MACHINE LEARNING IN SPARK

Константин Макарычев

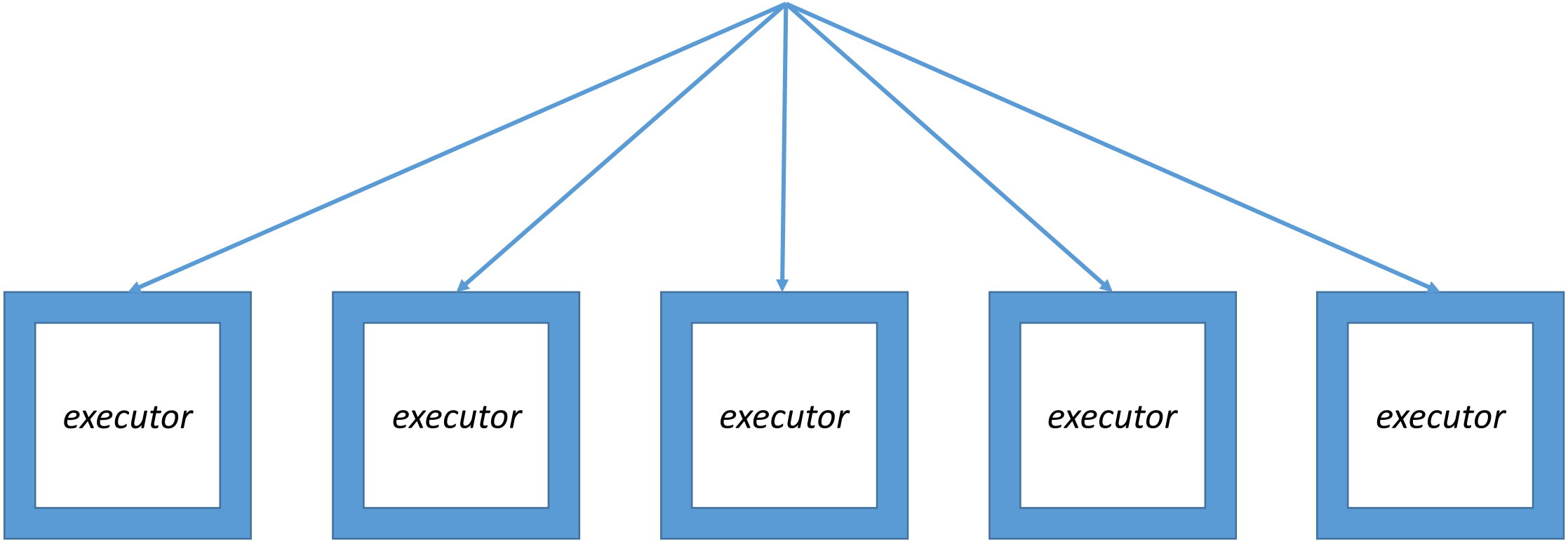
secon 2017

Big Data: Volume, Velocity, Variety

Apache Spark

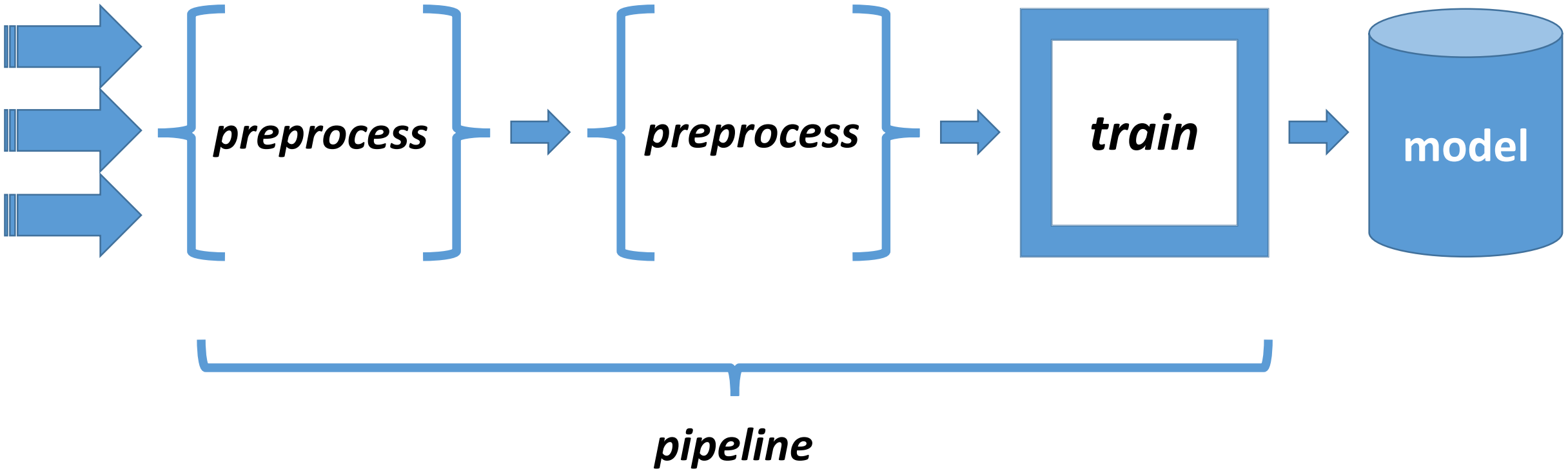
http://spark.apache.org/

```
val wordCounts = textFile
  .flatMap(line => line.split(" "))
  .map(word => (word, 1))
  .reduceByKey((a, b) => a + b)
```



