академия больших данных

mail.ru group

made

Инструменты визуализации

Дмитрий Музалевский

10.10.2020

План занятия

1. Hadoop Local
2. Hadoop Cloud (EMR)
3. Sagemaker
4. Apache Zeppelin
5. Polynote
6. Практика

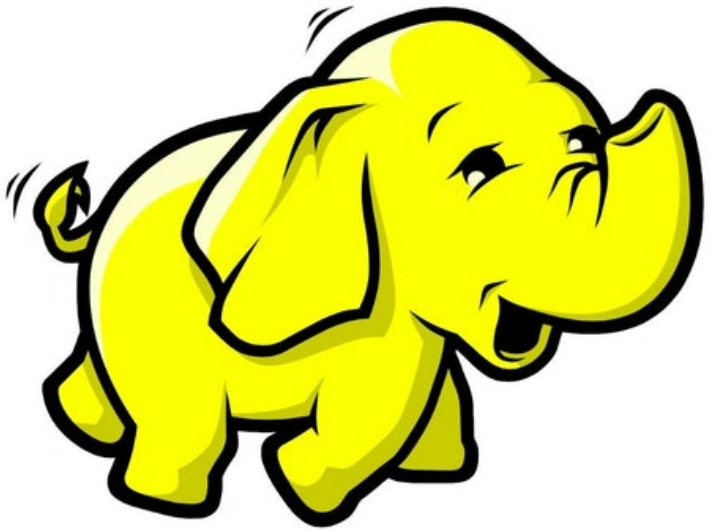
Hadoop Local Install (Mac)

Менеджер пакетов Brew:

- java -version
- (если не установлено): brew cask install homebrew/cask-versions/adoptopenjdk8

- brew install Hadoop

- cd /usr/local/cellar/hadoop/3.3.0/libexec/etc/hadoop



Hadoop Local Config (Mac)

$ open core-site.xml

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

$ open hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

$ open mapred-site.xml

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>mapreduce.application.classpath</name>
<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADO
OP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>
```

Hadoop Local Config (Mac)

$ open yarn-site.xml

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HO
ME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADO
OP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>
```

Hadoop Local Settings (Mac)

Удаление пароля:

- $ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
- $ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
- $ chmod 0600 ~/.ssh/authorized_keys

Формат NameNode:

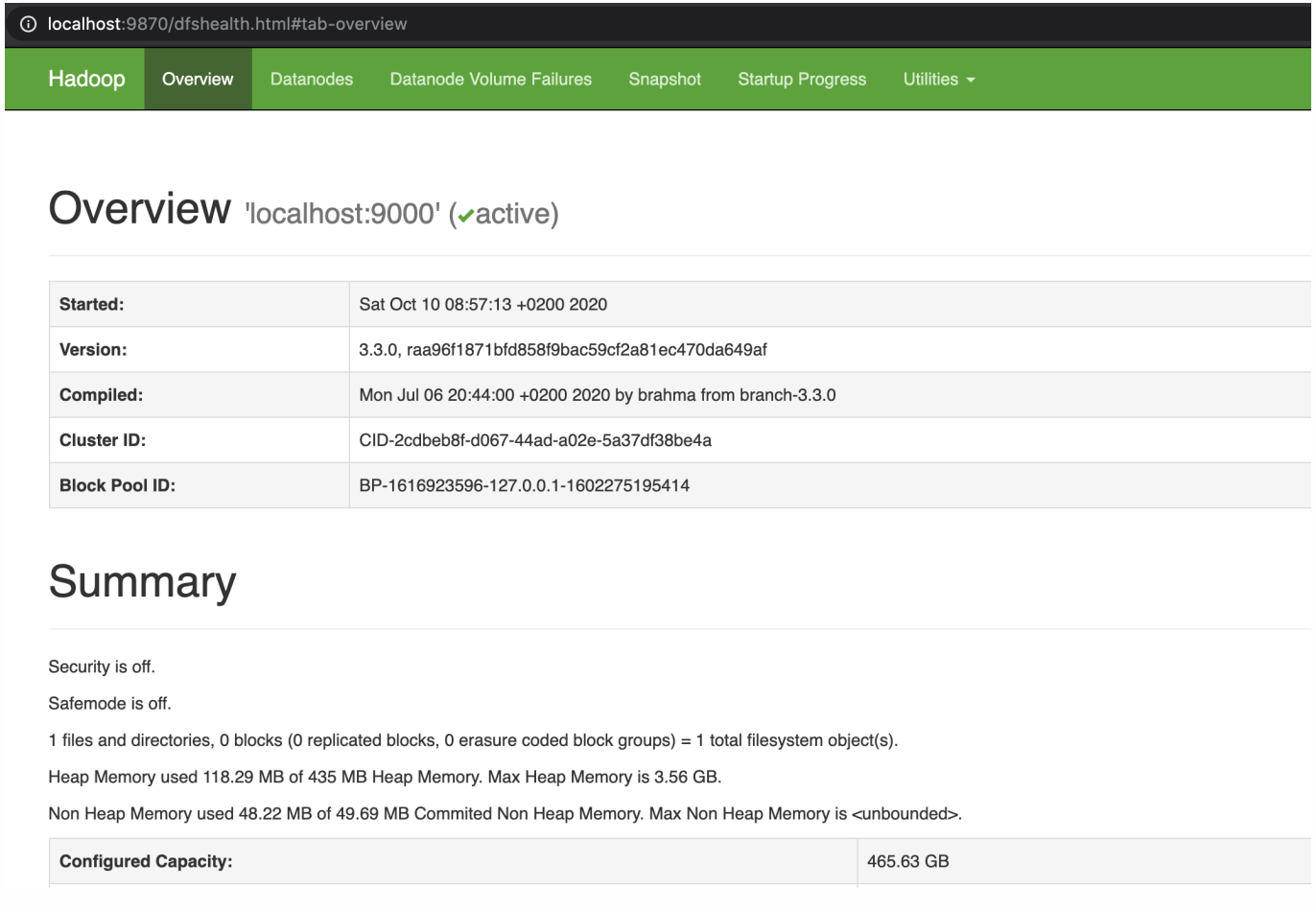
- $ cd /usr/local/cellar/hadoop/3.2.1/libexec/bin
- $ hdfs namenode -format

Hadoop Local Run (Mac)

