
node
python
Terraform
I.A.
OVHcloud
docker
Tech Talks
kubernetes
WA
hadoop
openstack
Machine Learning

OVHcloud Data Processing :

Le nouveau service pour valoriser vos données à l'aide du cloud

BELLLION Bastien

2020-06-25

Bastien BELLION

- Data Convergence team
- bastien.bellion@ovhcloud.com
- @BastienBellion

Data Convergence :

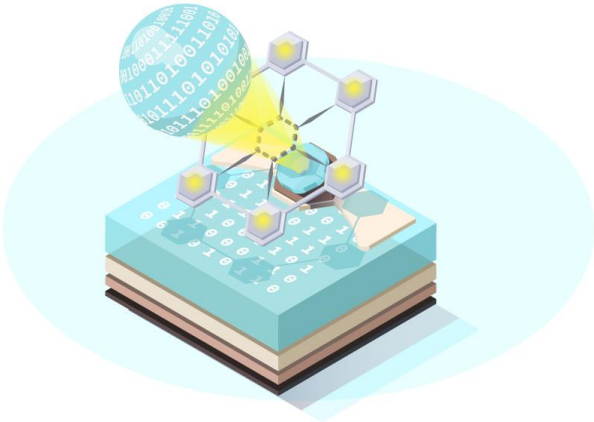
- Analytics Data Platform
- Data Processing

Content

- Big Data
- Cluster Computing
- Apache Spark
- Exemple 1: Lyrics Wordcount
- OVHcloud Data Processing platform
- Exemple 2: New-York Taxis
- Roadmap

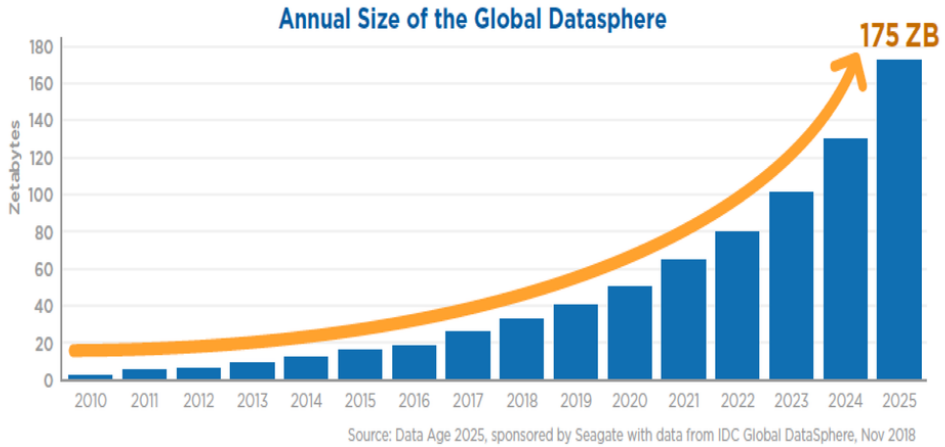
Why do we need to process data?

- Invent new products
- Make better decisions
- Prevent danger and fraud
- Forecasting events

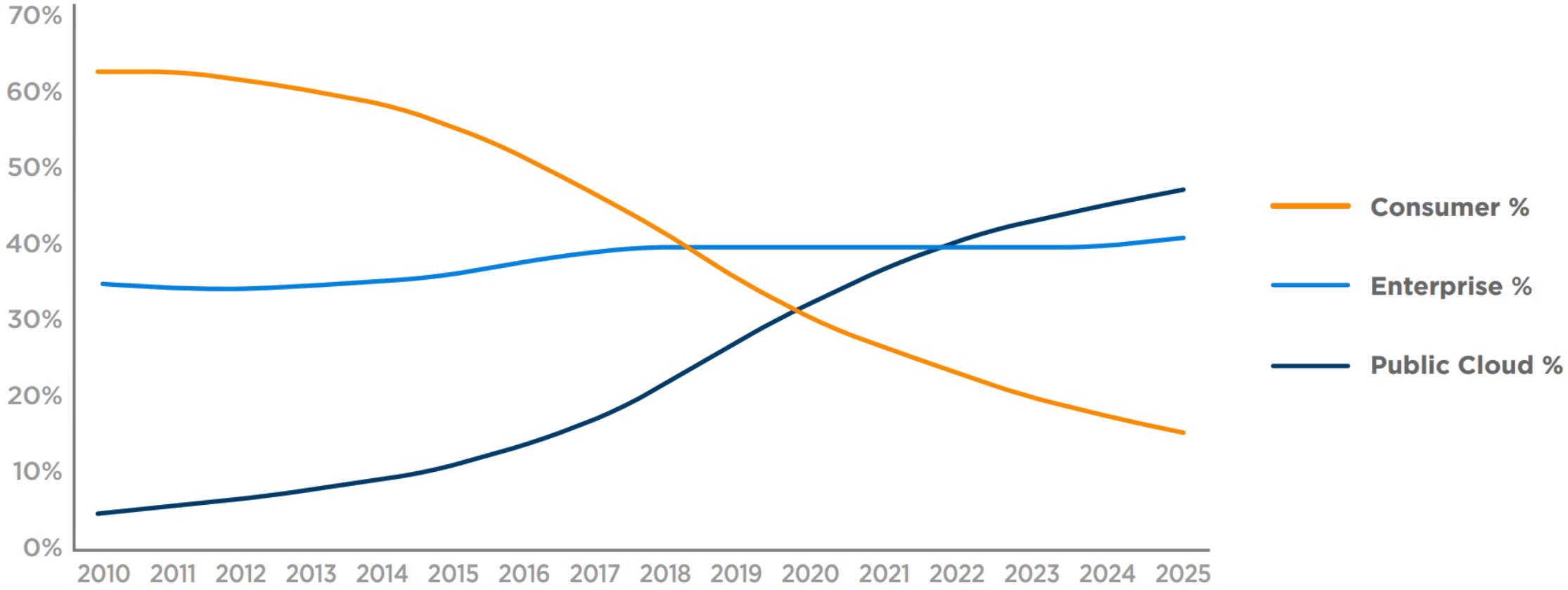


Flood of data

- 90 percent of data in the world has been generated in last 2 years.
- Autonomous cars: 20TB /hour
- 500 million tweets per day.
- 75 billion IoT devices by 2025
- Data exhaust

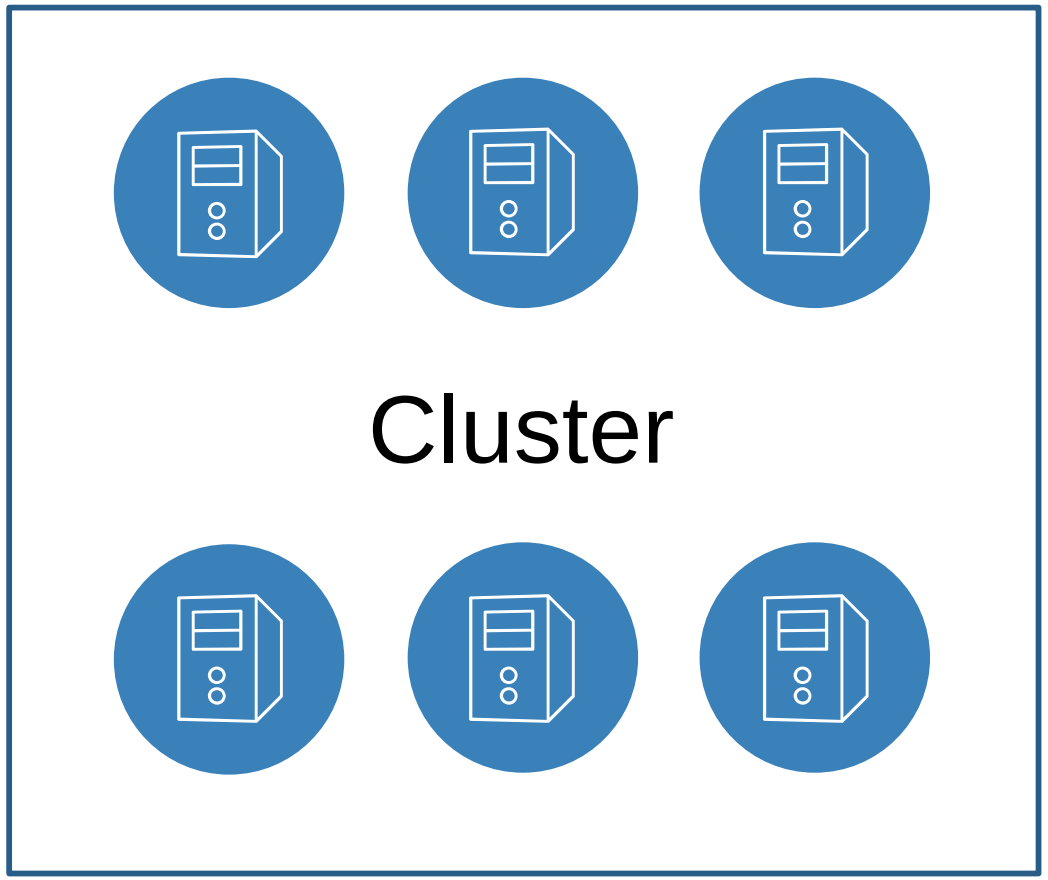


Where is the data stored?



