Introduction to Zeppelin

Lee moon soo, NFLabs
moon@apache.org

me

- Lee moon soo, aka. 'moon'

- moon@apache.org Committer and PPMC of Zeppelin

- moon@nflabs.com Co-founder of NFLabs

Agenda

- Why do you like Zeppelin?

- Why your project likes Zeppelin?

- Roadmap

- QnA

What is Zeppelin?

Let's see demo

How Zeppelin started

Data = Understanding

War is ninety percent information
- Napoleon Bonaparate

A price of light is less than the cost of darkness
- Arthur C. Nielsen , Founder of ACNielsen

What gets measured, gets managed
- Peter Drucker

In God we trust, all others must bring data
- W. Edwards Deming

For data analysis, we need tool

But couldn't find one i like
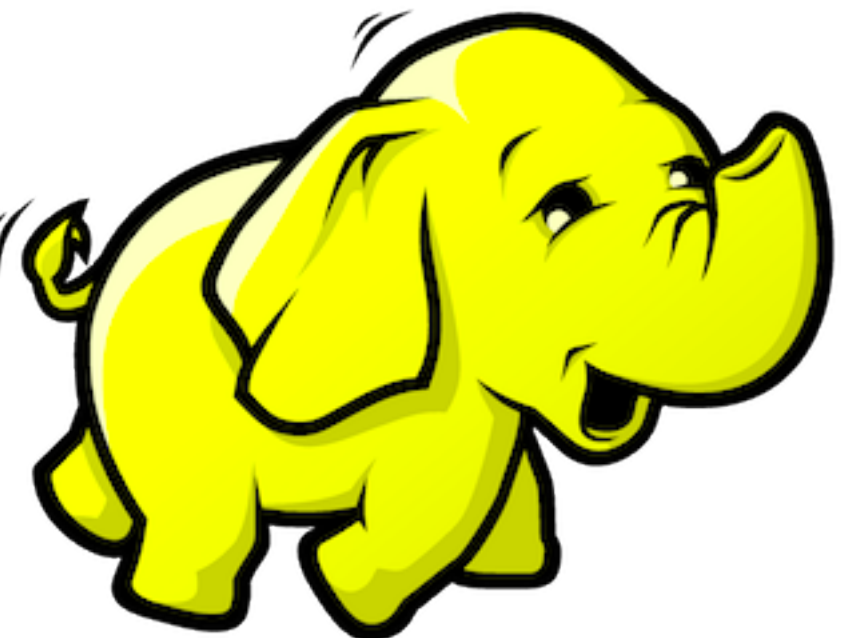
MLLib

Cloudera-ML

MRQL   Drill   Impala

Pig   Hive   tajo

Decided to make one

Really good one

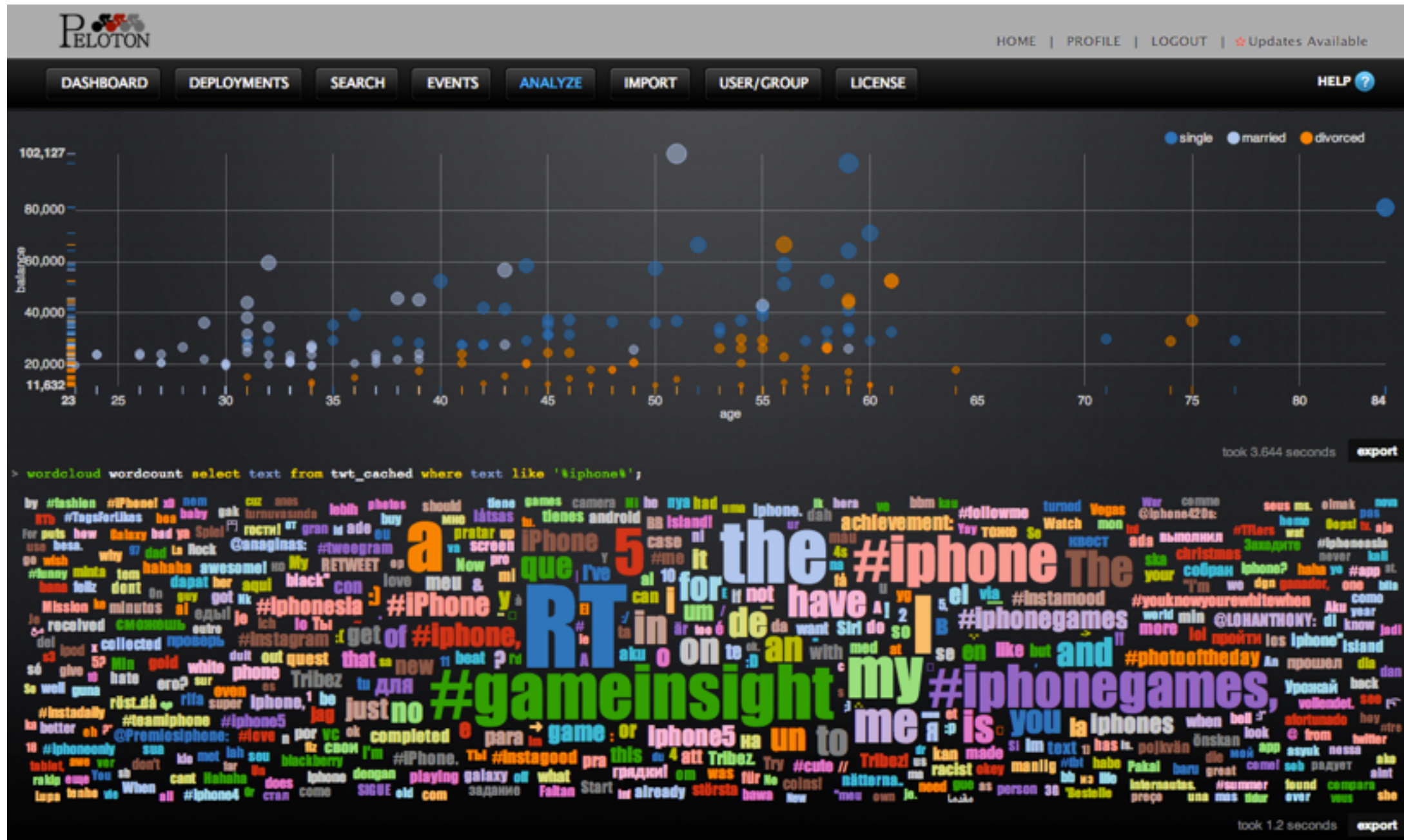# Good analytics environment
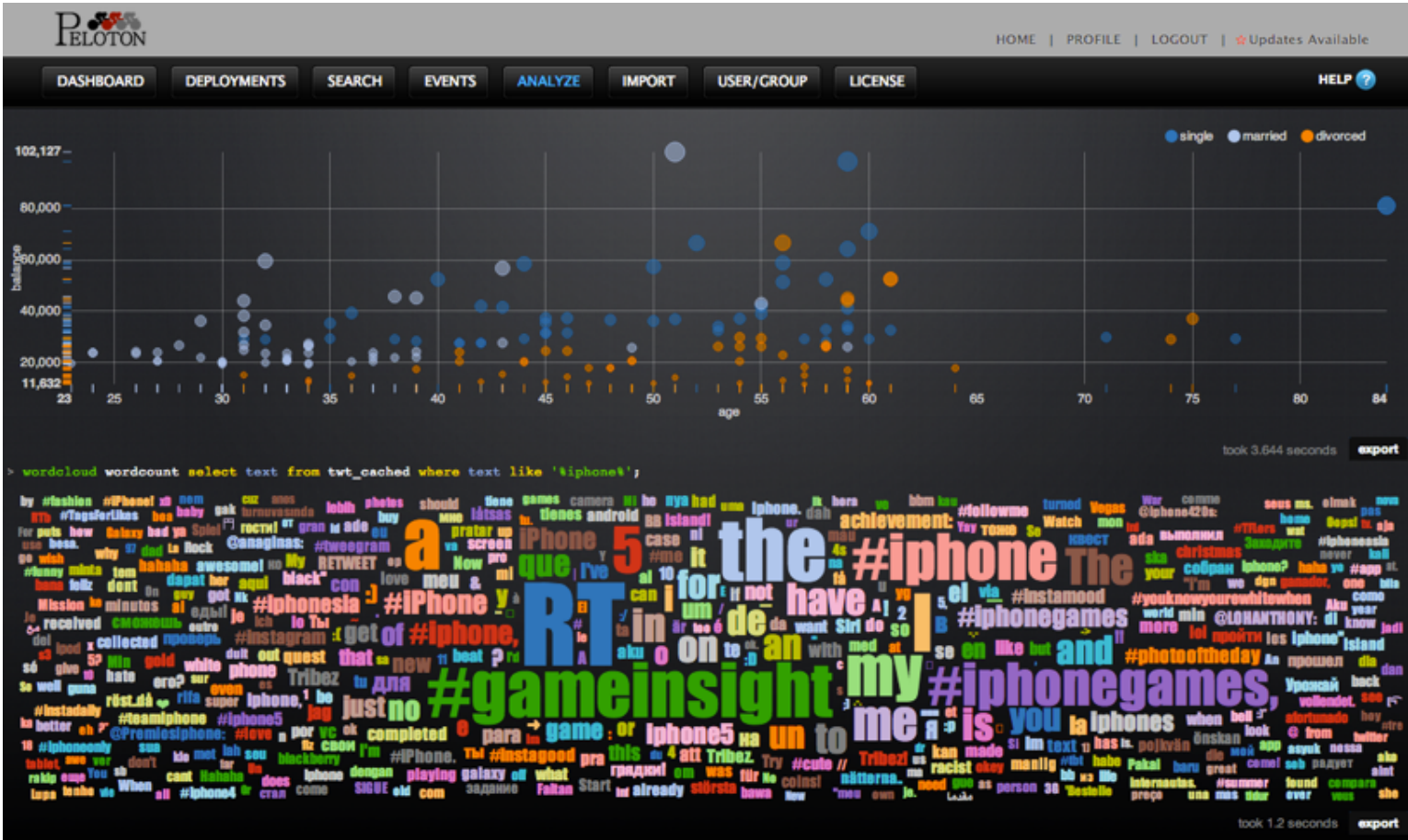
Analytical language
Many Libraries
Interactive
Visualization
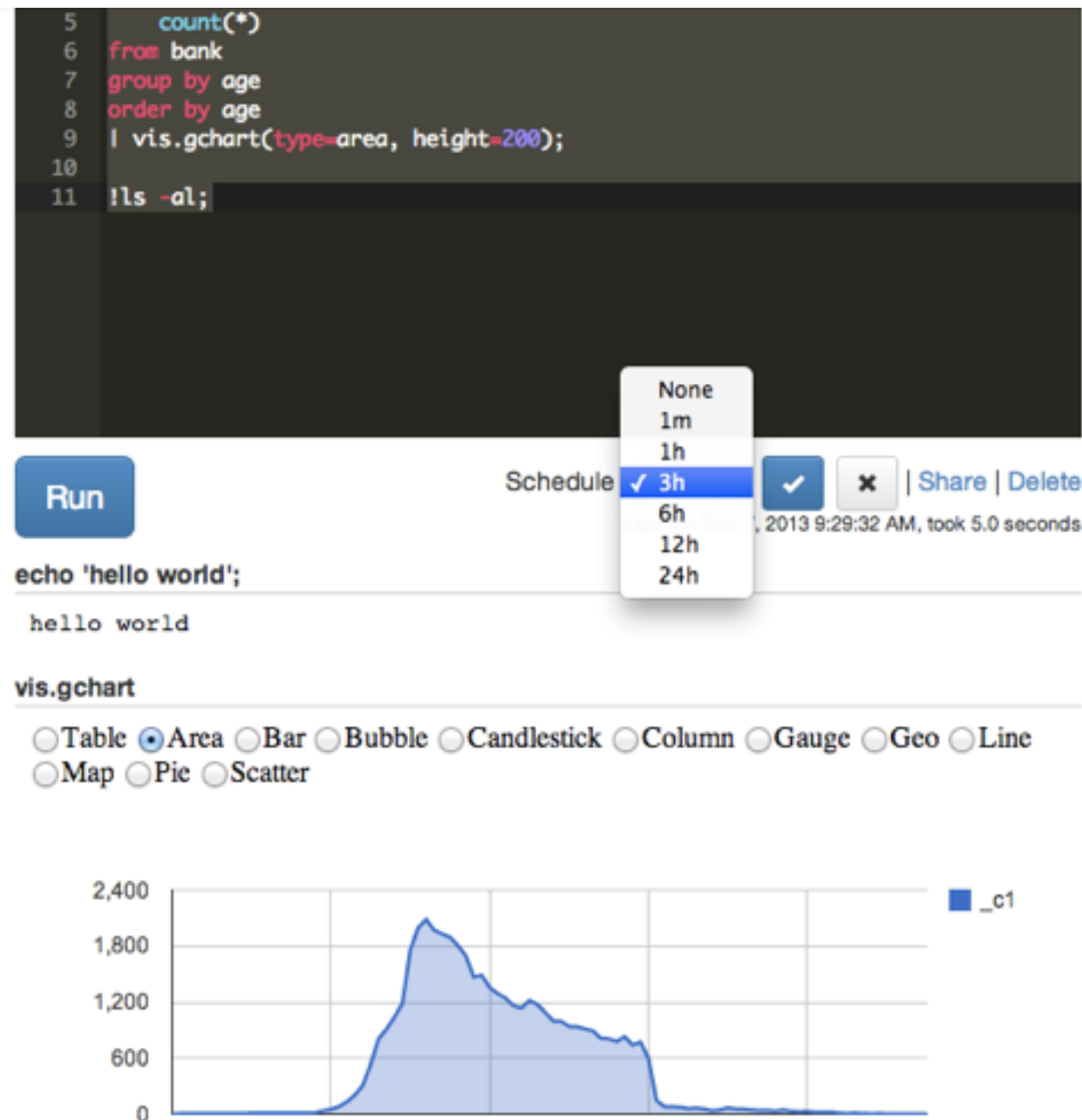Sharing