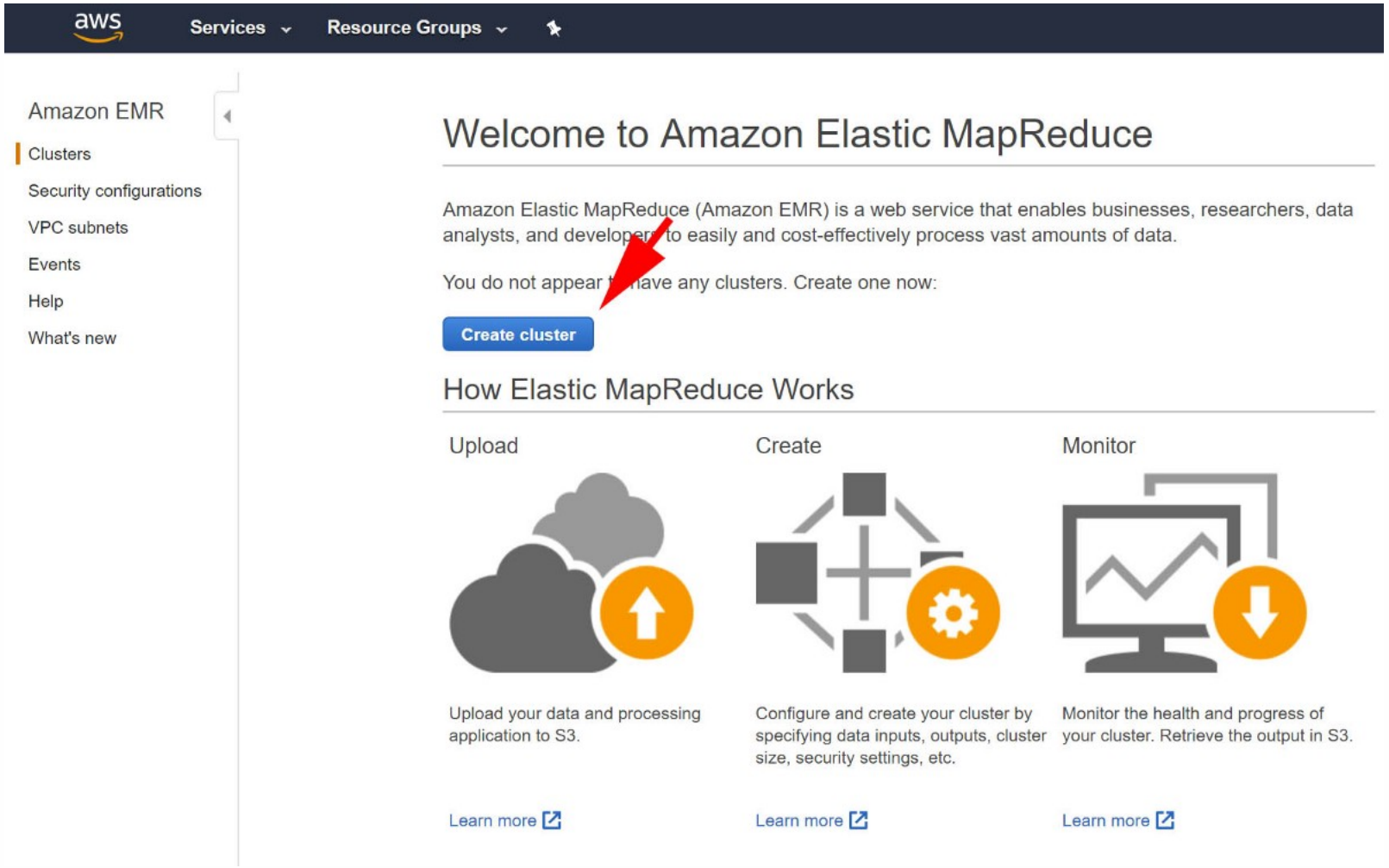
- $ cd /usr/local/cellar/hadoop/3.2.1/libexec/sbin
- $ ./start-all.sh
- $ jps

