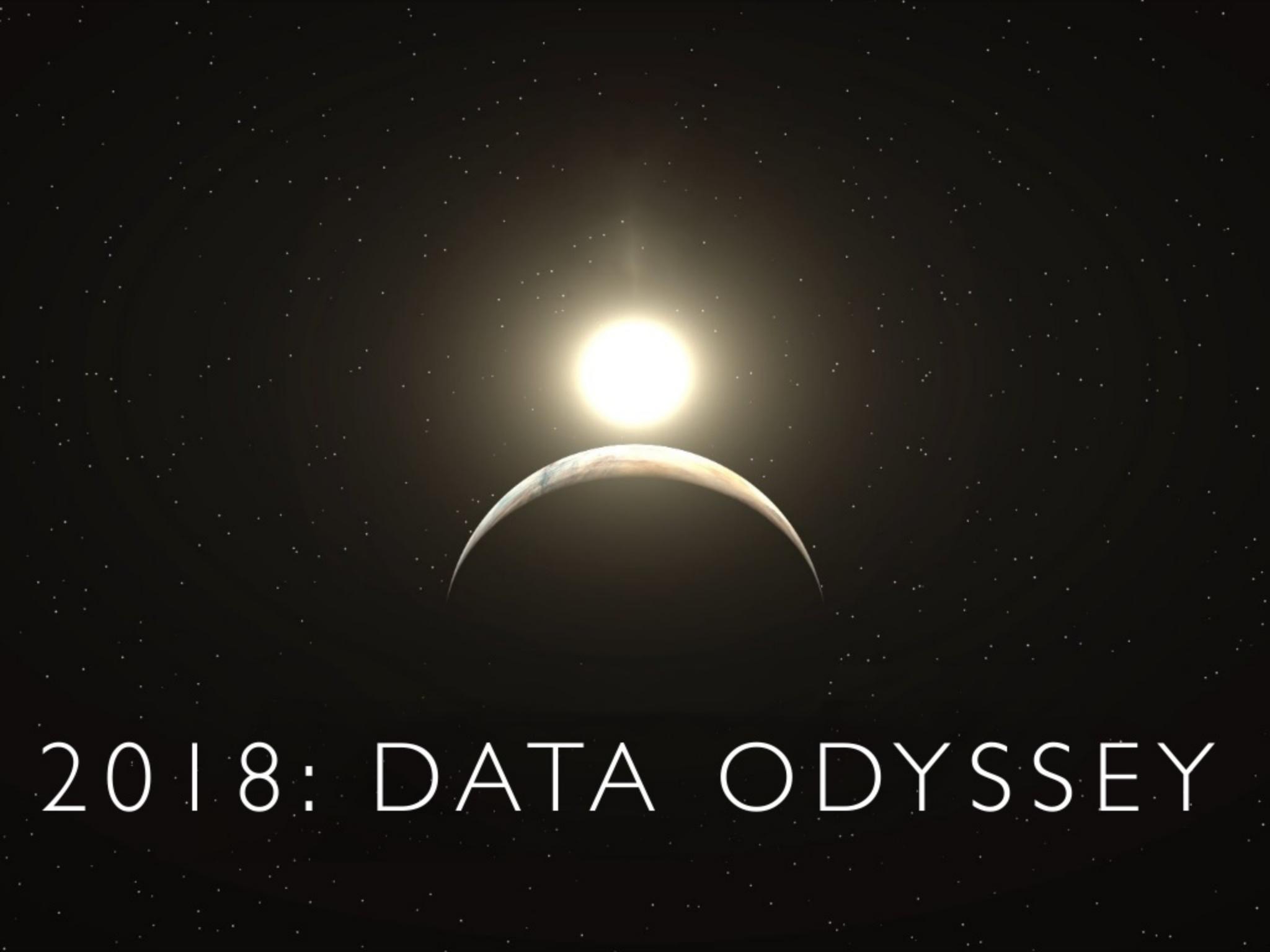




# **Designing a Horizontally Scalable Event-Driven Big Data Architecture with Apache Spark**

Ricardo Fanjul, letgo

#SAISExp2

The background of the image is a dark, star-filled space. A bright, circular light source, resembling the sun, is positioned at the top center. Below it, a large, semi-transparent planet or celestial body is visible, showing a curved horizon and some surface details. The overall atmosphere is mysterious and cosmic.

2018: DATA ODYSSEY



Ricardo Fanjul  
Data Engineer



**Post in a Snap**

in just seconds

Flag icon **Founded in 2015**

Up arrow icon **100MM+ downloads**

Up arrow icon **400MM+ listings**



## LETGO DATA PLATFORM IN NUMBERS

**1 billion**

Events Processed Daily

**50K**

Peaks of events per Second

**500GB**

Data daily

**600+**

Event Types

**200TB**

Storage (S3)