The First attempt 2012~2013

It's got graphic REPL, deployment, search, import tool

But failed, because

It wasn't widely used
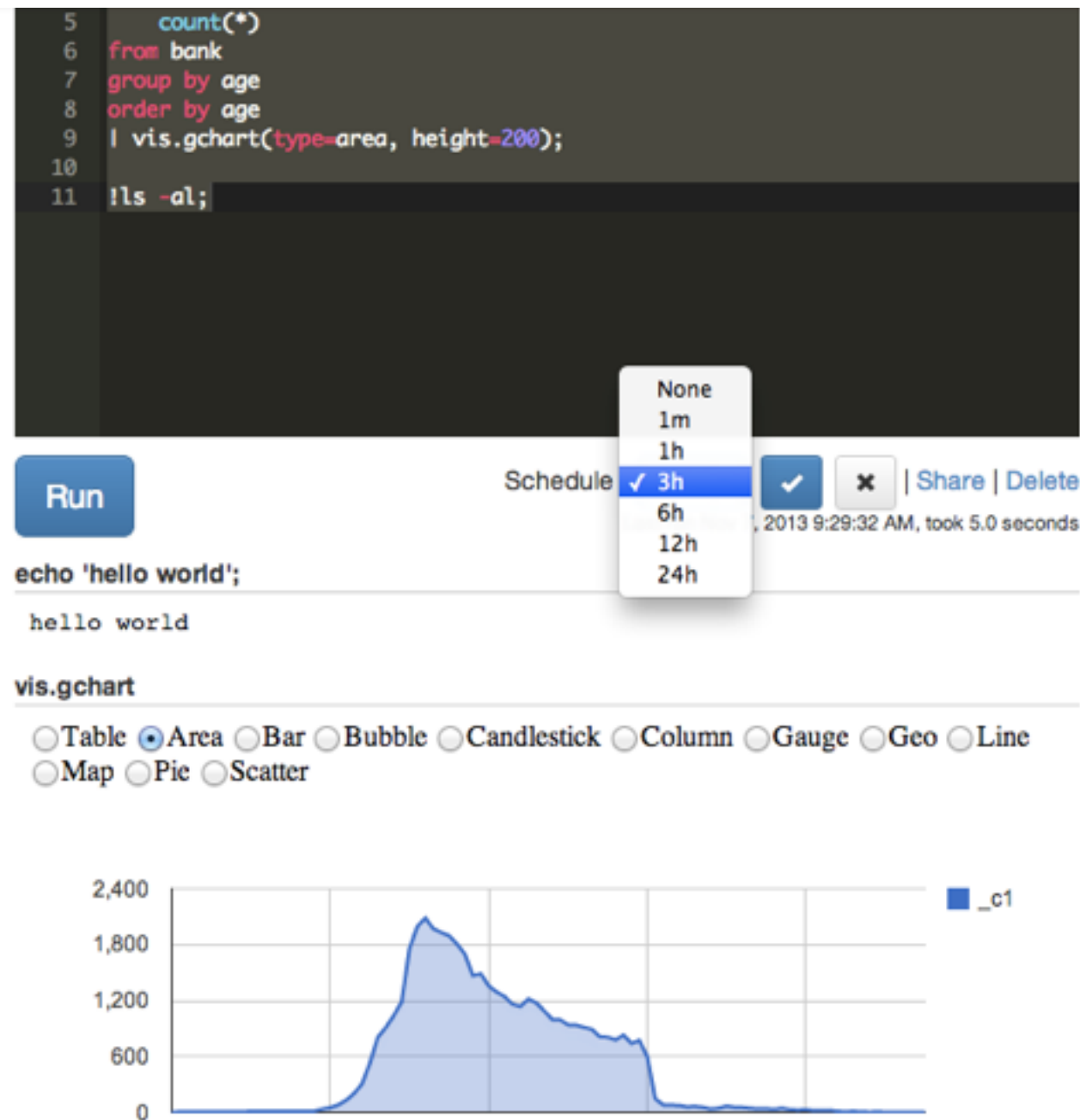It wasn't opensource