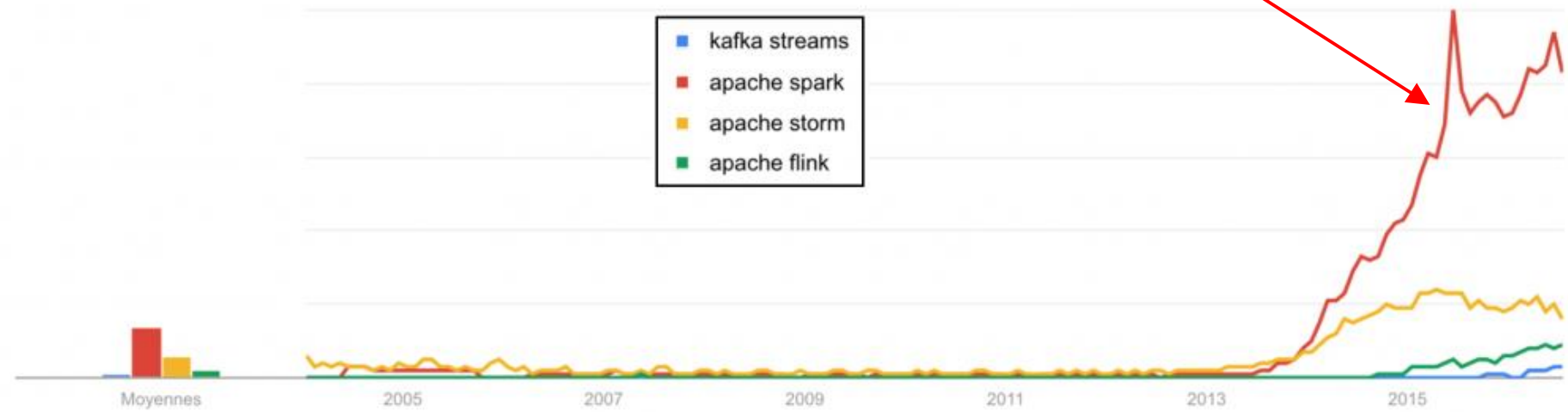
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Cluster computing



Selecting the right tool

- Open source
- Biggest open source community in big data
- Fast
- Java/Scala/Python/R
- Cluster



Big data tools – Google trend graph

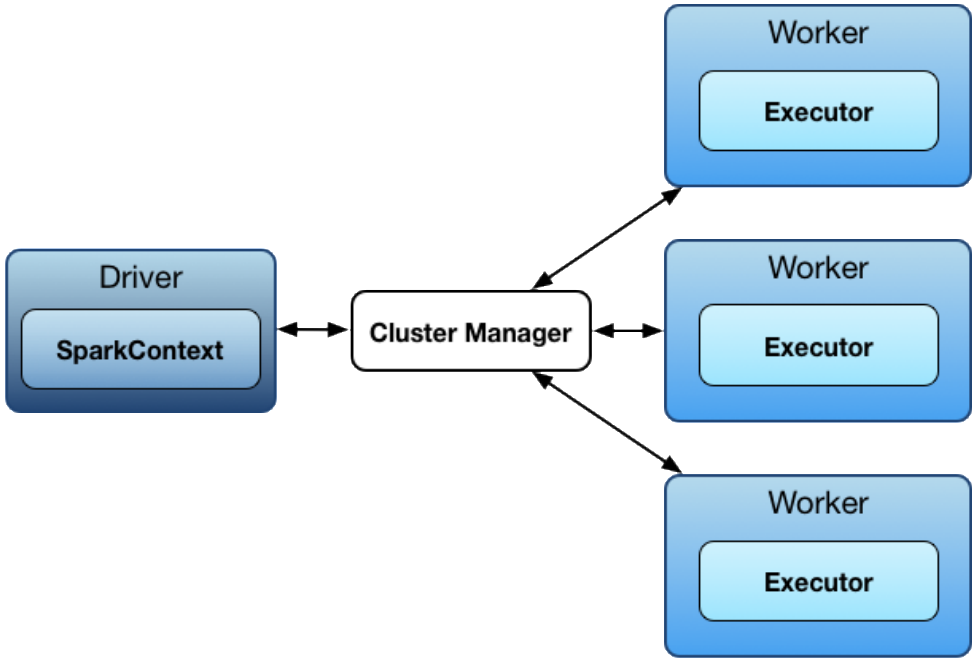
Who is using Spark ?

https://databricks.com/fr/session/netflixs-recommendation-ml-pipeline-using-apache-spark

https://labs.criteo.com/2018/01/spark-out-of-memory/

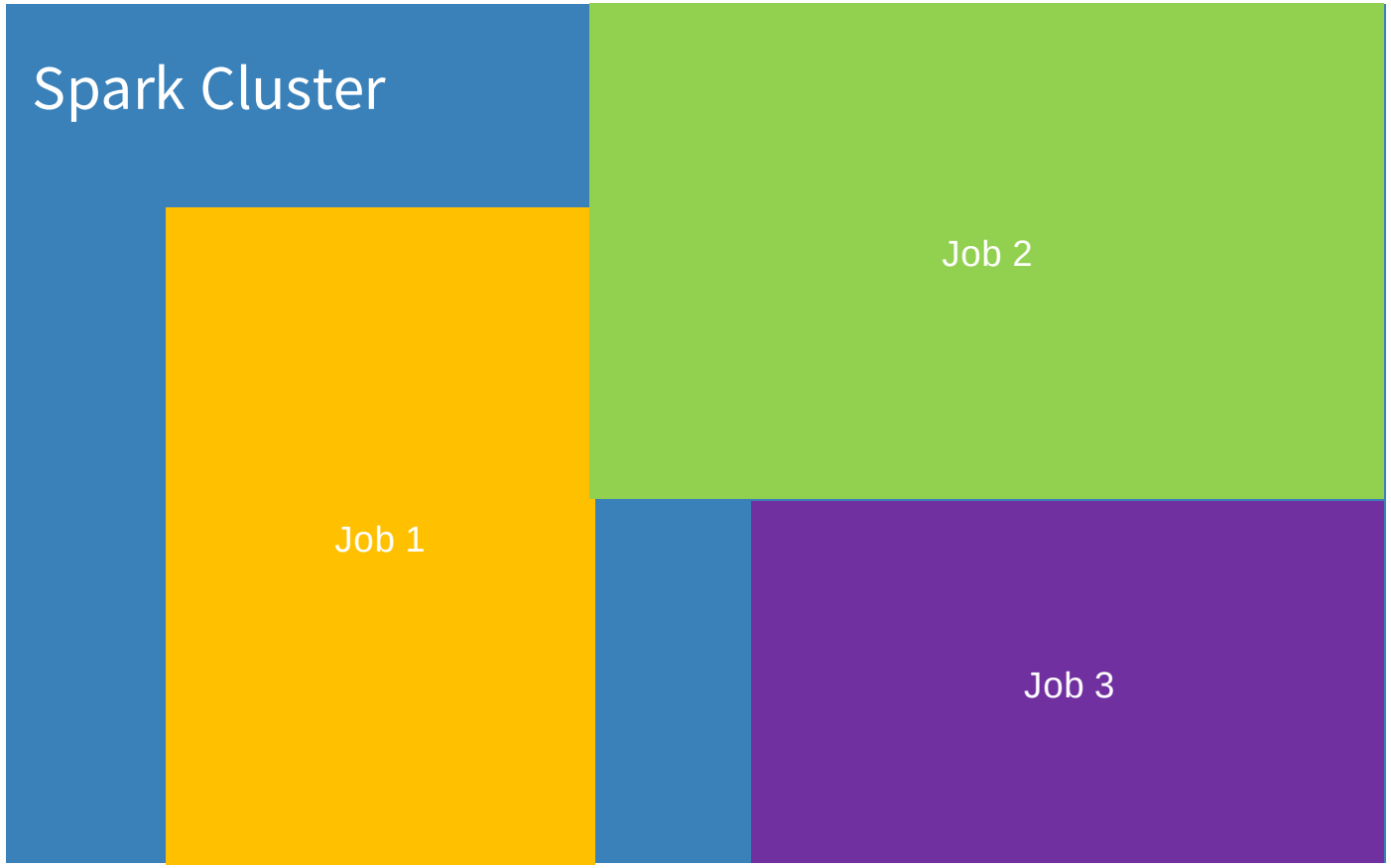
Spark


APACHE
Spark


Worker
Executor
Driver
SparkContext
Cluster Manager
Worker
Executor
Worker
Executor

Spark

```
spark-submit --class org.apache.spark.examples.SparkPi
             \ --driver-memory 8G
             \ --master local[2]
             \ --executor-memory 8G
             \ --executor-cores 2
             \ --num-executors 3
             \ spark-examples.jar
```



Example 1 : Lyrics Wordcount

Billboard hot 100 between 1965-2015

billboard

5100 song with for each:

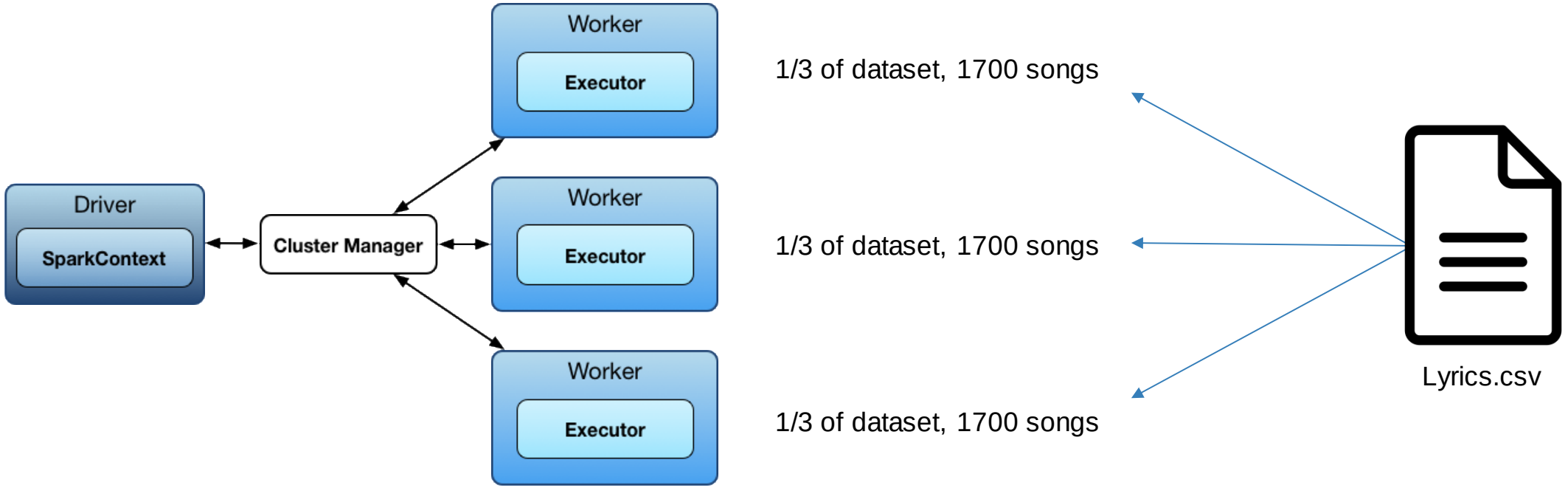
- Rank of the year
- Song name
- Artist
- Year
- Lyrics
- Source

```
"Rank","Song","Artist","Year","Lyrics","Source"
1,"wooly bully","sam the sham and the pharaohs",1965,"sam the sham miscellaneous wooly bully wooly bully sam the sham  the pharaohs  domingo samudio uno dos on
you someone really to pull the wool with you wooly bully wooly bully wooly bully wooly bully wooly bully lseven  the letter l and the number 7 when typed they
2,"i cant help myself sugar pie honey bunch","four tops",1965," sugar pie honey bunch you know that i love you i cant help myself i love you and nobody elsein
im a fool in love you seewanna tell you i dont love you tell you that were through and ive tried but every time i see your face i get all choked up insidewhen
you know that i love you i cant help myself i cant help myself ",1
3,"i cant get no satisfaction","the rolling stones",1965,"  ",1
4,"you were on my mind","we five",1965," when i woke up this morning you were on my mind and you were on my mind i got troubles whoaoh i got worries whoaoh i g
down in my shooooes yeah i got to ramble whoaoh i got to move on whoaoh i got to walk away my blues when i woke up this morning you were on my mind you were on
5,"youve lost that lovin feelin","the righteous brothers",1965," you never close your eyes anymore when i kiss your lips and theres no tenderness like before i
something beautifuls dyinyou lost that lovin feelin whoa that lovin feelin you lost that lovin feelin now its gone gone gone wohbaby baby id get down on my kne
it on backbring back that lovin feelin whoa that lovin feelin bring back that lovin feelin cause its gone gone gone and i cant go on wohbring back that lovin f
6,"downtown","petula clark",1965," when youre alone and life is making you lonely you can always go downtown when youve got worries all the noise and the hurry
things waiting for youdont hang around and let your problems surround you there are movie shows downtown maybe you know some little places to go to where they
nowdowntownand you may find somebody kind to help and understand you someone who is just like you and needs a gentle hand to guide them along so maybe ill see
7,"help","the beatles",1965,"help i need somebody help not just anybody help you know i need someone help when i was younger so much younger than today i never
independence seems to vanish in the haze but every now and then i feel so insecure i know that i just need you like ive never done before help me if you can im
down and i do appreciate you being round help me get my feet back on the ground wont you please please help me help me help me ooo",3
8,"cant you hear my heart beat","hermans hermits",1965,"carterlewis every time i see you lookin my way baby baby cant you hear my heartbeat in the car or walki
hear my heartbeat when you asked me to meet your ma i knew that baby wed be going far cant you hear the poundin of my heartbeat cause youre the one i love your
baby baby cant you hear my heartbeat baby baby cant you hear my heartbeat baby baby cant you hear my heartbeat",5
9,"crying in the chapel","elvis presley",1965," you saw me crying in the chapel the tears i shed were tears of joy i know the meaning of contentment i am happy
chapel just to sing and praise the lordyoull search and youll search but youll never find no way on earth to find peace of mindtake your troubles to the chapel
```

https://github.com/walkerkq/musiclyrics

Full written guide : https://docs.ovh.com/gb/en/data-processing/wordcount-spark/

Example 1 : Lyrics Wordcount



Example 1 : Lyrics Wordcount

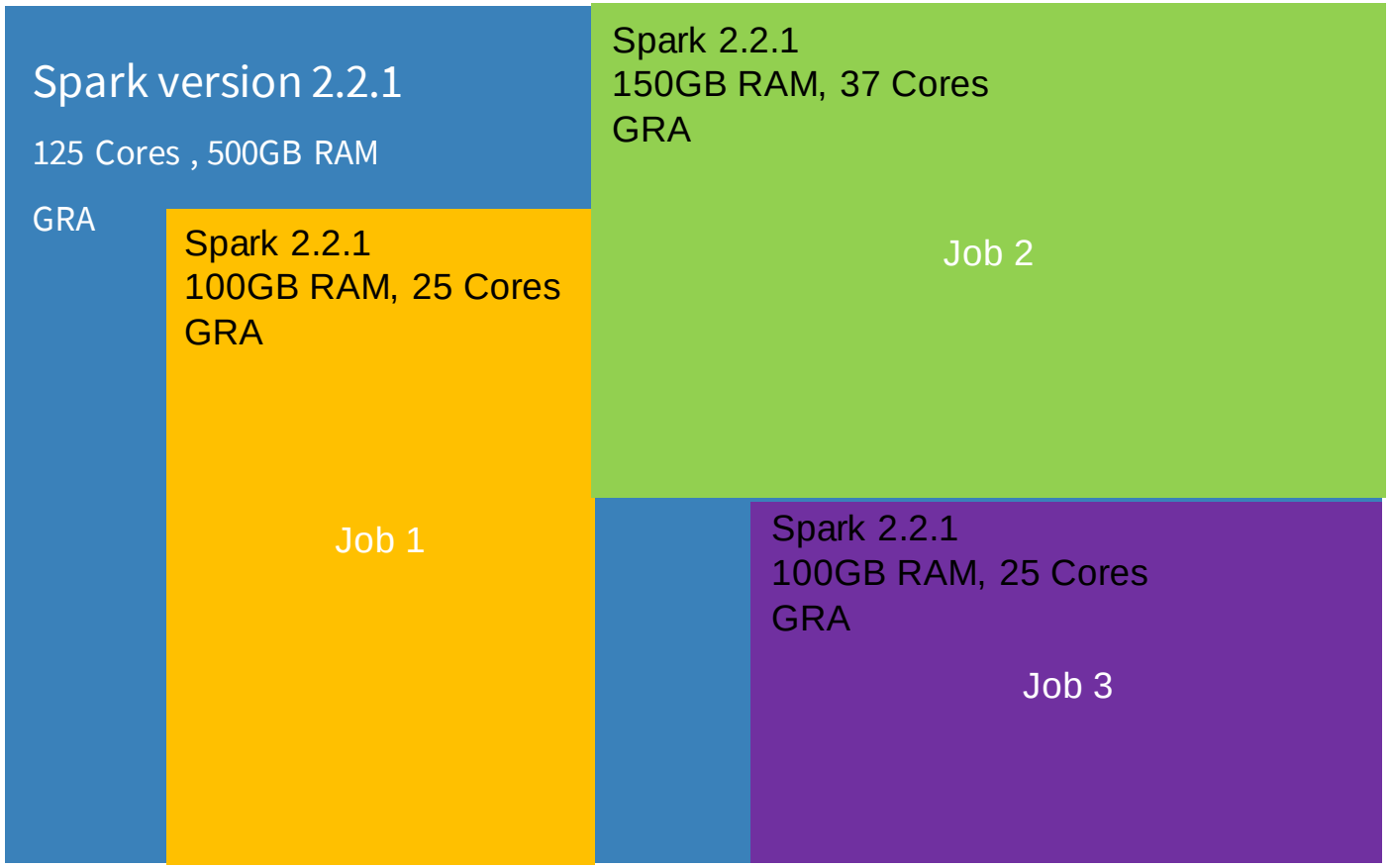
Work computer
4 cpu cores, 8 threads
16 Gb of RAM

Volume of processed data	Time elapsed one computer	Time elapsed spark cluster
7.59 Mb	0,5 s	60 s
759 Mb	5,1 s	100 s
7.59 Gb	48 s	186 s
76 Gb	490 s	358 s
760 Gb	4 820 s / 80 minutes	2 250 s / 37 minutes