http://localhost:9870

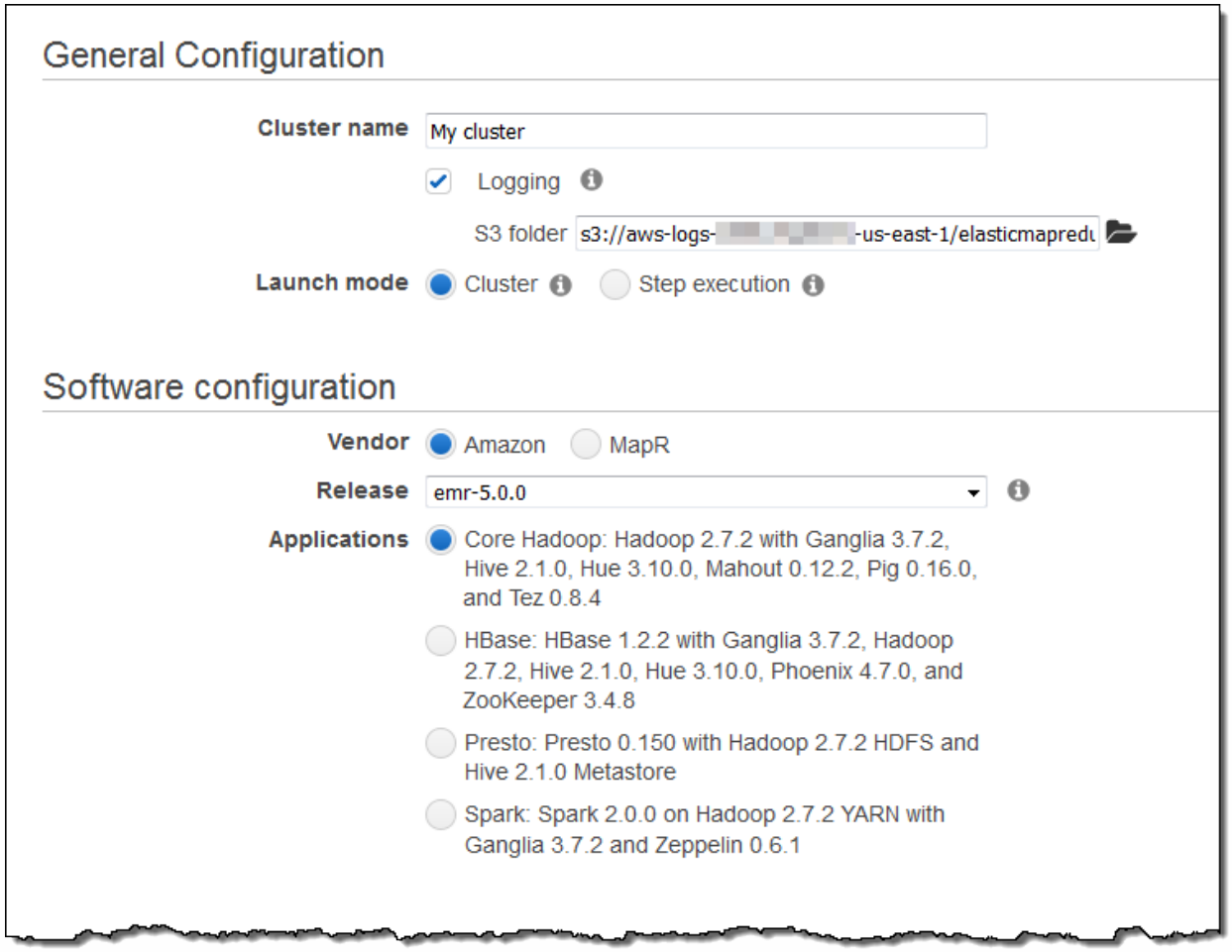
- $ ./stop-all.sh



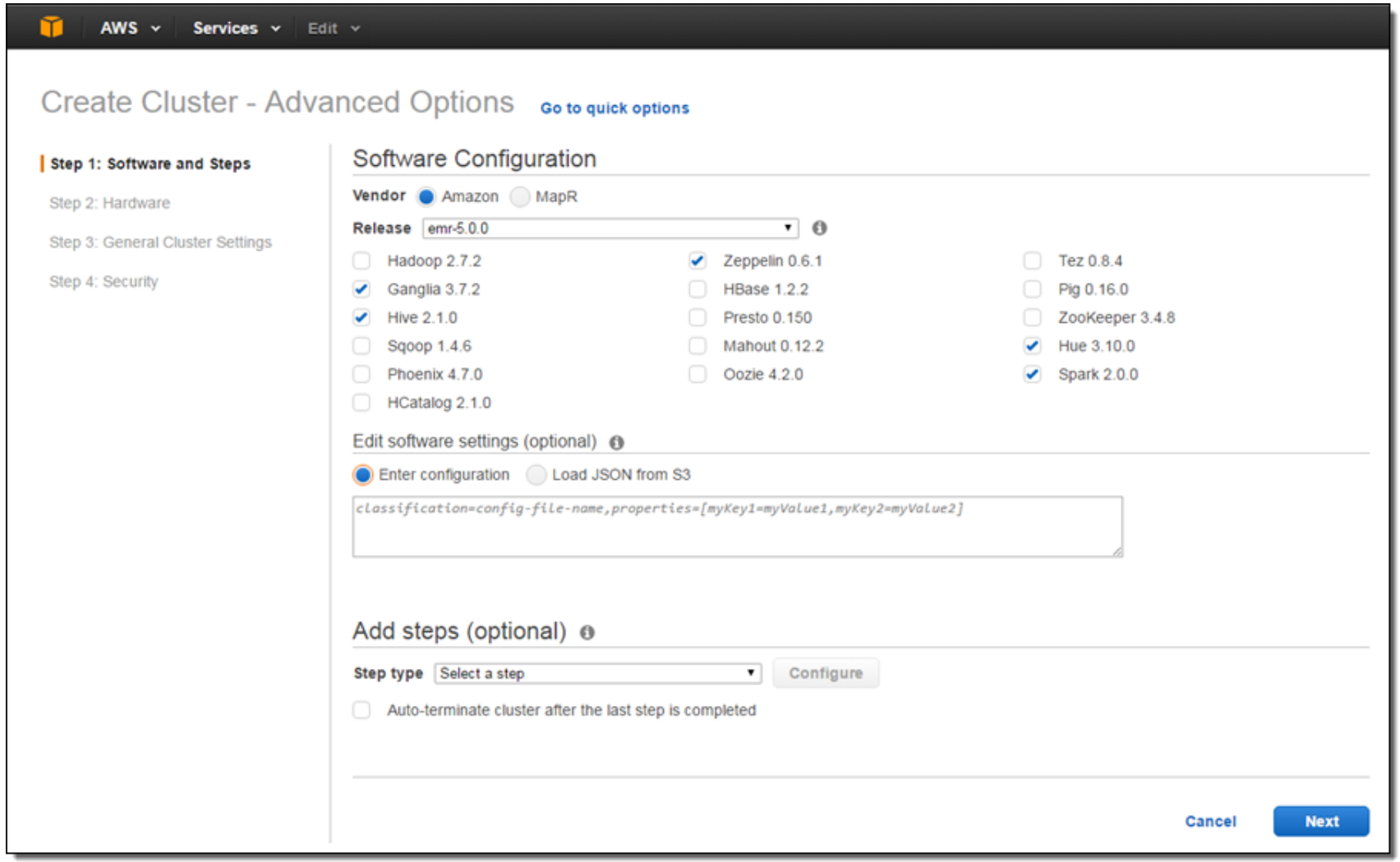
Создание Кластера



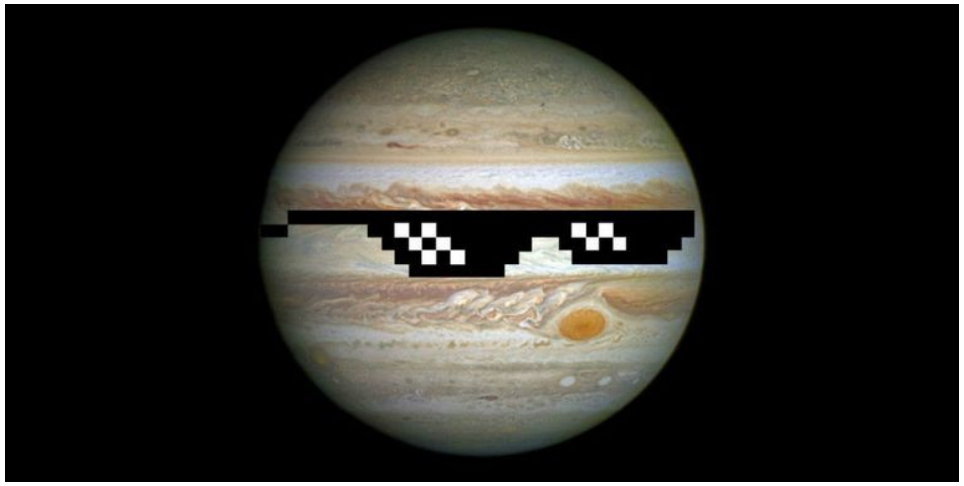
General Configuration



Advanced Options



Jupyter Notebook: Spark + Livy



Notebooks

Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, ar
Hadoop, Spark, and Livy. Notebooks run free of charge and are saved in Amazon S3 independently of clusters. St