Second attempt 2013~2014

Opensourced graphic REPL from commercial product

The first version of Zeppelin
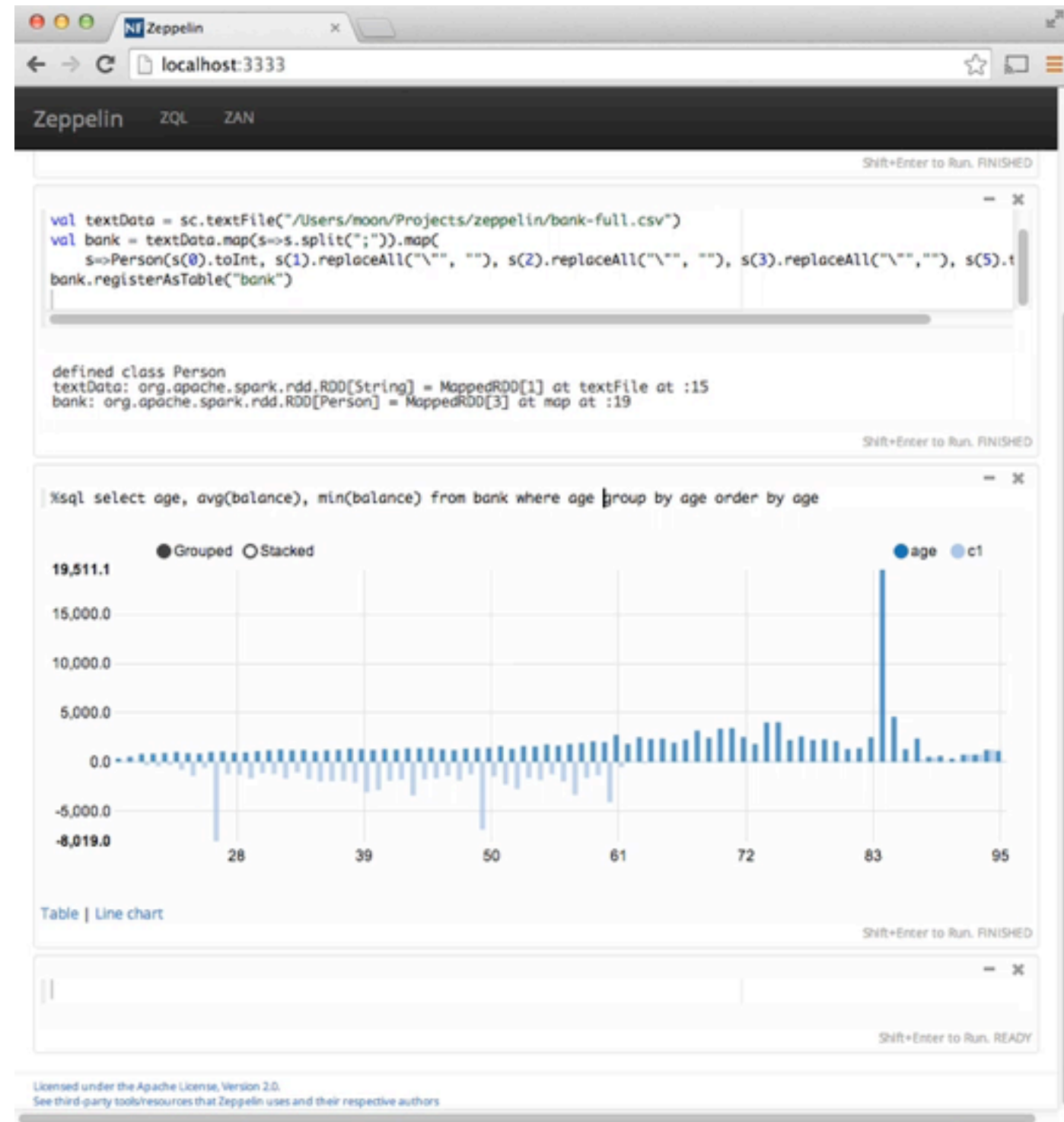
Second attempt 2013~2014

Not widely used

It was slow, difficult to use, …
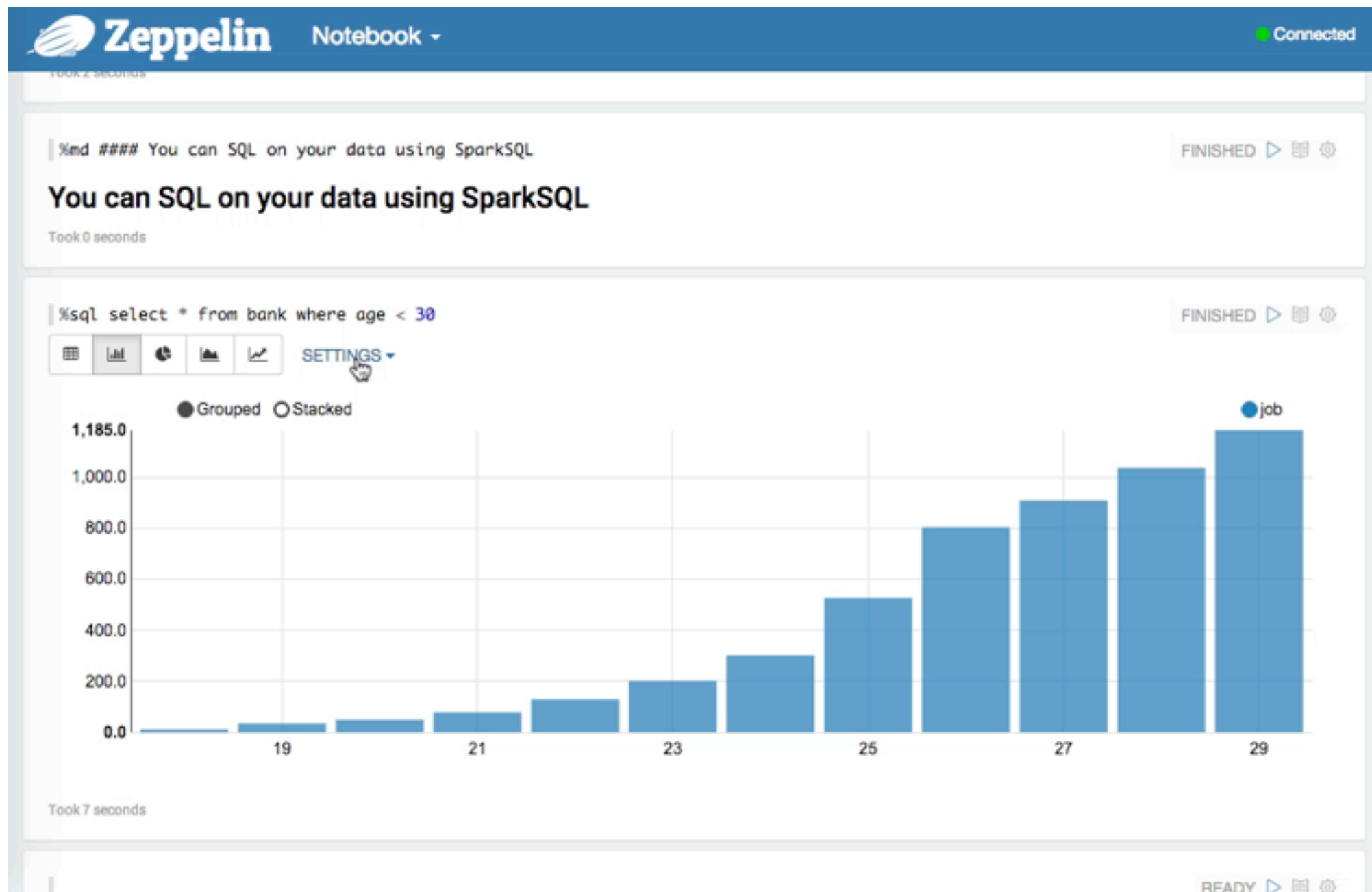
Third attempt 2014~