**< 1 sec**

NRT Processing Time



THE DAWN OF LETGO

THE DAWN OF LETGO

# CLASSICAL BI PLATFORM



THE DAWN OF LETGO

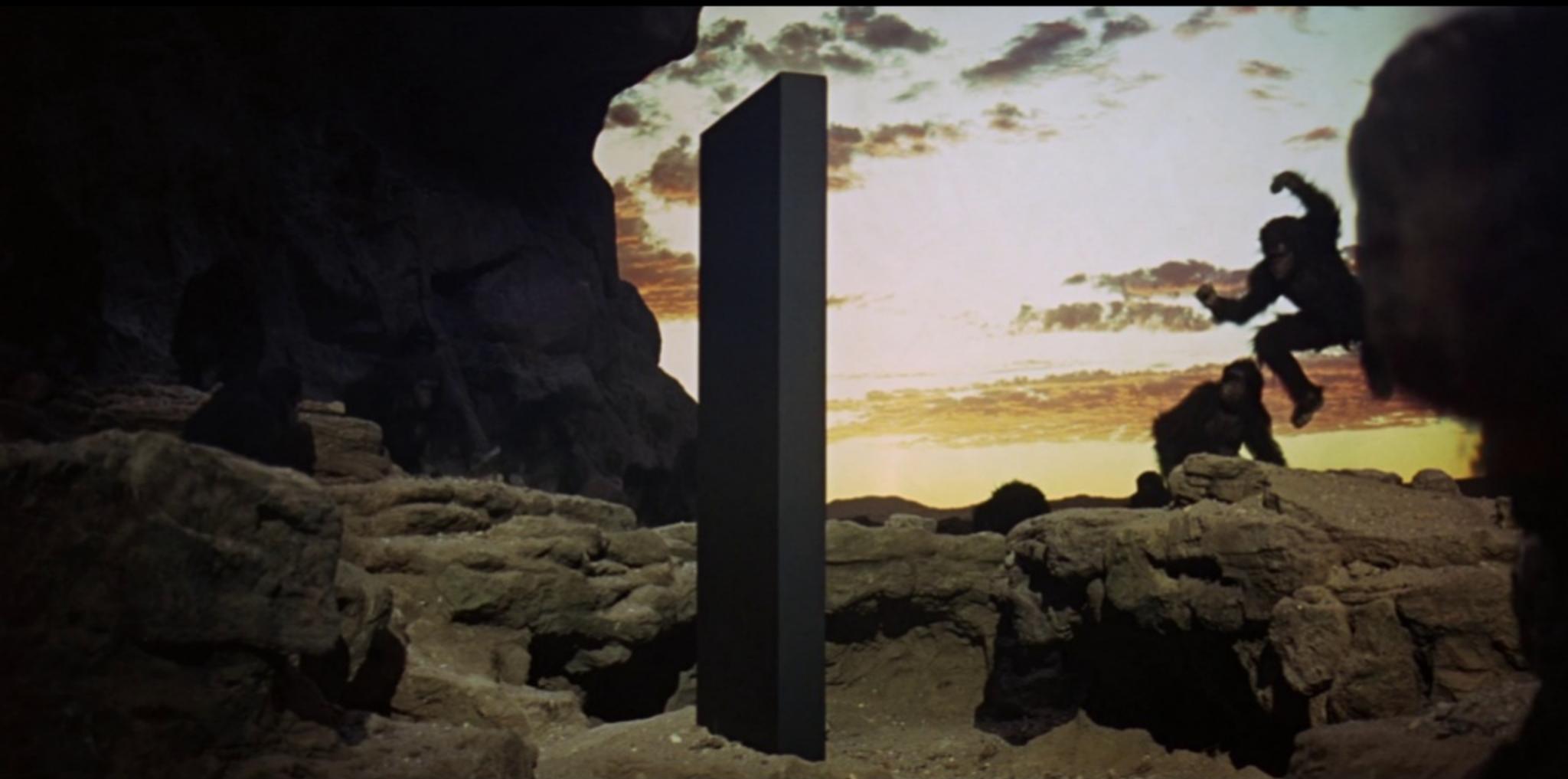
# CLASSICAL BI PLATFORM



amazon  
REDSHIFT

THE DAWN OF LETGO

# MOVING TO $\mu$ -SERVICES AND EVENTS

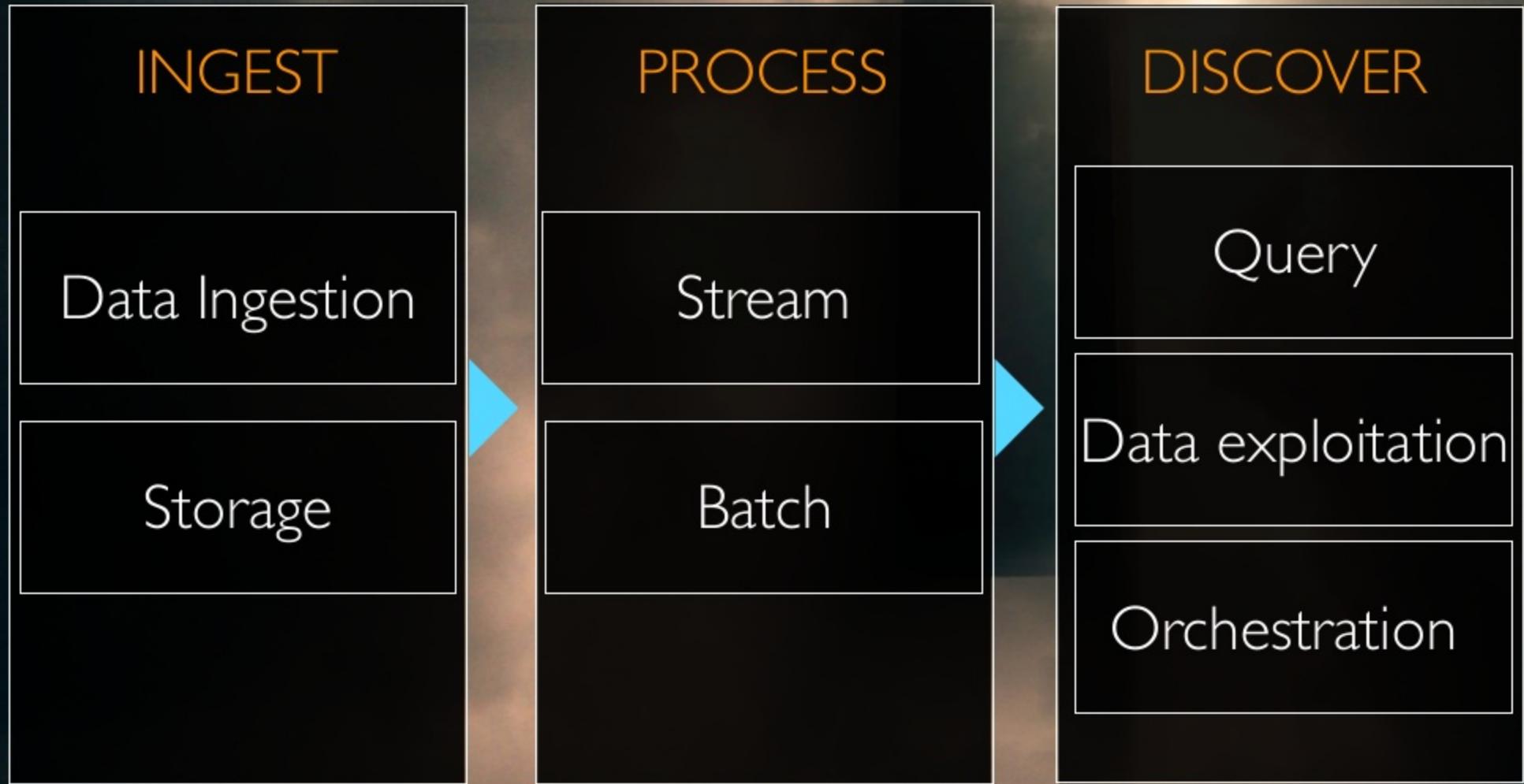


THE DAWN OF LETGO

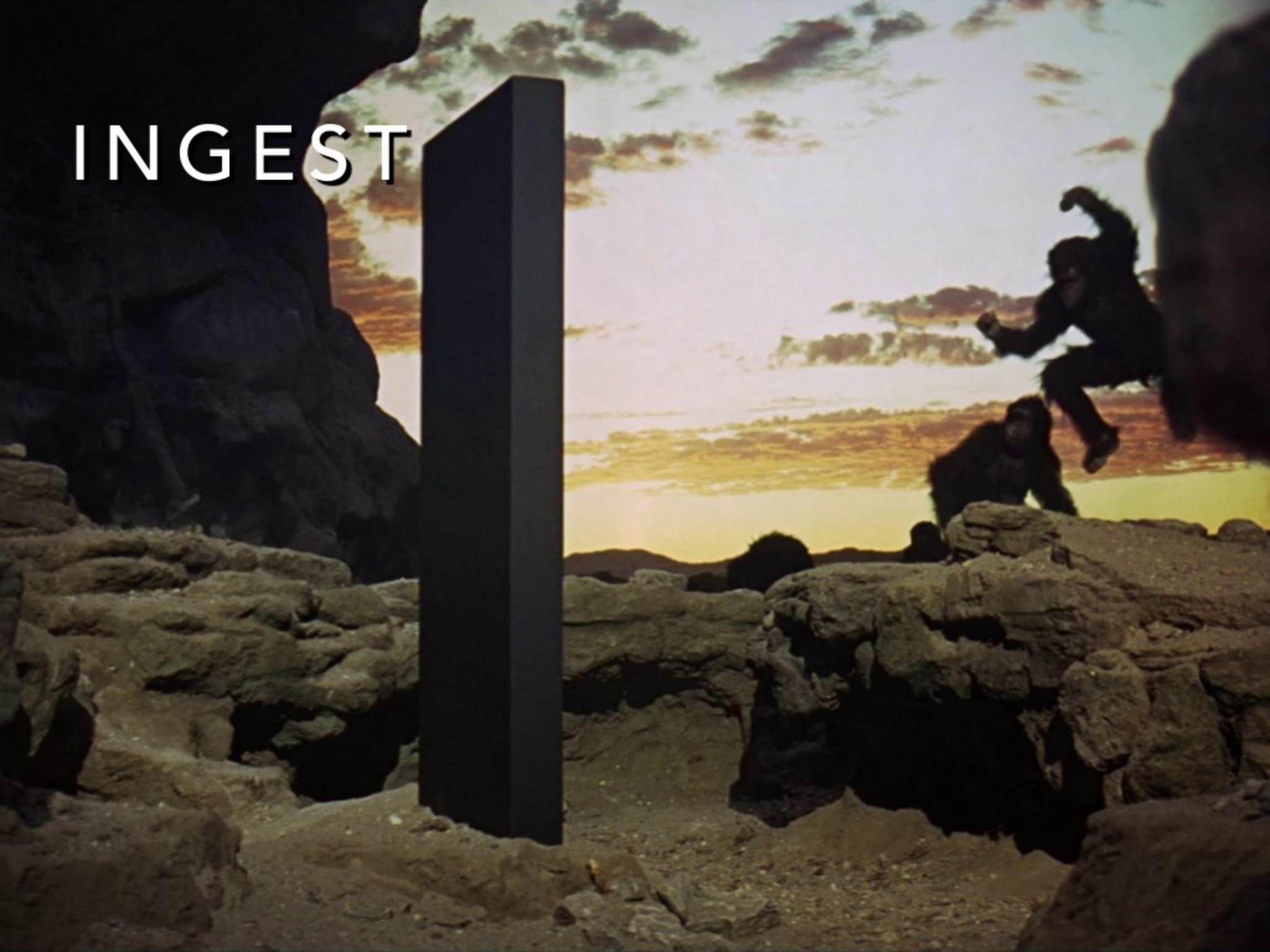
# MOVING TO $\mu$ -SERVICES AND EVENTS

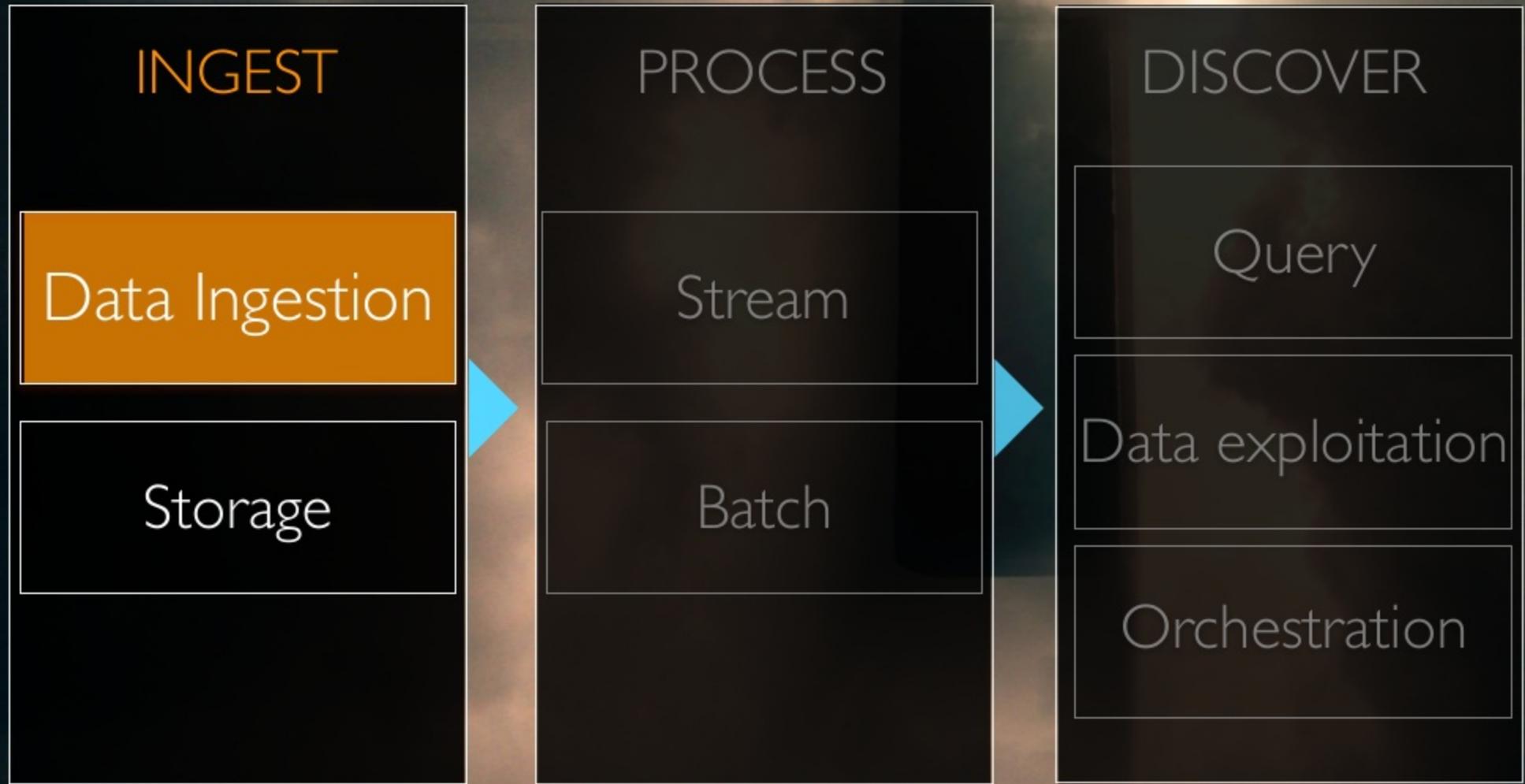
Tracking Events

Domain Events



# INGEST





DATA INGESTION

# OUR GOAL



**amazon**  
SQS



**kafka**

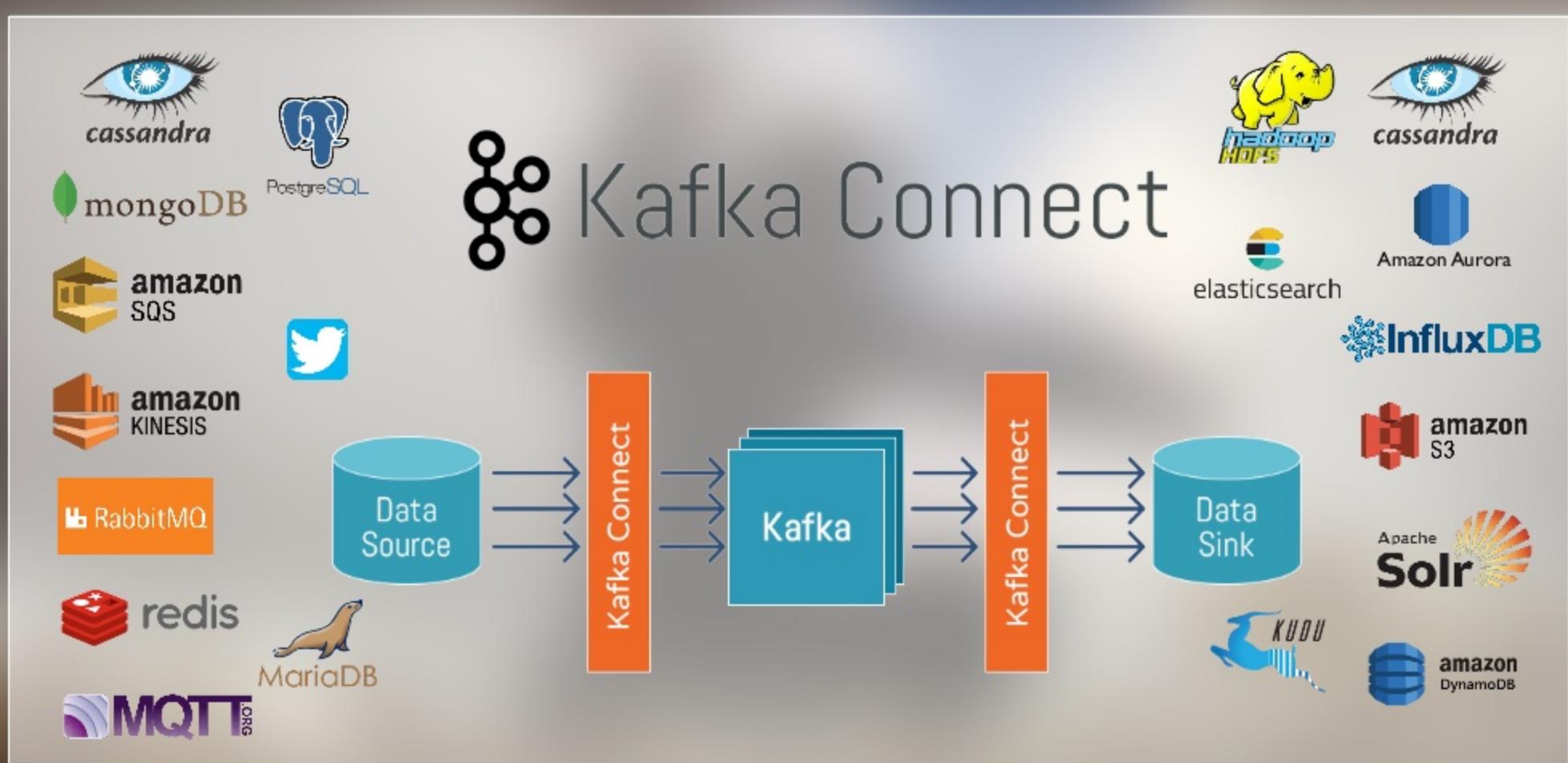
DATA INGESTION

# THE DISCOVERY



DATA INGESTION

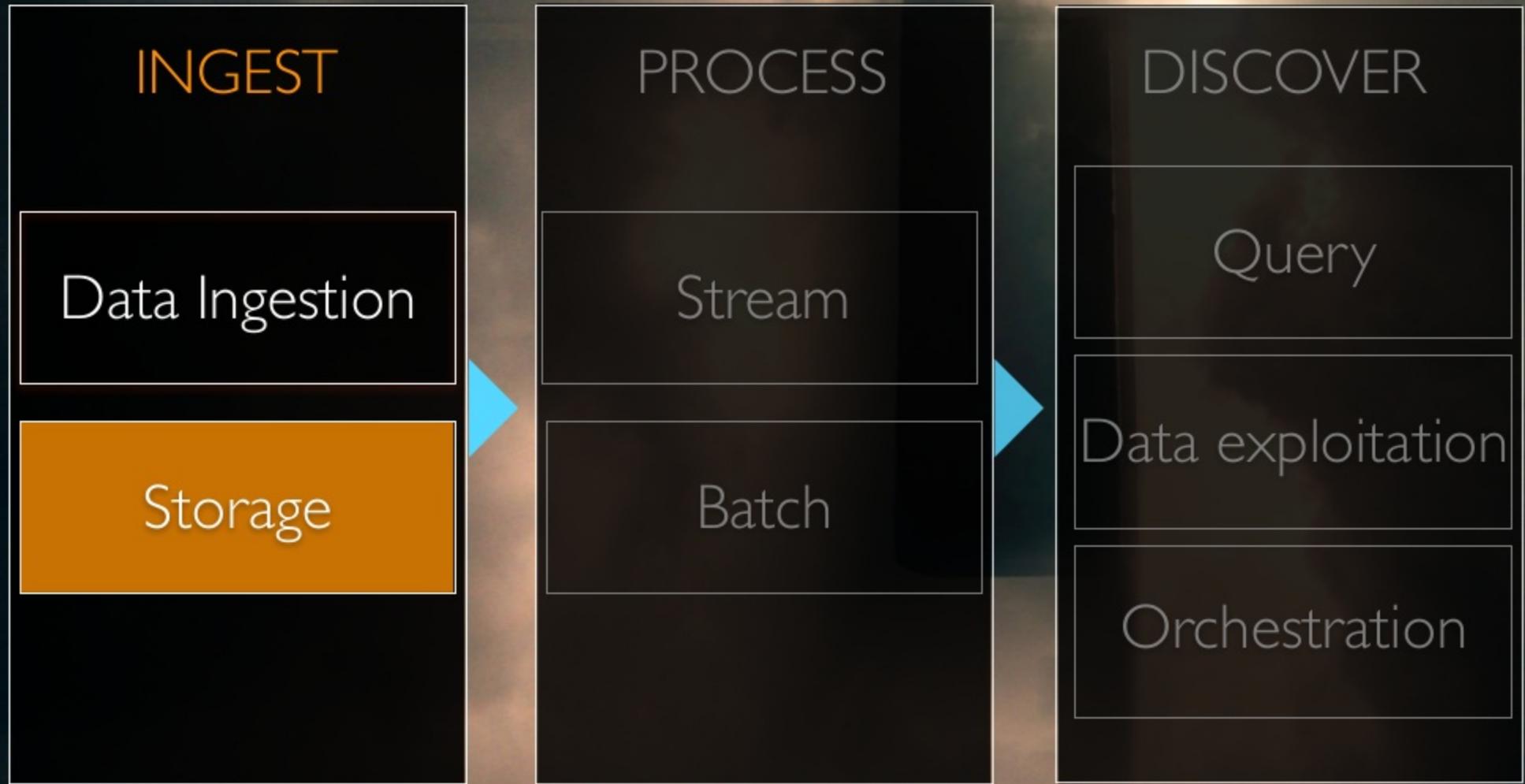
# KAFKA CONNECT





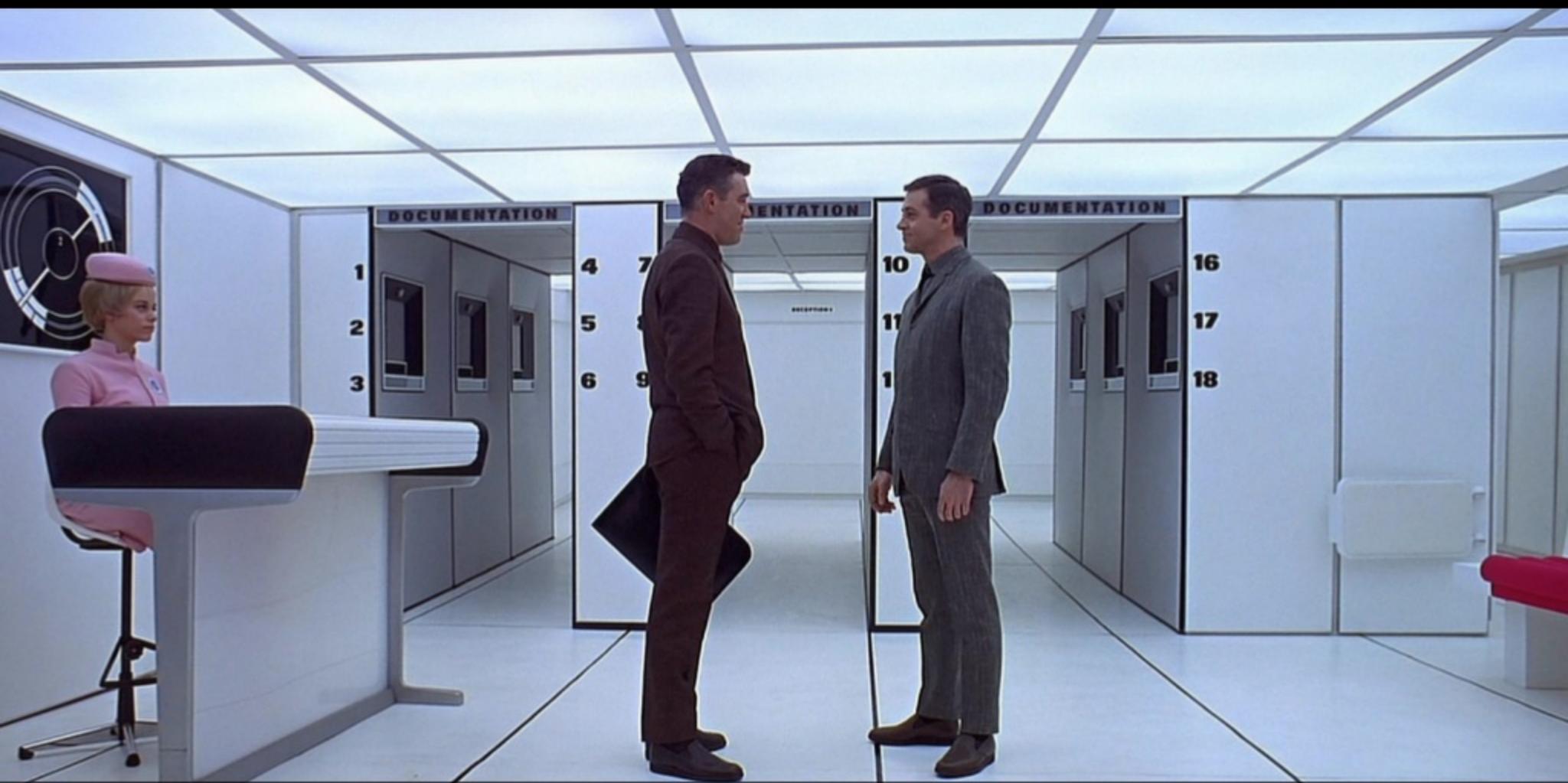


THE JOURNEY BEGINS



STORAGE

# BUILDING THE DATA LAKE



STORAGE

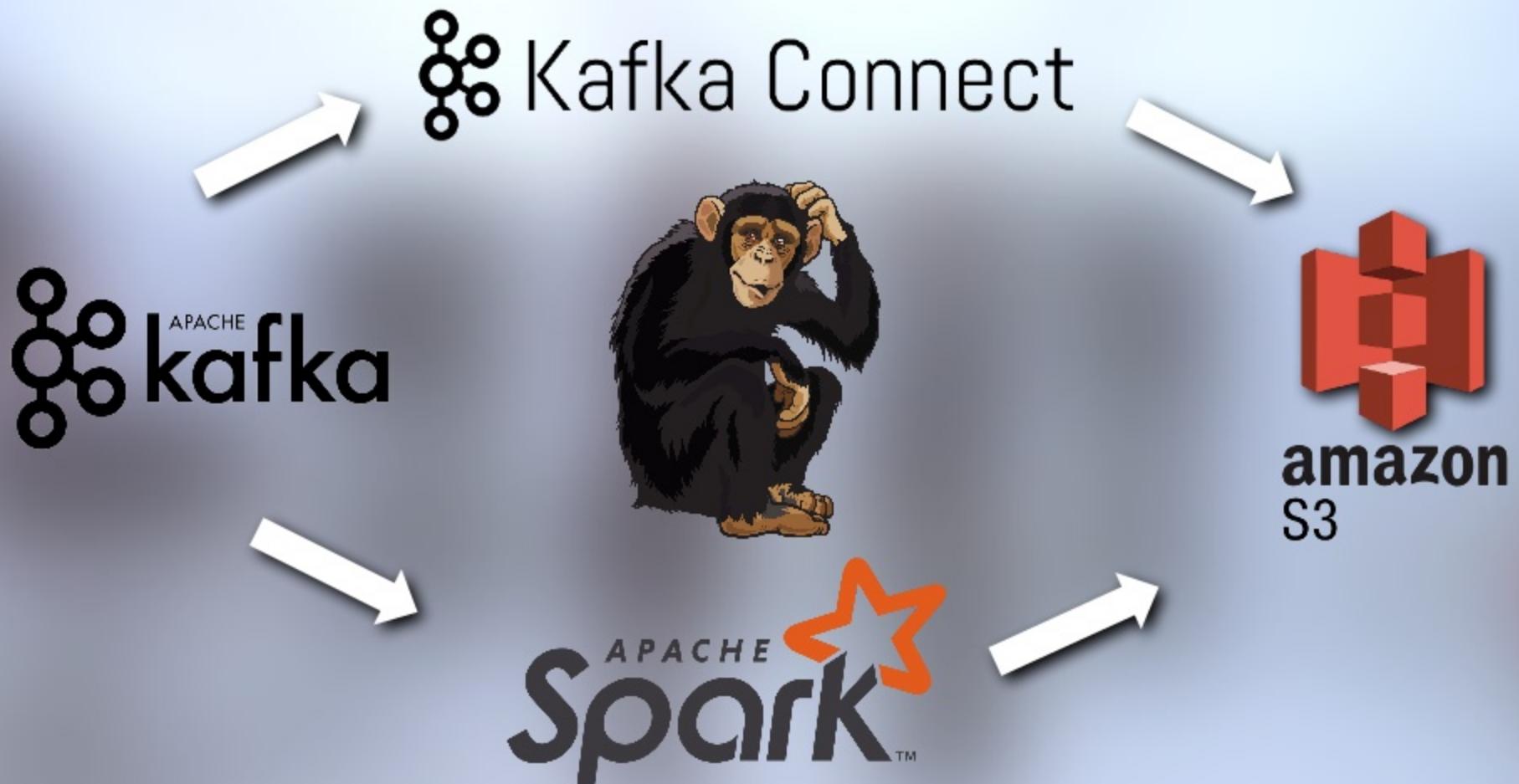
# BUILDING THE DATA LAKE



We want to store all events coming from Kafka to S3.

STORAGE

# BUILDING THE DATA LAKE



STORAGE

# SOMETIMES... SHIT HAPPENS



## ZERO GRAVITY TOILET

PASSENGERS ARE ADVISED TO  
READ INSTRUCTIONS BEFORE USE

- 1 The toilet is of the standard zero-gravity type. Depending on requirements, certain A colour options can be used, which of which are clearly marked on the toilet compartment. When operating options A, depress lever and a flush valve alternative will be triggered through the slot mechanism underneath. When an alarm function is activated by the alarm function switch on the base "G" will be heard. Press the alarm function ring one tenth before the maximum point with your hand if it fails.
- 2 The toilet is very simple for use. The flusher chamber is activated by the small switch on the left. When switched, toilet can very easily be in initial position. If the toilet always remains in initial position, then the flusher chamber is in maximum position. In this case, activate by pressing the other button.
- 3 The controls for option 1 are located on the control unit. The red colour switch places the watermeter on the position. The green colour switch places the watermeter on the position. The yellow colour switch places the watermeter on the position. The blue colour switch places the watermeter on the position. The orange colour switch places the watermeter on the position.