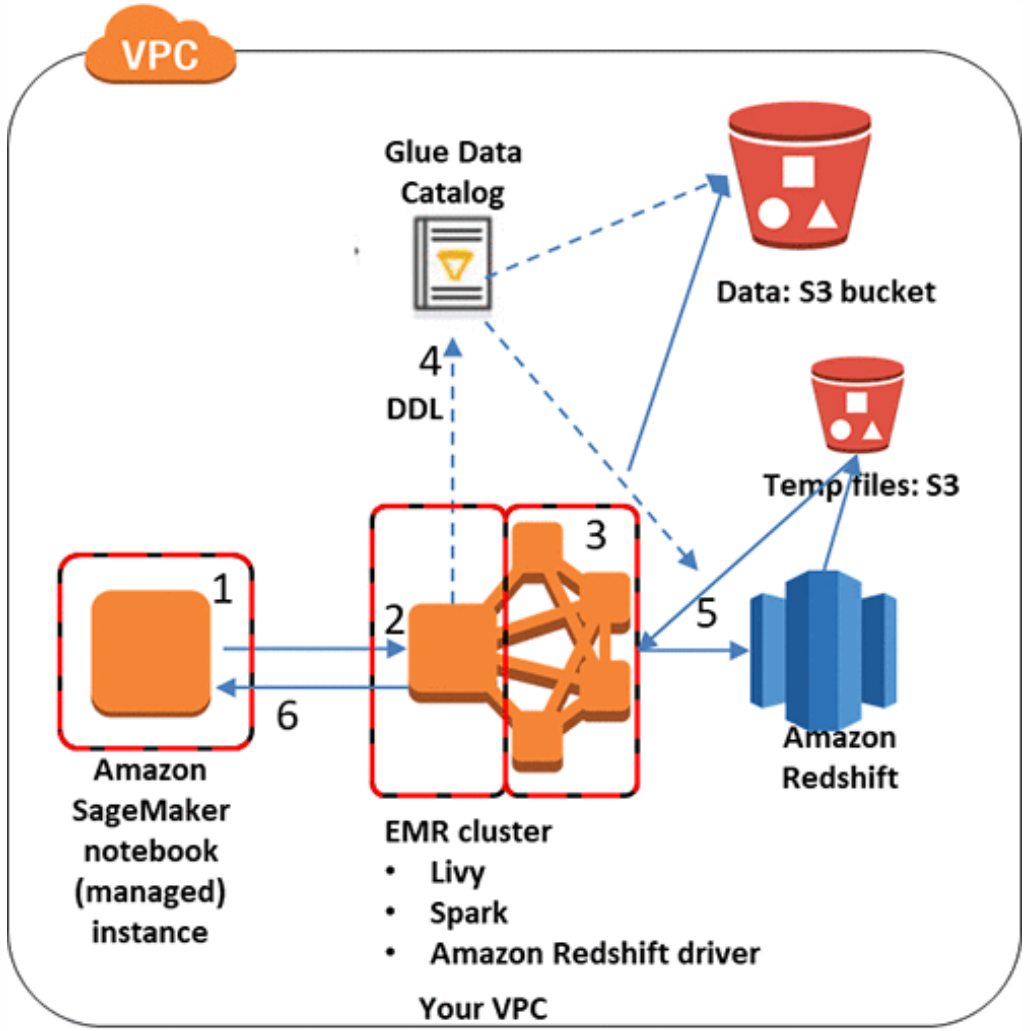
Create notebook | View details | Open in JupyterLab | Open in Jupyter | Start | Stop

Filter: All notebooks | Filter notebooks ... | 1 notebook (all loaded)

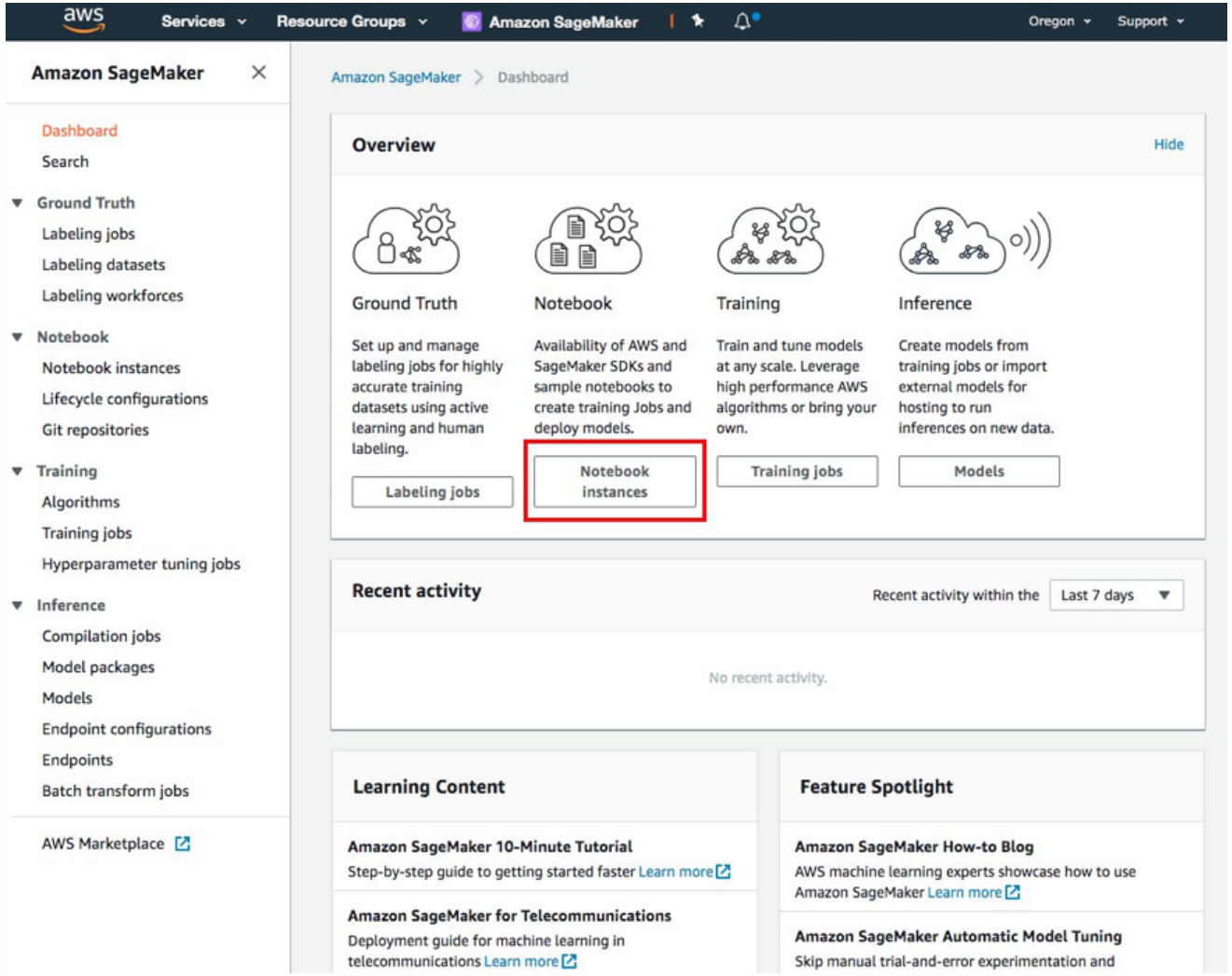
Name	Status	Cluster
Notebook	Ready	j-23WKIE4KI8HPW