After few weeks of study, decided to rewrite Zeppelin

Graphic REPl -> Notebook with Apache Spark integration

Third attempt 2014~

Next week, beautifulized

Why do you like Zeppelin?

Web Based

web framework

Visualization
d3.js

Language

Package management
bower

Build

Notebook

Code

```
val a=1
```
FINISHED

Result

```
a: Int = 1
```
Took 24 seconds

Code

```
println(s"""%table
name\tage
sun\t60
moon\t50""")
```
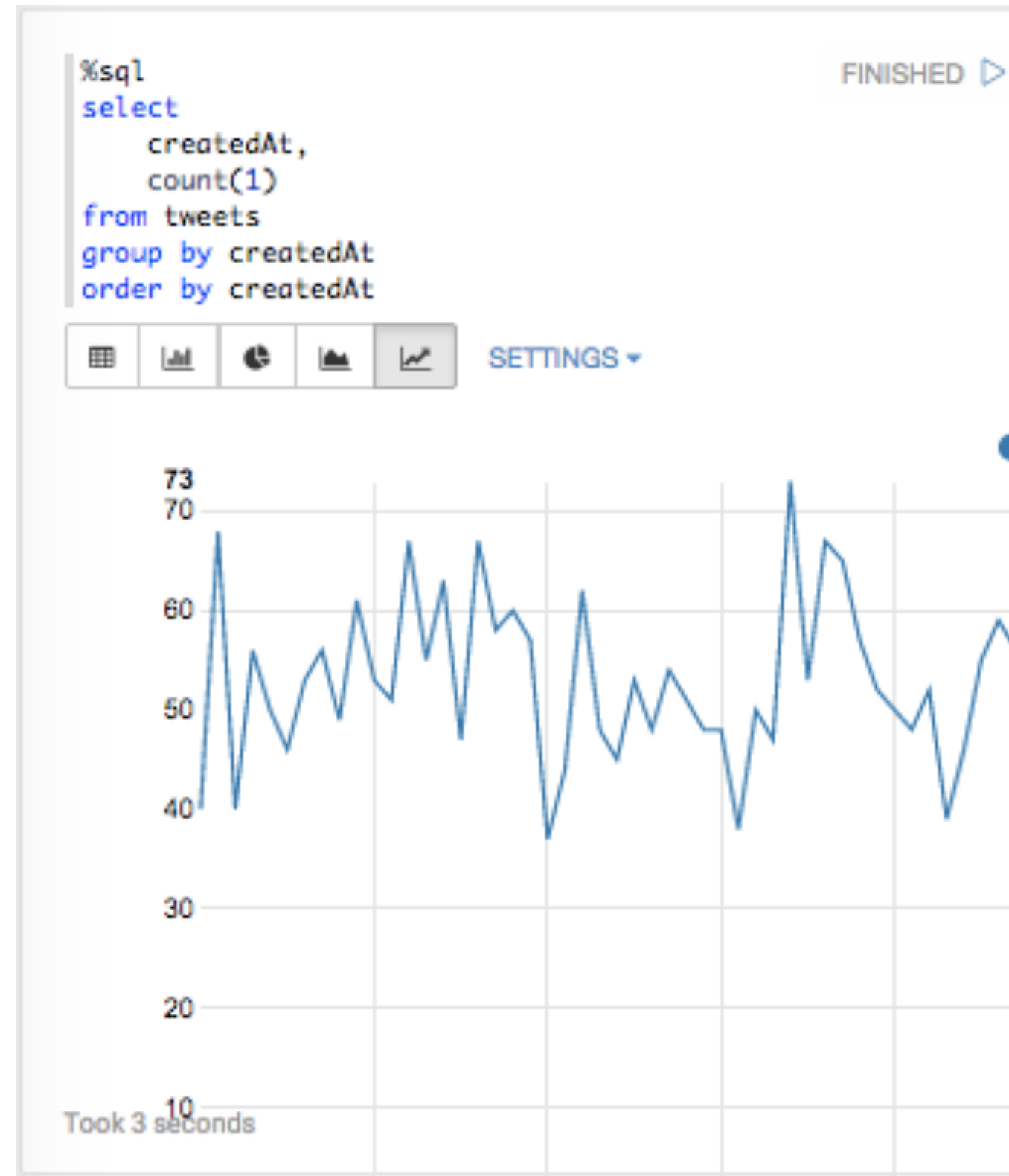FINISHED