SQL, Streaming, GraphX, **MLlib**

Machine Learning: training + serving


preprocess
preprocess
train
model
pipeline

apache spark	1
hadoop mapreduce	0
spark machine learning	1

[apache, spark]	1
[hadoop, mapreduce]	0
[spark, machine, learning]	1

[apache, spark]	1
[hadoop, mapreduce]	0
[spark, machine, learning]	1

[105, 495], [1.0, 1.0]	1
[6, 638, 655], [1.0, 1.0, 1.0]	0
[105, 72, 852], [1.0, 1.0, 1.0]	1

[105, 495], [1.0, 1.0]	1
[6, 638, 655], [1.0, 1.0, 1.0]	0
[105, 72, 852], [1.0, 1.0, 1.0]	1

0	72	-2.7138781446090308
0	94	0.9042505436914775
0	105	3.083567089049664 5

...

0	495	3.2071722417080766
0	722	0.9042505436914775

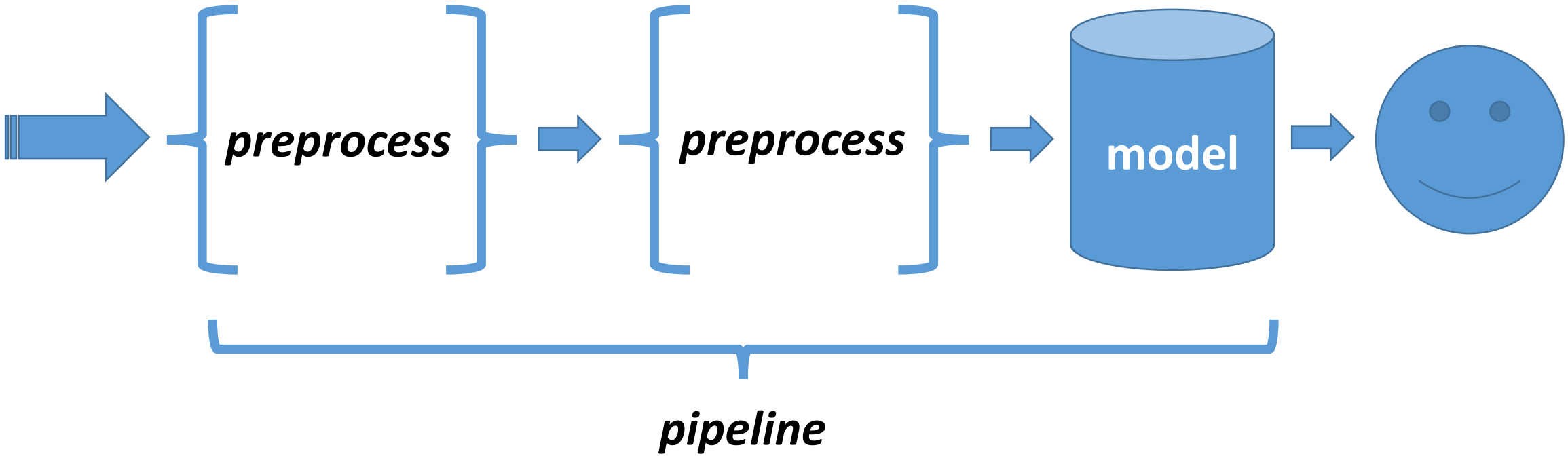
```
val tokenizer = new Tokenizer()
  .setInputCol("text")
  .setOutputCol("words")

val hashingTF = new HashingTF()
  .setNumFeatures(1000)
  .setInputCol(tokenizer.getOutputCol)
  .setOutputCol("features")

val lr = new LogisticRegression()
  .setMaxIter(10)
  .setRegParam(0.001)

val pipeline = new Pipeline()
  .setStages(Array(tokenizer, hashingTF, lr))

val model = pipeline.fit(training)
model.write.save("/tmp/spark-model")
```