OVHcloud instances
10 workers
2 cpu cores per workers
8 Gb of RAM per workers

Manual Spark cluster problems

- Take time and knowledge to setup
- Cluster Management

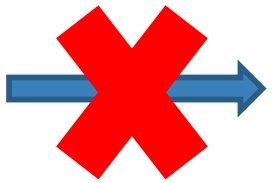


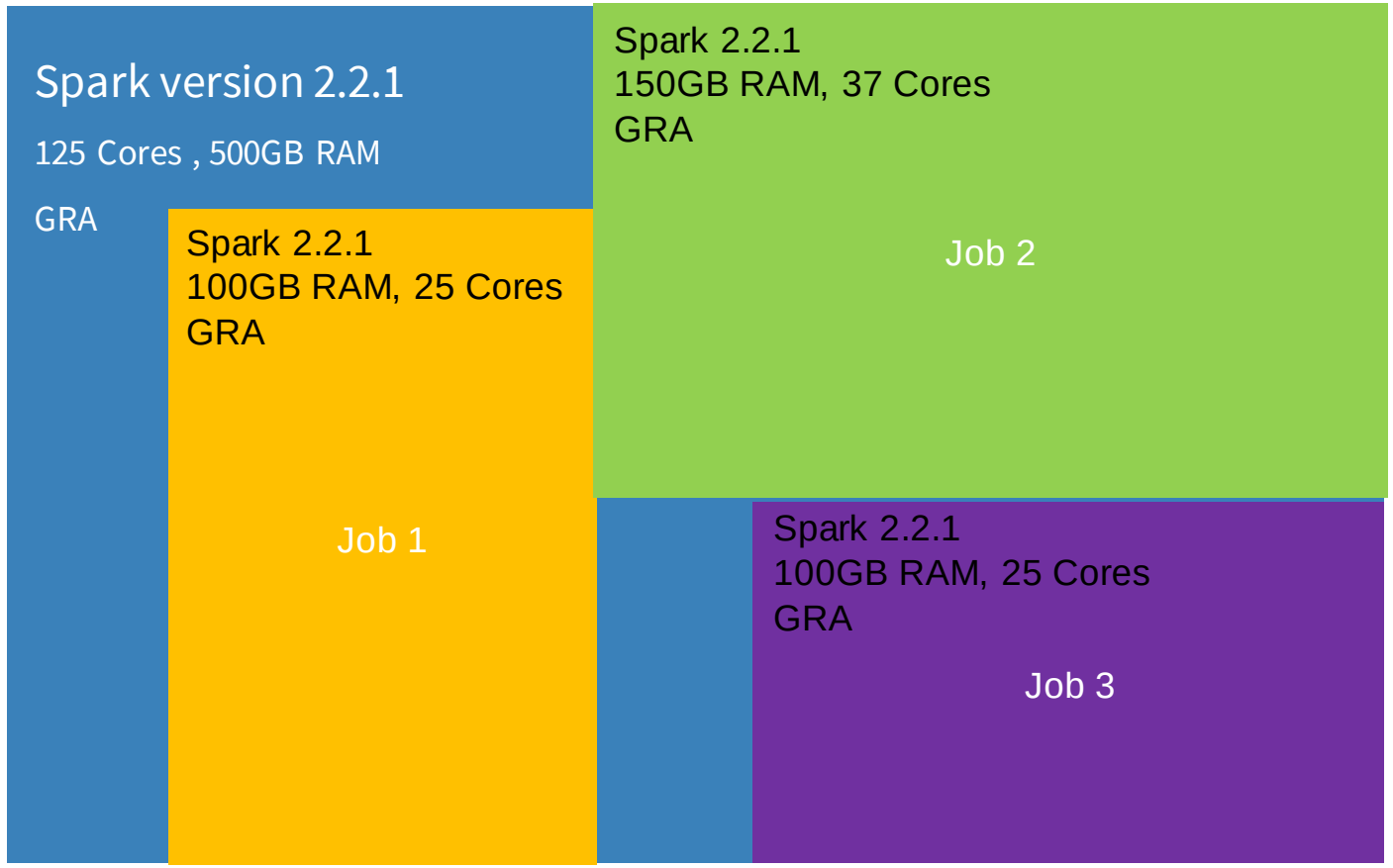
Manual Spark cluster problems

- Take time and knowledge to setup
- Cluster Management

- Spark version bounded

Spark 3.0.0
50GB RAM, 10 Cores
GRA Job 4

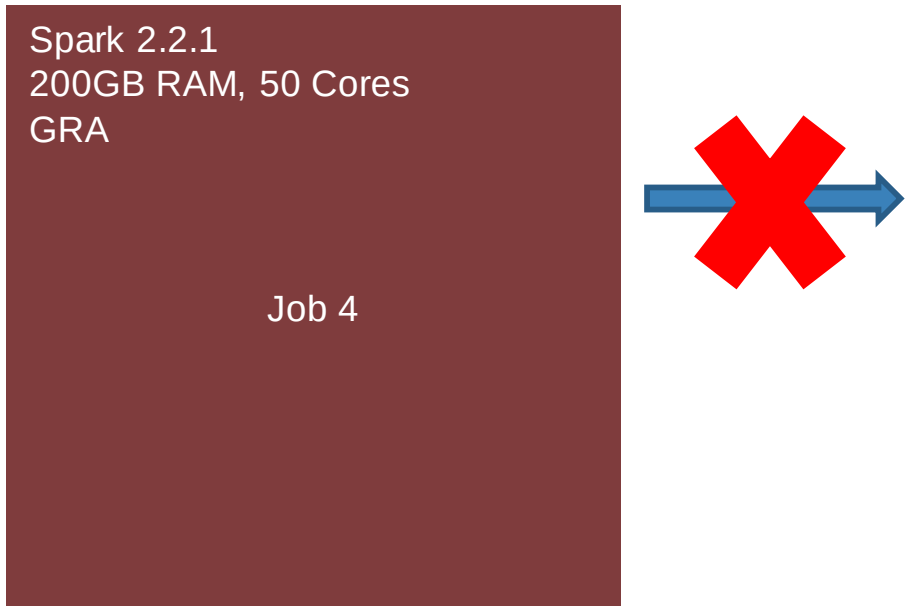


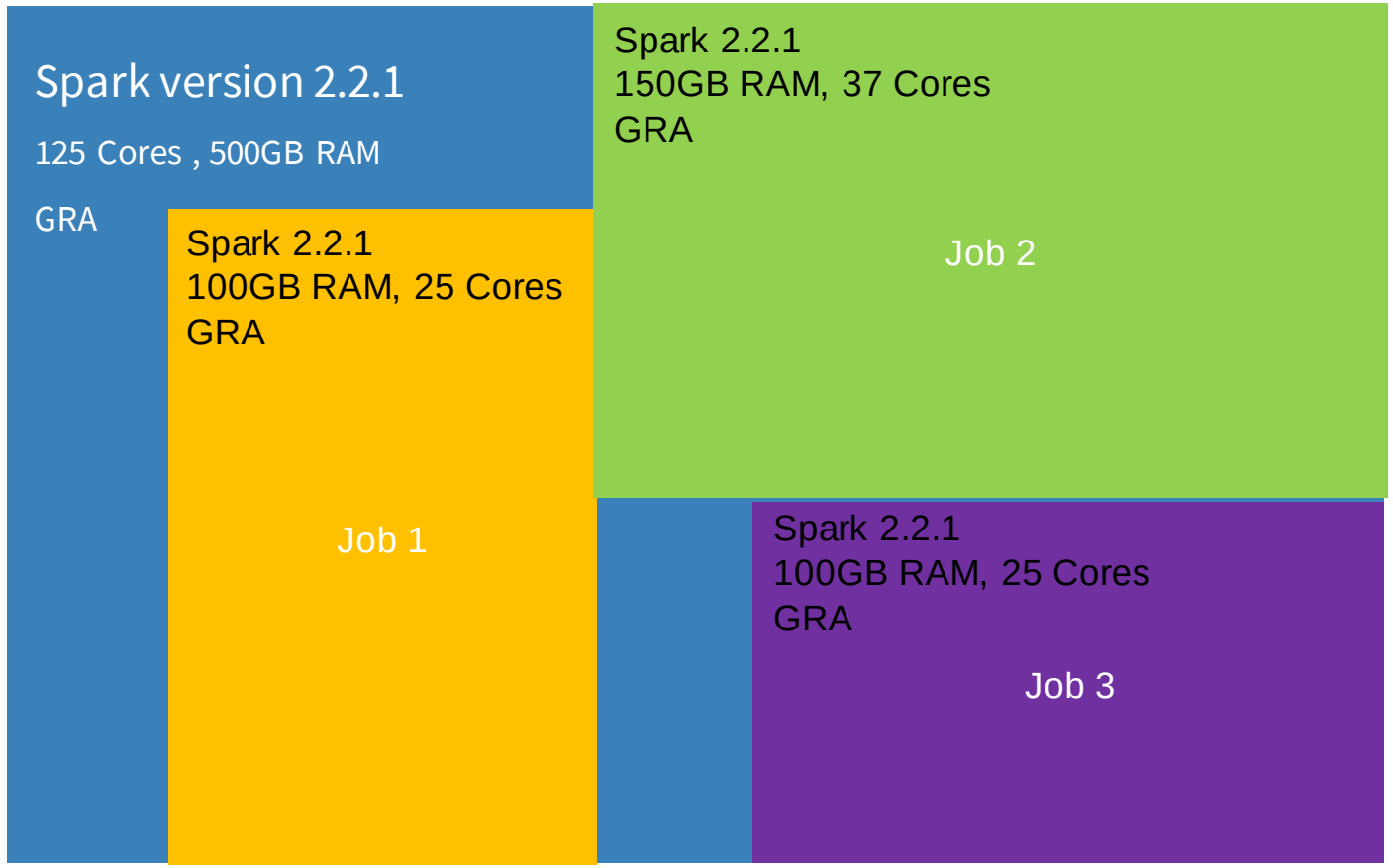


Manual Spark cluster problems

- Take time and knowledge to setup
- Cluster Management

- Spark version bounded
- Limited ressources



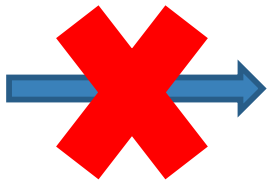


Manual Spark cluster problems

- Take time and knowledge to setup
- Cluster Management

- Spark version bounded
- Limited ressources
- Region bounded

Spark 2.2.1
50GB RAM, 10 Cores
BHS Job 4



Spark version 2.2.1
125 Cores , 500GB RAM
GRA

Spark 2.2.1
100GB RAM, 25 Cores
GRA
Job 1

Spark 2.2.1
150GB RAM, 37 Cores
GRA
Job 2

Spark 2.2.1
100GB RAM, 25 Cores
GRA
Job 3

Manual Spark cluster problems

- No Cluster Management
- Start a cluster in seconds

- Dedicated resources
- No Unused resources

- Any Spark version
- Any region

Spark 3.0.0
200GB RAM, 50 Cores
BHS

Job 4

Spark 2.4.5
4 nodes, 100GB RAM
GRA

Job 1

Spark 2.4.3
5 nodes, 200GB RAM
BHS

Job 2

Spark 3
2 nodes, 50GB RAM
GRA

Job 3

OVHcloud Data Processing

A single cluster per job	→	To focus on what matters
Be in the cloud	→	To scale up and down quickly
Bin different regions	→	Data policy compliance and be close to data
Distributed Computing Technology	→	To distribute processing task to several computers automatically