Result

name	age
sun	60
moon	50

결과가 텝과 뉴라인으로 구분된 테이블 형식이면
자동으로 테이블로 포메팅

Data Visualize

Notebook

Pivot

Dynamic Form

Dynamic Form Creation


```
%sql
select age, count(1) value
from bank
where age < ${maxAge=30}
group by age
order by age
```
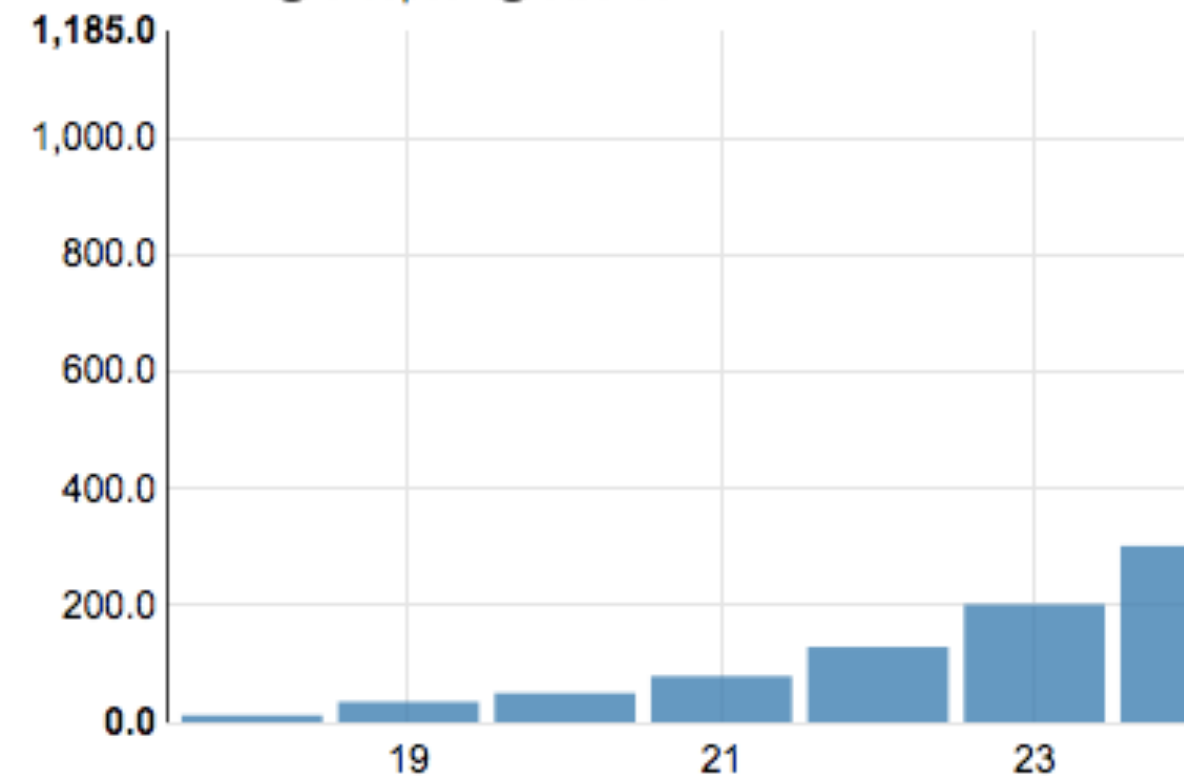
maxAge

```
30
```

Grouped Stacked

1,185.0

1,000.0

800.0

600.0

400.0

200.0

0.0

19   21   23

Sharing

Synchronized

%sql
select
    createdAt,
    count(1)
from tweets
group by createdAt
order by createdAt

