& Enterprise

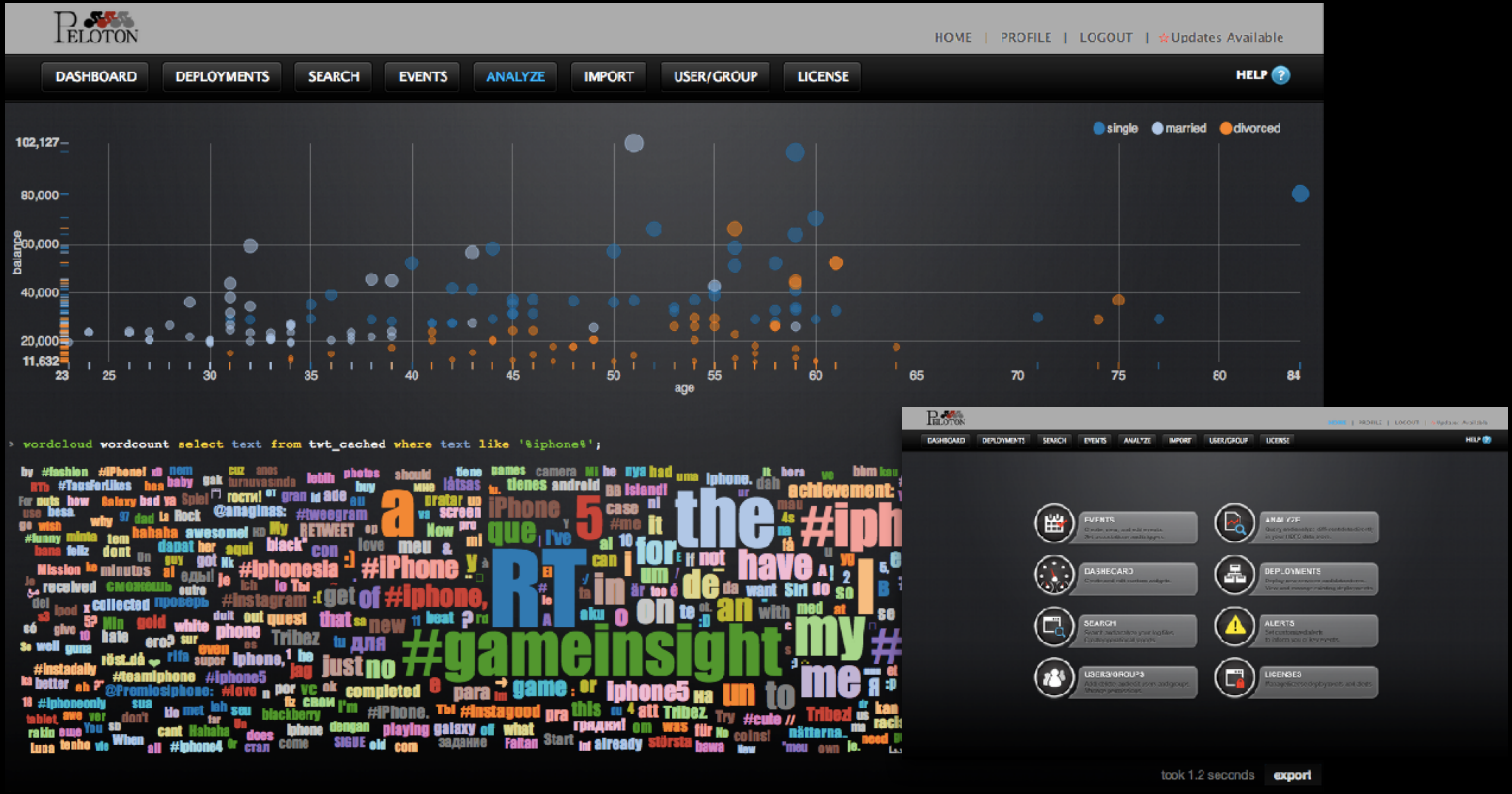
Jongyoul Lee

PMC of Apache Zeppelin

Software Development Engineer at NFLabs

Zeppelin

2012. 12   Data analytics solution based on AMP Lab Spark/Shark

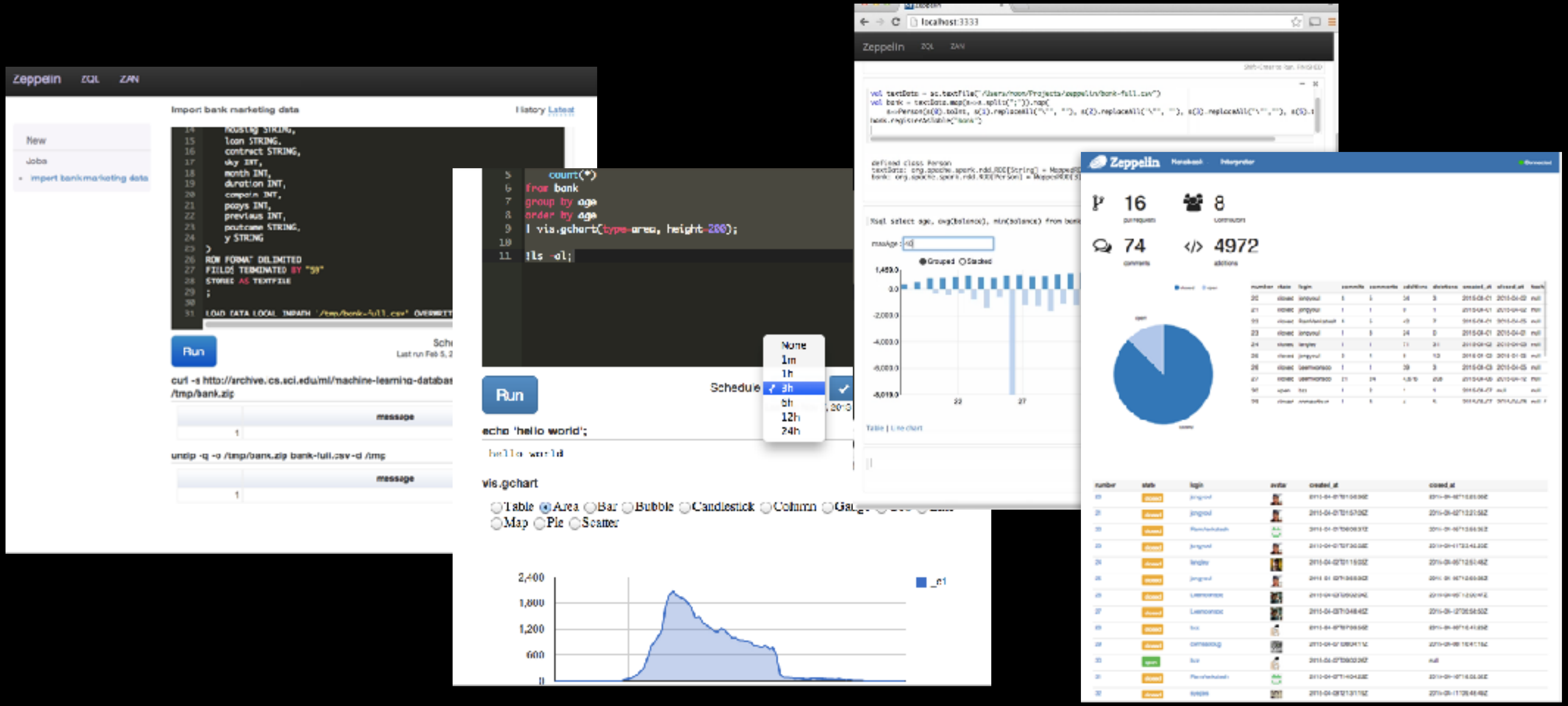


```
> wordcloud wordcount select text from twt_cached where text like '%iphone%';
```

Zeppelin

2012. 12　Data analytics solution based on AMP Lab Spark/Shark

2013. 10　Opensource interactive analytics feature as ‘Zeppelin’



Import bank marketing data

```
    housing STRING,
    loan STRING,
    contract STRING,
    day INT,
    month INT,
    duration INT,
    campaign INT,
    pdays INT,
    previous INT,
    poutcome STRING,
    y STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ";"
STORED AS TEXTFILE
;

LOAD DATA LOCAL INPATH '/tmp/bank-full.csv' OVERWRITE
```

curl -s http://archive.ics.uci.edu/ml/machine-learning-databas /tmp/bank.zip

	message
1	

unzip -q -o /tmp/bank.zip bank-full.csv -d /tmp

	message
1	

```
    count(*)
from bank
group by age
order by age
| vis.gchart(type=area, height=200);

!ls -al;
```

echo 'hello world';

```
hello world
```

```
val textData = sc.textFile("/Users/moon/Projects/zeppelin/bank-full.csv")
val bank = textData.map(s=>s.split(";")).map(
    s=>Person(s(0).toInt, s(1).replaceAll("\"", ""), s(2).replaceAll("\"", ""), s(3).replaceAll("\"",""), s(5).t
bank.registerAsTable("bank")
```

```
defined class Person
textData: org.apache.spark.rdd.RDD[String] = MappedRDD
bank: org.apache.spark.rdd.RDD[Person] = MappedRDD[3]
```