OVHcloud Data Processing

OVHcloud Data Processing

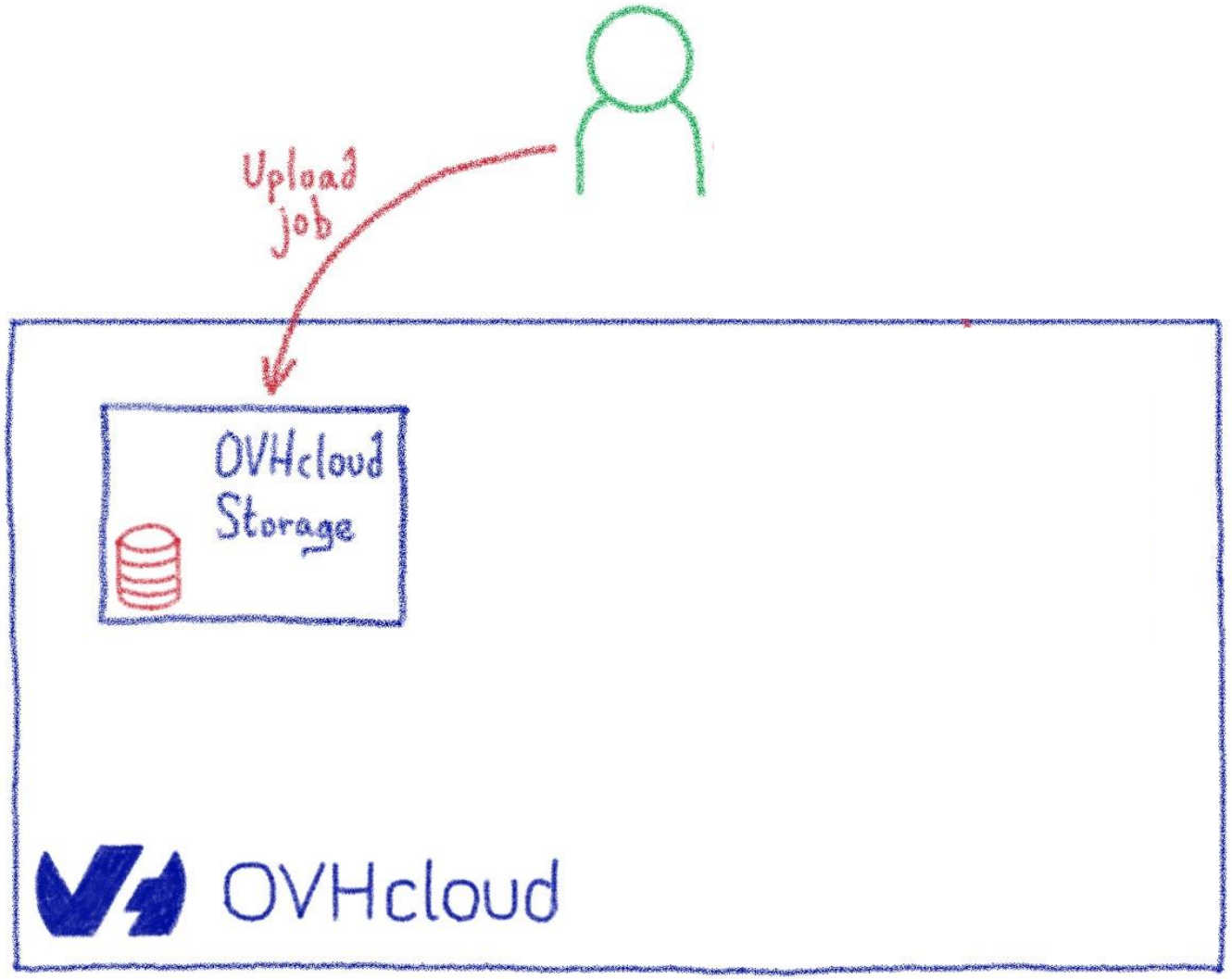
A single cluster per job → To focus on what matters

Be in the cloud → To scale up and down quickly

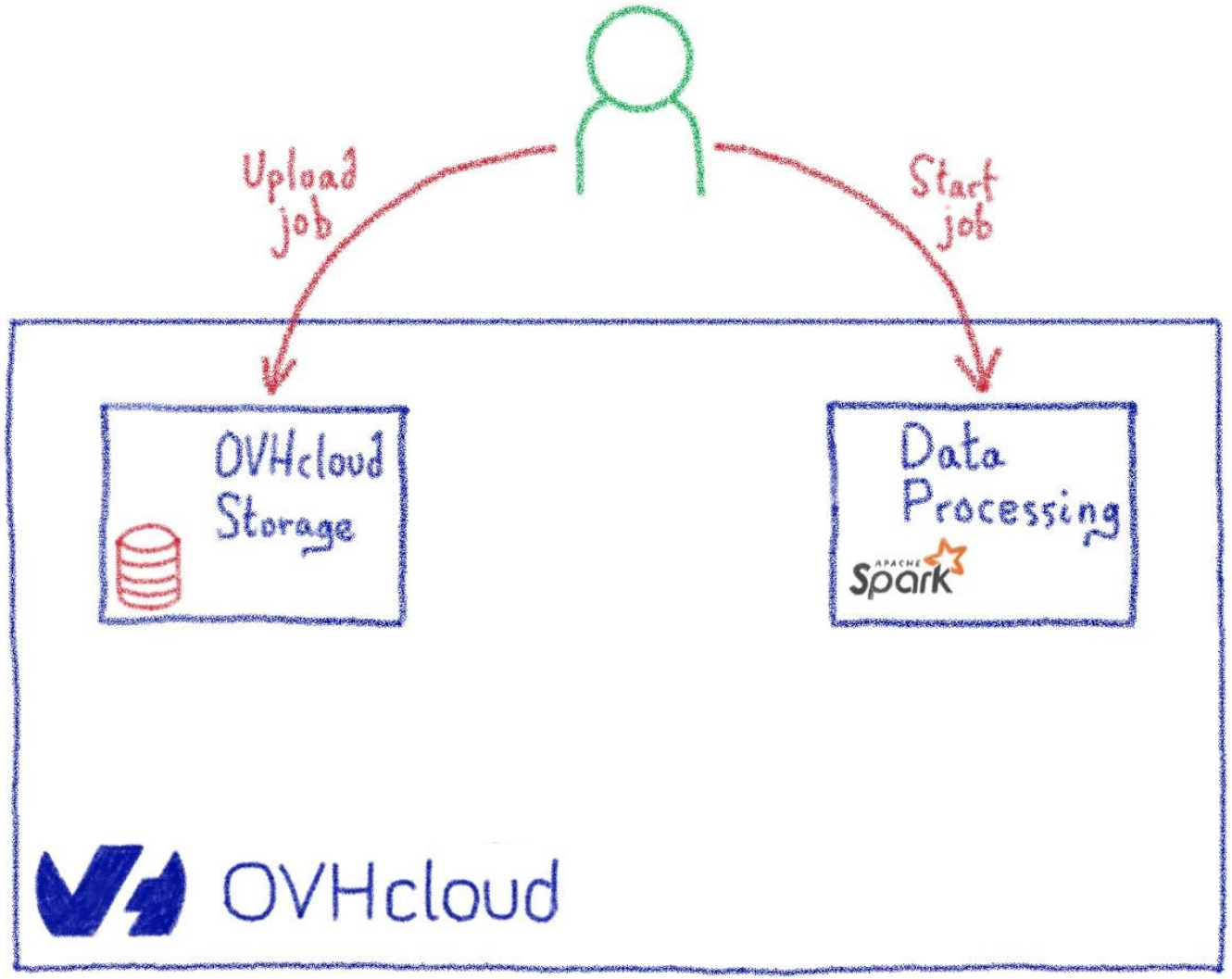
Bin different regions → Data policy compliance and be close to data

Distributed Computing Technology → To distribute processing task to several computers automatically