FINISHED

SETTINGS ▾

73
70

c1

73
70

60

60

50

50

40

40

30

30

20

20

10
Took 3 seconds

10
Took 3 seconds
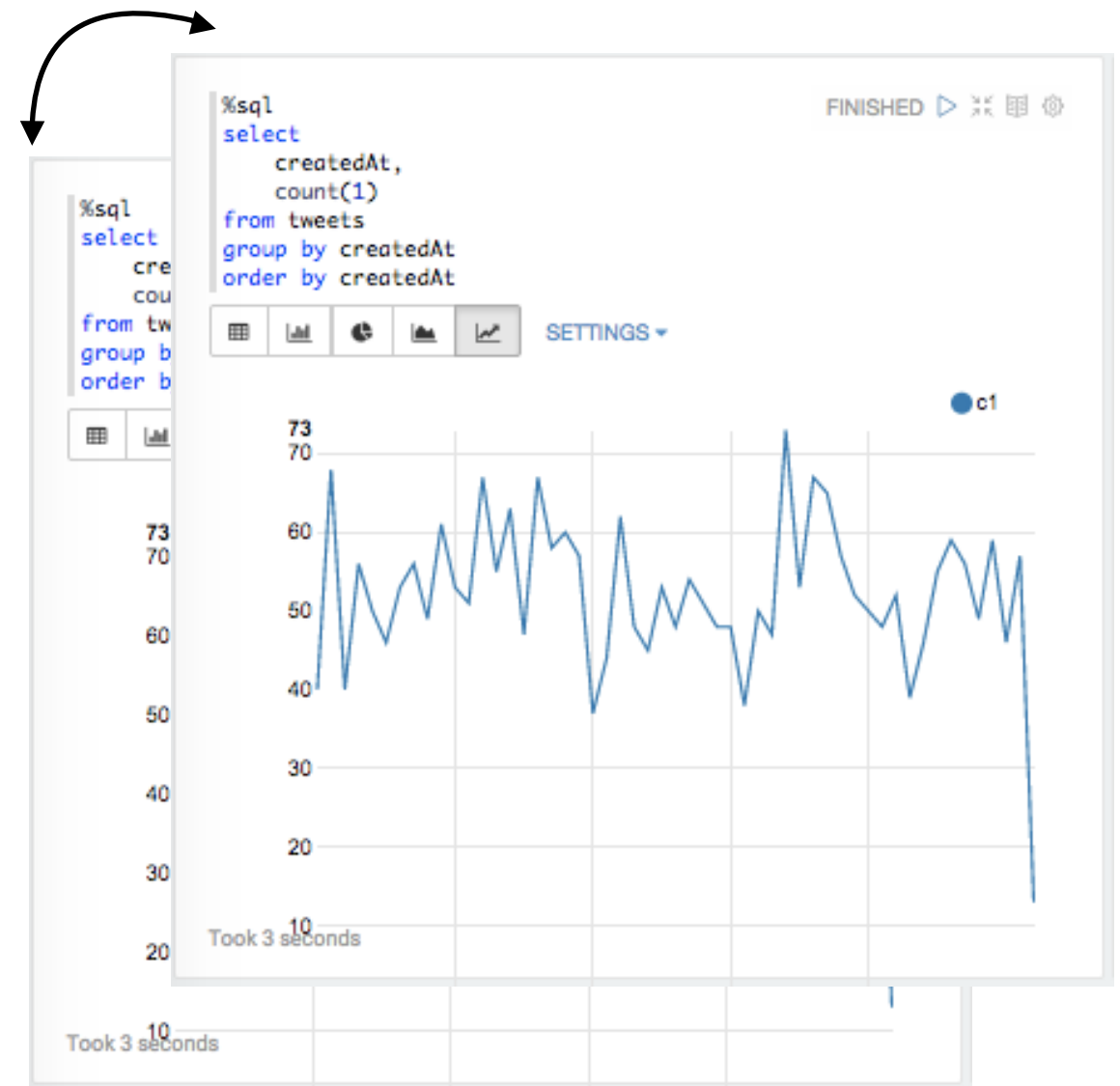
REST

Web socket

Zeppelin simplifies data analysis

Why do your project likes Zeppelin?
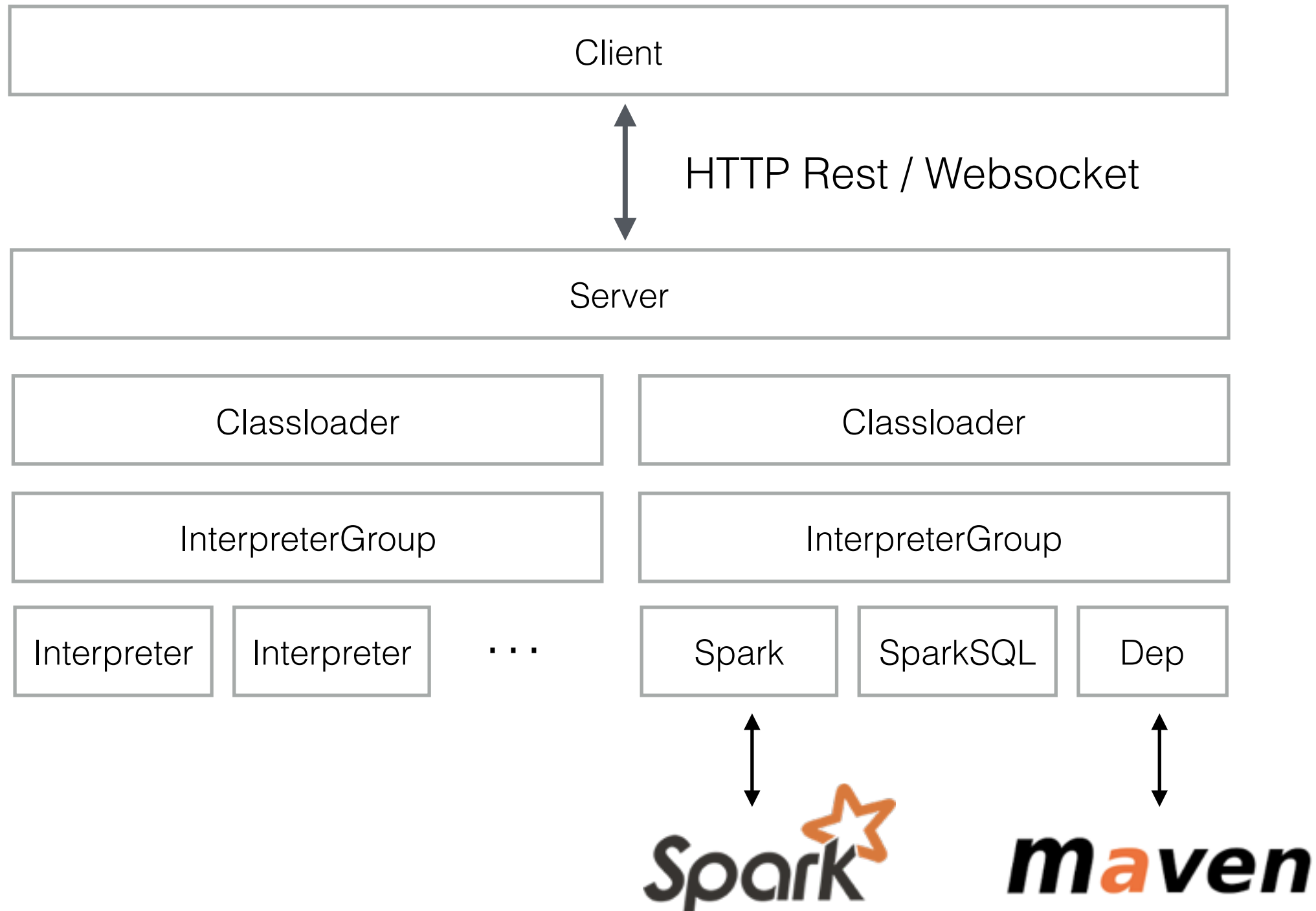
Easy to extend