%sql select age, avg(balance), min(balance) from bank

2013. 10

2014. 08

Zeppelin

2012. 12	Data analytics solution based on AMP Lab Spark/Shark
2013. 10	Opensource interactive analytics feature as ‘Zeppelin’
2014. 12	ASF incubation

Incubation Status http://incubator.apache.org/projects/zeppelin.html

Zeppelin

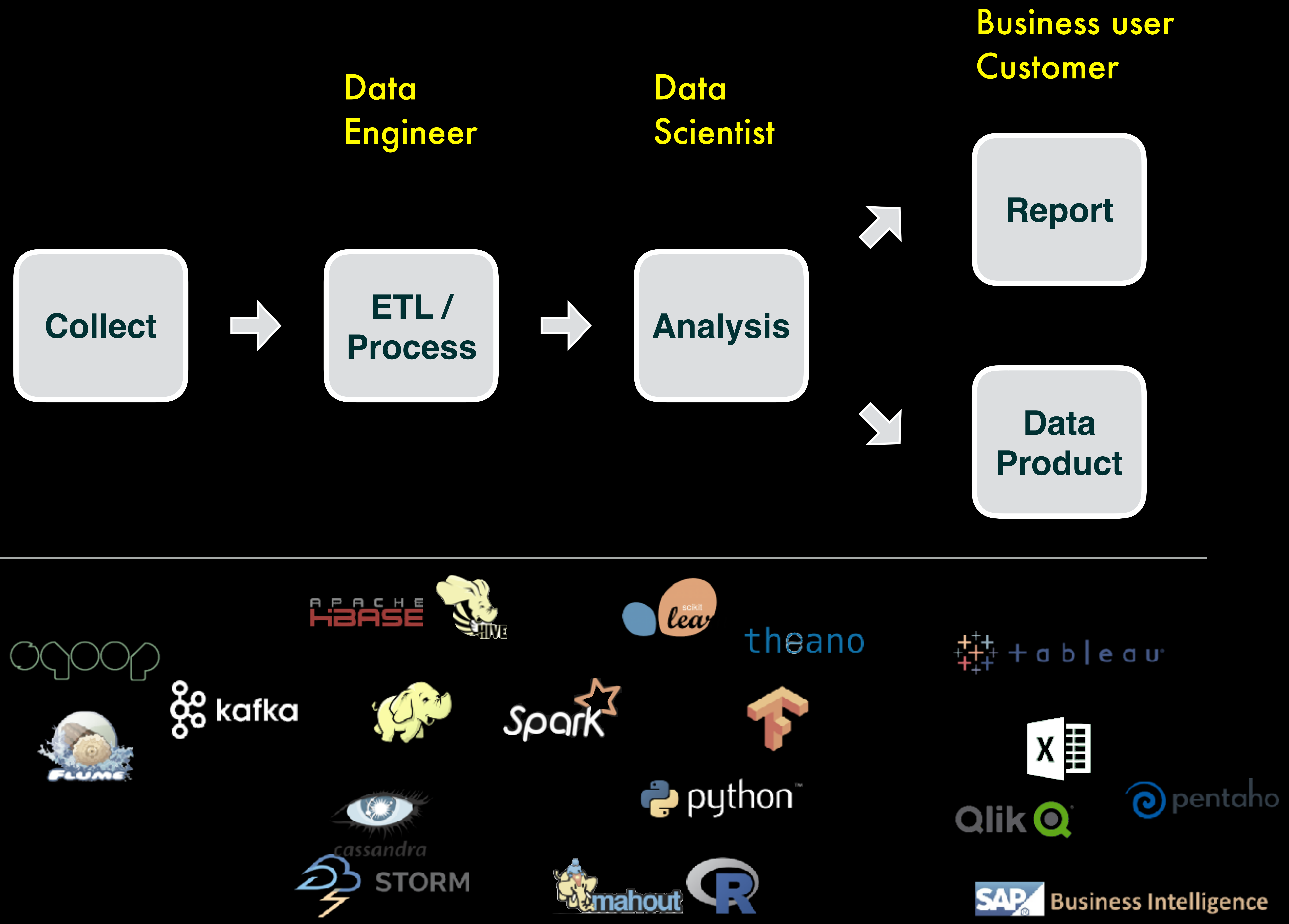
2012. 12	Data analytics solution based on AMP Lab Spark/Shark
2013. 10	Opensource interactive analytics feature as ‘Zeppelin’
2014. 12	ASF incubation