Sagemaker Infrastructure

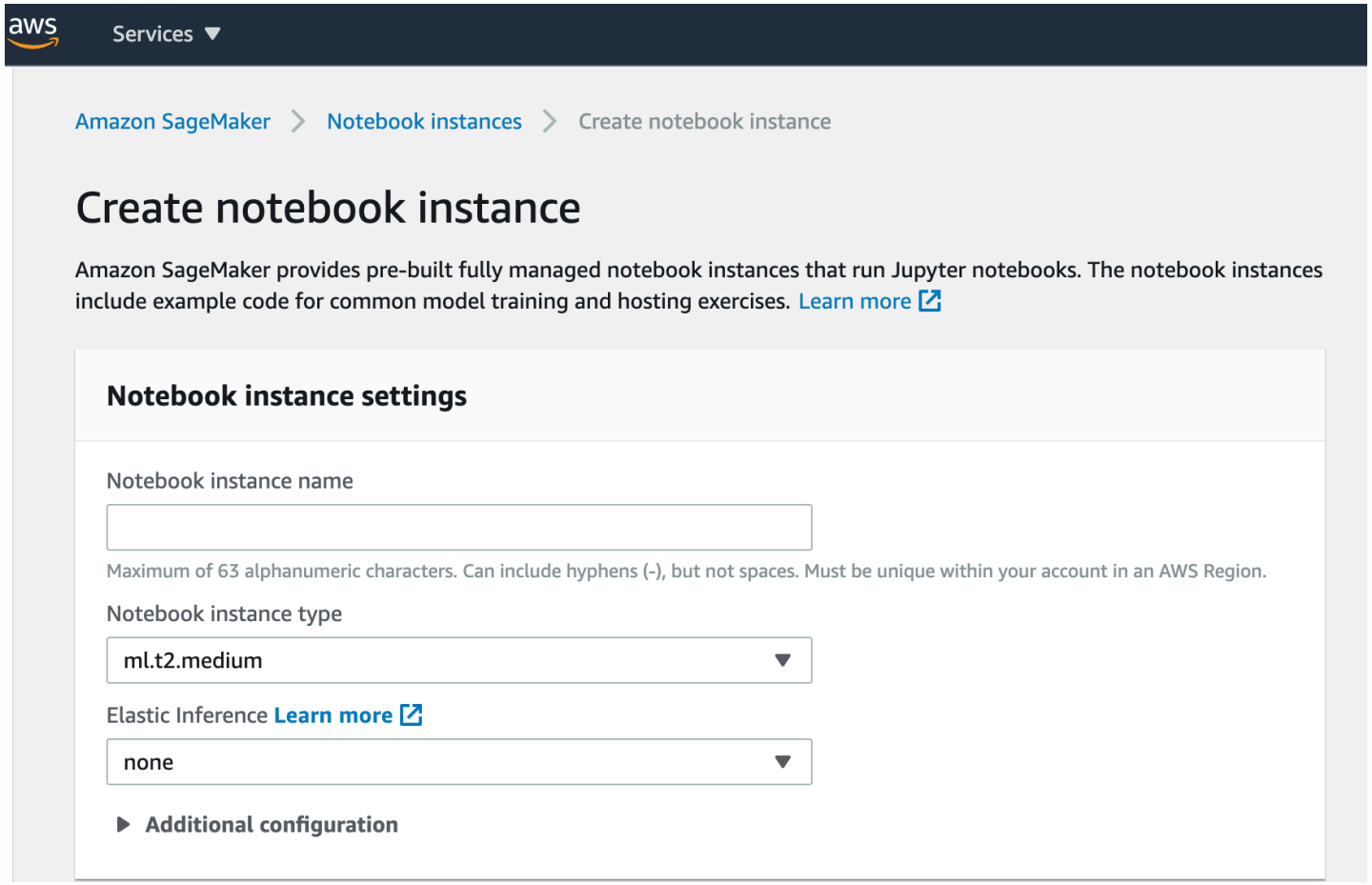
Amazon SageMaker



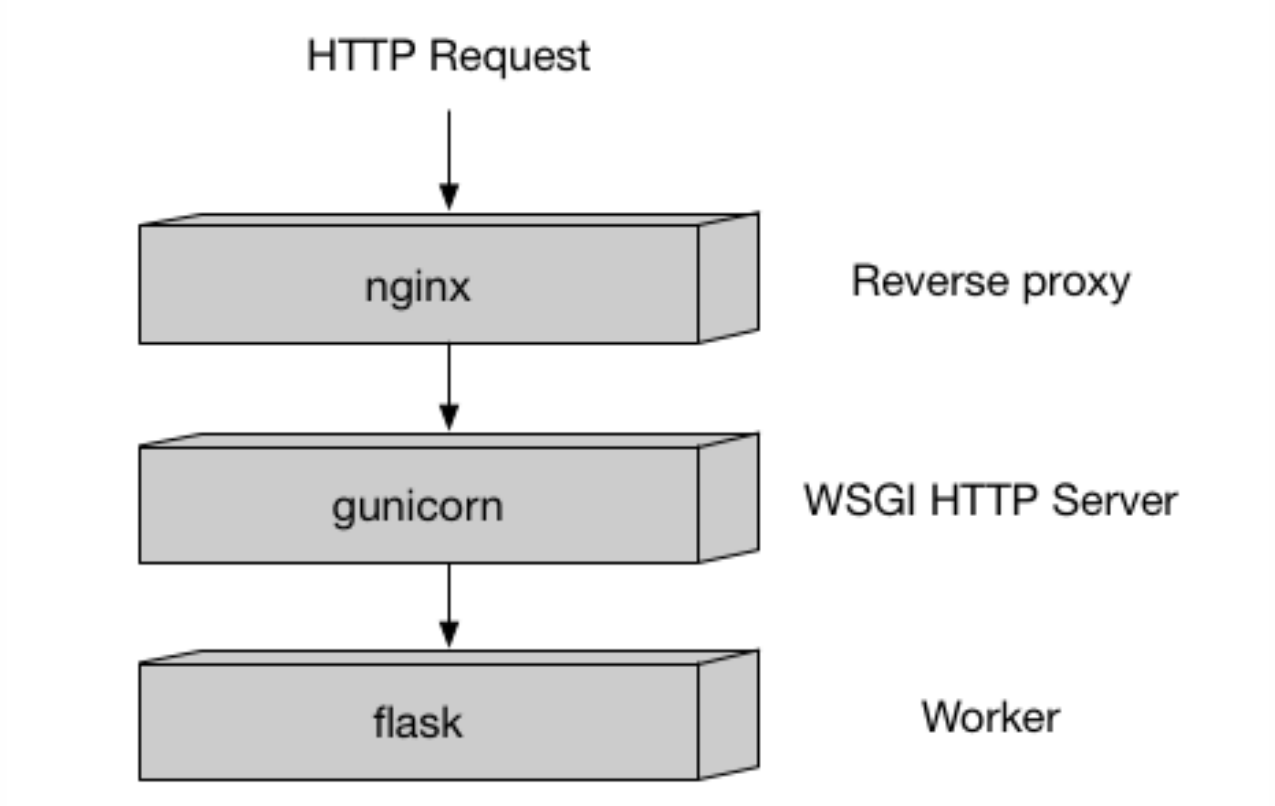
Notebook Instances



Notebook Instances



Sagemaker Estimator



Zeppelin Config

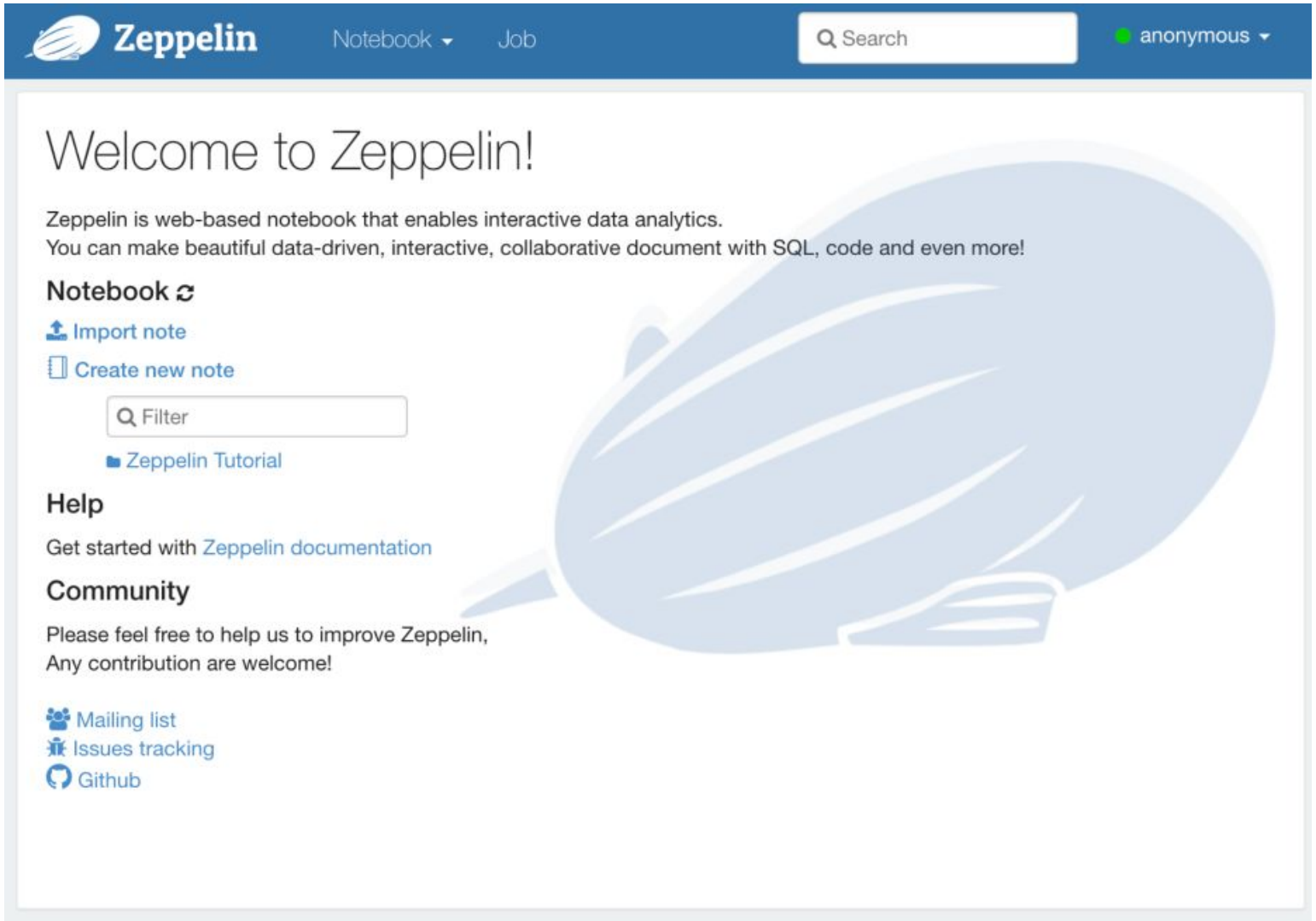
Name of interface	URI
YARN ResourceManager	http://*master-public-dns-name*:8088/
YARN NodeManager	http://*coretask-public-dns-name*:8042/
Hadoop HDFS NameNode	http://*master-public-dns-name*:50070/
Hadoop HDFS DataNode	http://*coretask-public-dns-name*:50075/
Spark HistoryServer	http://*master-public-dns-name*:18080/
Zeppelin	http://*master-public-dns-name*:8890/
Hue	http://*master-public-dns-name*:8888/
Ganglia	http://*master-public-dns-name*/ganglia/
HBase UI	http://*master-public-dns-name*:16010/