Interpreters

Spark
PySpark
SparkSQL
Hive
Mysql (JDBC)
Markdown
Shell

Zeppelin Interpreter Architecture

Zeppelin Interpreter Architecture

Client

↕ HTTP Rest / Websocket

Server

↕ Thrift

Classloader

InterpreterGroup

| Interpreter | Interpreter | · · · |

InterpreterGroup

| Interpreter | Interpreter | · · · |

Seprate JVM process
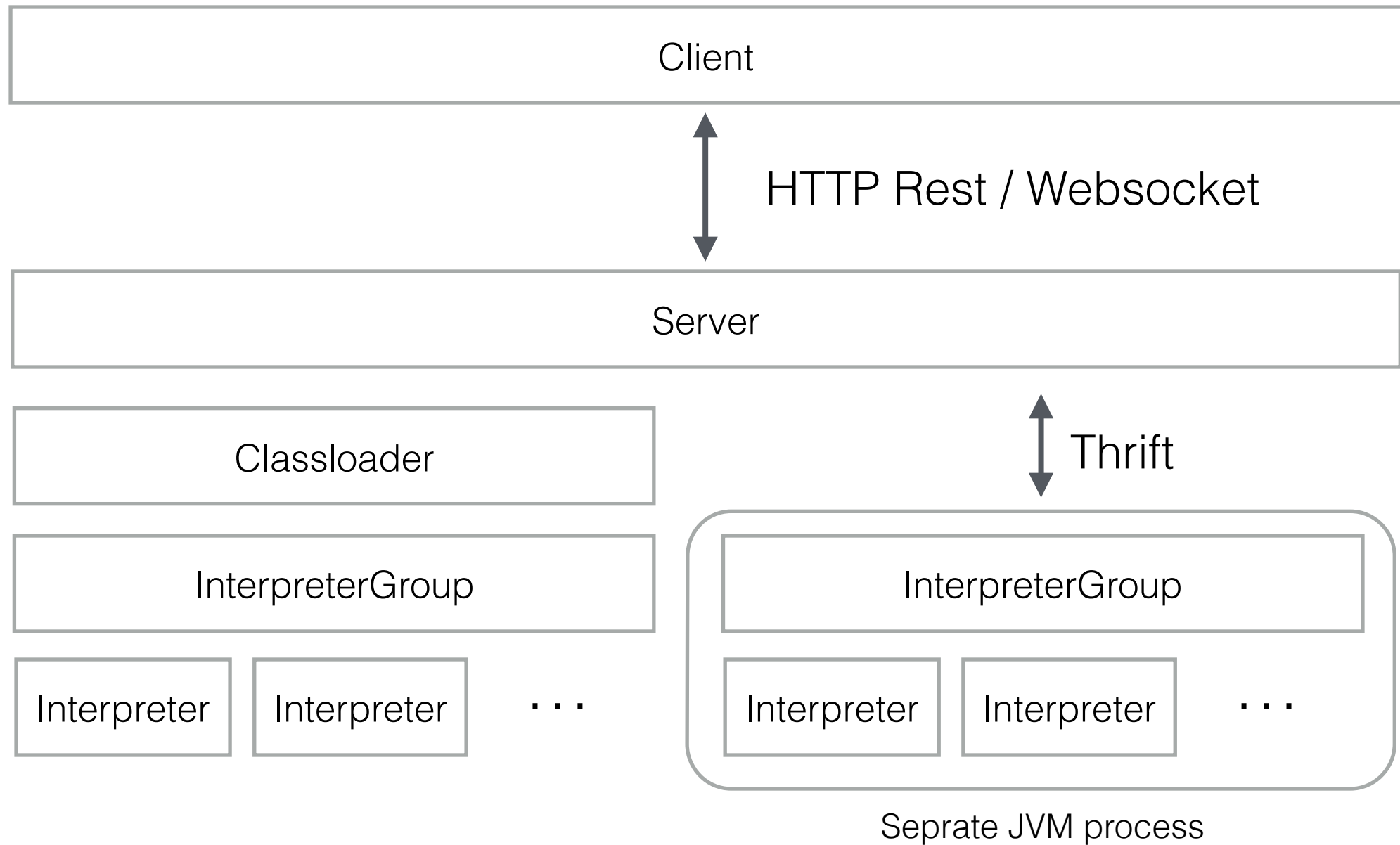
Implementing new Interpreter

Must have

```java
public abstract void open();

public abstract void close();

public abstract InterpreterResult interpret(String st, InterpreterContext context);
```

Good to have

```java
public abstract void cancel(InterpreterContext context);

public abstract int getProgress(InterpreterContext context);

public abstract List<String> completion(String buf, int cursor);
```

More controls

```java
public abstract FormType getFormType();

public Scheduler getScheduler();
```

Roadmap

- Integration with more distributed processing framework

  - Flink, Ignite, Tajo, etc..

- Output message streaming
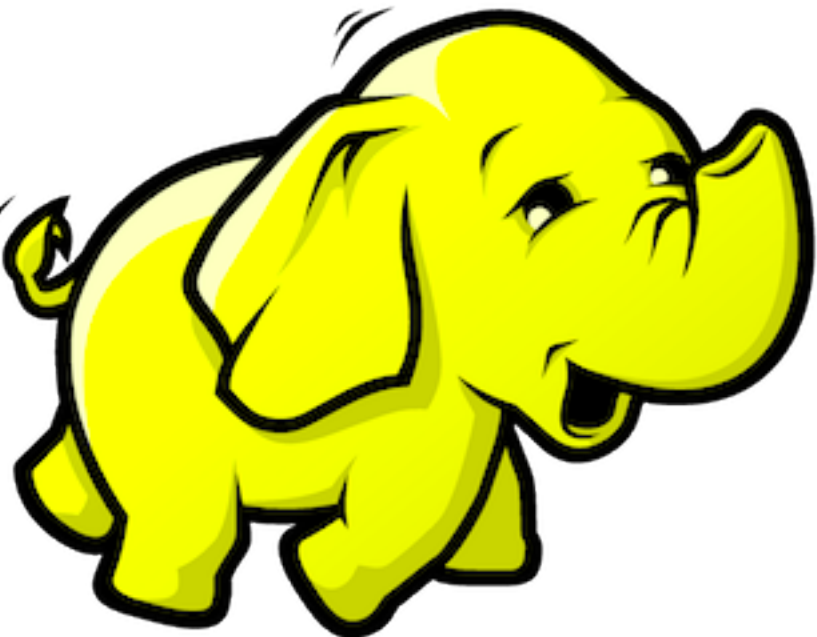
- Ability to create rich GUI

MLLib

Cloudera-ML mahout

Zeppelin

OOZiE

Cassandra H·BASE

accumulo

MRQL   Drill   Impala

Pig   Hive   tajo

Flume

Sqoop

Spark   Map Reduce

HDFS

QnA

Thanks

Lee moon soo
moon@apache.org
moon@nflabs.com