2016. 10

157 Contributors world wide
2071 Stars on github repo
6 Releases

One of the most popular project in ASF

Life cycle of big data



Business user
Customer

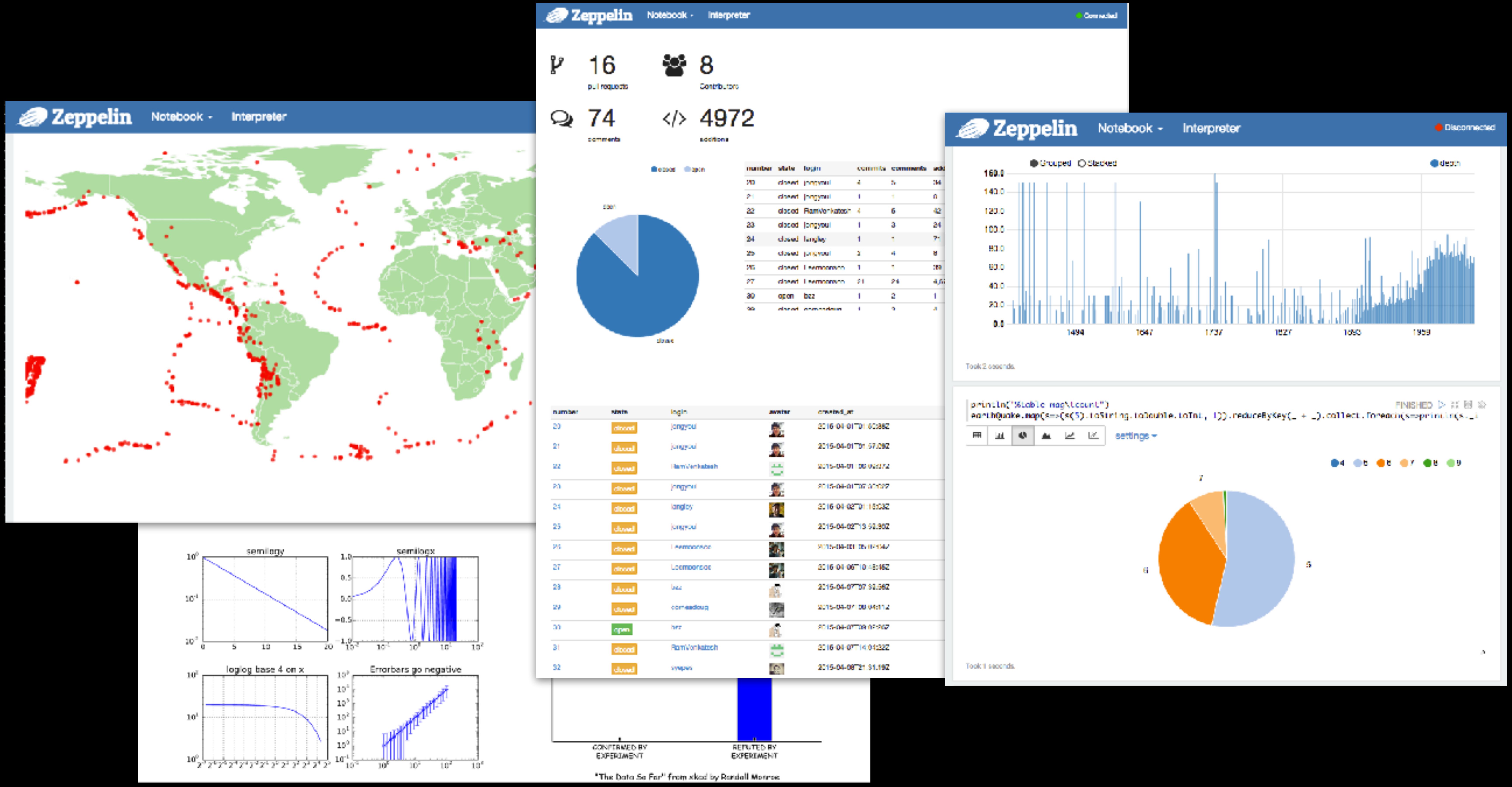
Data
Engineer

Data
Scientist

Collect → ETL / Process → Analysis → Report
Analysis → Data Product

Zeppelin

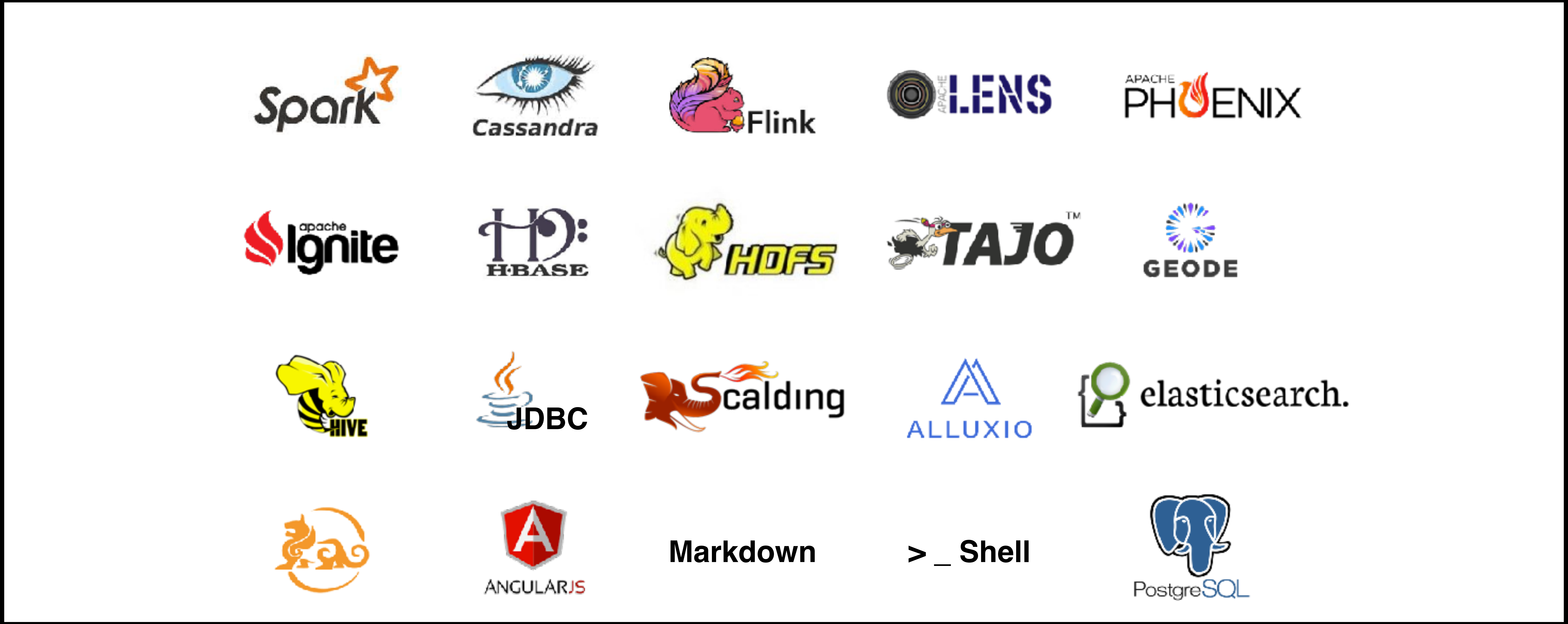
A web-based notebook that enables interactive data analytics. You can make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.



Zeppelin

Interpreter : pluggable layer for language / processing backend integration

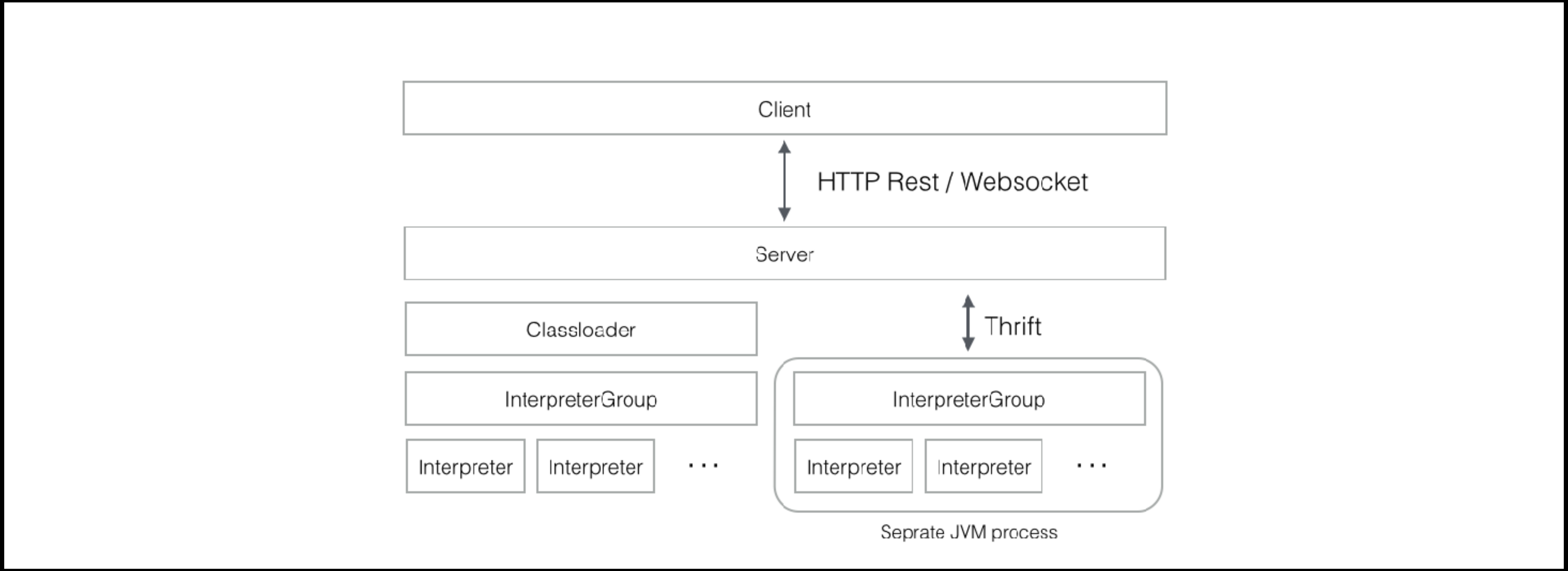
20+ interpreters are supported officially



2016. 03. Interpreters in Zeppelin source tree. Does not include 3rd party interpreters

Zeppelin

Interpreter : pluggable layer for language / processing backend integration



Zeppelin

Interpreter : Easy to extend

```
               public abstract class Interpreter {

                 public void open();
Must have      {  public void close();
                 public InterpreterResult interpret(String st, InterpreterContext context);

                 public void cancel(InterpreterContext context);
Good to have   {  public int getProgress(InterpreterContext context);
                 public List<String> completion(String buf, int cursor);

Advanced       {  public FormType getFormType();
                 public Scheduler getScheduler();

               }
```

Zeppelin

Notebook Repo : pluggable layer for notebook persistence

5+ Notebook repos are supported officially

commons VFS™

git

S3

Microsoft Azure

ZeppelinHub

2016. 03. Notebook repos in Zeppelin source tree. Does not include 3rd party interpreters

Zeppelin

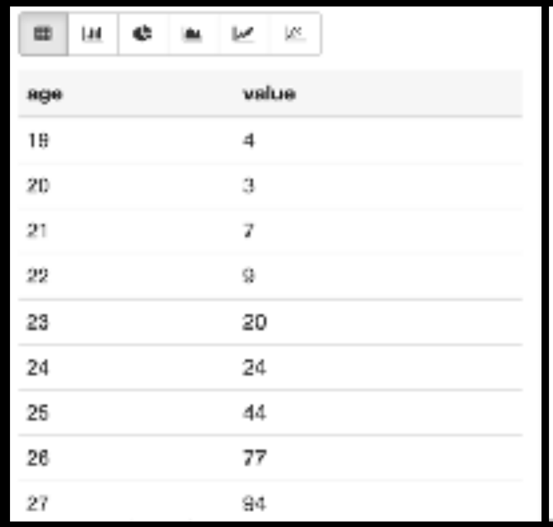
Notebook Repo : Easy to extend

```java
public interface NotebookRepo {

  public List<NoteInfo> list() throws IOException;
  public Note get(String noteId) throws IOException;
  public void save(Note note) throws IOException;
  public void remove(String noteId) throws IOException;
  public void checkpoint(String noteId, String checkPointName) throws IOException;
  public void close();

}
```

