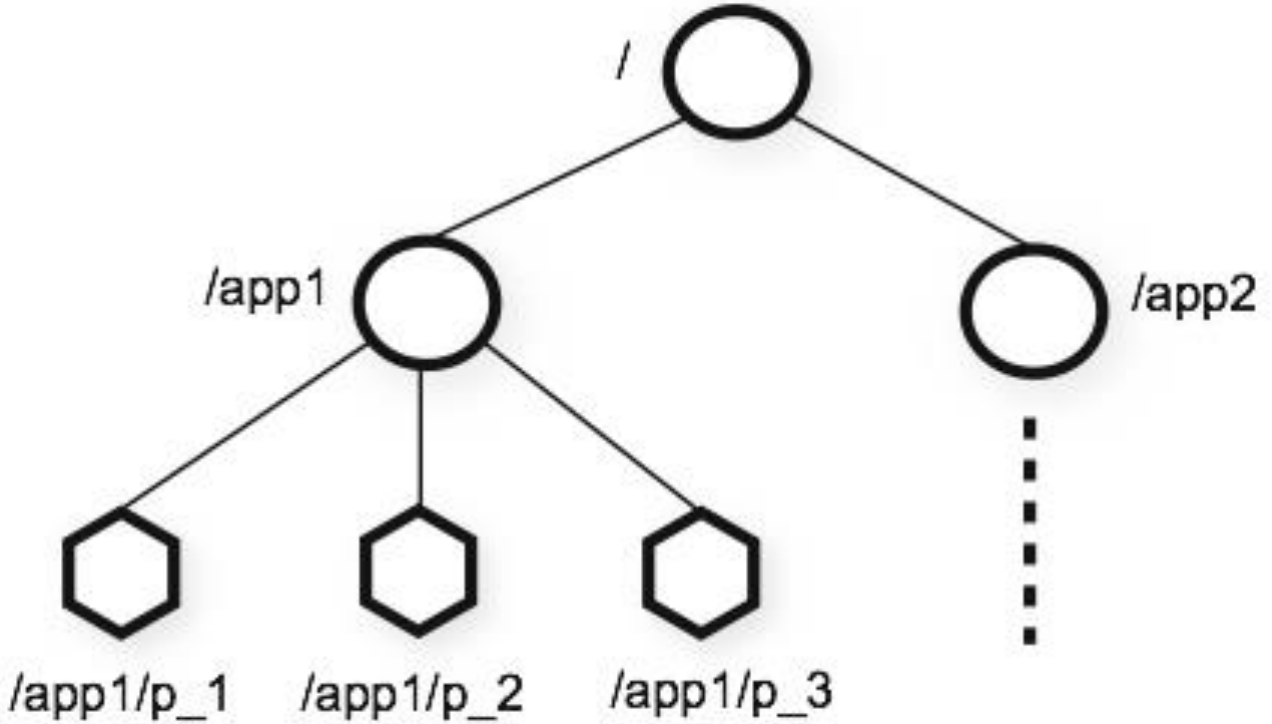
Kerberos

APACHE KNOX

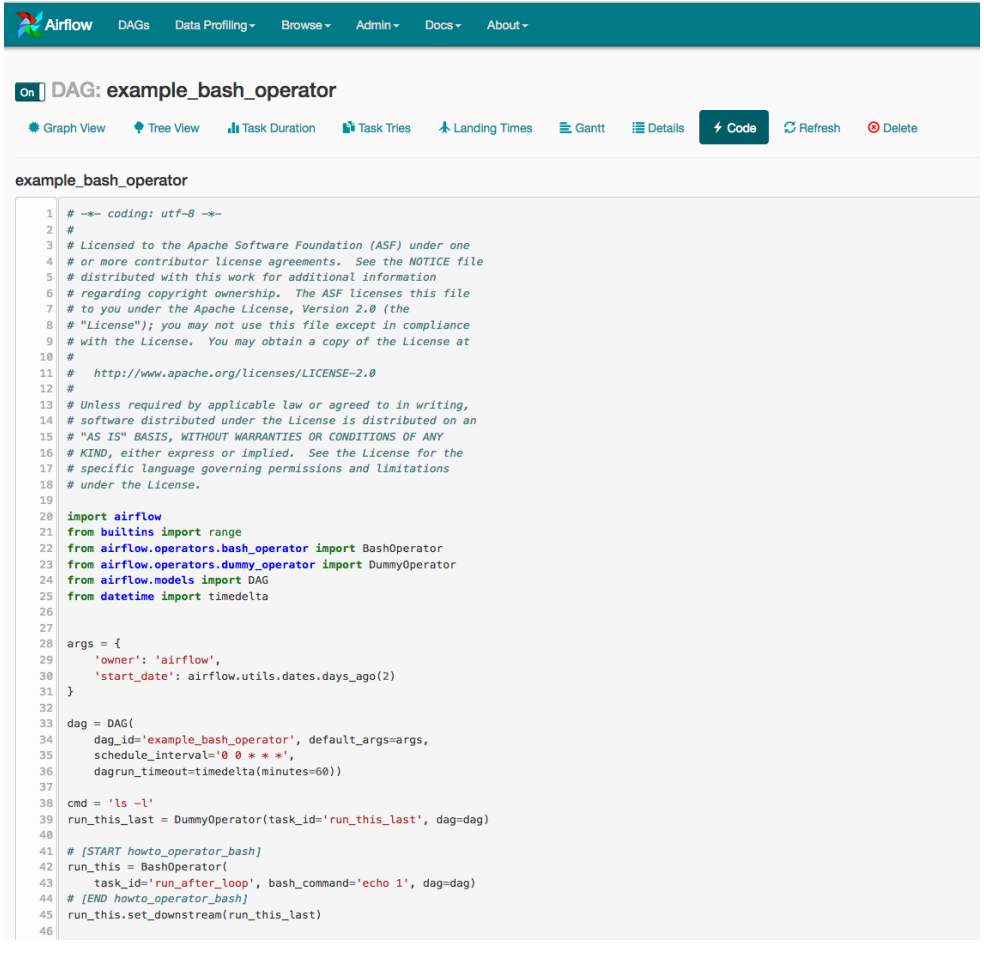


Orchestration