- 4 You can have the function if the green light is on over the door. Press the "A" function, one of the function buttons. It will appear that the door is closed. Then the "G" function and button to the right of the door and function of the door will appear. These green light light goes on or will not go off after about 10 seconds. When these doors behind you.
- 5 If you are the passenger, first washroom and place all your clothes in the washroom and then go to the washroom. If you are the passenger, first washroom and then go to the washroom. If you are the passenger, first washroom and then go to the washroom. If you are the passenger, first washroom and then go to the washroom.
- 6 The function of the toilet is to flush the toilet when the toilet is full. The function of the toilet is to flush the toilet when the toilet is full. The function of the toilet is to flush the toilet when the toilet is full. The function of the toilet is to flush the toilet when the toilet is full.
- 7 If the red light goes on over the door, then the toilet is in use. When the green light is on over the door, then the toilet is not in use. When the yellow light is on over the door, then the toilet is not in use. When the blue light is on over the door, then the toilet is not in use. When the orange light is on over the door, then the toilet is not in use.

#SAISExp2

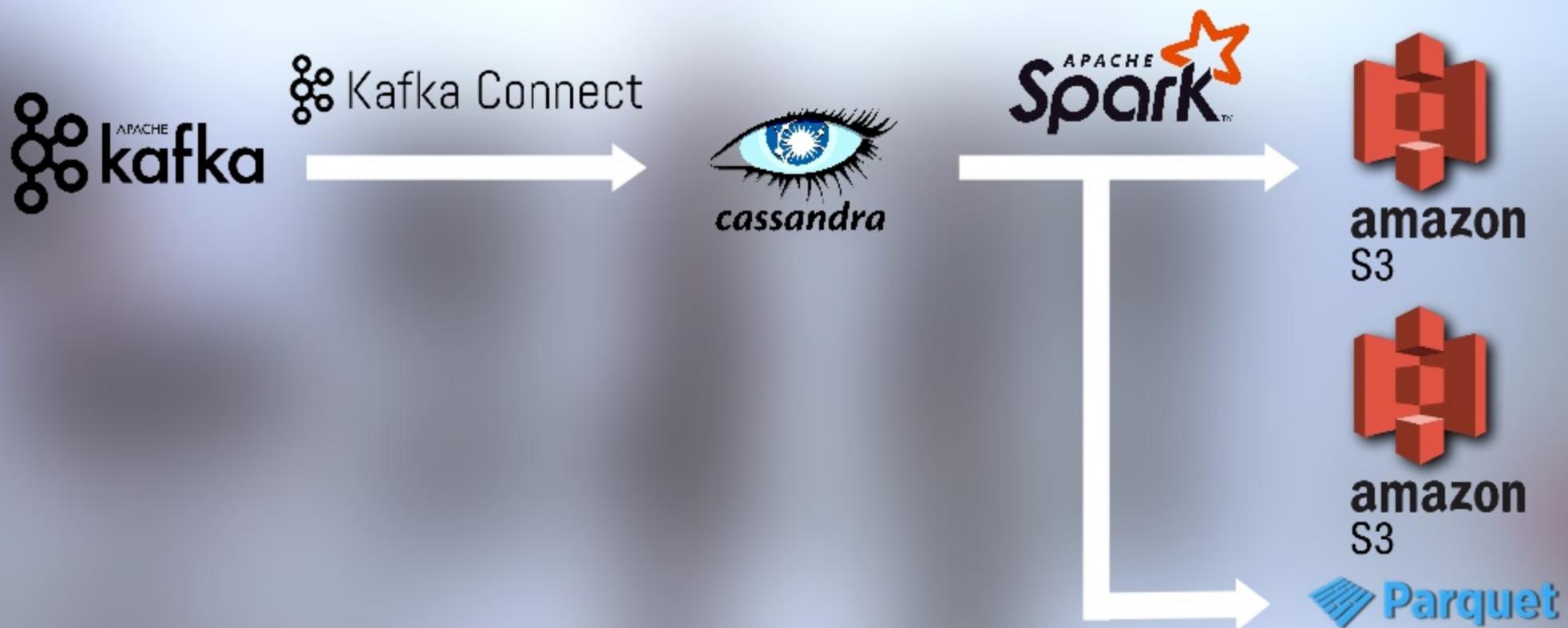
STORAGE

# DUPLICATED EVENTS



STORAGE

# DUPLICATED EVENTS

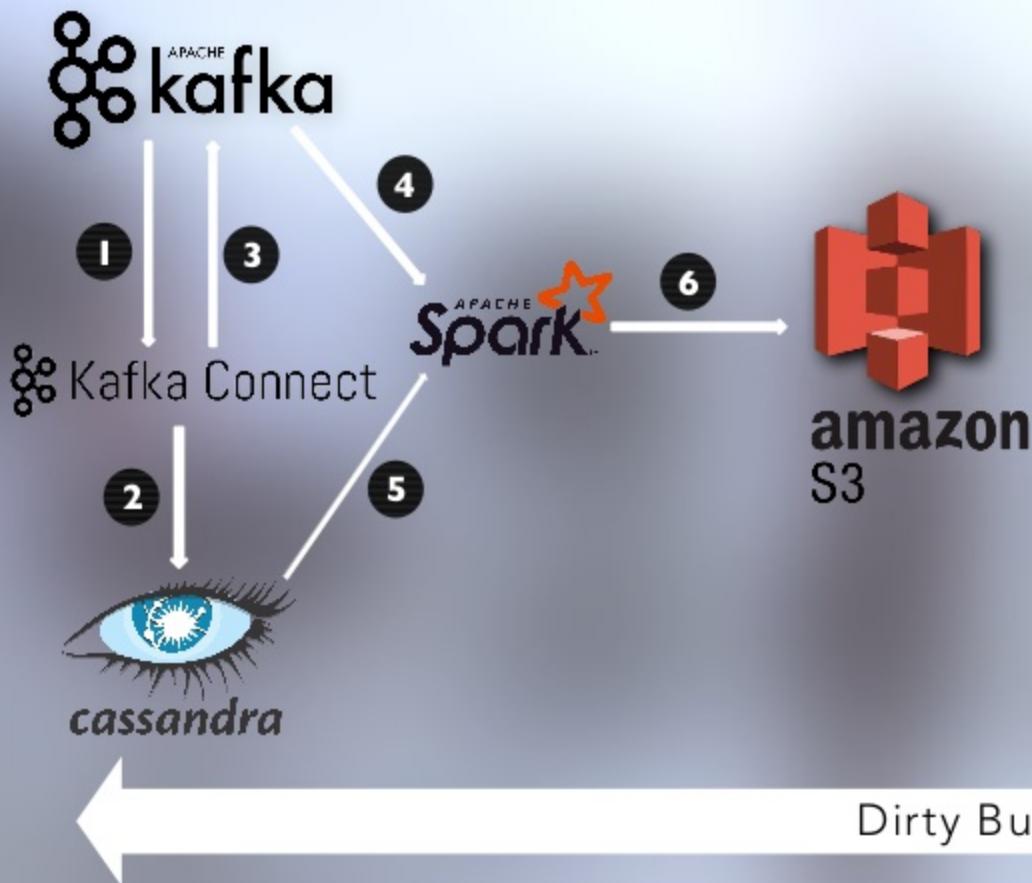


STORAGE

# (VERY) LATE EVENTS



## (VERY) LATE EVENTS



- 1 Read batch of events from Kafka.
- 2 Write each event to Cassandra.
- 3 Write “dirty hours” to compact topic: Key=(event\_type, hour).
- 4 Read dirty “hours” topic.
- 5 Read all events with dirty hours.
- 6 Store in S3

STORAGE

# S3 PROBLEMS



STORAGE

# S3 PROBLEMS



## SOME S3 BIG DATA PROBLEMS:

1. Eventual consistency
2. Very slow renames

STORAGE

# S3 PROBLEMS: EVENTUAL CONSISTENCY



STORAGE

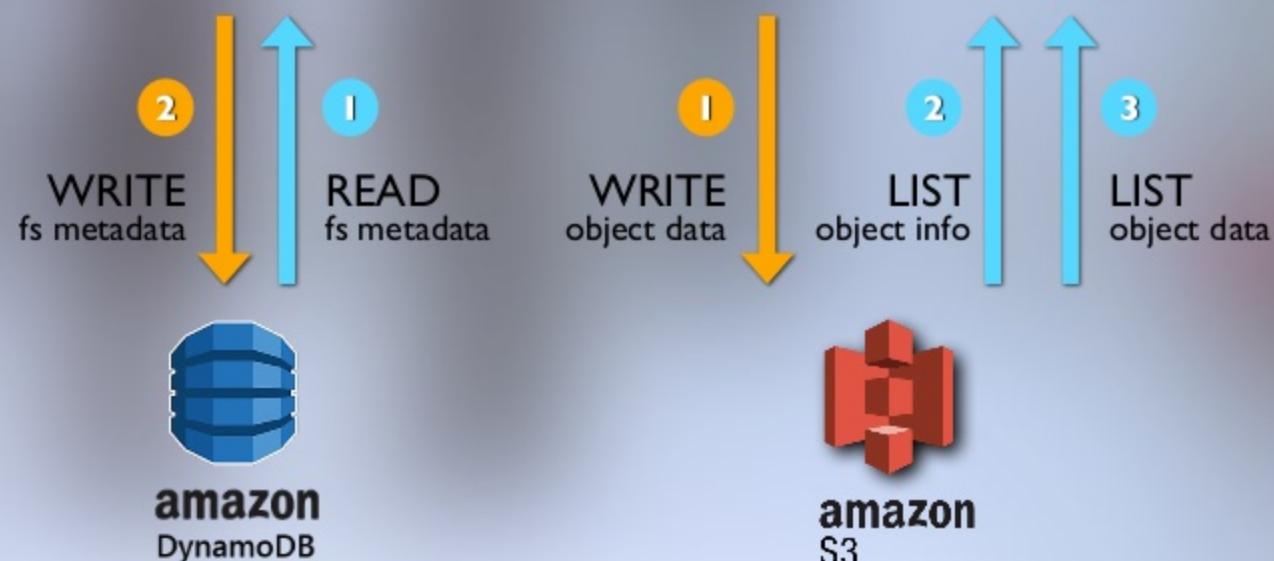
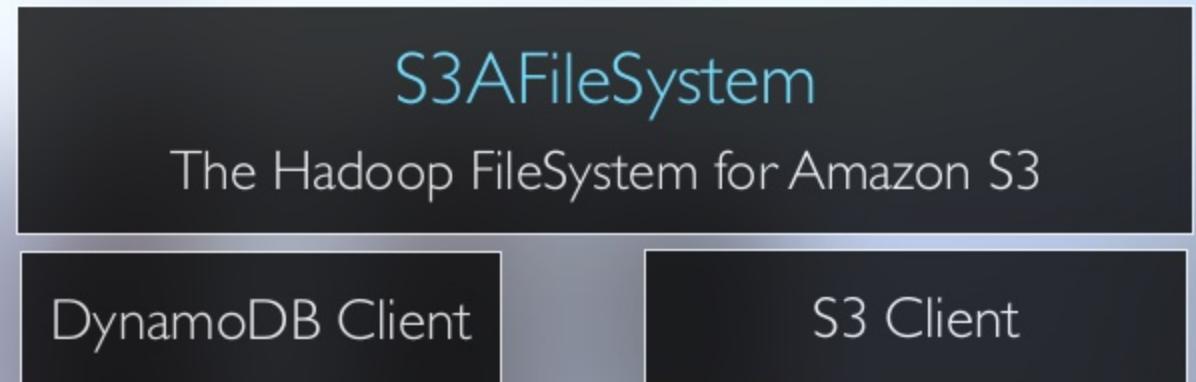
# S3 PROBLEMS: EVENTUAL CONSISTENCY



S3GUARD



FileSystem  
Operations



## STORAGE

# S3 PROBLEMS: SLOW RENAMES



### Summary

RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks
Active(2)	4	107.1 KB / 23.1 GB	0.0 B	8	0	0	8
Dead(10)	4	128.3 KB / 95.5 GB	0.0 B	80	0	1436	1436
Total(12)	8	235.4 KB / 118.6 GB	0.0 B	88	0	1444	1444

### Executors

¿Job freeze?

Type a prefix and press Enter to search. Press ESC to clear.