Zeppelin

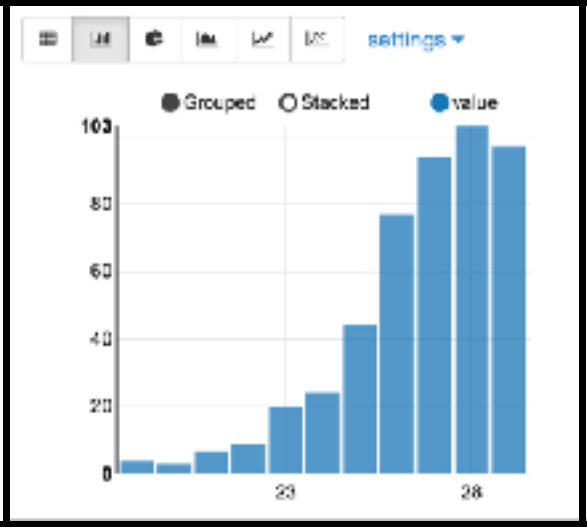
Visualizations : 6 Built-in visualizations comes with pivot

Table

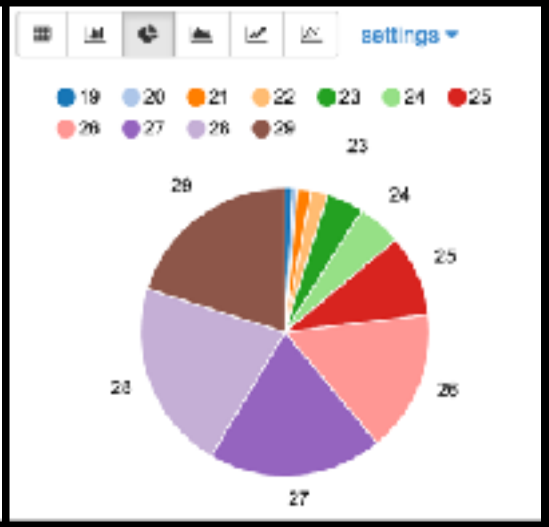


age	value
19	4
20	3
21	7
22	9
23	20
24	24
25	44
26	77
27	84

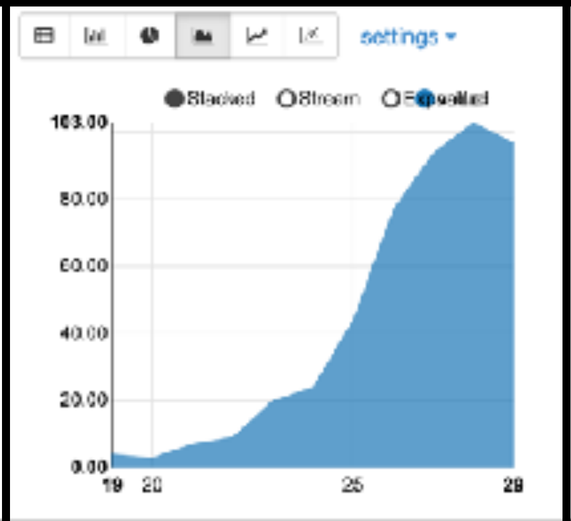
Bar



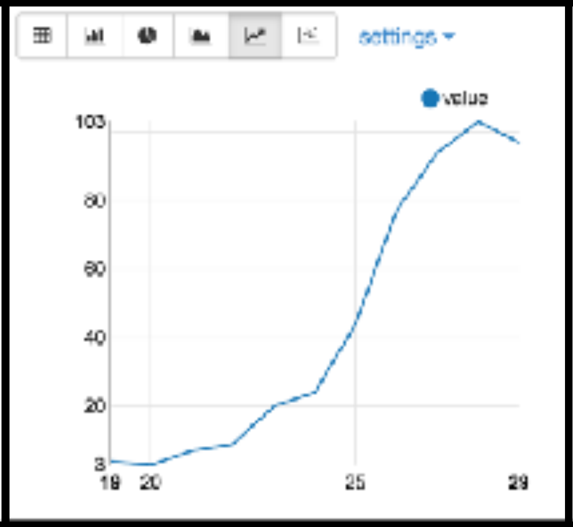
Pie



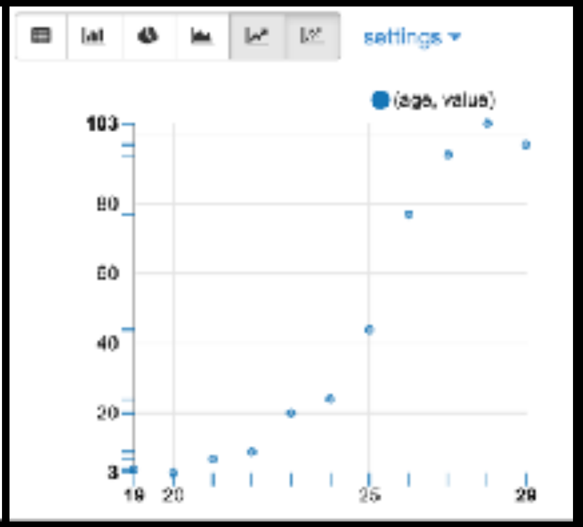
Area



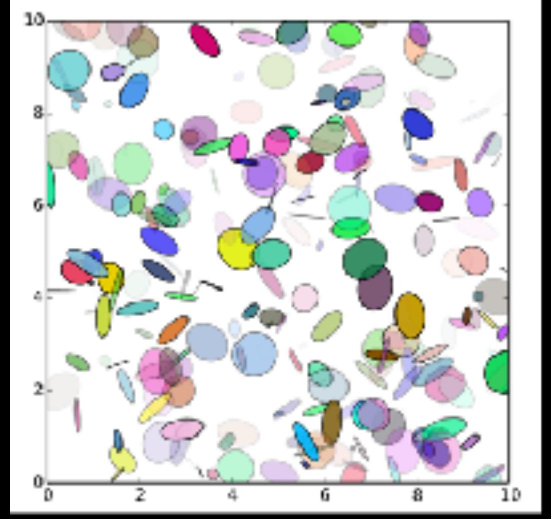
Line

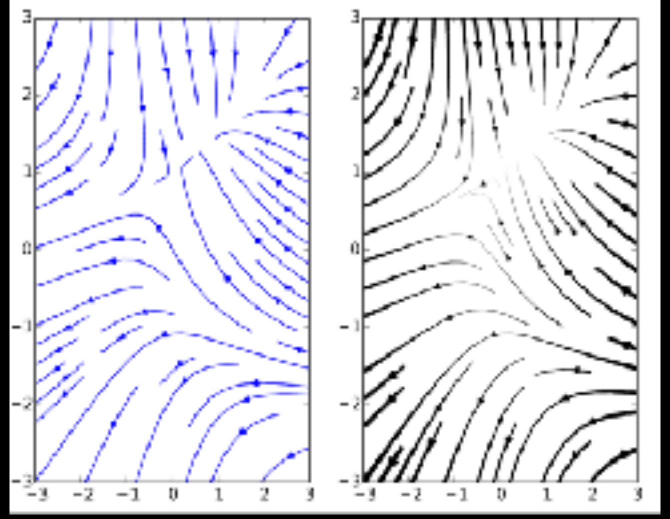


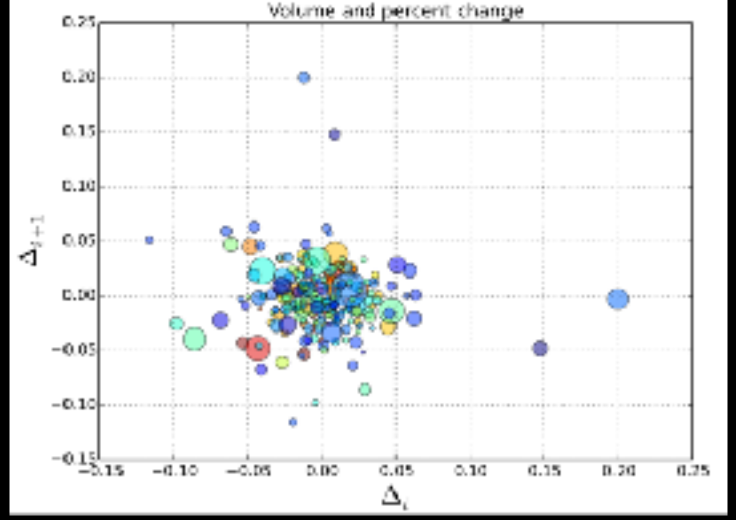
Scatter

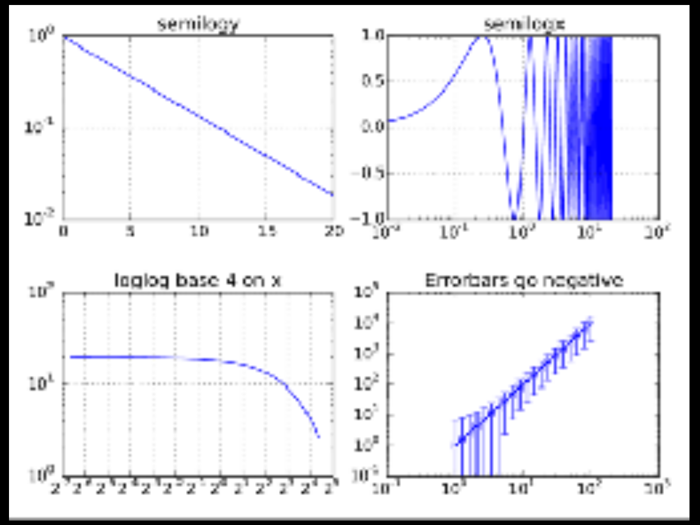


Free to draw any customized visualizations inside of notebook









...