Apache ZooKeeper™



Orchestration

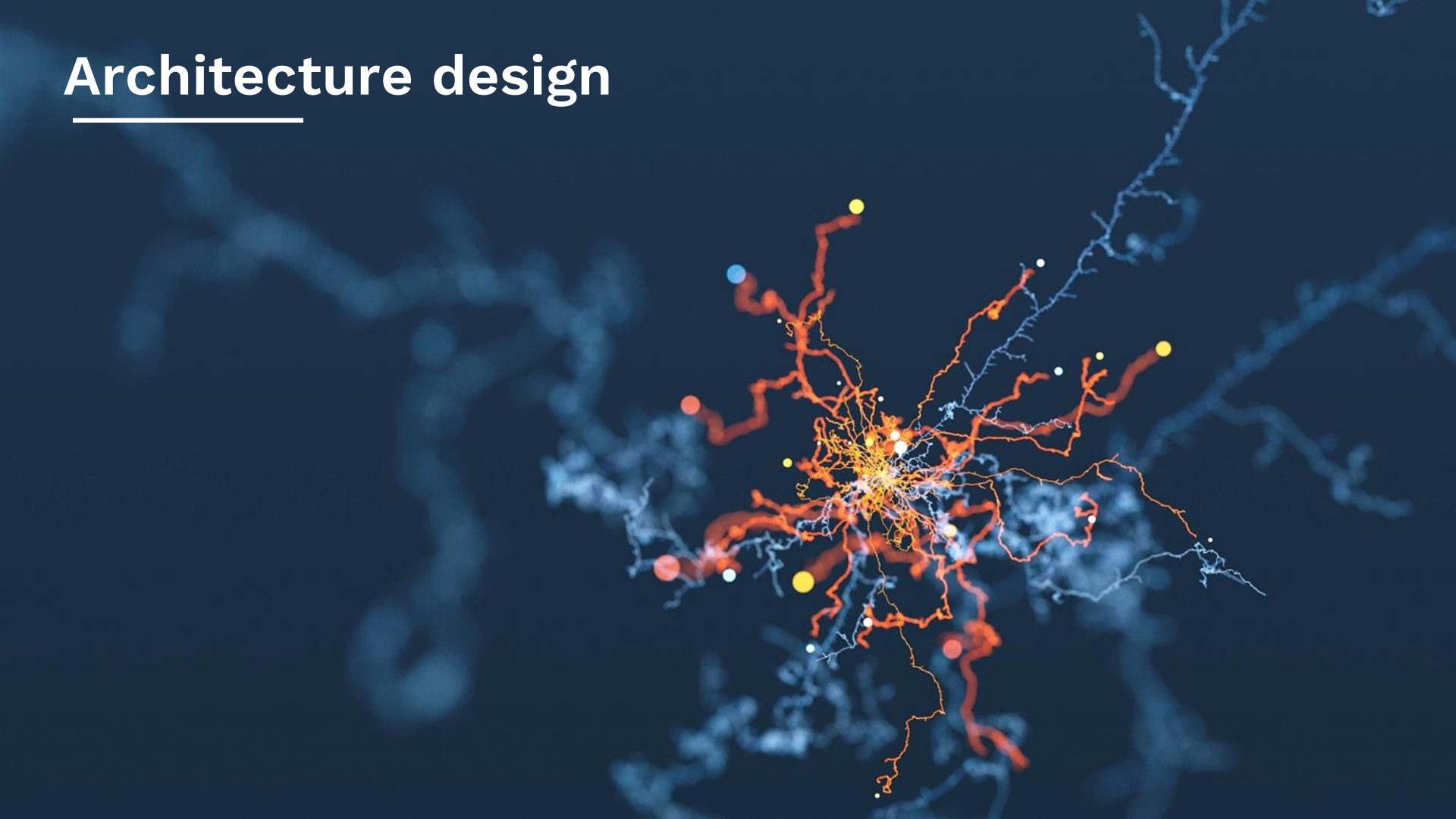


Overview