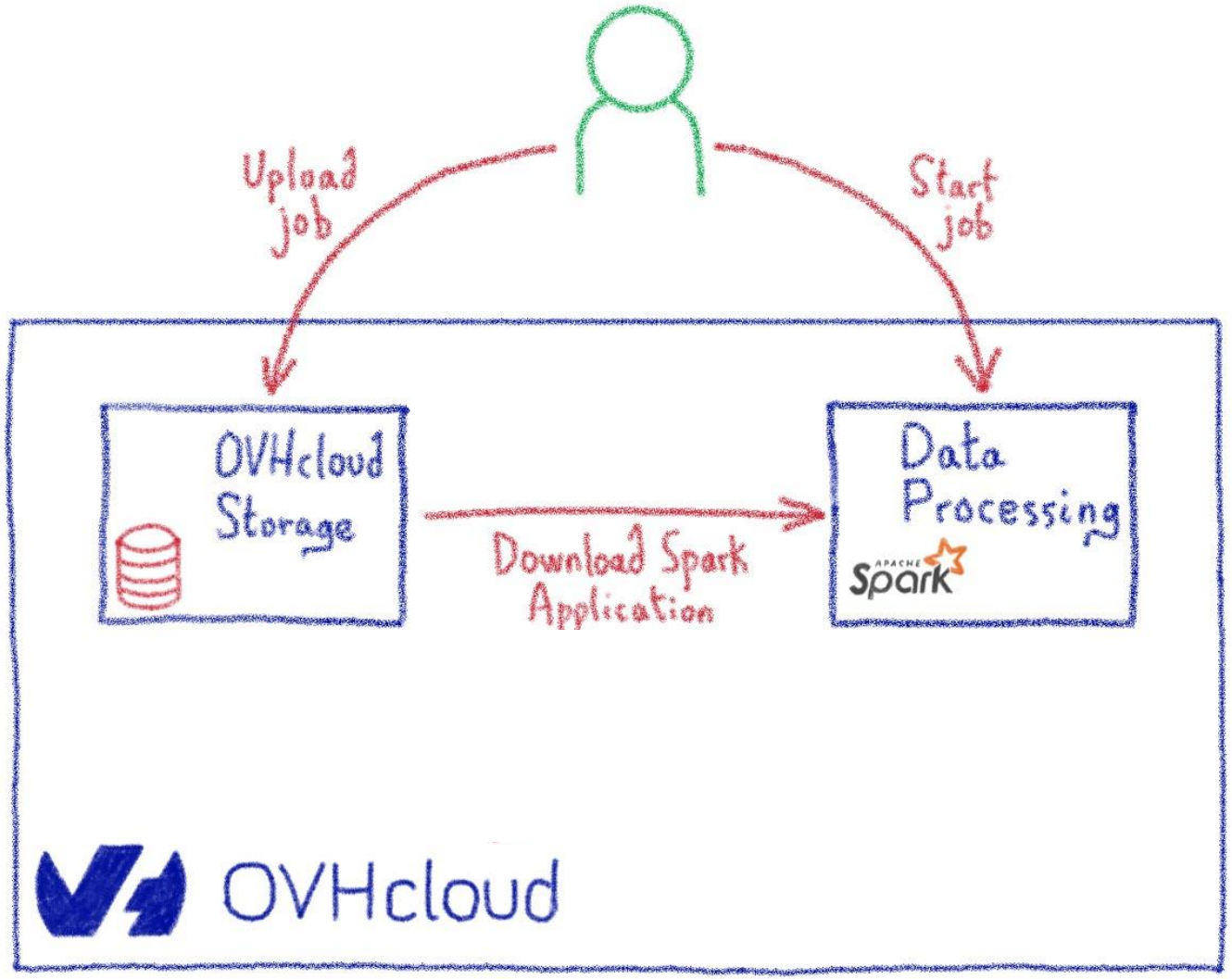
OVHcloud Data Processing



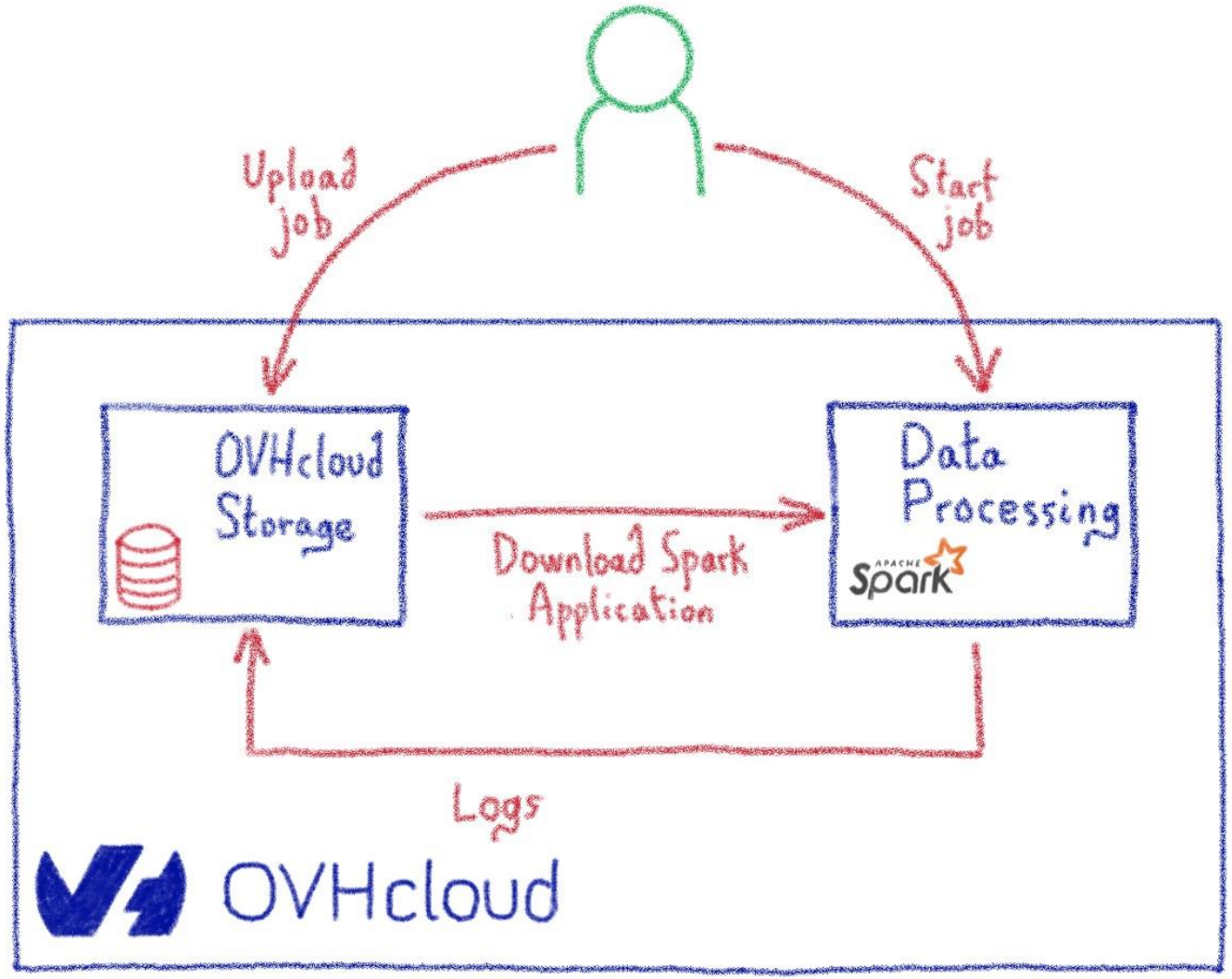
OVHcloud Data Processing



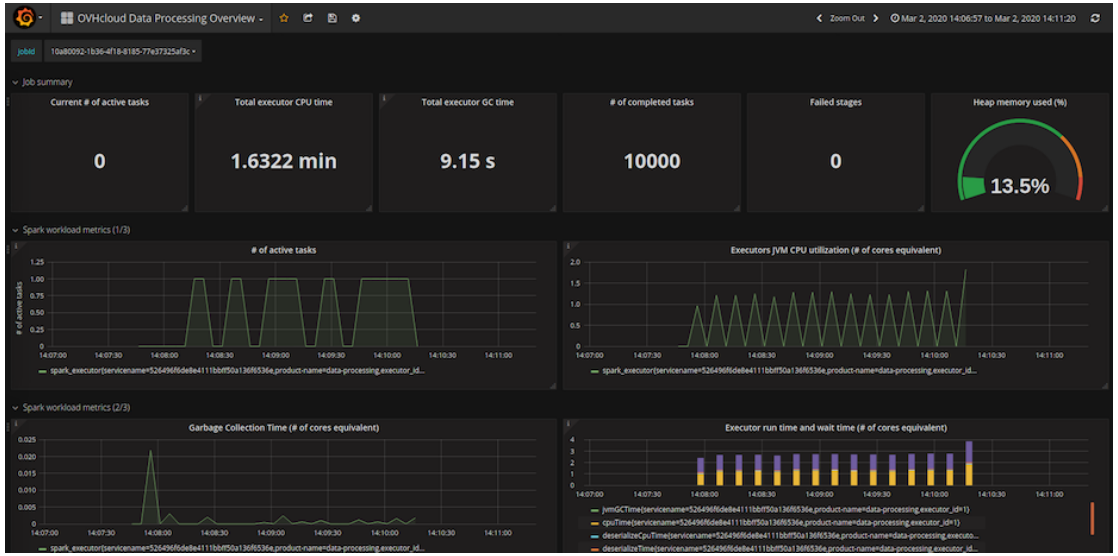
OVHcloud Data Processing

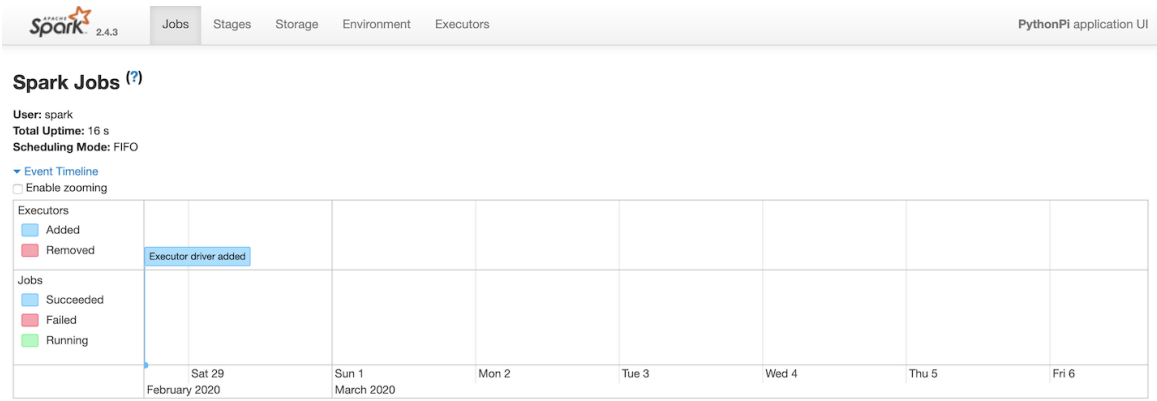


OVHcloud Data Processing



OVHcloud Data Processing





OVHcloud Data Processing

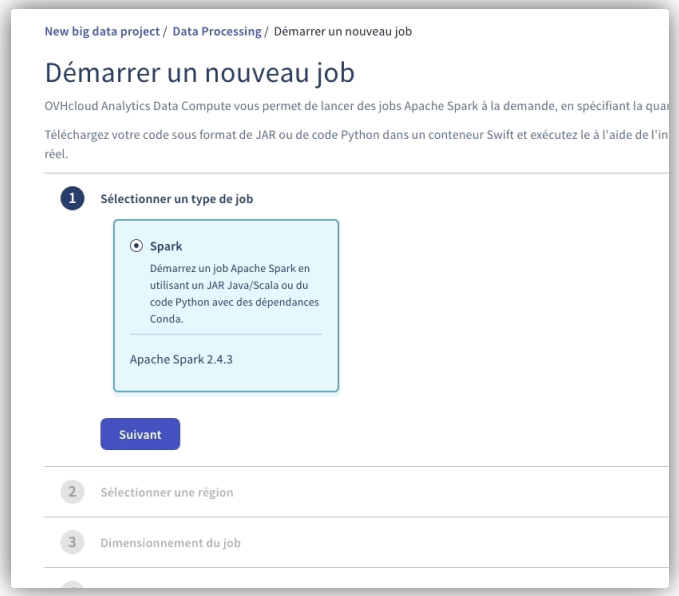
Engine:
- Apache Spark 2.4.3
- Apache Spark 3.0.0 (Soon)