preprocess
preprocess
model
pipeline

```
val test = spark.createDataFrame(Seq(
  ("spark hadoop"),
  ("hadoop learning")
)).toDF("text")

val model = PipelineModel.load("/tmp/spark-model")

model.transform(test).collect()
```

./bin/spark-submit …

Mark Roddy
@digitallogic

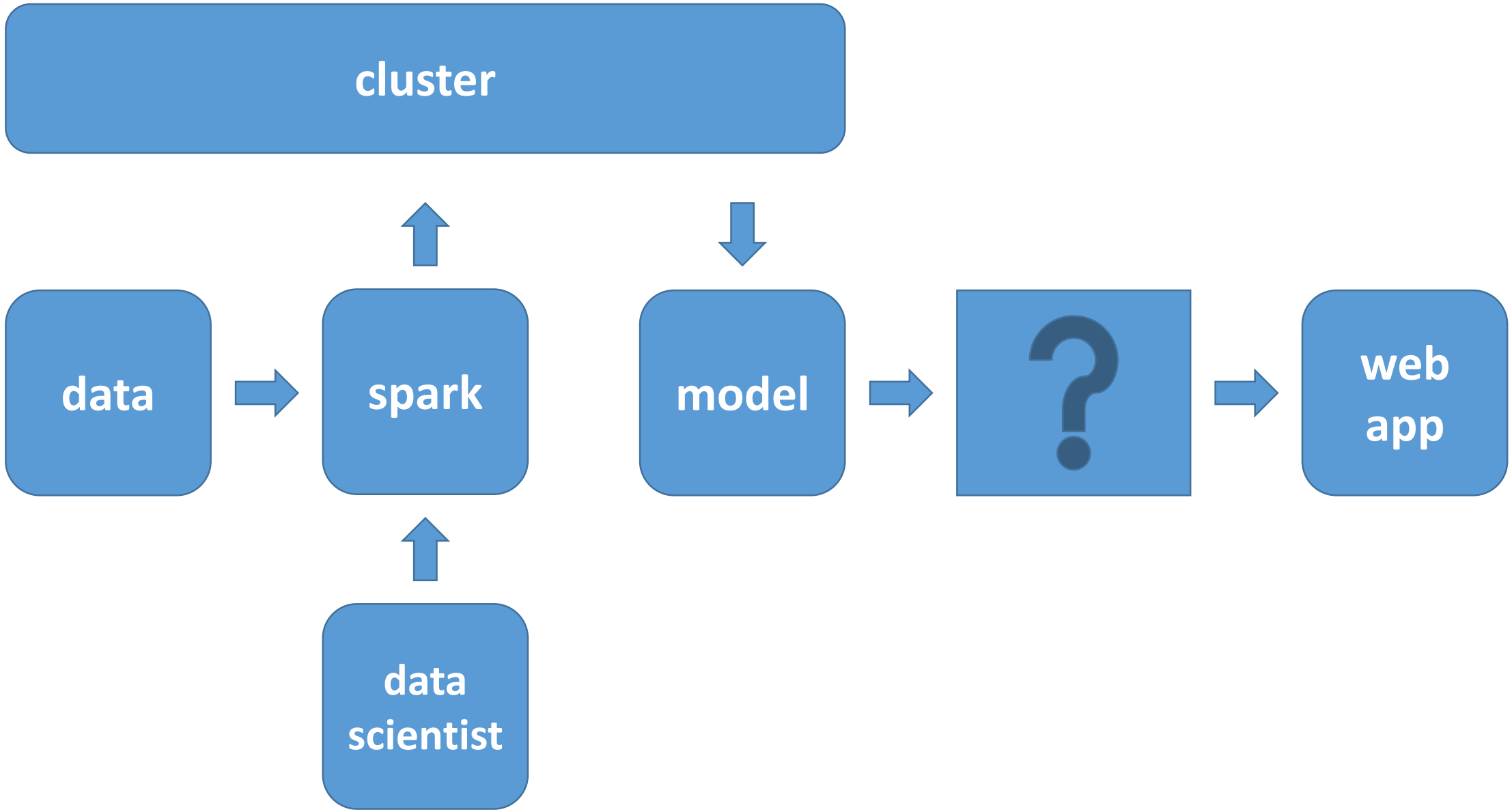
Follow

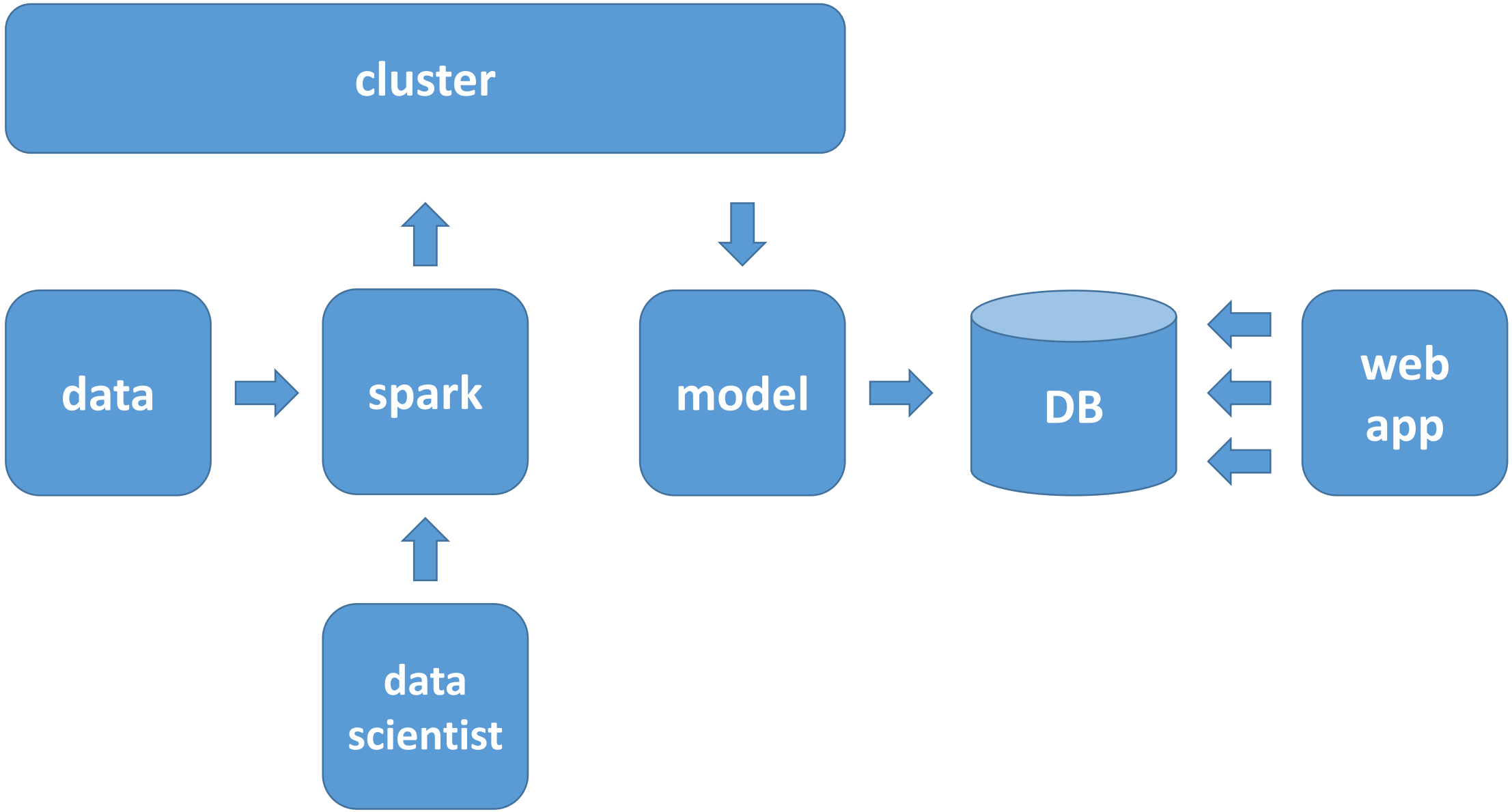
@cdubhland @John4man Hell is being handed 2k python script w/no comments that queries db on a laptop and being told "make this work in prod"