Upload Create folder More

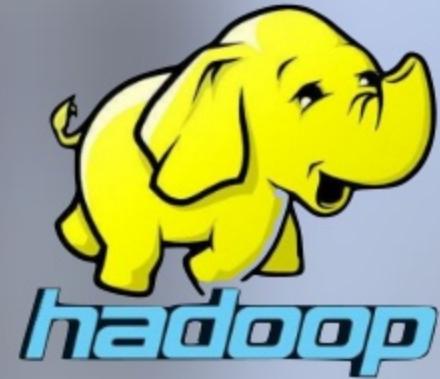
- Name ↑ ↓
- \_temporary
- year=2017

# S3 PROBLEMS: SLOW RENAMES



New **Hadoop 3.1** S3A committers:

- Directory
- Partitioned
- Magic





PROCESS



## INGEST

Data Ingestion

Storage

## PROCESS

Stream

Batch

## DISCOVER

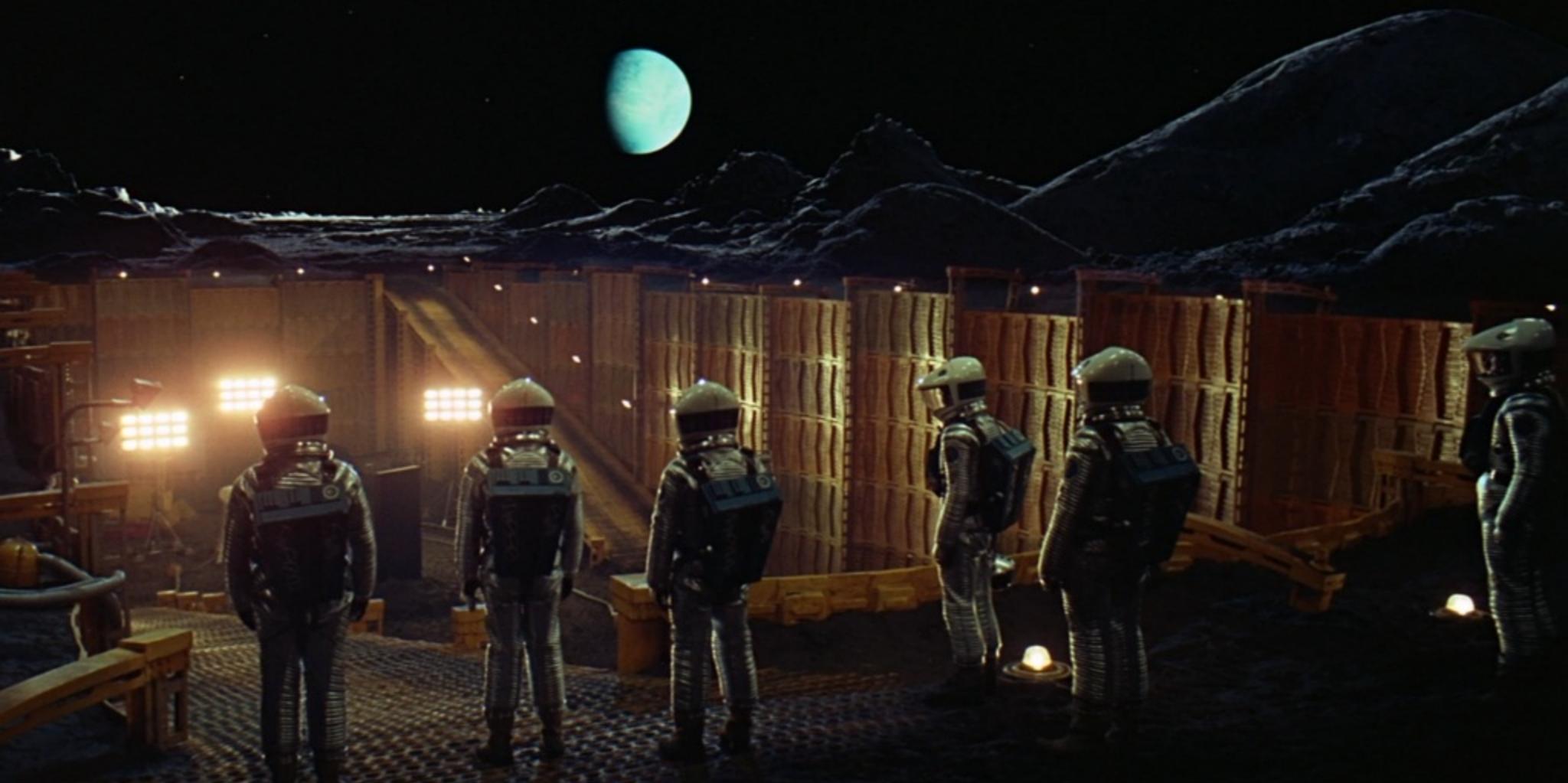
Query

Data exploitation

Orchestration

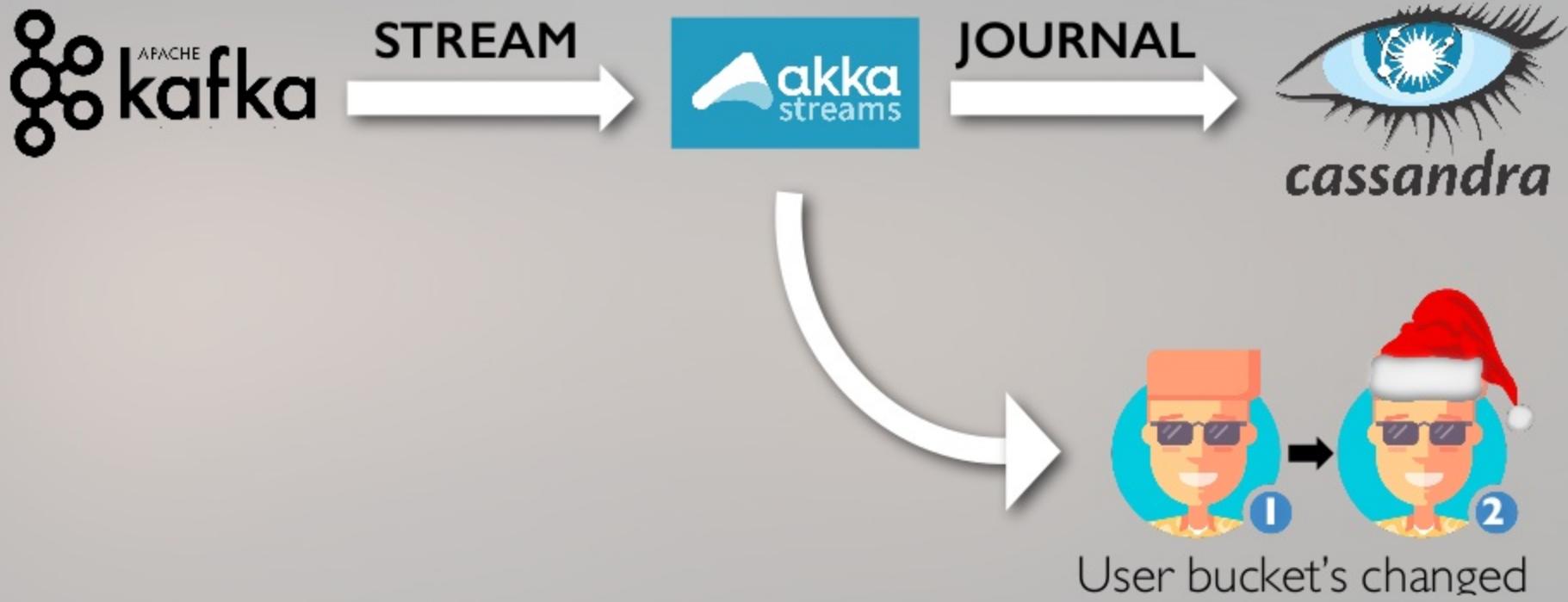
STREAM

# REAL TIME USER SEGMENTATION



STREAM

# REAL TIME USER SEGMENTATION



STREAM

# REAL TIME PATTERN DETECTION



Is it still available?

Is the price negotiable?

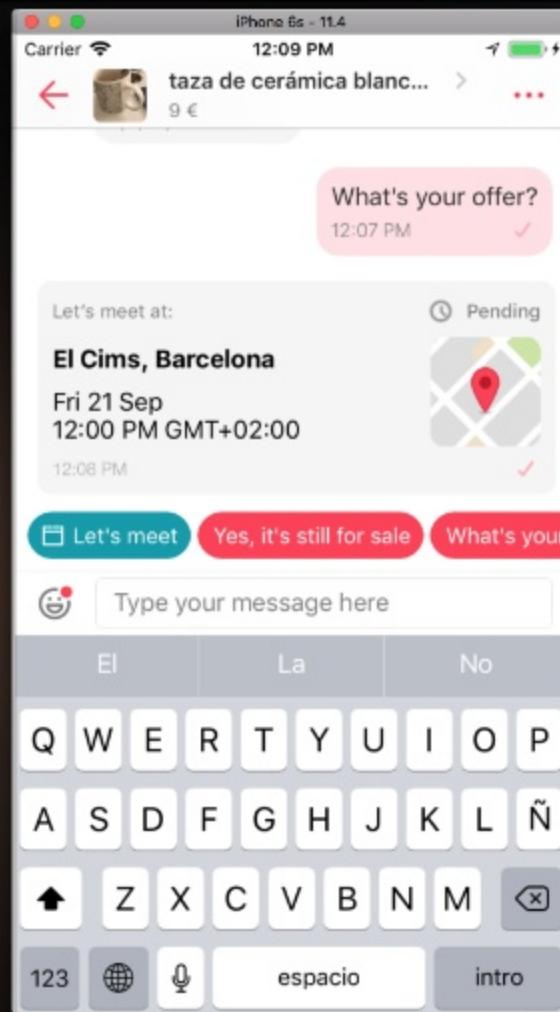
What condition is it in?

I offer you....\$

Could we meet at.....?

STREAM

# REAL TIME PATTERN DETECTION

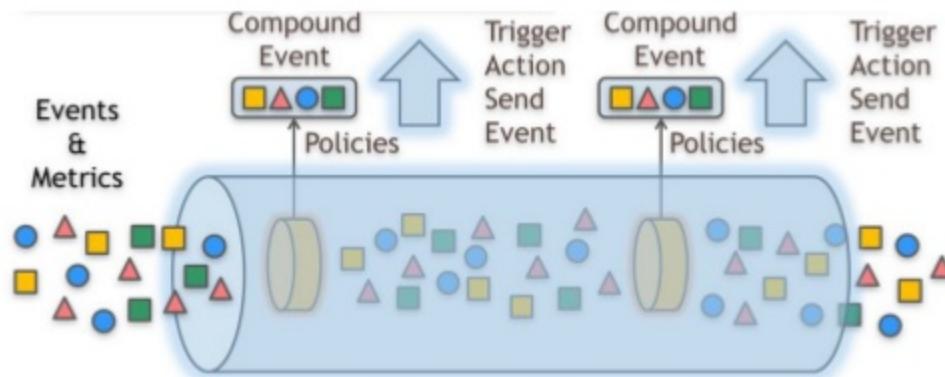


Structured data

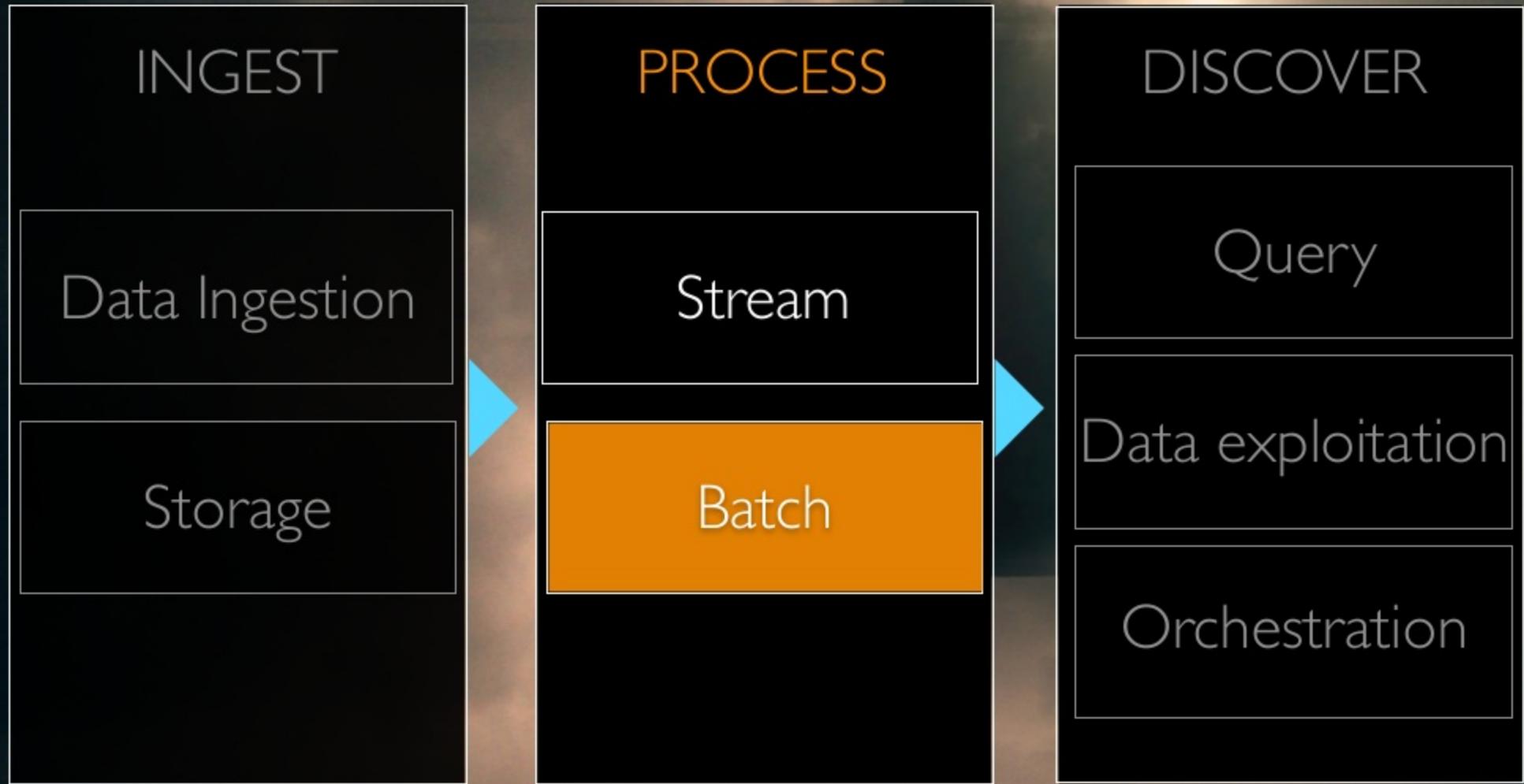
```
{  
  "type": "meeting_proposal",  
  "properties": {  
    "location_name": "Letgo HQ",  
    "geo": {  
      "lat": "41.390205",  
      "lon": "2.154007"  
    },  
    "date": "1511193820350",  
    "meeting_id": "23213213213"  
  }  
}
```

STREAM

# REAL TIME PATTERN DETECTION



Meeting proposed + meeting accepted = emit accepted-meeting event  
Meeting proposed + nothing in X time = “You have a proposal to meet”