2 Helium

Makes Zeppelin fly!

Platform for data analytics application that makes visualization pluggable and more.

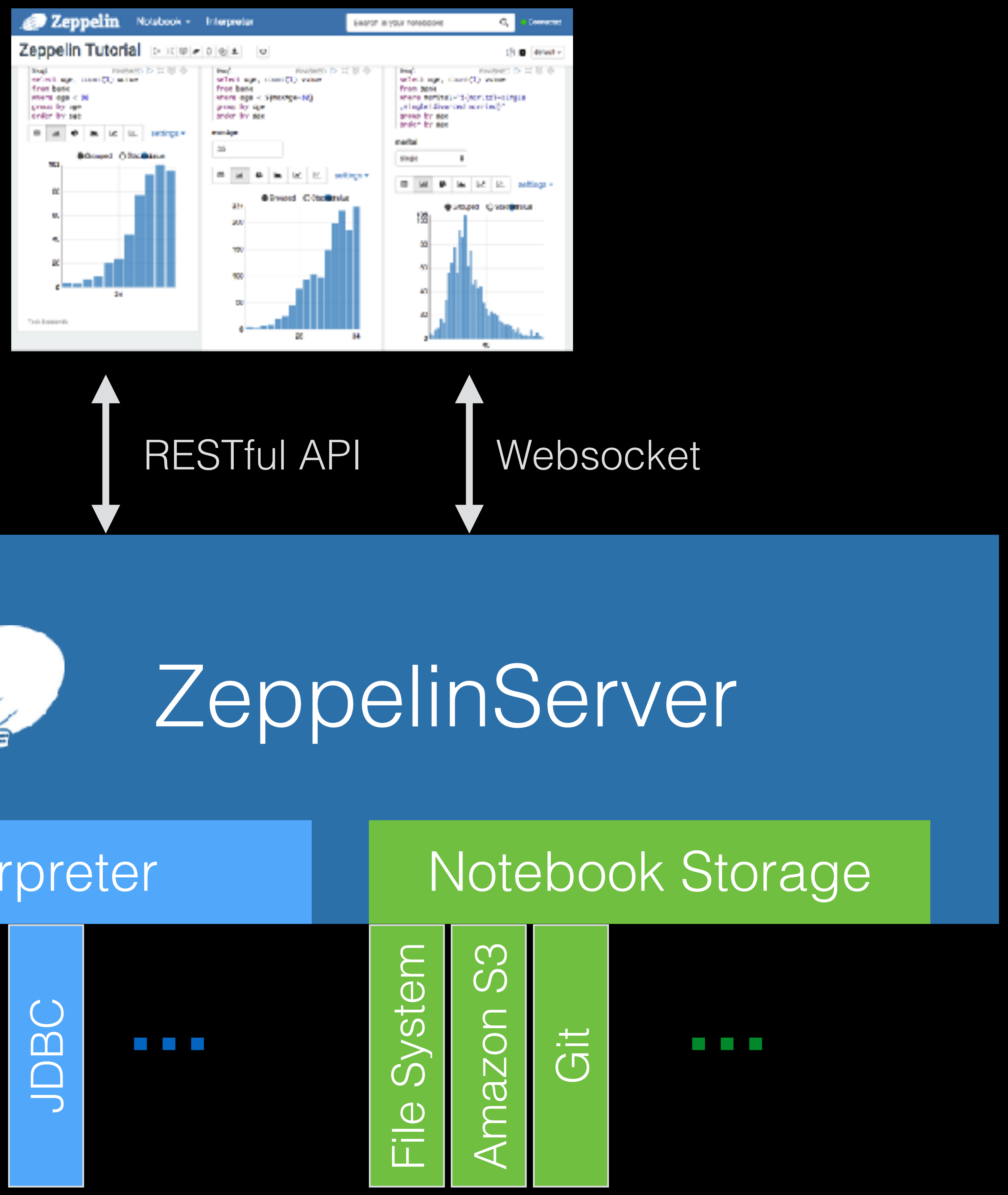
Proposal

https://cwiki.apache.org/confluence/display/ZEPPELIN/Helium+proposal

Umbrella issue

http://issues.apache.org/jira/browse/ZEPPELIN-533

He^{2}lium



RESTful API

Websocket

ZeppelinServer

Interpreter

Spark

Flink

Geode

JDBC

...

Notebook Storage

File System

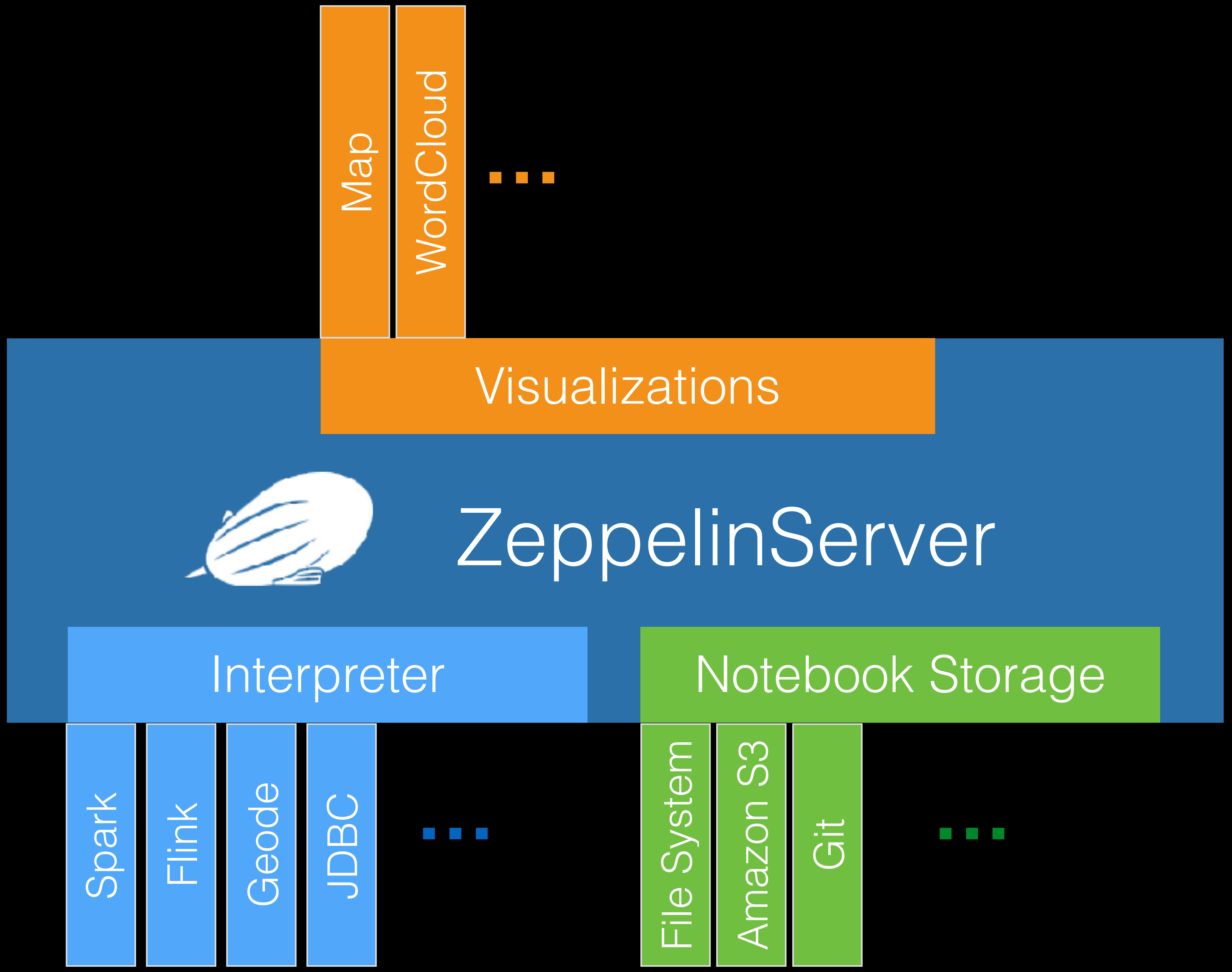
Amazon S3

Git

...