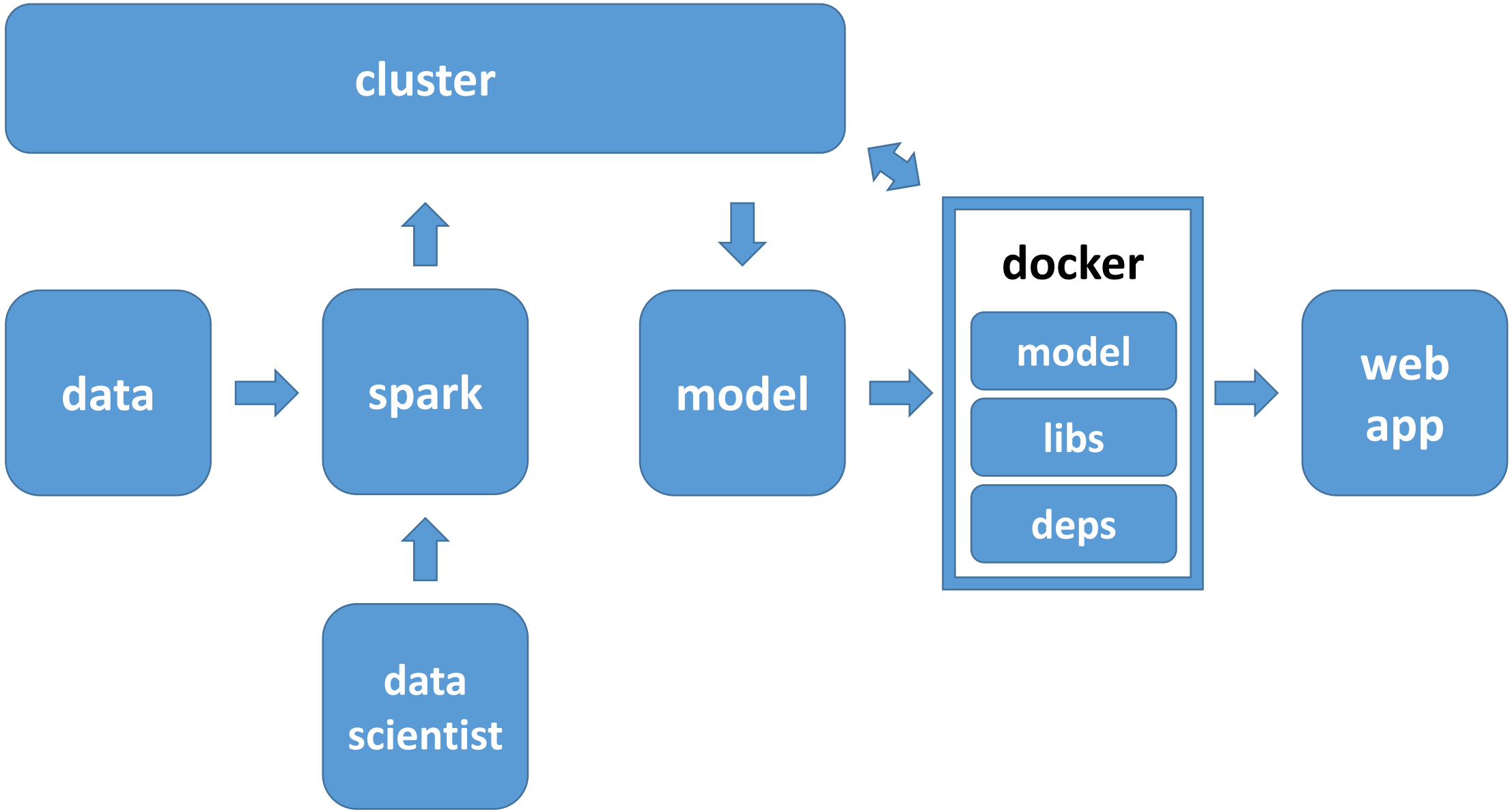
Trey Causey
@treycausey

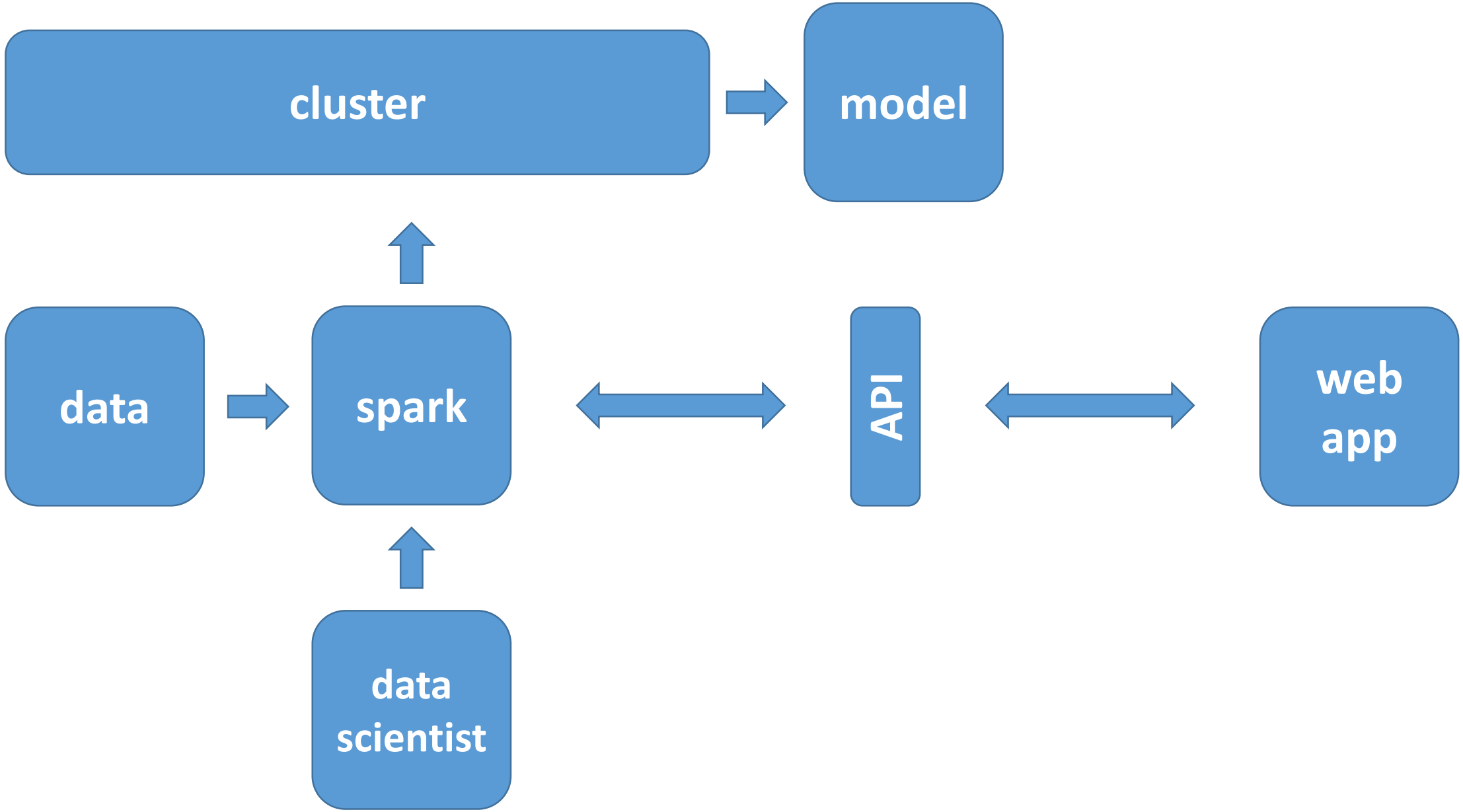
Follow

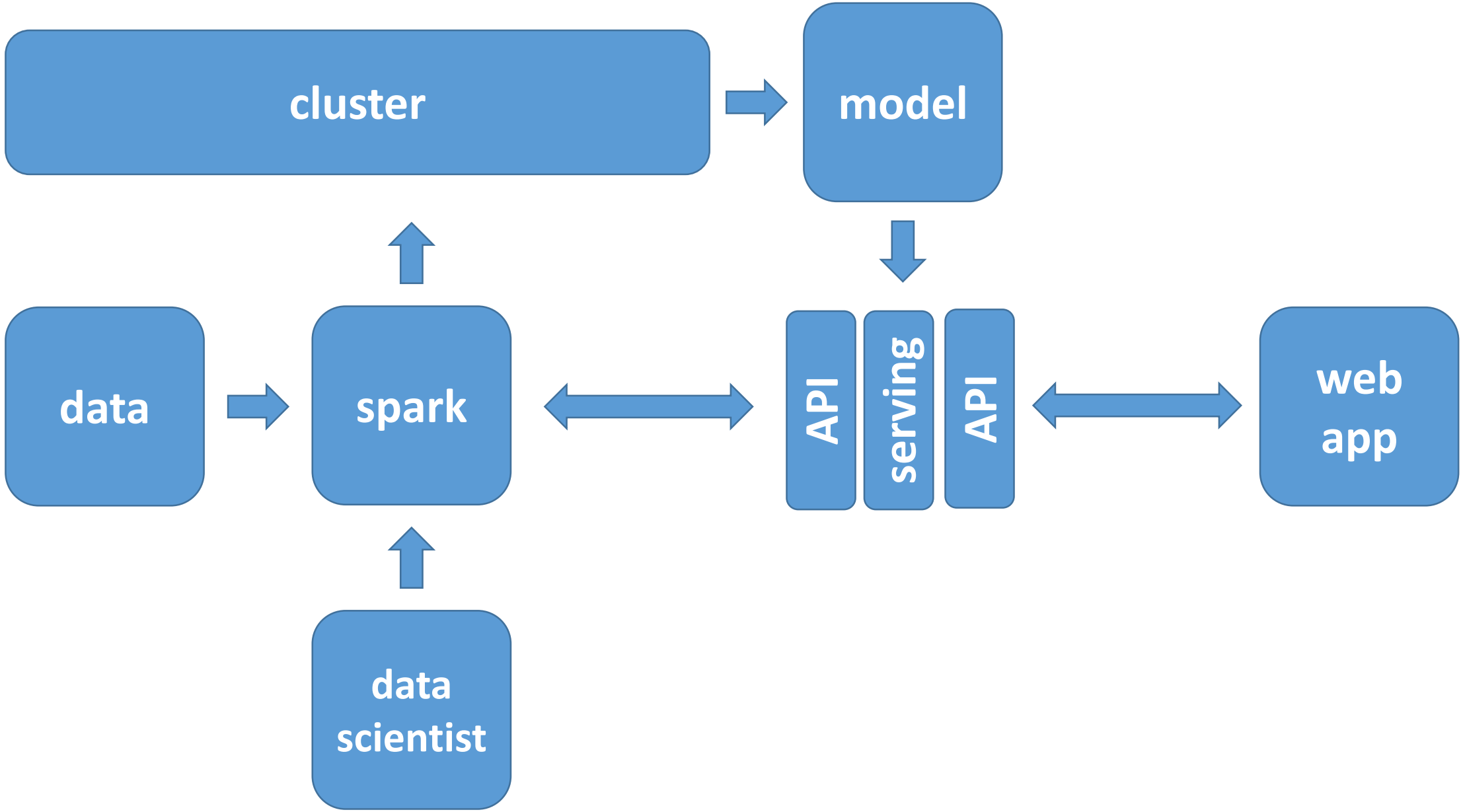
@digitallogic @cdubhland @John4man sudo chmod a+x script.py, edit crontab, bingo you're in production.


cluster
data
spark
model
?
web app
data scientist


cluster
data
spark
model
DB
web
app
data
scientist


cluster
data
spark
model
docker
model
libs
deps
web app
data scientist


cluster
model
data
spark
API
web app
data scientist


cluster
model
data
spark
API
serving
API
web app
data scientist

Hydrosphere Mist

https://github.com/hydrospheredata/mist