Zeppelin SSH

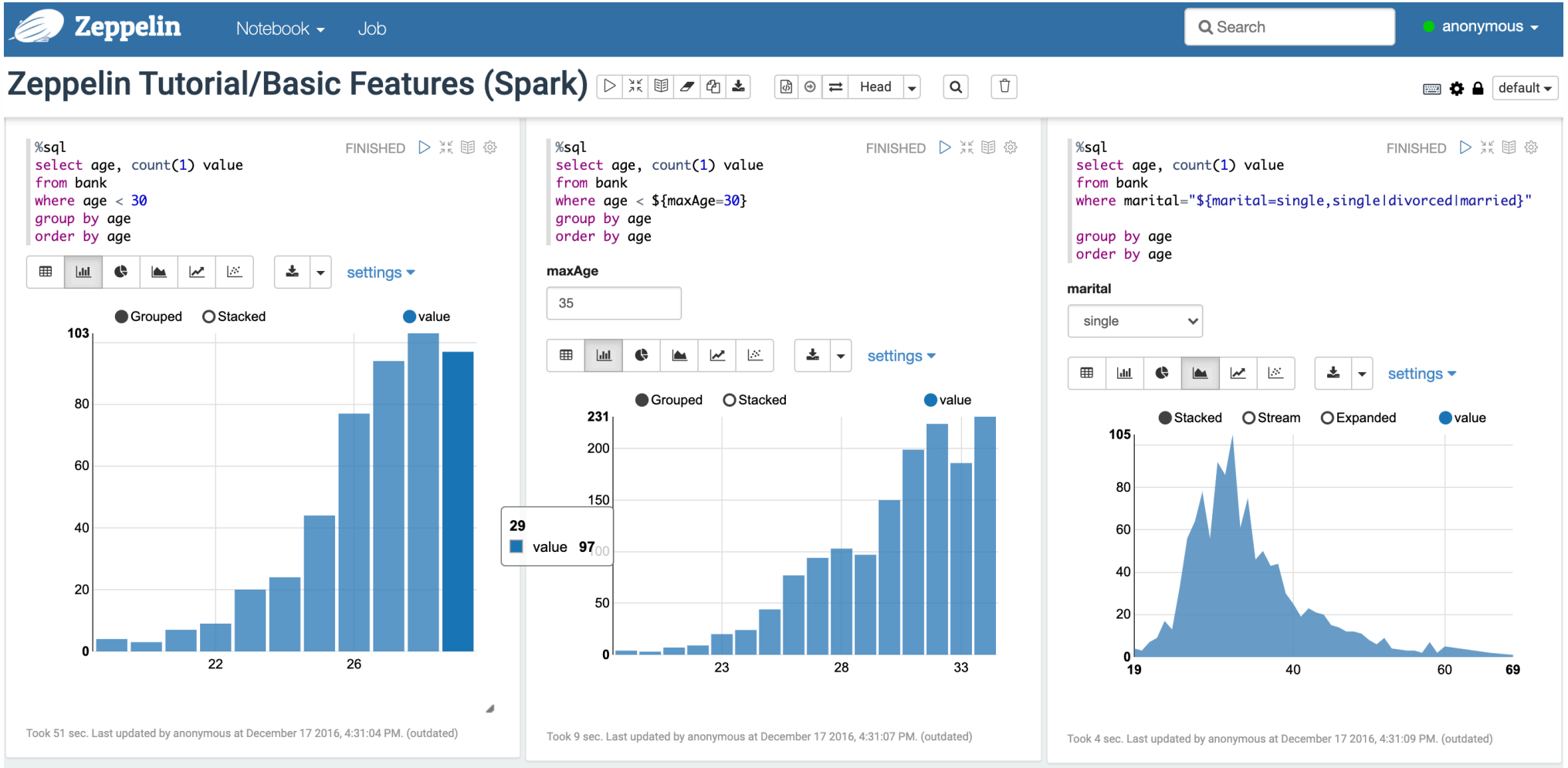
ssh -L 8890:127.0.0.1:8890 -i Test.pem hadoop@ec2-XXX.compute-1.amazonaws.com

```
https://aws.amazon.com/amazon-linux-2/
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file
or directory

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M::::::M         M::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M:::::::M       M:::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M::::::::M     M::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-2-175 ~]$
```