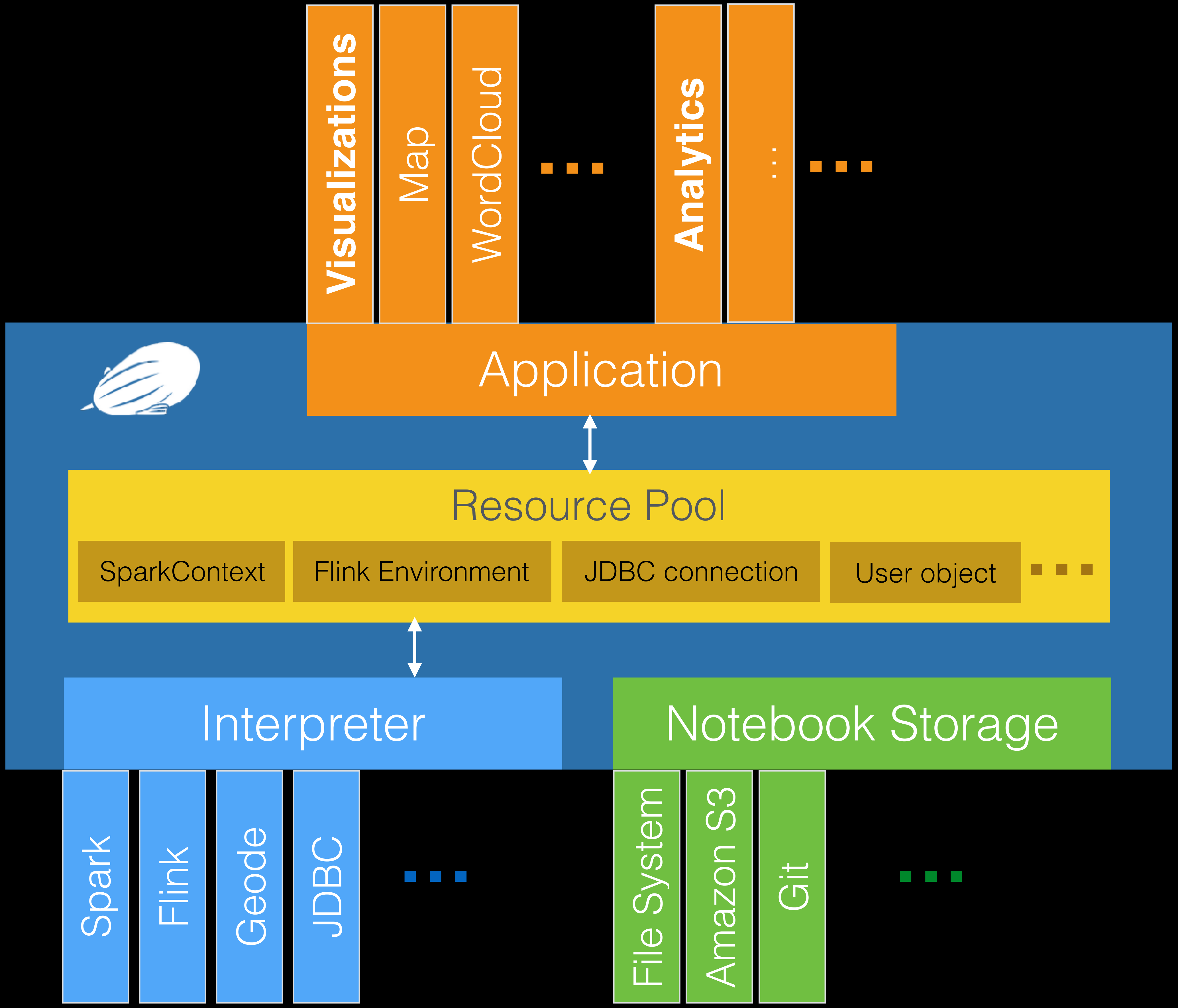
Interpreters and Notebook storage are pluggable

He^2 lium



We want visualization be pluggable

He lium



Extend pluggable visualization to pluggable analytics application

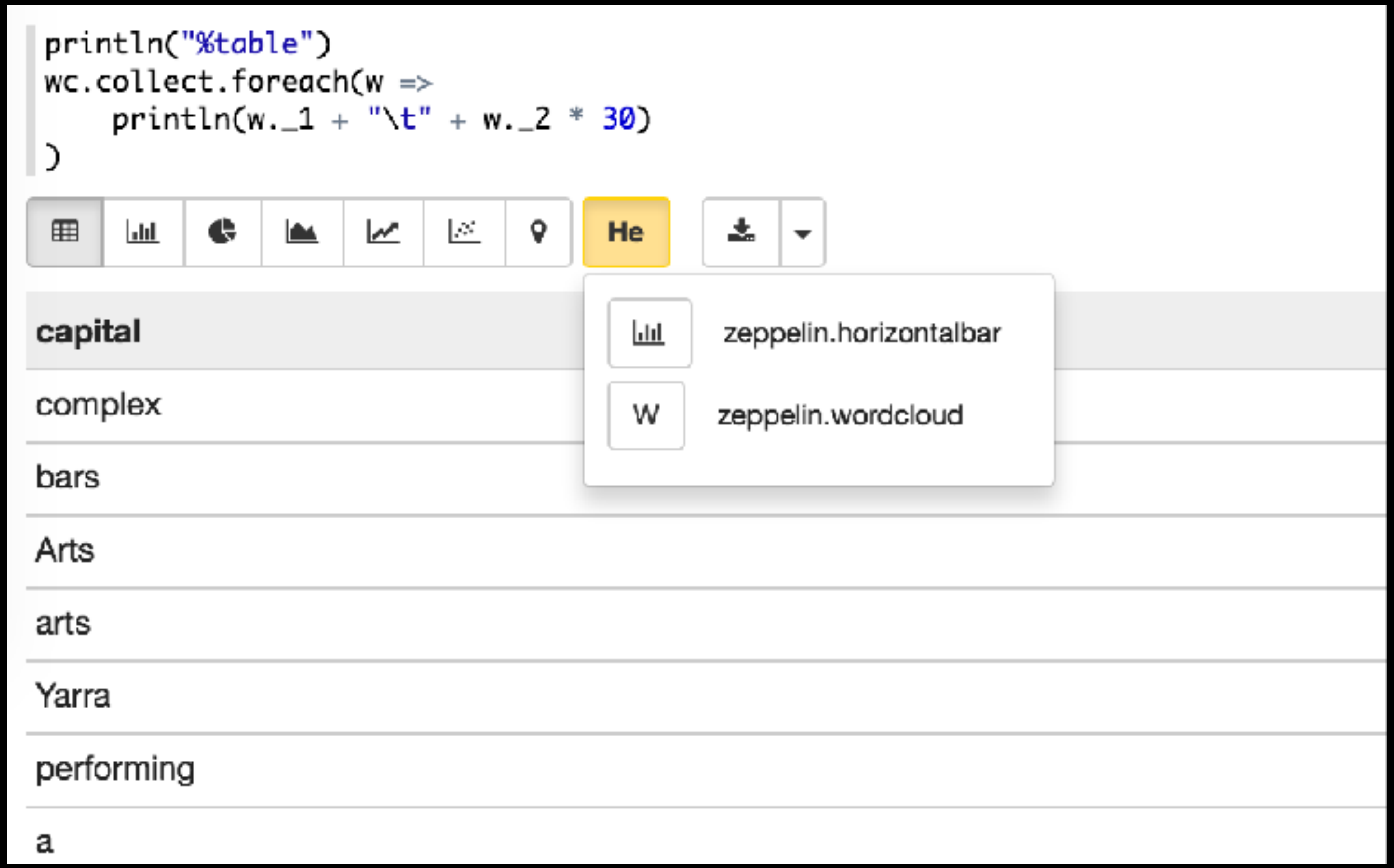
Helium

Helium Application: Easy to extend

```
public abstract class Application {

  public Application(ApplicationContext context);

  public abstract void run(ResourceSet args);

  public abstract void unload();

}
```