Supported languages :
- Java 8
- Scala 2.12
- Python v2.7+ and v3.4+

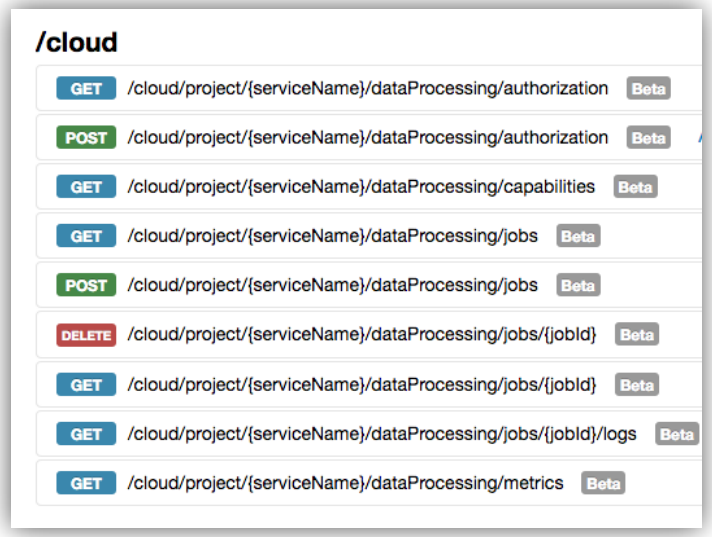
Available regions :
- Europe
- North America (Soon)

Job size :
- Up to 60 GB of RAM per Executor
- Up to 16 Cores of CPU per Executor
- Up to 10 Executors per job

1 Via Control Panel



2 Via API OVHcloud



3 Via CLI (Spark-Submit)

```
$ ./ovh-spark-submit \
--project-id yourProjectId \
--upload ./spark-examples.jar \
--class org.apache.spark.examples.SparkPi \
--driver-cores 1 \
--driver-memory 4G \
--executor-cores 1 \
--executor-memory 4G \
--num-executors 1 \
swift://odp/spark-examples.jar 1000
```

- Auto upload your application code
- Open source

https://github.com/ovh/data-processing-spark-submit

Example 2 : New-York Taxis

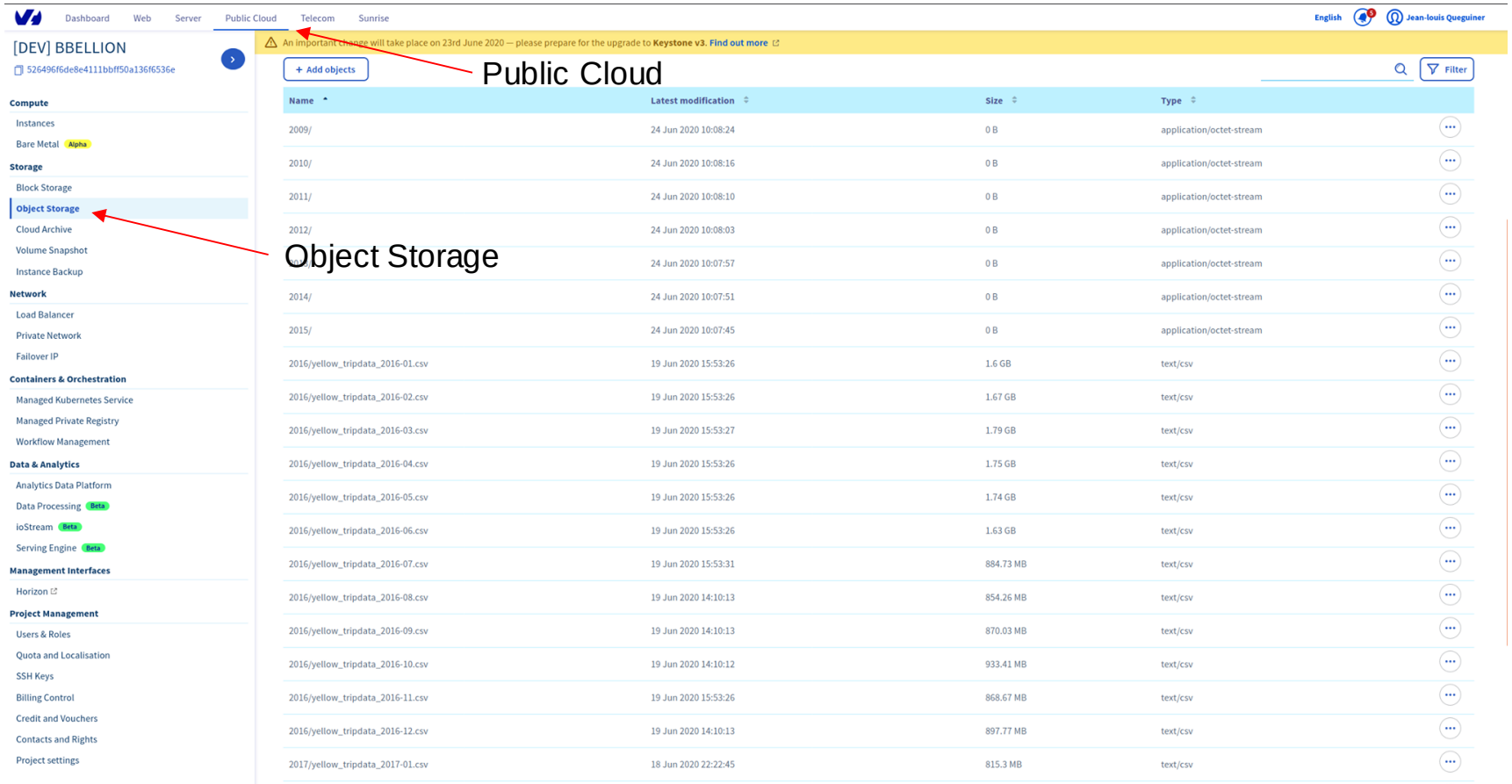
- 10 years worth of taxis ride
- +100 Go of data

- Date
- Number of people per ride
- Price
- Tips
- Pickup and dropoff zone

```
1,2019-01-01 00:46:40,2019-01-01 00:53:20,1,1.50,1,N,151,239,1,7,0.5,0.5,1.65,0,0.3,9.95,
1,2019-01-01 00:59:47,2019-01-01 01:18:59,1,2.60,1,N,239,246,1,14,0.5,0.5,1,0,0.3,16.3,
2,2018-12-21 13:48:30,2018-12-21 13:52:40,3,.00,1,N,236,236,1,4.5,0.5,0.5,0,0,0.3,5.8,
2,2018-11-28 15:52:25,2018-11-28 15:55:45,5,.00,1,N,193,193,2,3.5,0.5,0.5,0,0,0.3,7.55,
2,2018-11-28 15:56:57,2018-11-28 15:58:33,5,.00,2,N,193,193,2,52,0,0.5,0,0,0.3,55.55,
2,2018-11-28 16:25:49,2018-11-28 16:28:26,5,.00,1,N,193,193,2,3.5,0.5,0.5,0,5.76,0.3,13.31,
2,2018-11-28 16:29:37,2018-11-28 16:33:43,5,.00,2,N,193,193,2,52,0,0.5,0,0,0.3,55.55,
1,2019-01-01 00:21:28,2019-01-01 00:28:37,1,1.30,1,N,163,229,1,6.5,0.5,0.5,1.25,0,0.3,9.05,
1,2019-01-01 00:32:01,2019-01-01 00:45:39,1,3.70,1,N,229,7,1,13.5,0.5,0.5,3.7,0,0.3,18.5,
```