Zeppelin WebUI



Zeppelin WebUI

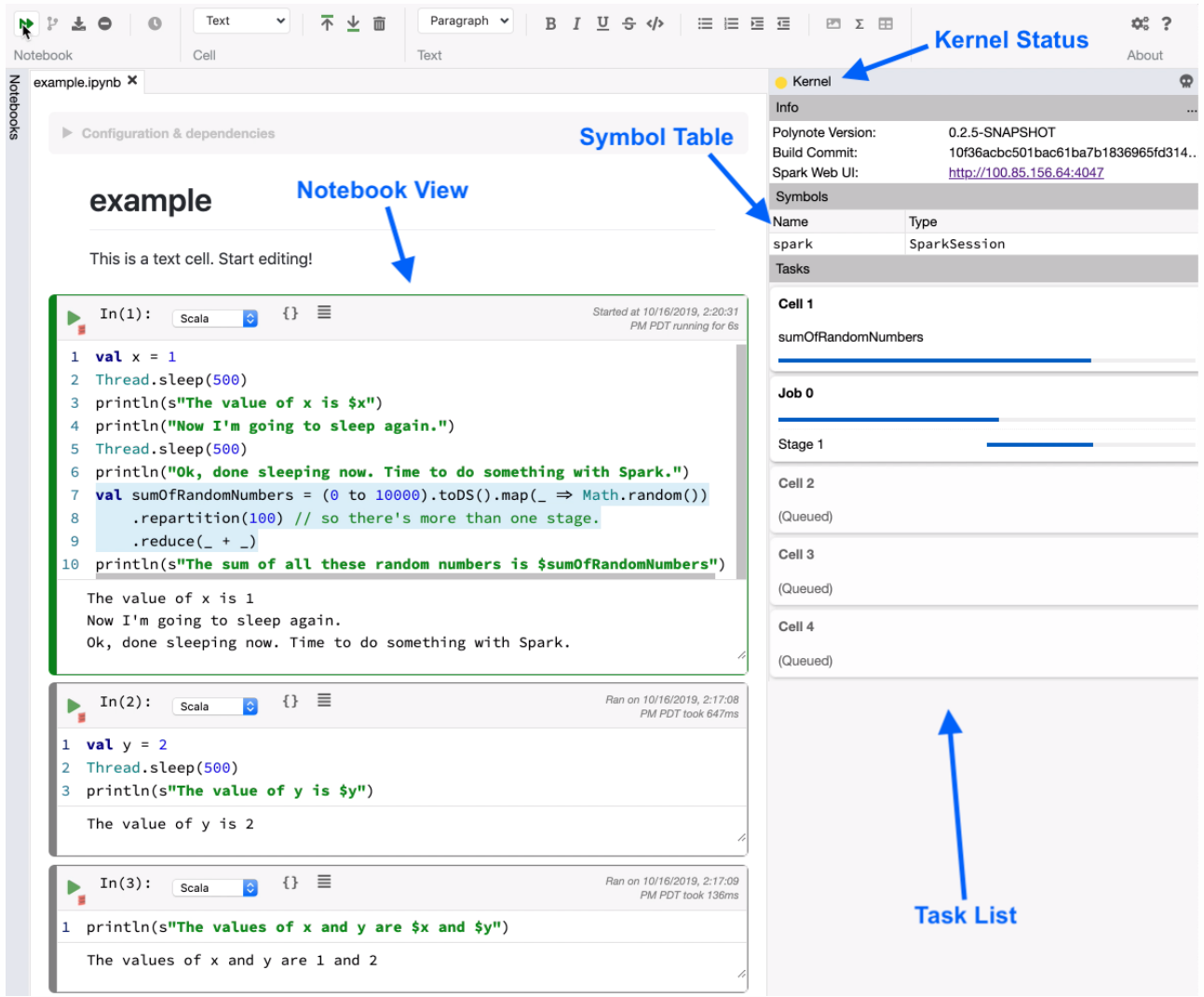


Polynote Install

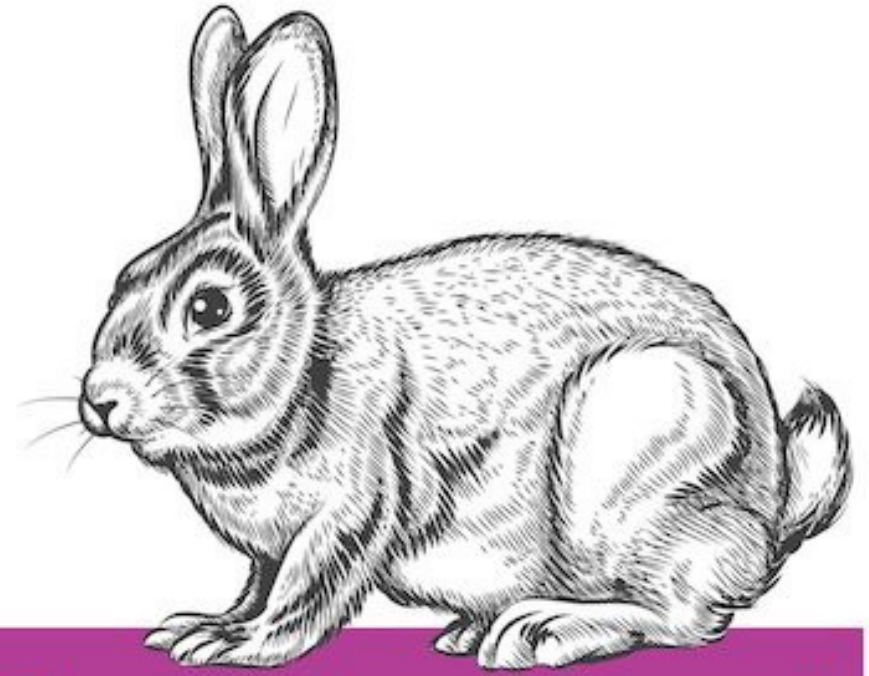
- python3 -m ensurepip
- pip3 install jep jedi pyspark virtualenv numpy pandas –user

- wget https://github.com/polynote/polynote/releases/download/0.2.13/polynote-dist.tar.gz
- tar -zxvpf polynote-dist.tar.gz

Polynote


Kernel Status
Symbol Table
Notebook View
Task List
Notebook
Cell
Text
About
example.ipynb
Notebooks
Configuration & dependencies
example
This is a text cell. Start editing!
In(1):
Scala
Started at 10/16/2019, 2:20:31 PM PDT running for 6s
1 val x = 1
2 Thread.sleep(500)
3 println(s"The value of x is $x")
4 println("Now I'm going to sleep again.")
5 Thread.sleep(500)
6 println("Ok, done sleeping now. Time to do something with Spark.")
7 val sumOfRandomNumbers = (0 to 10000).toDS().map(_ => Math.random())
8 .repartition(100) // so there's more than one stage.
9 .reduce(_ + _)
10 println(s"The sum of all these random numbers is $sumOfRandomNumbers")
The value of x is 1
Now I'm going to sleep again.
Ok, done sleeping now. Time to do something with Spark.
In(2):
Scala
Ran on 10/16/2019, 2:17:08 PM PDT took 647ms
1 val y = 2
2 Thread.sleep(500)
3 println(s"The value of y is $y")
The value of y is 2
In(3):
Scala
Ran on 10/16/2019, 2:17:09 PM PDT took 136ms
1 println(s"The values of x and y are $x and $y")
The values of x and y are 1 and 2
Kernel
Info
Polynote Version: 0.2.5-SNAPSHOT
Build Commit: 10f36acbc501bac61ba7b1836965fd314..
Spark Web UI: http://100.85.156.64:4047
Symbols
Name Type
spark SparkSession
Tasks
Cell 1
sumOfRandomNumbers
Job 0
Stage 1
Cell 2
(Queued)
Cell 3
(Queued)
Cell 4
(Queued)

Depending on a vague popularity contest



Choosing Based on GitHub Stars

You Only Live Once

O RLY?

@ThePracticalDev