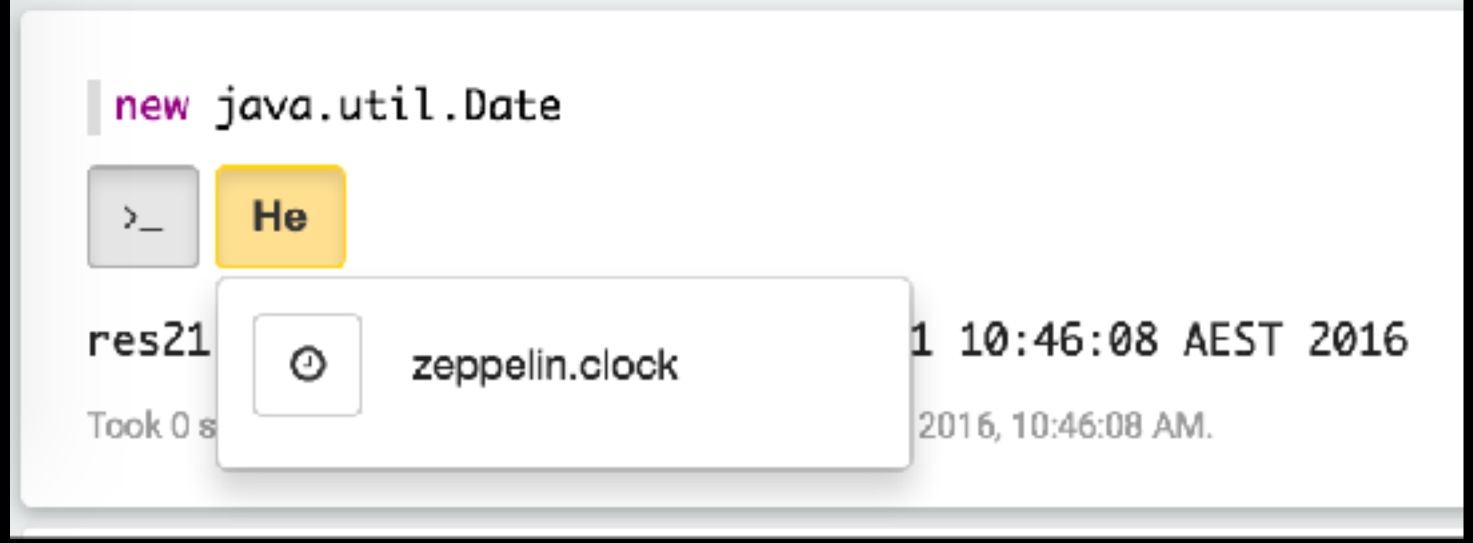
He^2 lium

Launcher: Suggest application according to data type in ResourcePool



```
println("%table")
wc.collect.foreach(w =>
    println(w._1 + "\t" + w._2 * 30)
)
```

capital
complex
bars
Arts
arts
Yarra
performing
a



```
new java.util.Date
```

```
res21                                    1 10:46:08 AEST 2016
Took 0 s                                 2016, 10:46:08 AM.
```

& Enterprise

More than 1000 employers use Apache Zeppelin

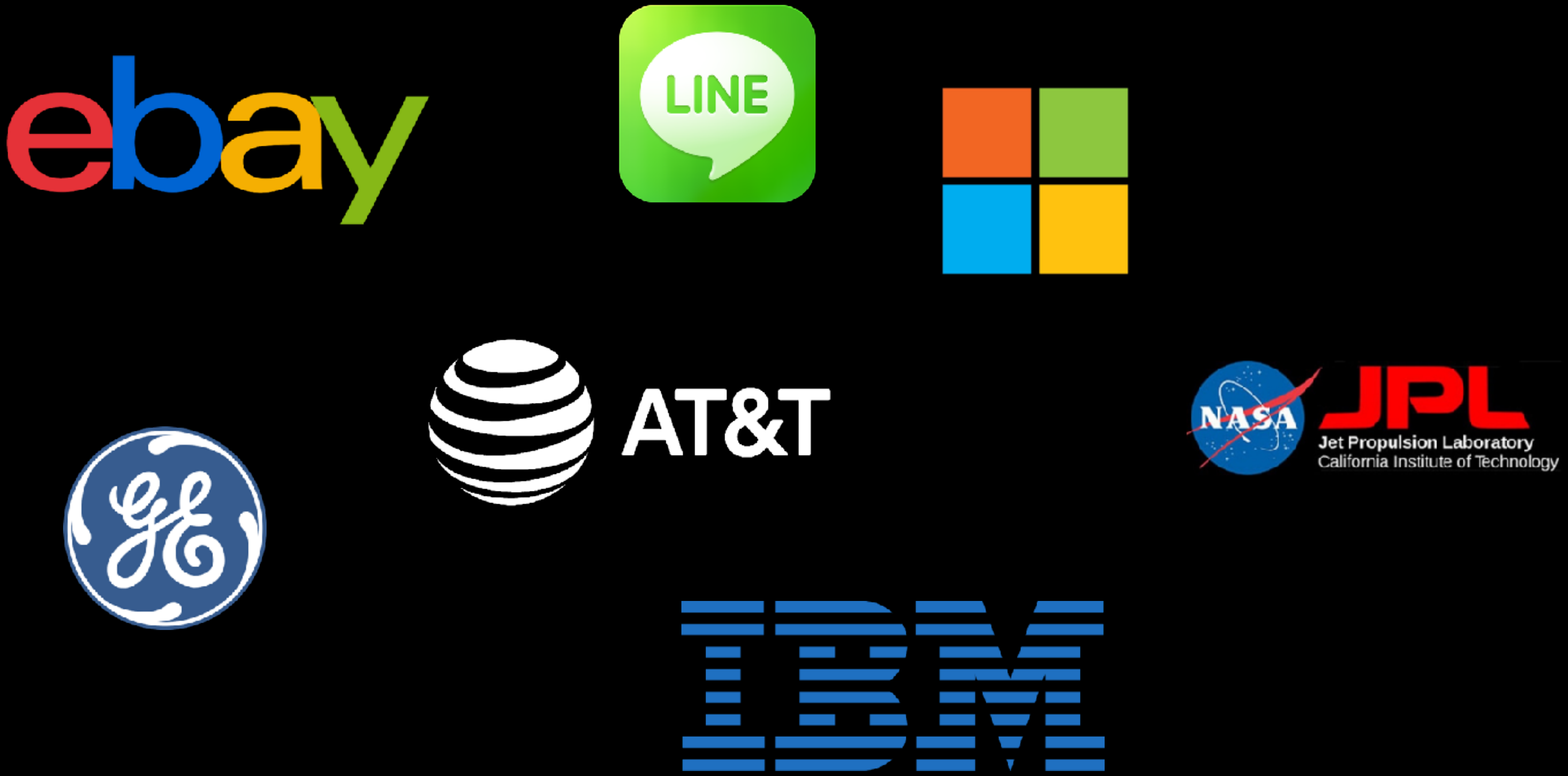
NETFLIX

Supports Apache Zeppelin as an internal service
Recommendation team uses Apache Zeppelin

CERN

Monitors their infrastructures via Apache Zeppelin

& Enterprise

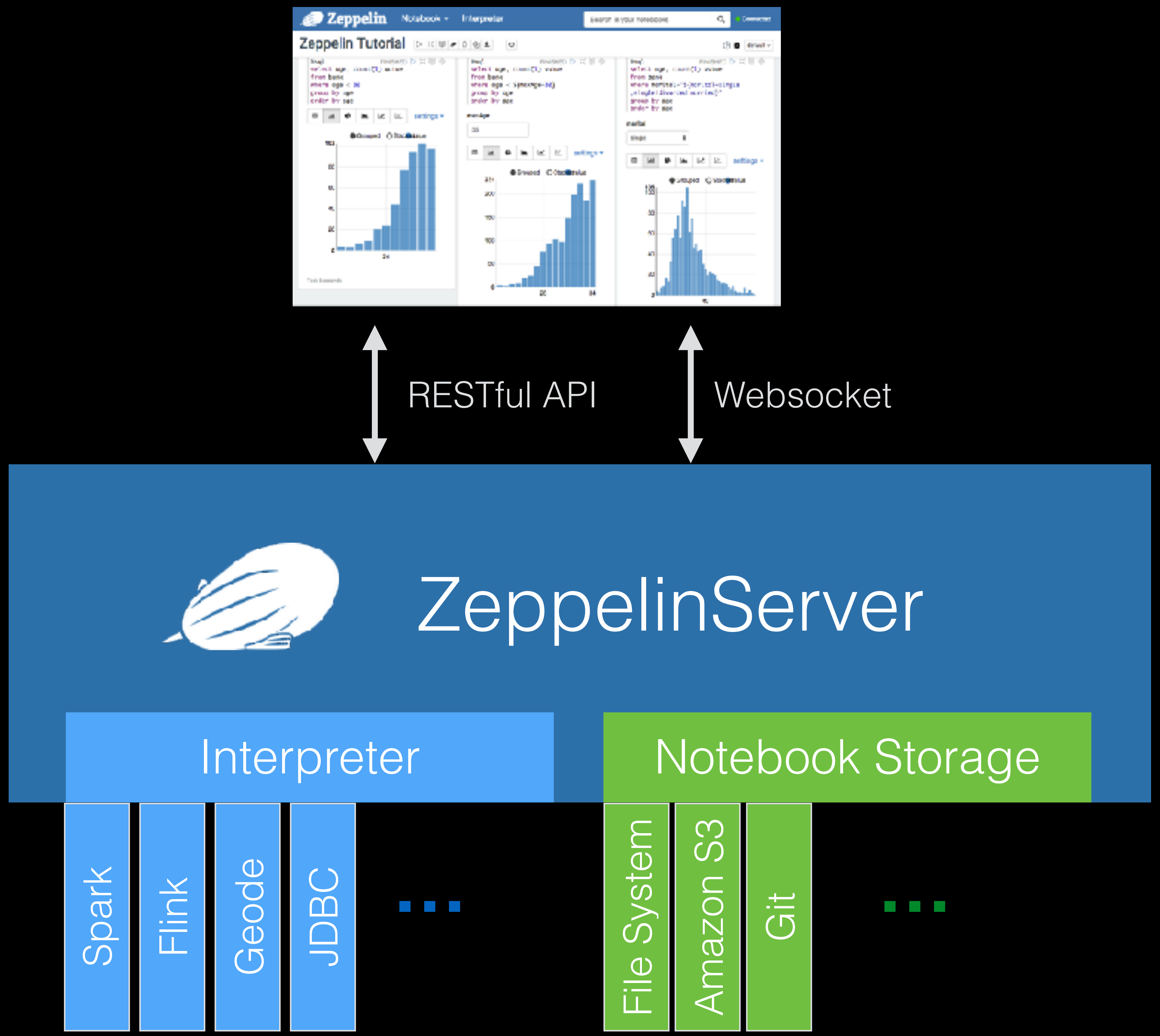

ebay
LINE
AT&T
NASA
JPL
Jet Propulsion Laboratory
California Institute of Technology
GE
IBM