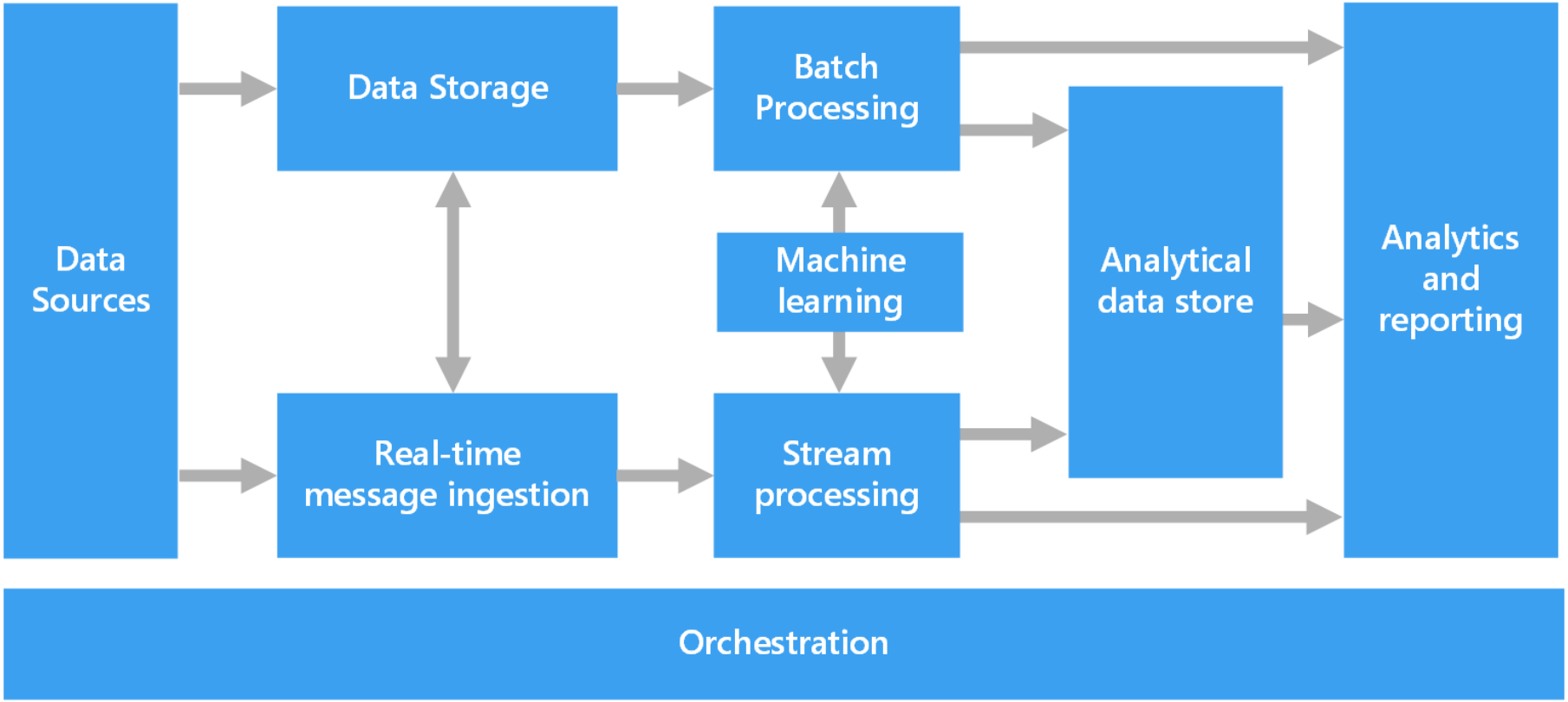

Acquisition
Transport
Storage
Processing
Servicing
Security
Orchestration


RELAX