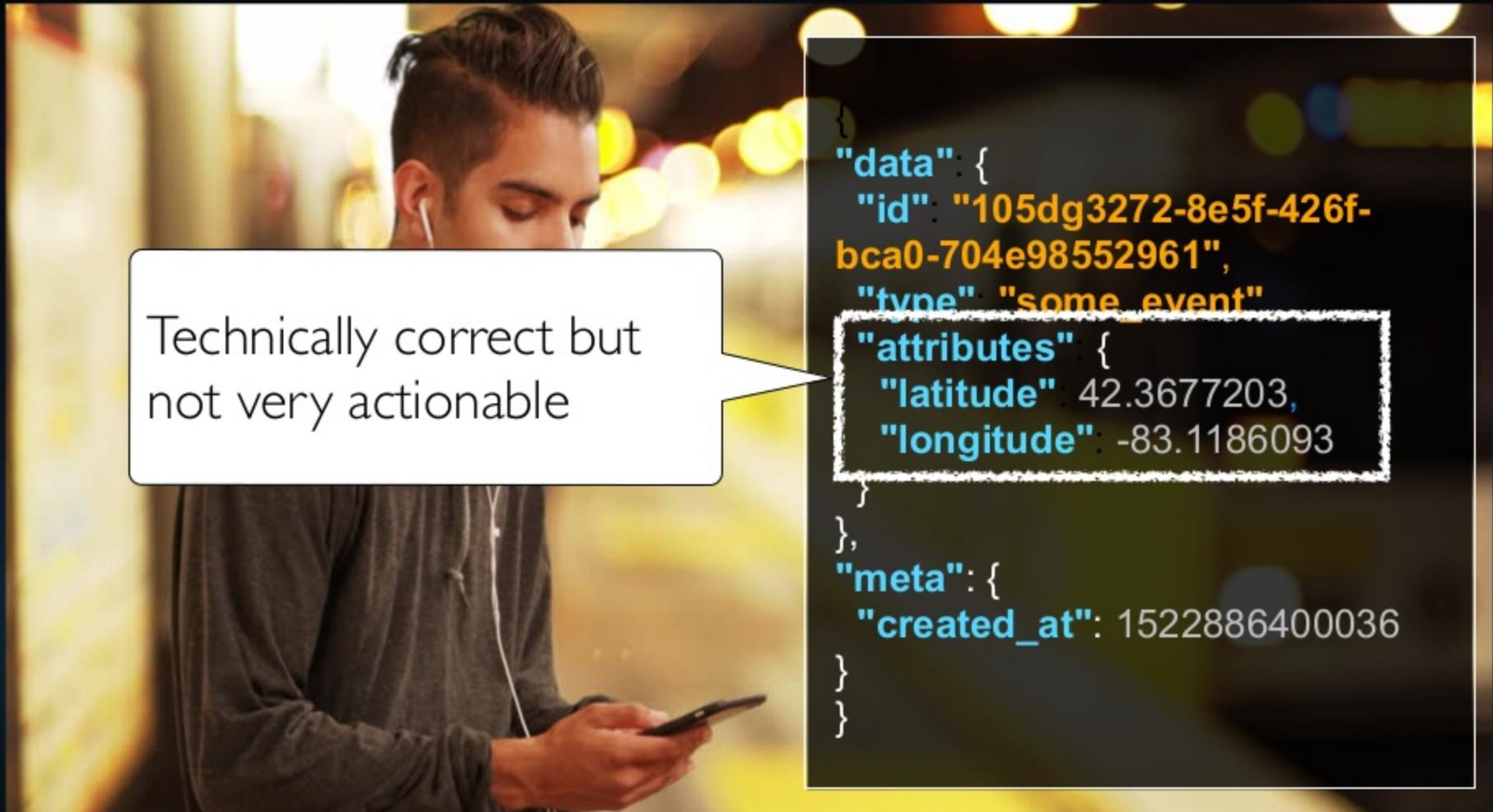


BATCH

# GEODATA ENRICHMENT



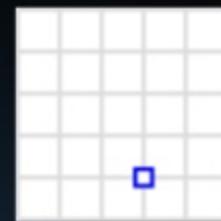
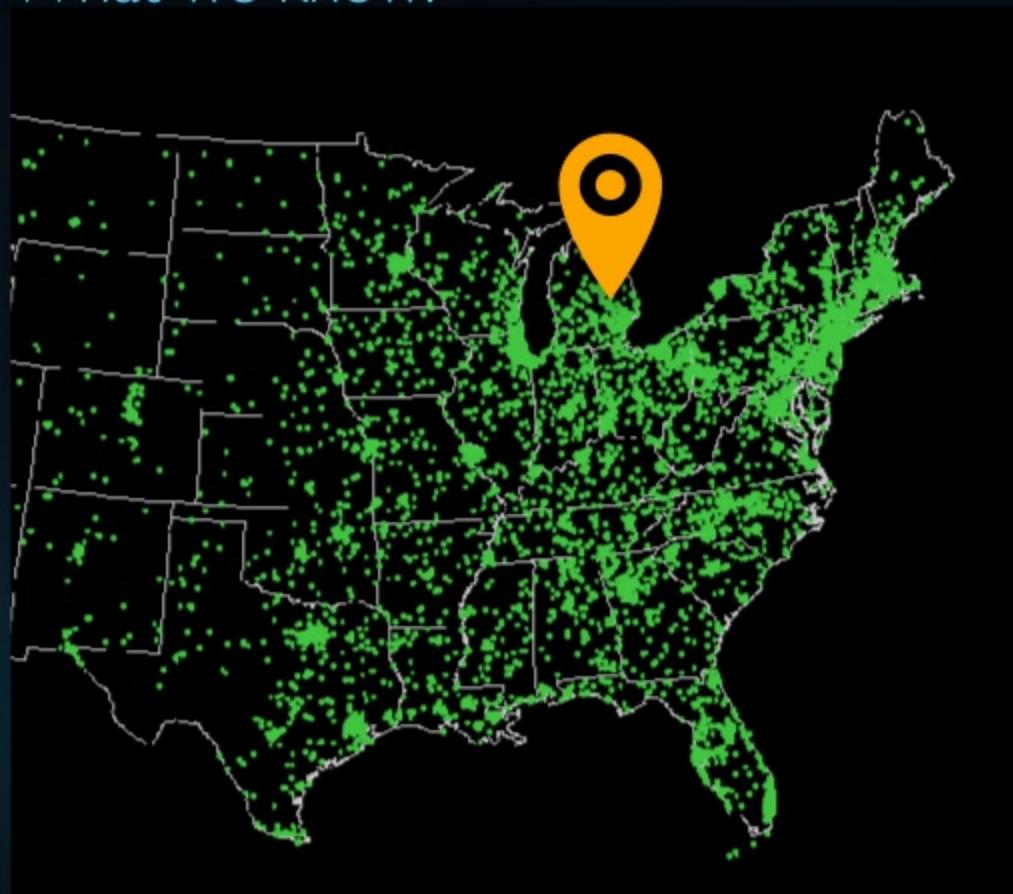
# GEODATA ENRICHMENT



BATCH

# GEODATA ENRICHMENT

What we know:



(42.3677203, -83.1186093)

- City: Detroit
- Postal code: 48206
- State: Michigan
- DMA: Detroit
- Country: US

BATCH

# GEODATA ENRICHMENT

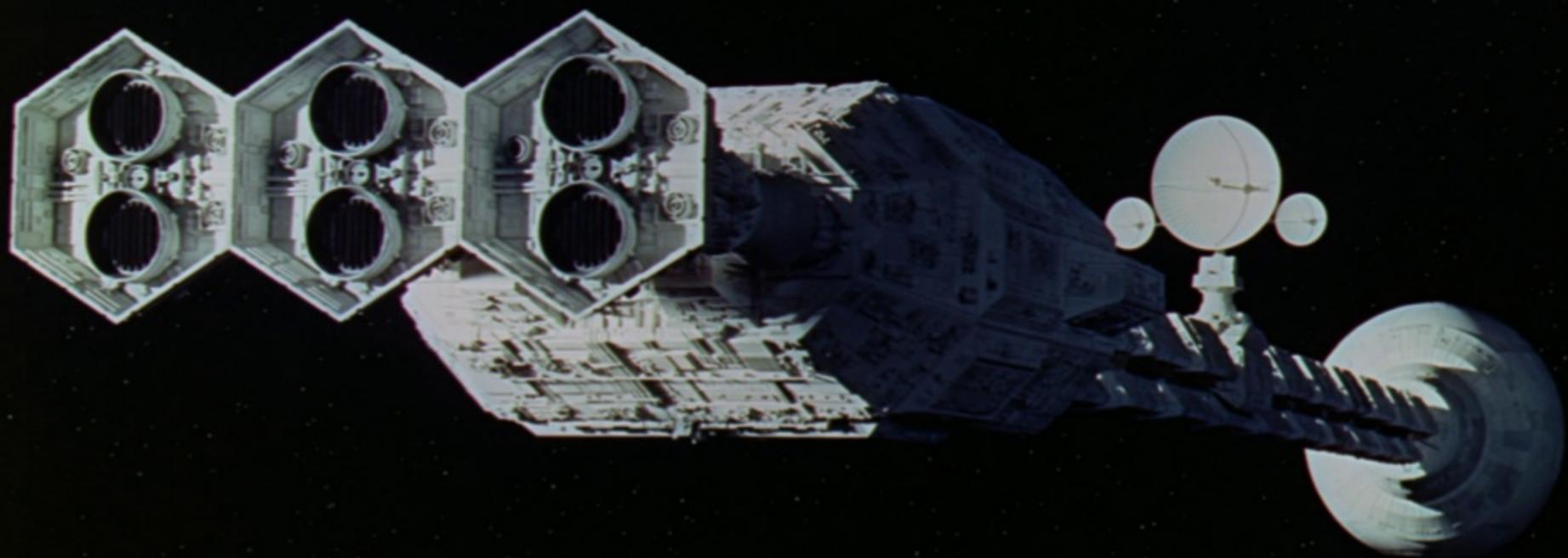
How we do it:

- Populating JTS indices from WKT polygon data
- Custom Spark SQL UDF



```
SELECT geodata.dma_name  
      geodata.dma_number AS dma_number  
      geodata.city     AS city  
      geodata.state    AS state  
      geodata.zip_code AS zip_code  
FROM  (  
        SELECT  
          geodata(longitude, latitude) AS geodata  
        FROM  ....  
      )
```

# DISCOVER



## INGEST

Data Ingestion

Storage

## PROCESS

Stream

Batch

## DISCOVER

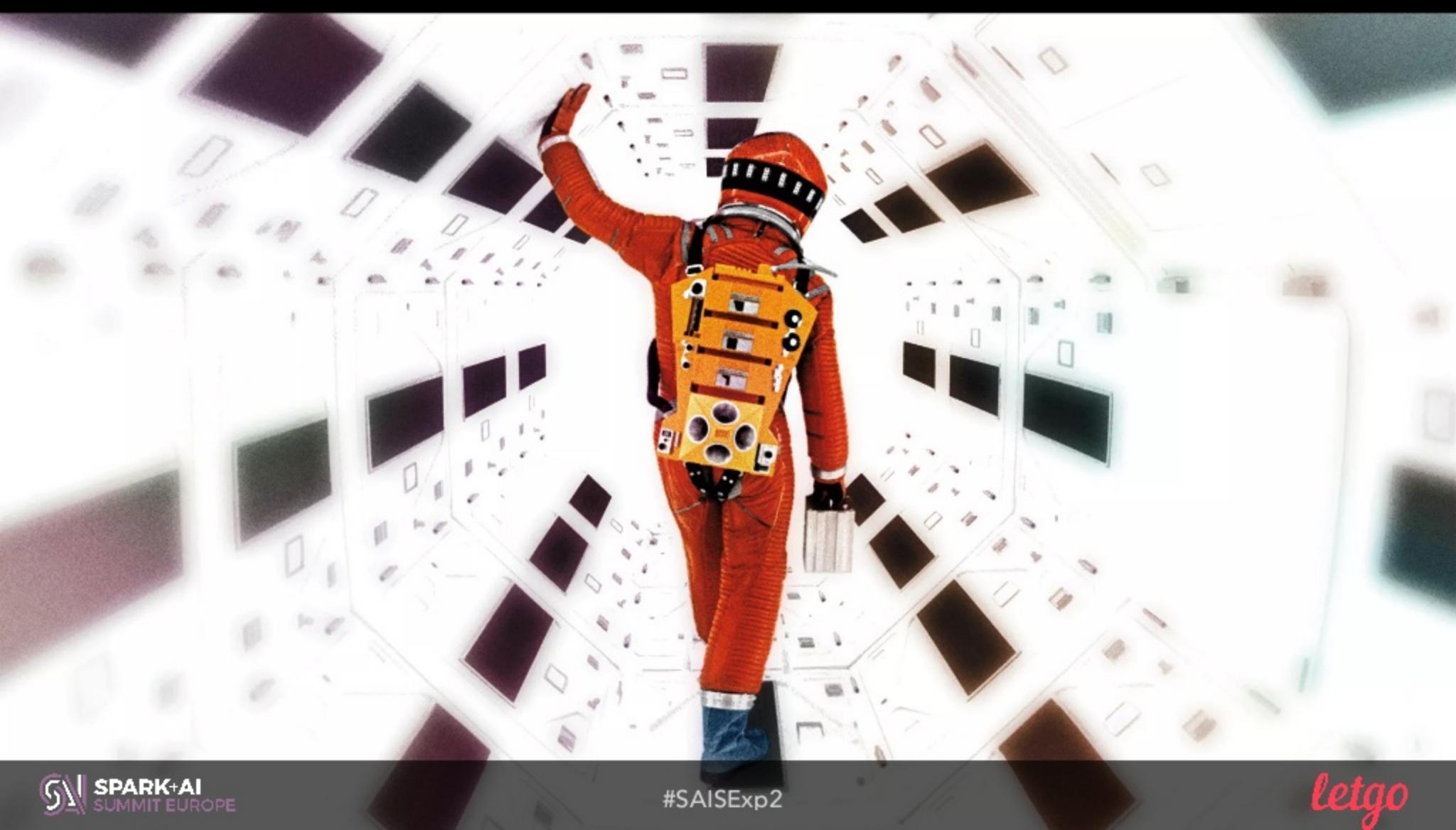
Query

Data exploitation

Orchestration

QUERY

# QUERYING DATA



QUERY

# QUERYING DATA



QUERY

# QUERYING DATA



cassandra



MariaDB



amazon  
S3



kafka



Amazon Aurora



redis



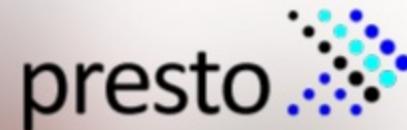
amazon  
REDSHIFT

QUERY

# QUERYING DATA



**HAWQ**



Amazon Athena



**amazon**  
REDSHIFT  
SPECTRUM

QUERY

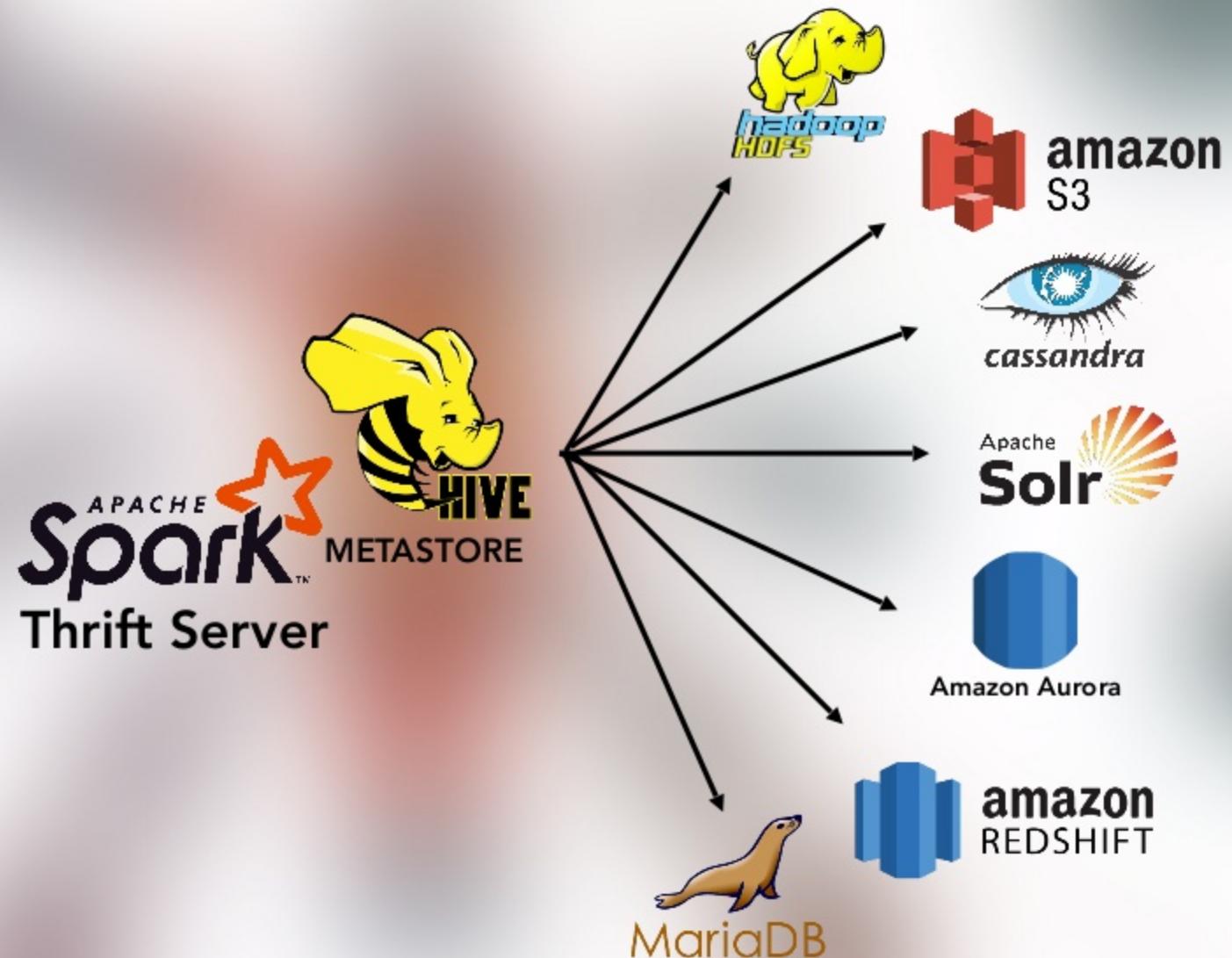
# QUERYING DATA

**READ TABLES EVERYWHERE**



QUERY

# QUERYING DATA



QUERY

# QUERYING DATA



```
CREATE TABLE IF NOT EXISTS
database_name.table_name(
some_column STRING,
...
dt DATE
)
USING json
PARTITIONED BY ('dt')
```



amazon  
S3



```
CREATE EXTERNAL TABLE IF NOT EXISTS
database_name.table_name(
some_column STRING,
dt DATE
)
PARTITIONED BY ( dt )
USING PARQUET
LOCATION 's3a://bucket-name/database_name/table_name'
```



```
CREATE TEMPORARY VIEW table_name
USING org.apache.spark.sql.cassandra
OPTIONS (
table "table_name",
keyspace "keyspace_name")
```



amazon  
REDSHIFT

```
CREATE TABLE IF NOT EXISTS database_name.table_name
using com.databricks.spark.redshift
options (
dbtable 'schema.redshift_table_name',
tempdir 's3a://redshift-temp/',
url 'jdbc:redshift://xxxx.redshift.amazonaws.com:5439/letgo?
user=xxx&password=xxx',
forward_spark_s3_credentials 'true')
```

QUERY

# QUERYING DATA

CREATE TABLE ...  
STORED AS...