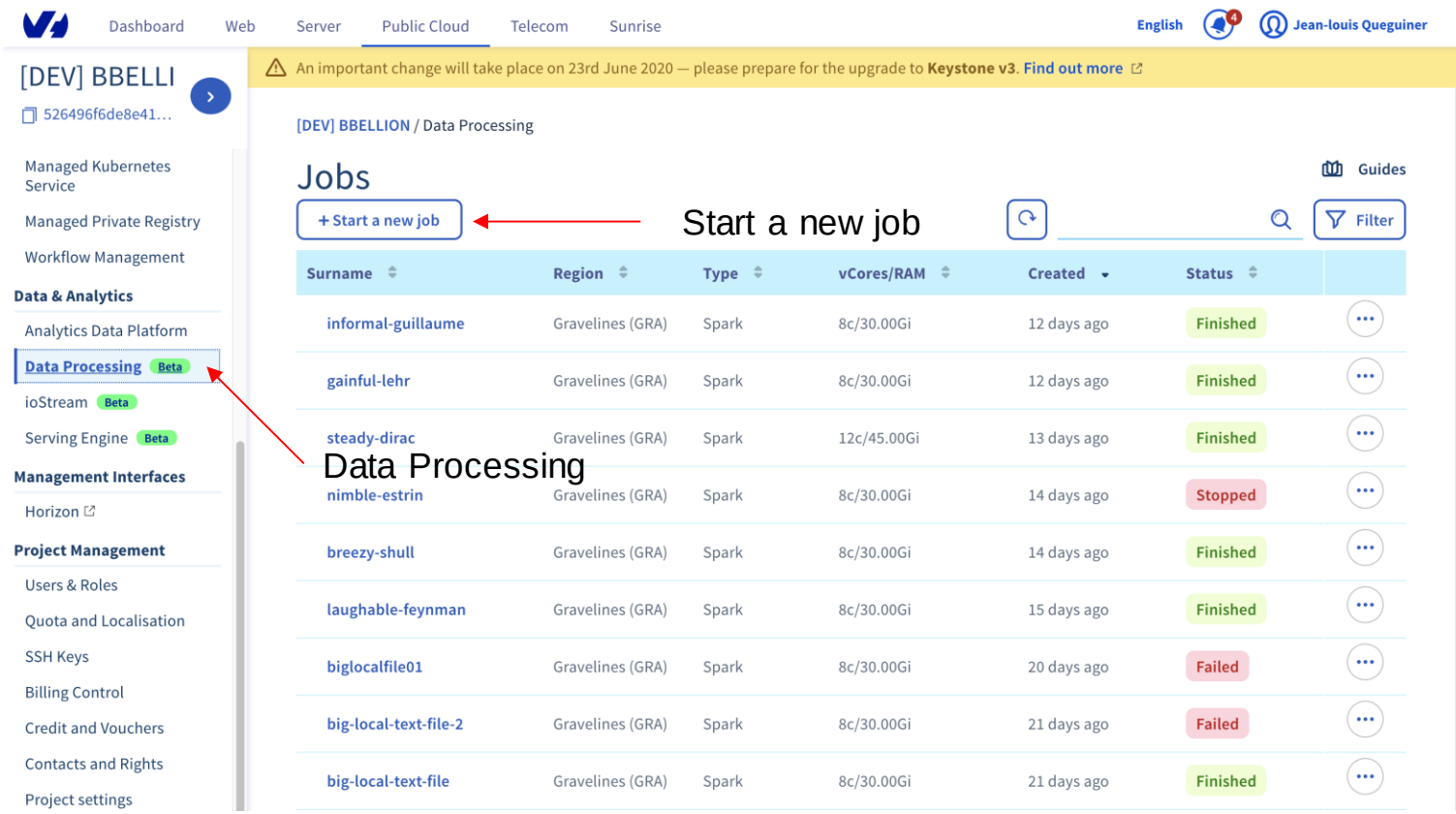
Example 2 : New-York Taxis

First step: make the data accessible



Example 2 : New-York Taxis

Second step: submit your job



Example 2 : New-York Taxis

Second step: submit your job

1 Select a job type

Spark

Start an Apache Spark job using a JAR Java/Scala file, or Python code with Conda dependencies.

Apache Spark 2.4.3

Next

2 Select a region

All locations | Western Europe

Gravelines (GRA)

Next

Example 2 : New-York Taxis

Second step: submit your job

3 **Job sizing**

Your Spark job is managed by a master node (driver) distributing tasks to the computing nodes (executor). Select the number of cores and the quantity of memory allocated to your nodes, as well as the number of nodes you want.

Driver template

4 virtual cores - 15 GiB memory

Executor template

4 virtual cores - 15 GiB memory

Number of executors

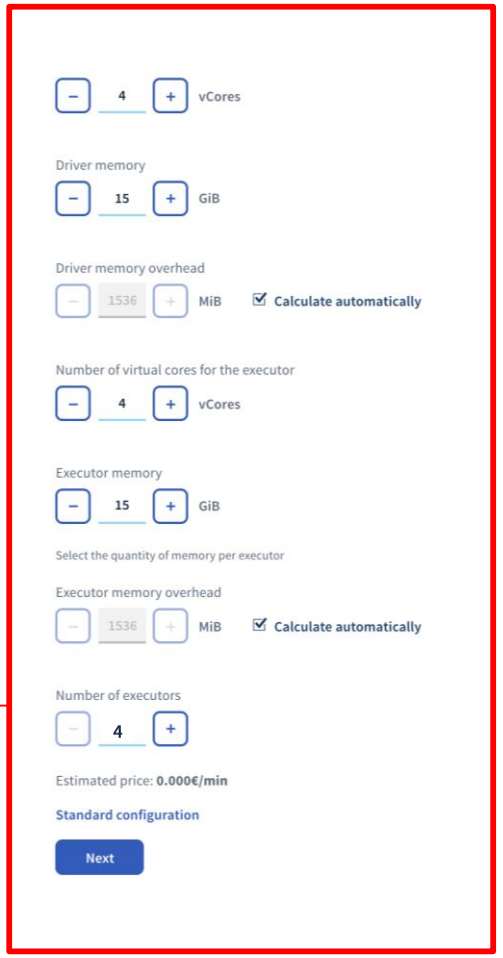
− 4 +

Select the number of executors for your job

Estimated price: **0.000€/min**

Advanced configuration

Next



Example 2 : New-York Taxis

Second step: submit your job

4 **Configure your job**

Job name

new-york_taxis

Minimum 3 characters

Swift container

odp

Job type

Python

Python file

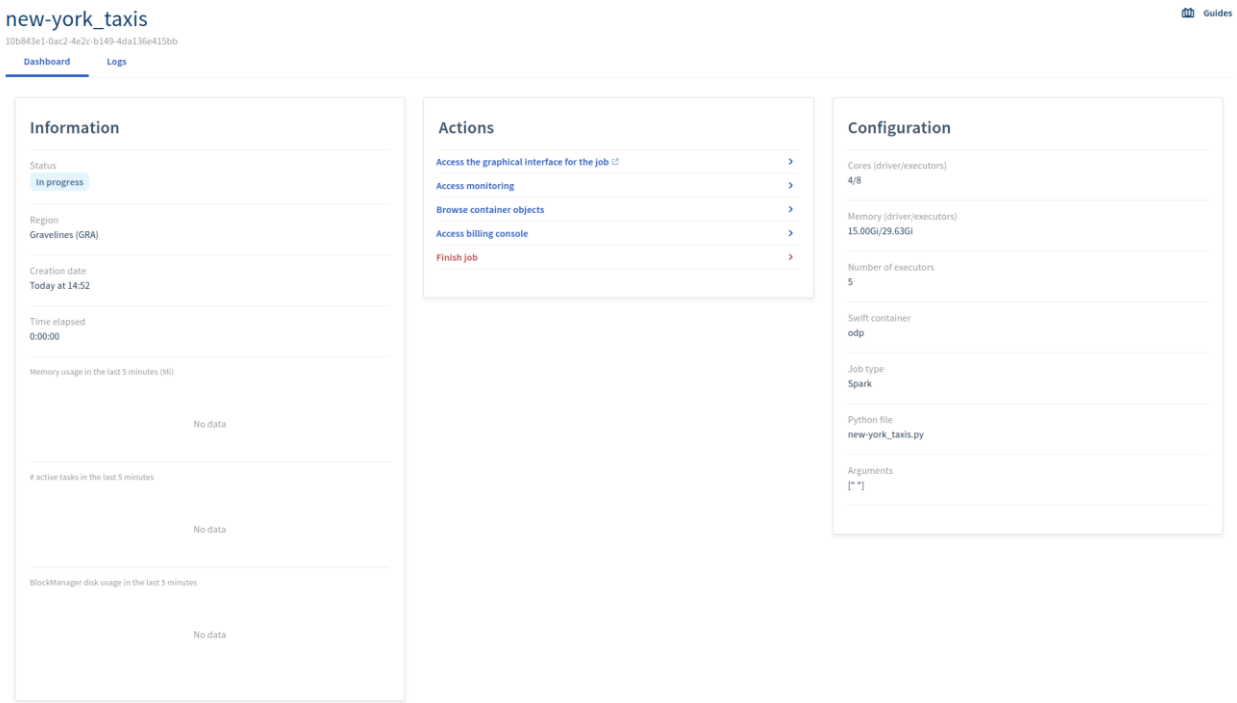
new-york_taxis.py

Pathway to your JAR file within your object container

Arguments (optional)

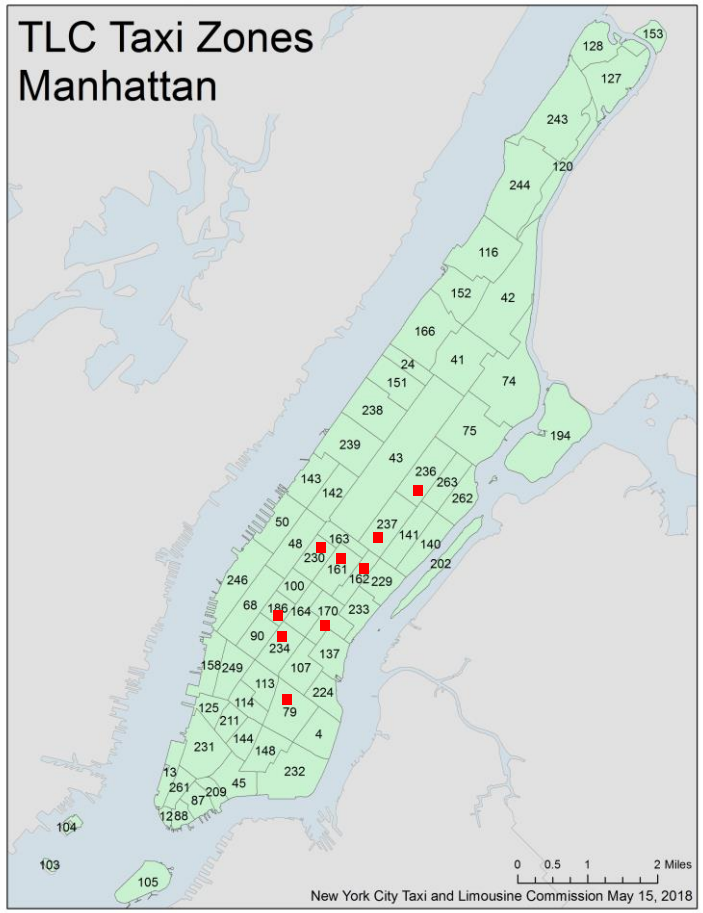
Press the <Enter> key to add an argument

Submit job



Example 2 : New-York Taxis

```
161  :  387702
170  :  338603
230  :  325090
162  :  322163
237  :  294053
186  :  291651
234  :  289236
236  :  287215
48   :  278305
79   :  254939
```



Bill

I submit 1 job, with
- 1 driver (4 vCore / 15GB RAM)
- 5 executors (5 x 8 vCore / 30GB RAM)

I also used OVHcloud Object Storage
Job lasted **15 minutes**

Payment : (44 vCores + 165 GB RAM) for **15 minutes**
= 1,33€
+ classic price of Object Storage (storage and data transfer)

Roadmap

Launched (April)	ETA = July 2020	S2 2020

Development → Lab → GA → Soon

Free ! Full working:
- Product itself
- Control panel/API/CLI
- Documentation

Added:
- Bug fix
- Billing

Backlog:
- More Regions
- More flavors/prices
- More Engines
- More Spark versions

Thank You !

https://gitter.im/ovh/data-processing

https://www.ovh.com/blog/ovhcloud-techs-talk-fr-s01e09/