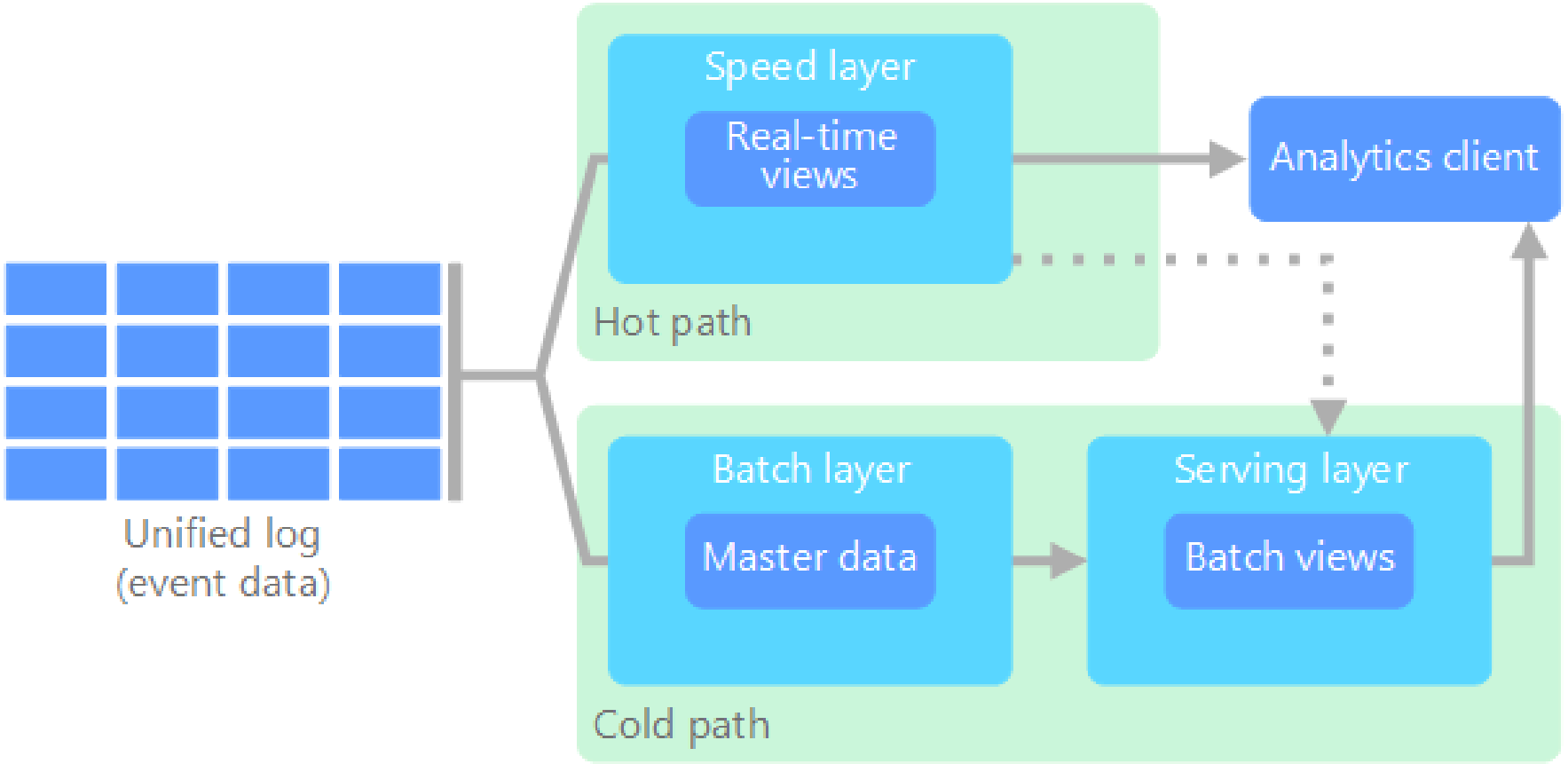
Architecture design



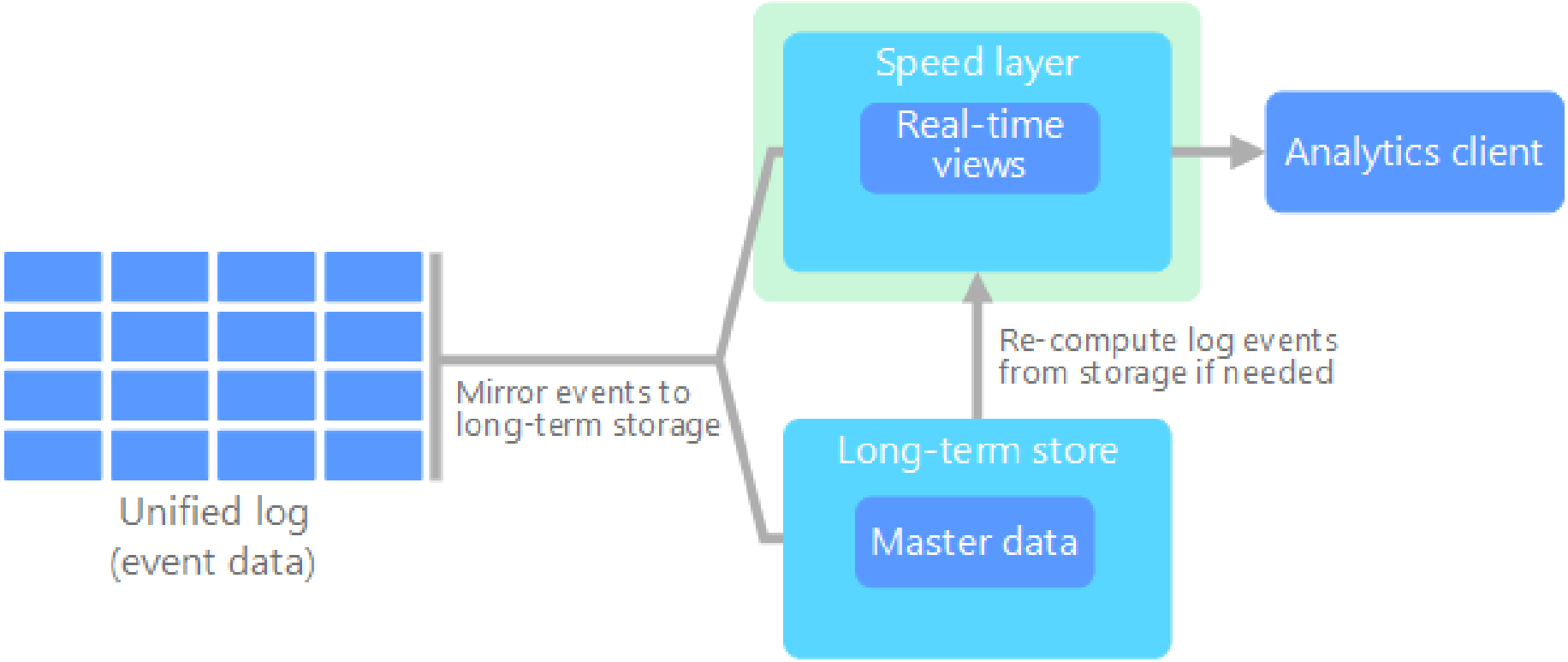
Multiple architectures



Lambda architecture



Kappa architecture



CONCLUSION

THANKS

@andfanilo

@andfanilo

andfanilo@gmail.com