VS

CREATE TABLE ...  
USING [parquet,json,csv...]



70%

Higher performance!

## QUERY

# QUERYING DATA: BATCHES WITH SQL

1

Creating the  
table

2

Inserting data

## QUERY

# QUERYING DATA: BATCHES WITH SQL

1

Creating the  
table

```
CREATE EXTERNAL TABLE IF NOT EXISTS database some_name(  
    user_id STRING,  
    column_b STRING,  
    ...  
)  
USING PARQUET  
PARTITIONED BY (dt` STRING)  
LOCATION 's3a://example/some_table'
```

## QUERY

# QUERYING DATA: BATCHES WITH SQL

2

Inserting data

```
INSERT OVERWRITE TABLE database.some_name PARTITION(dt)
SELECT
    user_id,
    column_b,
    dt
FROM other_table
...
```

## QUERY

# QUERYING DATA: BATCHES WITH SQL

Viewing 1 to 200

Name	Last modified	Size	Storage class
part-00000-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:38 AM GMT+0100	6.5 MB	Standard
part-00001-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:39 AM GMT+0100	10.6 MB	Standard
part-00002-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:39 AM GMT+0100	10.5 MB	Standard
part-00003-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:40 AM GMT+0100	8.9 MB	Standard
part-00004-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:41 AM GMT+0100	9.3 MB	Standard
part-00005-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:42 AM GMT+0100	7.5 MB	Standard
part-00006-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:42 AM GMT+0100	10.8 MB	Standard
part-00007-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:43 AM GMT+0100	7.2 MB	Standard
part-00008-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:44 AM GMT+0100	9.6 MB	Standard
part-00009-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:45 AM GMT+0100	9.6 MB	Standard
part-00010-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:46 AM GMT+0100	8.7 MB	Standard
part-00011-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000			
part-00012-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000			

Problem?



## QUERY

# QUERYING DATA: BATCHES WITH SQL

Name	Last modified	Size	Storage class
part-00000-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:38 AM GMT+0100	6.5 MB	Standard
part-00001-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:39 AM GMT+0100	10.6 MB	Standard
part-00002-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:39 AM GMT+0100	10.5 MB	Standard
part-00003-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:40 AM GMT+0100	8.9 MB	Standard
part-00004-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:41 AM GMT+0100	9.3 MB	Standard
part-00005-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:42 AM GMT+0100	7.5 MB	Standard
part-00006-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:42 AM GMT+0100	10.8 MB	Standard
part-00007-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:43 AM GMT+0100	7.2 MB	Standard
part-00008-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:44 AM GMT+0100	9.6 MB	Standard
part-00009-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:45 AM GMT+0100	9.6 MB	Standard
part-00010-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:46 AM GMT+0100	8.7 MB	Standard
part-00011-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:47 AM GMT+0100	9.2 MB	Standard
part-00012-d0f97fca-b705-49e1-b3d1-b8ab2af5b174-c000	Mar 5, 2018 10:58:48 AM GMT+0100	12.9 MB	Standard

200 files because default value of  
“spark.sql.shuffle.partition”

QUERY

# QUERYING DATA: BATCHES WITH SQL



```
INSERT OVERWRITE TABLE database.some_name PARTITION(dt)
SELECT
    user_id,
    column_b,
    dt
FROM other_table
...
```



## QUERY

# QUERYING DATA: BATCHES WITH SQL



### **DISTRIBUTE BY (dt):**

Only one file not Sorted



### **CLUSTERED BY (dt, user\_id, column\_b):**

Multiple files



### **DISTRIBUTE BY (dt) SORT BY (user\_id, column\_b):**

Only one file sorted by user\_id, column\_b.

Good for joins using this properties.

## QUERY

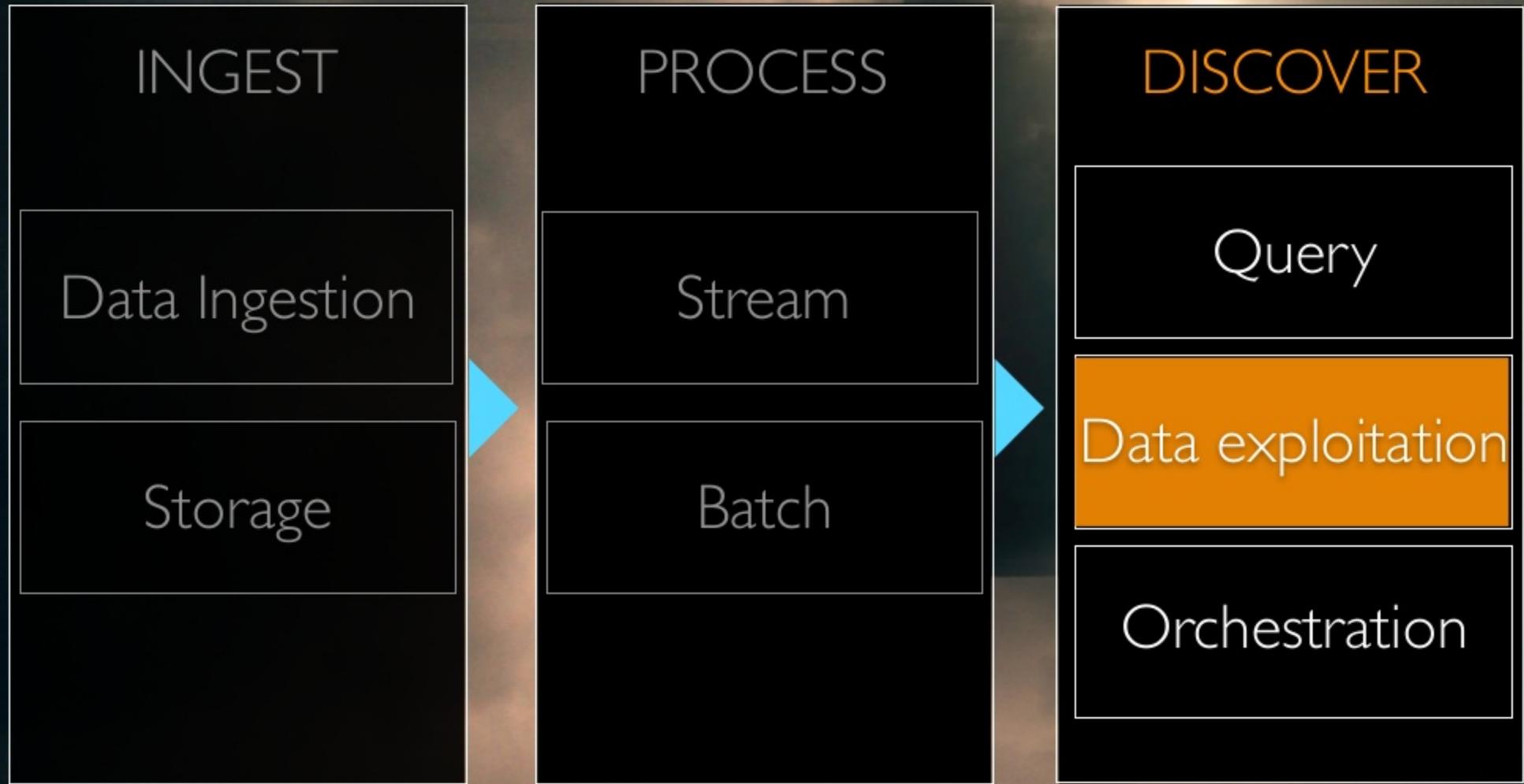
# QUERYING DATA: BATCHES WITH SQL

```
INSERT OVERWRITE TABLE database.some_name  
PARTITION(dt)  
SELECT  
user_id,  
column_b,  
dt  
FROM other_table  
...  
DISTRIBUTE BY (dt) SORT BY (user_id)
```

## QUERY

# QUERYING DATA: BATCHES WITH SQL

Name	Last modified	Size	Storage class	Viewing 1 to 1
part-00026-d86907ed-dbc5-4c5f-a399-51a72852f960.c000	Mar 5, 2018 12:35:38 PM GMT+0100	2.5 GB	Standard	Viewing 1 to 1



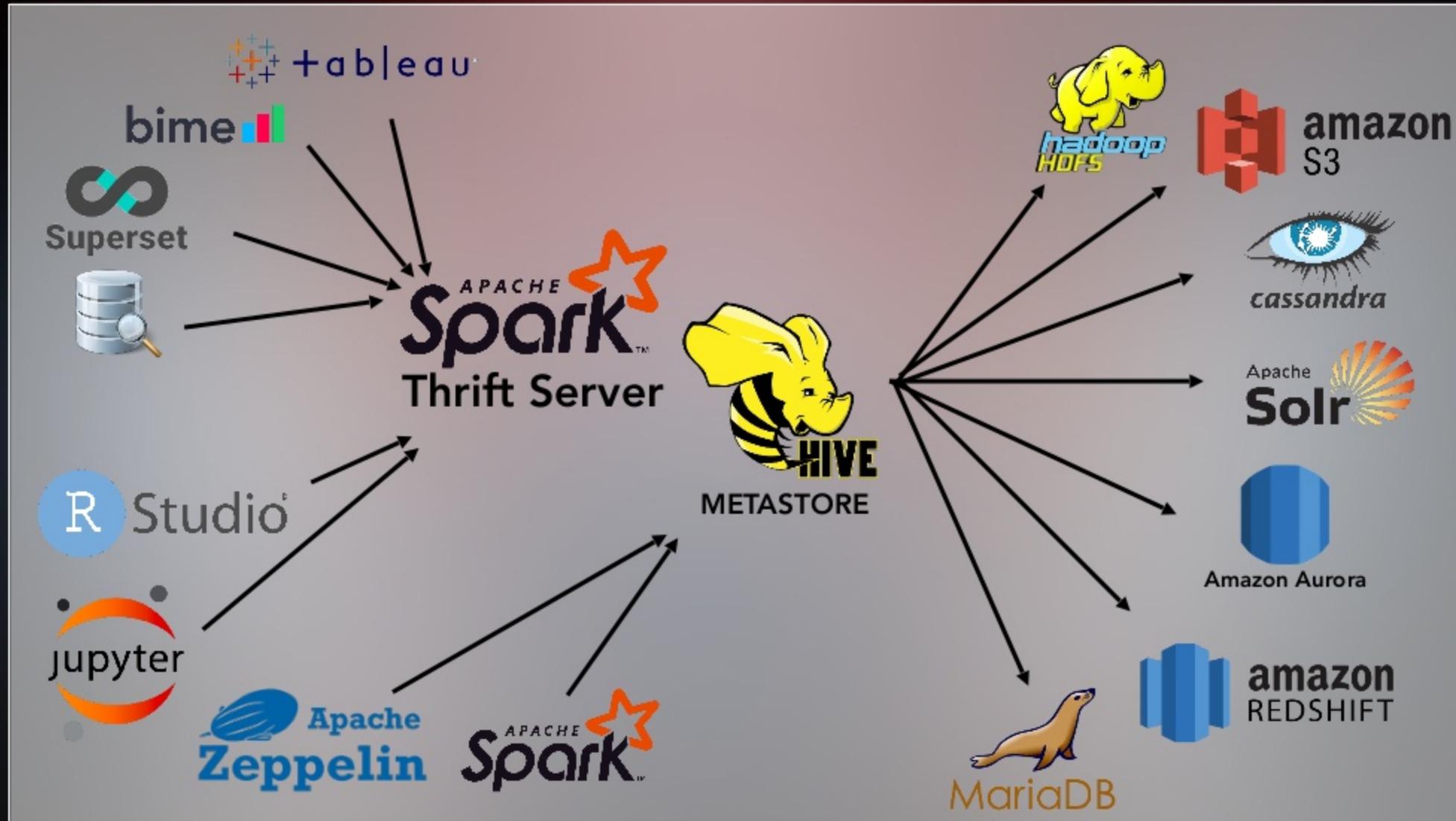
DATA EXPLOITATION

# OUR ANALYTICAL STACK



DATA EXPLOITATION

# OUR ANALYTICAL STACK



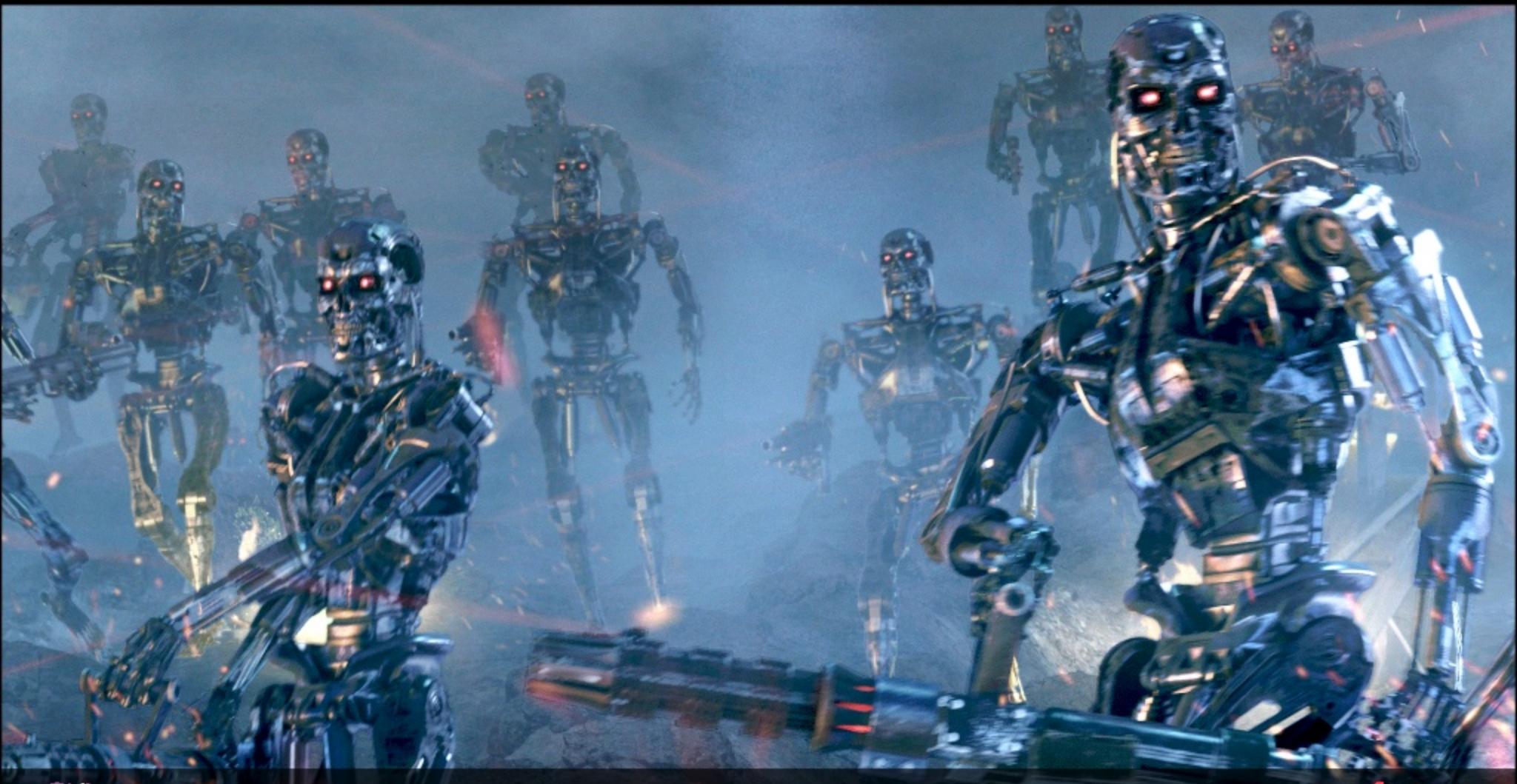
DATA EXPLOITATION

# DATA SCIENTISTS TEAM?



DATA EXPLOITATION

# DATA SCIENTISTS AS I SEE THEM



DATA EXPLOITATION

DATA SCIENTISTS SINS

COMPUTER  
MALFUNCTION

# DATA SCIENTISTS SINS

Hadoop Overview Datanodes Snapshot Startup Progress Utilities ▾

128 MB 1 de 3.701 ▲ ▼

Browse Directory

/user/hive/warehouse/db/tmp\_user\_retention [Go!](#)

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	admin	hive	0 B	Wed Mar 14 12:56:56 +0100 2018	3	128 MB	_SUCCESS
-rw-r--r--	root	hive	1.8 MB	Tue Mar 13 17:34:13 +0100 2018	3	128 MB	part-00000-2b43d4fd-8faa-432c-8e00-c94a9e6aaecd-c000.snappy.parquet
-rw-r--r--	root	hive	2.71 MB	Wed Mar 14 11:15:29 +0100 2018	3	128 MB	part-00000-4de6c2b1-a21c-4774-a674-354b95b836ec-c000.snappy.parquet
-rw-r--r--	root	hive	2.55 MB	Tue Mar 13 20:20:36 +0100 2018	3	128 MB	beb305e7-197c-4f3a-8300-000000000000.snappy.parquet
-rw-r--r--	root	hive	1.99 MB	Tue Mar 13 18:16:14 +0100 2018	3	128 MB	197c-4f3a-8300-000000000000.snappy.parquet
-rw-r--r--	root	hive	2.89 MB	Tue Mar 13 22:55:18 +0100 2018	3	128 MB	part-00000-6dffde8b8-d5c5-4c51-bc05-532c26997e.snappy.parquet

Too many small files!

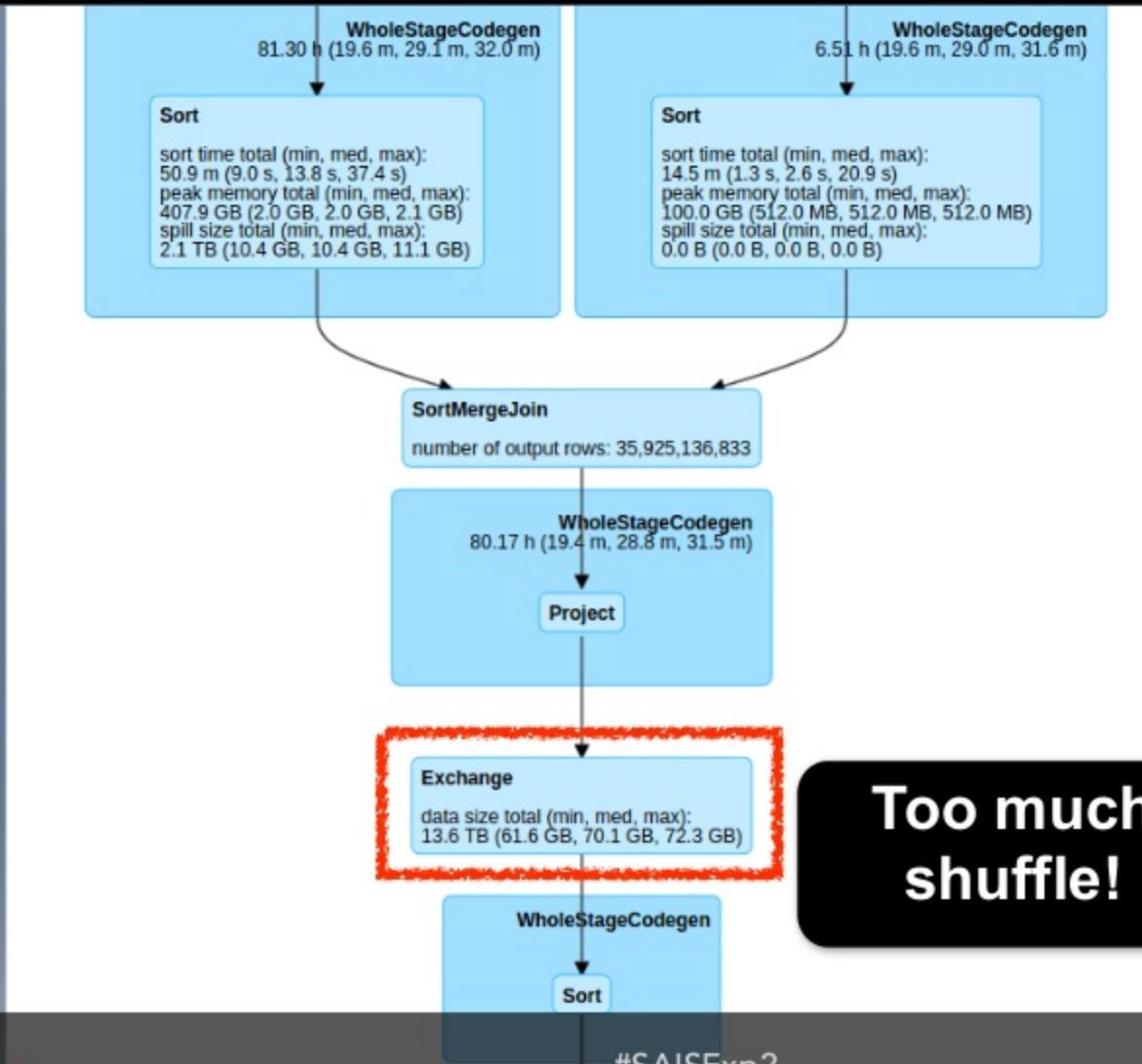


DATA EXPLOITATION

# DATA SCIENTISTS SINS



# DATA SCIENTISTS SINS



## INGEST

Data Ingestion

Storage

## PROCESS

Stream

Batch

## DISCOVER

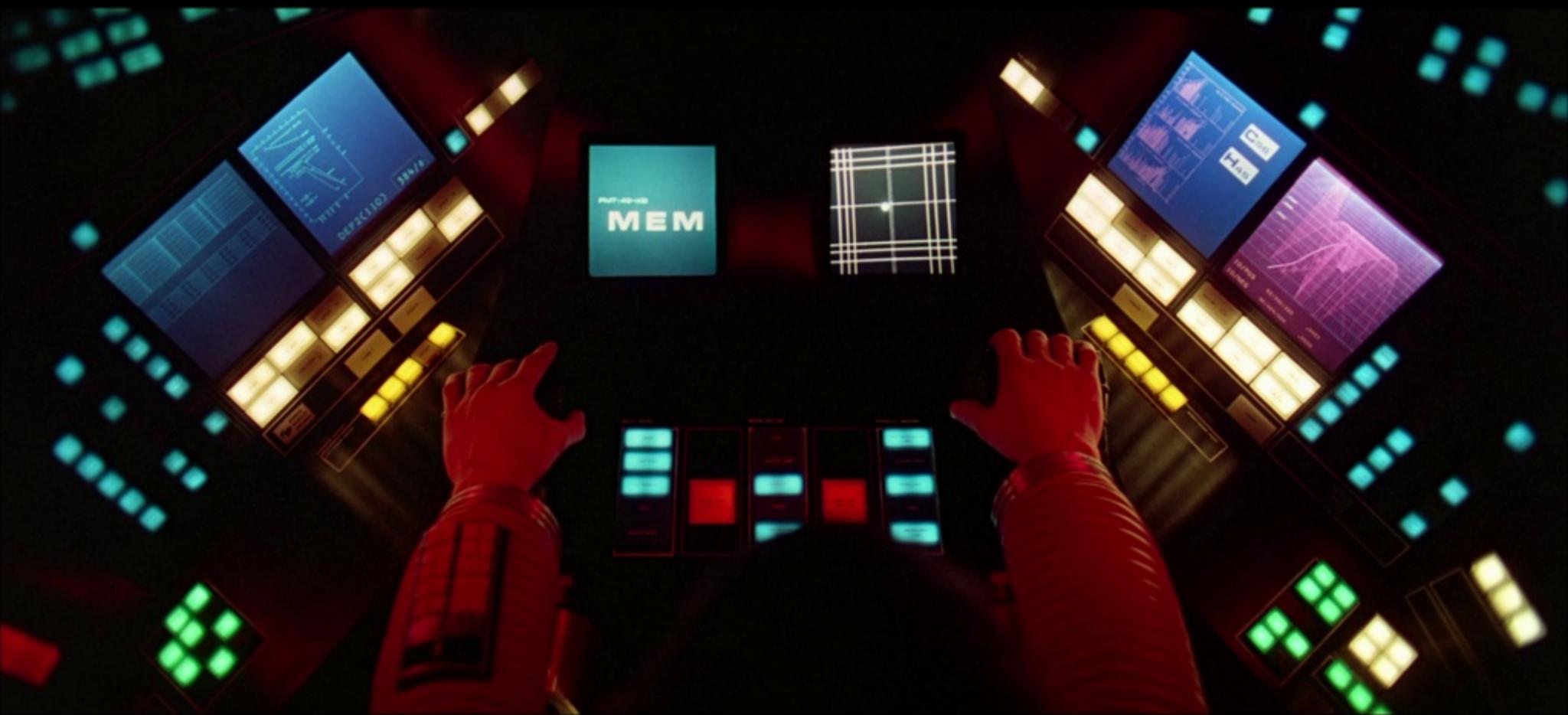
Query

Data exploitation

Orchestration

ORCHESTRATION

# AIRFLOW



# ORCHESTRATION

# AIRFLOW



## ORCHESTRATION

# AIRFLOW

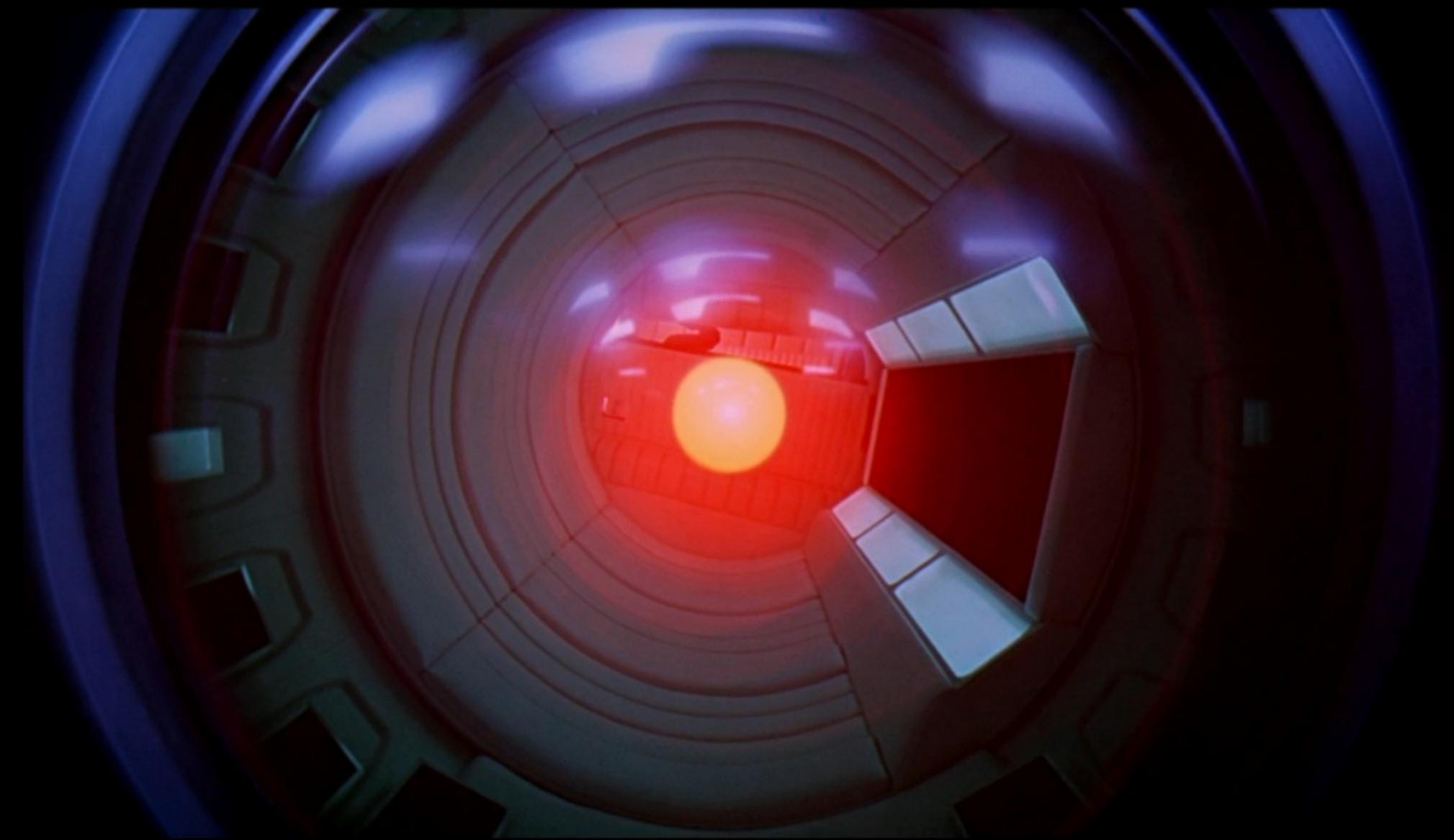


```
127 )
128
129     create_table_user_skwid_distribution = HiveOperatorLetgo(
130         task_id='create_table_user_skwid_distribution',
131         dag=dag,
132         sql="personalization/create_table_user_skwid_distribution.sql",
133         parameters={
134             's3_location': path.join(location, 'user_skwid_distribution')
135         }
136     )
137
138     insert_into_recently_active_users = HiveOperatorLetgo(
139         task_id='insert_into_recently_active_users',
140         dag=dag,
141         sql="personalization/insert_into_recently_active_users.sql",
142         parameters={
143             'current_ts': "{{ ts_nodash }}",
144             'lookback': lookback
145         }
146     )
```

I'm happy!!



MOVING TO STATELESS  
CLUSTER



I'M SORRY RICARDO. I'M  
AFRAID I CAN'T DO THAT.

# MOVING TO STATELESS CLUSTER PLATFORM LIMITATIONS

cloudera MANAGER

Clusters ▾ Hosts ▾ Diagnostics ▾ Audits Charts ▾ Administration ▾

Search Support ▾

Home

Status All Health Issues Configuration ⚡ 7 All Recent Commands Ad

**BIPRO HADOOP** (CDH 5.10.0, Packages) ▾

- Hosts ⚡ 7
- HDFS-2
- Hive
- YARN (MR2...)
- ZooKeeper-2

Cloudera Management Service

- Cloudera M...

Charts ▾

Cluster CPU

30m 1h 2h 6h 12h 1d 7d 30d

Hourly

BIPRO HADOOP Host CPU Usage Across Hosts 2.1%

Cluster Network IO

bytes / second

1.9G/s 954M/s 0

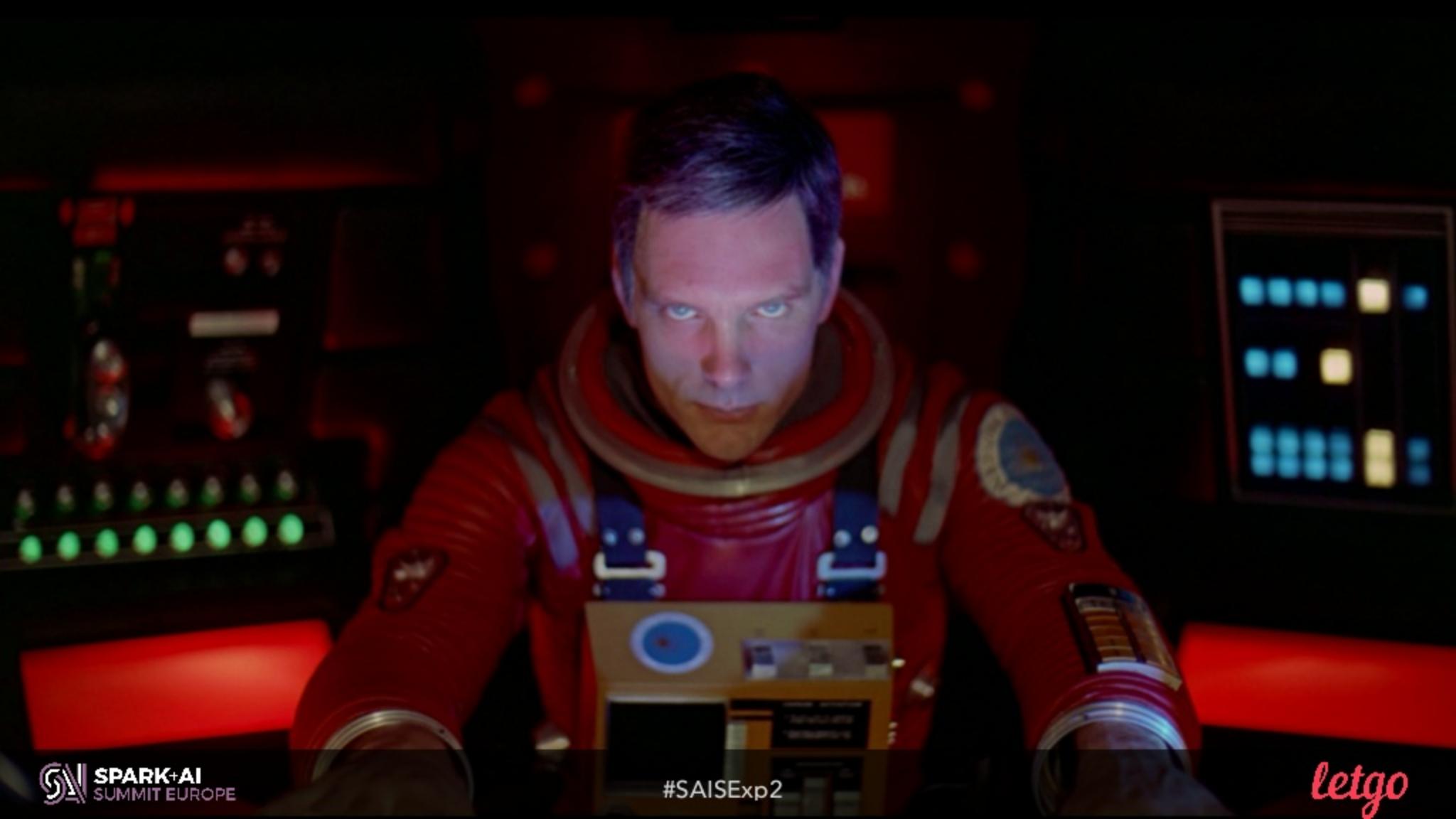
Jul 12 Mon 13 Tue 14 Wed 15 Thu 16 Fri 17 Sat 18

Total Bytes Received Across Network Interfaces 5.5M/s Total Bytes Transmitted Across Network Interfaces 4.1M/s

Hourly

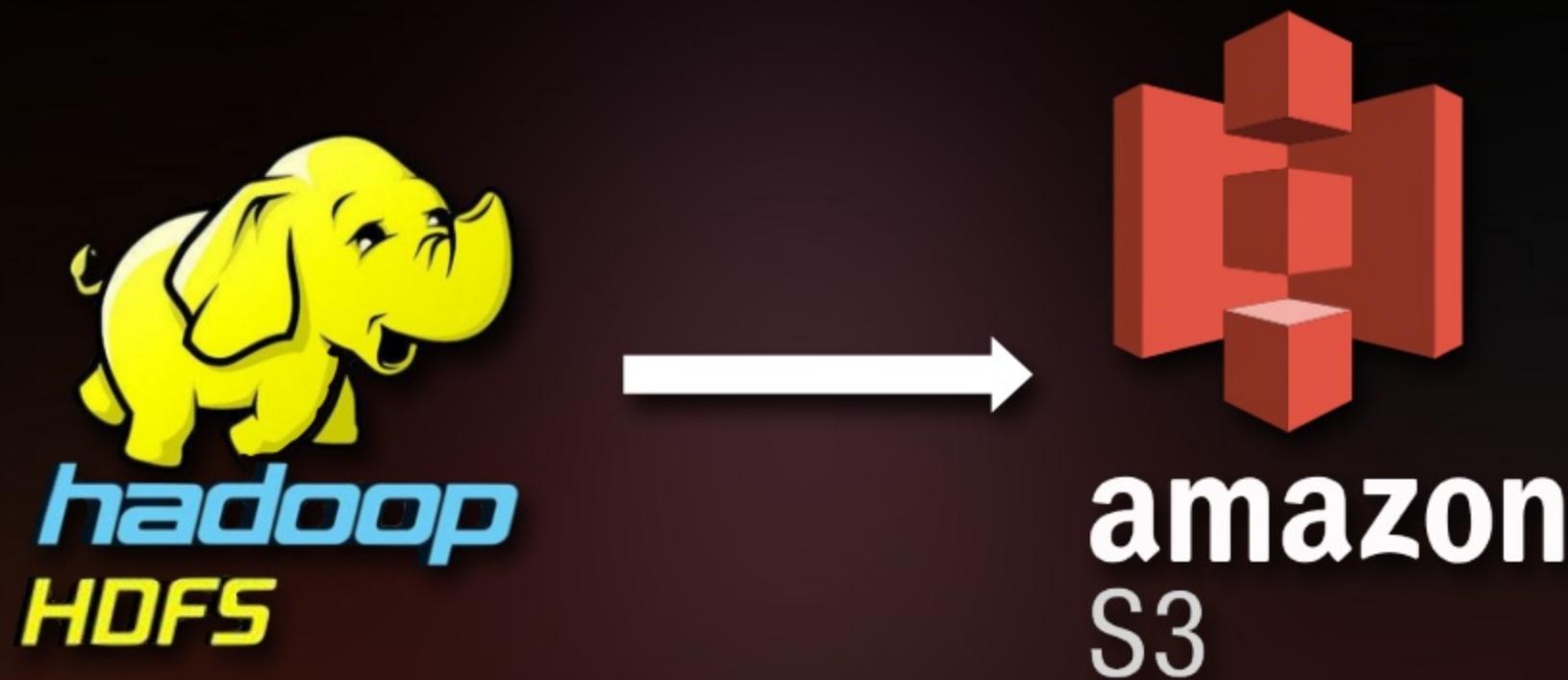
MOVING TO STATELESS CLUSTER

# PLANNING THE SOLUTION



MOVING TO STATELESS CLUSTER

# PLANNING THE SOLUTION



MOVING TO STATELESS CLUSTER

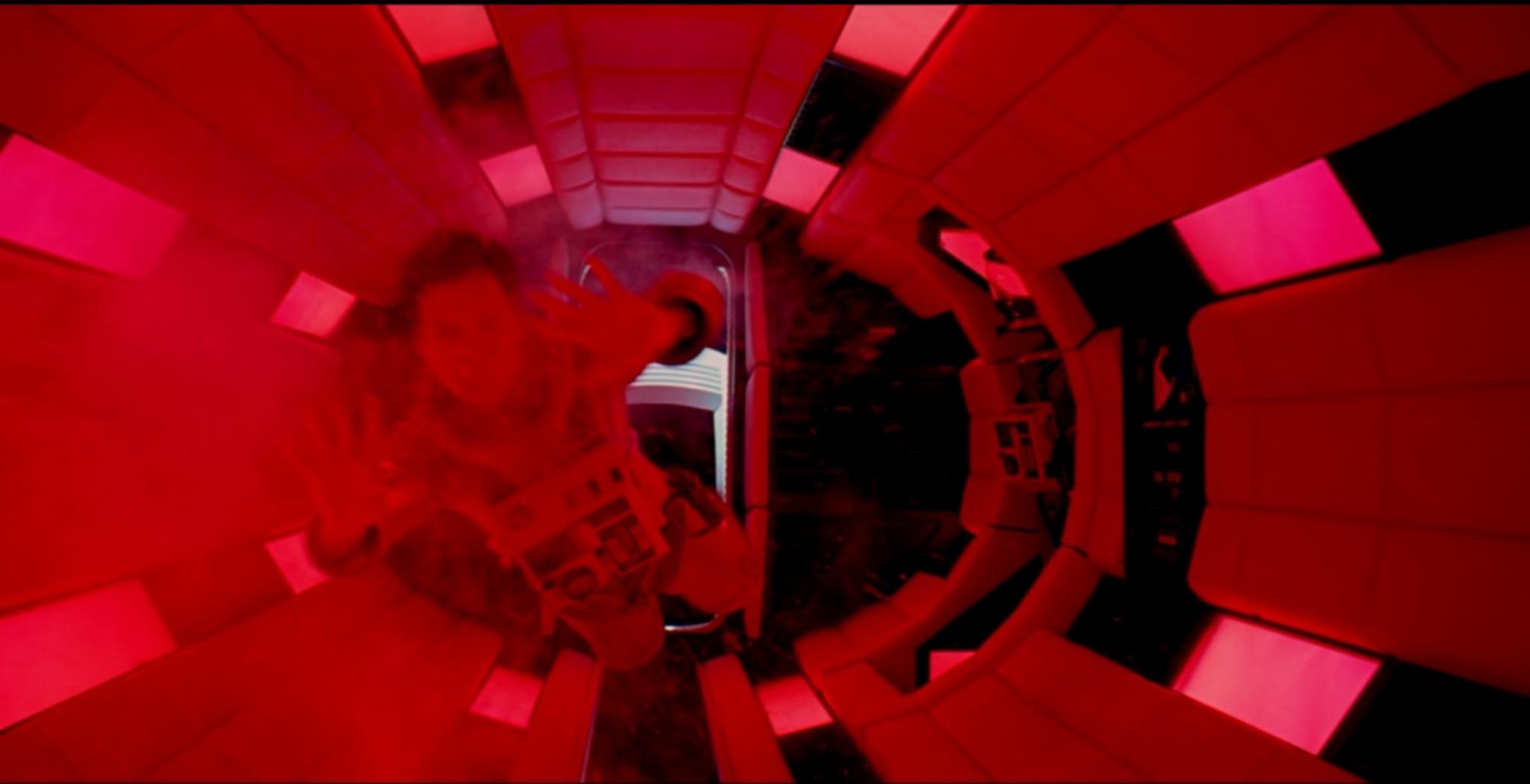
# PLANNING THE SOLUTION



## Cloudbreak

MOVING TO STATELESS CLUSTER

# A LONG AND WINDING PROCESS



# MOVING TO STATELESS CLUSTER A LONG AND WINDING PROCESS

DATA Software project

71 Issues in this epic

There are no ticks

Issue ID	Description	Status	Labels	Actions
DATA-331	Research Hive metastore deployment options	DONE		
DATA-301	List of errors made in Cloudera cluster	DONE		
DATA-520	Test node labels configuration in HDP3 cluster	DONE		
DATA-385	Create DNSs for Hadoop components	DONE		
DATA-39	Create a IAM role for production environment	DONE		
DATA-431	Migrate Kafka offsets from HDFS to S3	DONE		
DATA-453	Execute Spark Streaming Job in DEV environment	DONE		
DATA-23	HDP 3.0 / Hadoop 3.0	DONE		
DATA-440	Add spark.yarn.archive for spark-defaults.conf in puppet manifest	DONE		
DATA-456	Spark job history and logs are not working properly	DONE		
DATA-455	Configure inbound rules for Security groups in DEV environment	DONE		
DATA-448	Run some Airflow/STS job in DEV	DONE		

# MOVING TO STATELESS CLUSTER NEW CAPABILITIES OF THE PLATFORM



# MOVING TO STATELESS CLUSTER NEW CAPABILITIES OF THE PLATFORM

The screenshot shows the Hortonworks Cloudbreak interface. On the left is a sidebar with the following menu items:

- Clusters
- Credentials
- Blueprints
- Cluster Extensions
- Recipes
- Management Packs
- External Sources
- Authentication Configuration
- Database Configurations
- Image Catalogs
- Proxy Configurations
- History

The main area is titled "Clusters" and displays two clusters:

Cluster Name	Type	Status
WALLE	hadoop cluster	Running
HAL9000	hadoop cluster	Terminating

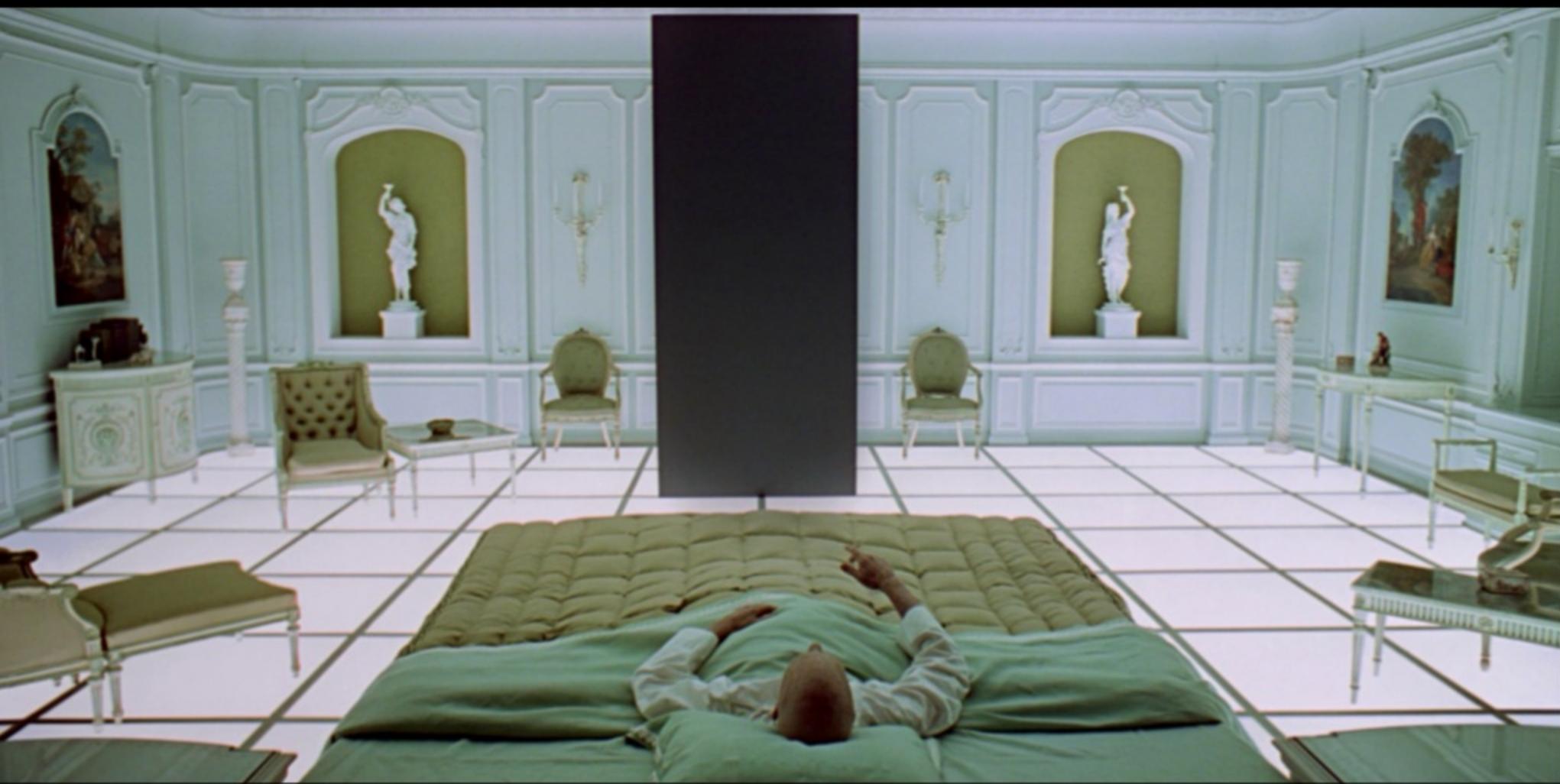
Both clusters are listed as "HDP 3.0.0.0-1634". Below each cluster summary are columns for "NODES" (13) and "UPTIME" (3m for WALLE, - for HAL9000).

At the top right of the main area, there are status indicators for "Autoscale" and "Cloudbreak", the version "2.7.1", and a notification bell with a "25" badge. A "CREATE CLUSTER" button is also present.



# JUPITER AND BEYOND THE INFINITE

JUPITER AND BEYOND THE INFINITE  
SOMETHING WONDERFUL WILL  
HAPPEN





yo presents  
**iEVE**

<https://youtu.be/CInMDMuSFwc>

"The only way to discover the limits of the possible  
is to go beyond them into the impossible."

-ARTHUR C. CLARKE



THE FUTURE...?

DO YOU WANT TO JOIN US?