& Enterprise

- Company complains
  - Why security works …
  - Why authentication works …
  - Why Zeppelin stores my password as plain …
  - Why two user use same Spark …
  - Why I wait while other run somethings

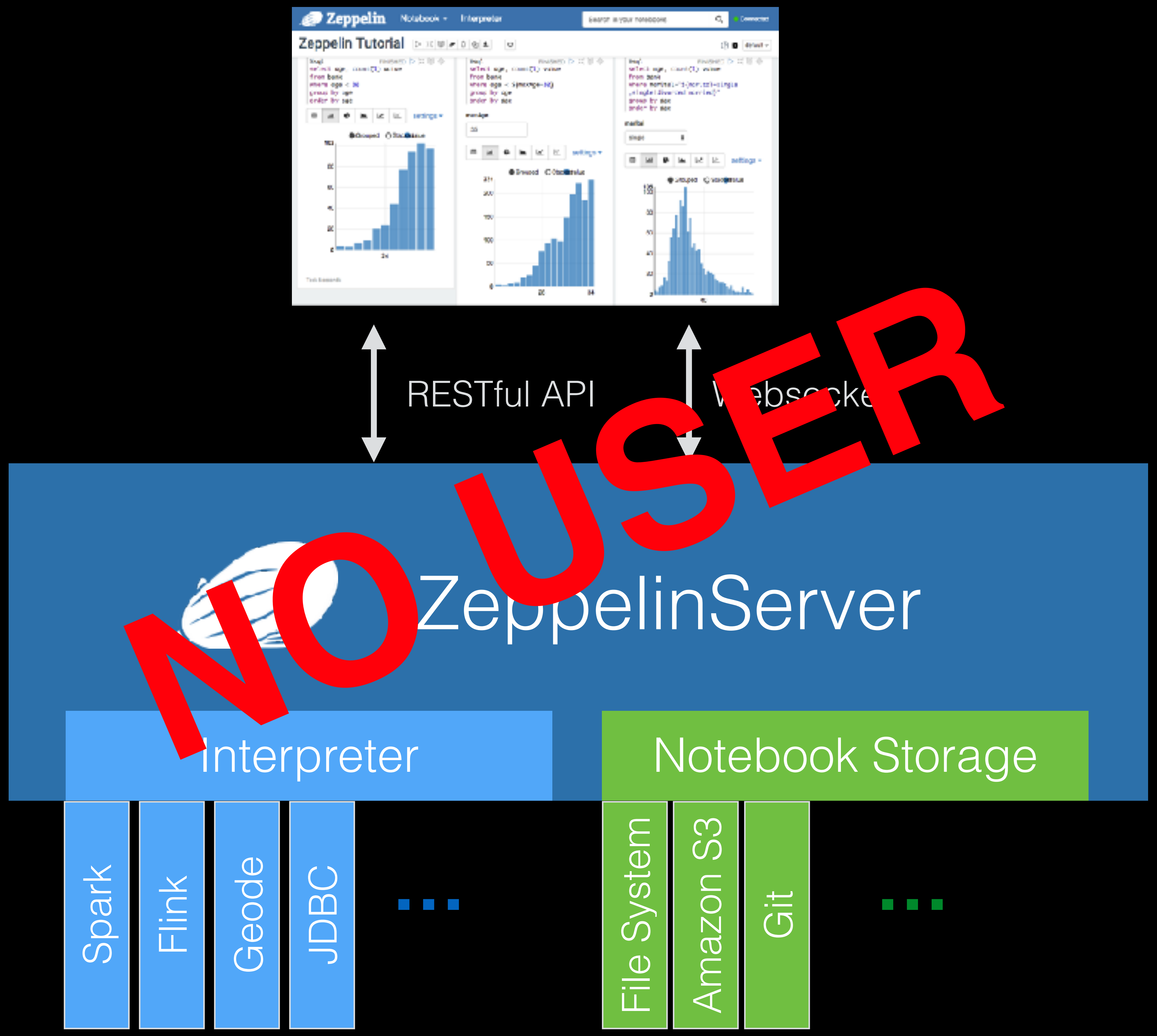
& Enterprise

Multi-tenancy



& Enterprise

Multi-tenancy



& Enterprise

History

~ 0.6

- NOTHING!!!

0.6.x

- Authentication & Authorization
- Note level permission
- Note level isolation
- Partially supported by Livy

& Enterprise

Future

0.7.0

- Enterprise Support
  - Multi users environment
  - Impersonation on Spark/JDBC interpreter
  - Job management
- Interpreter
  - Improvement on JDBC/Python interpreter
- Frontend performance improvement
- Pluggable visualization

& Enterprise

Future

0.7.0

- Enterprise Support
  - **Multi users environment**
  - Impersonation on Spark/JDBC interpreter
  - Job management
- Interpreter
  - Improvement on JDBC/Python interpreter
- Frontend performance improvement
- Pluggable visualization

Shared, Isolated, Scoped

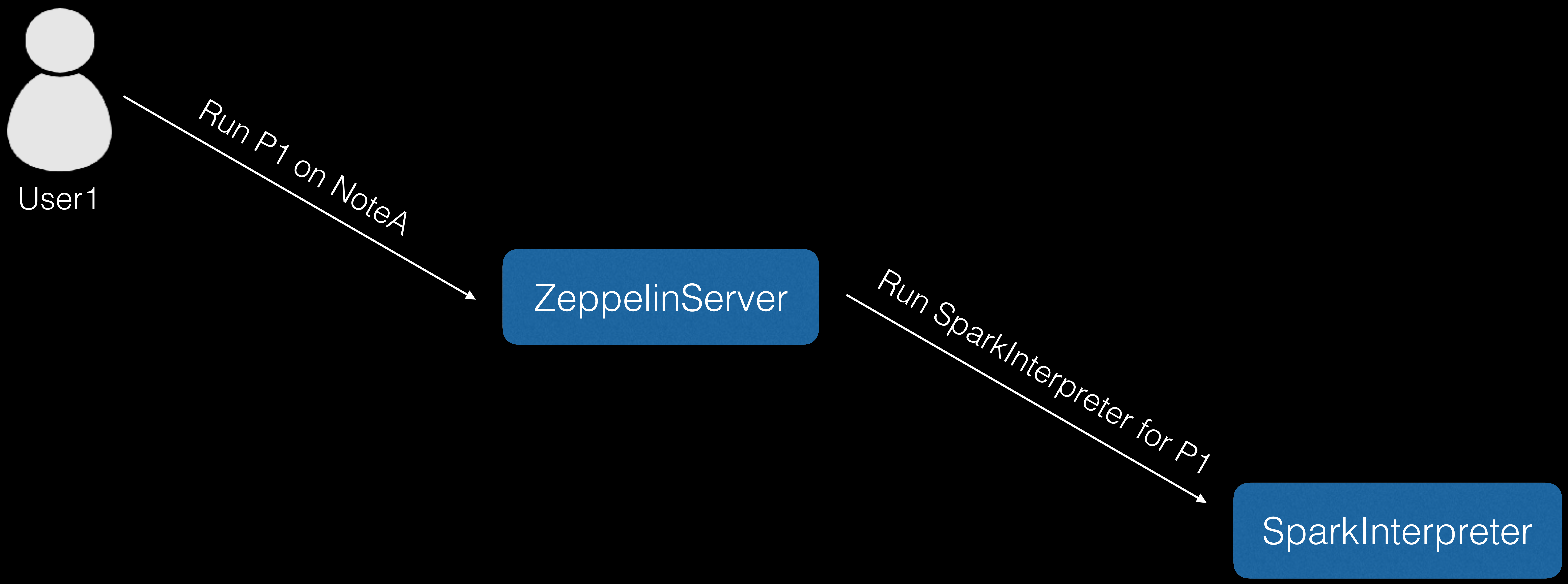
& Enterprise

Multi-tenancy

	SHARED	ISOLATED	SCOPED
PROCESS	1	N	1
THREADS	1	1	N

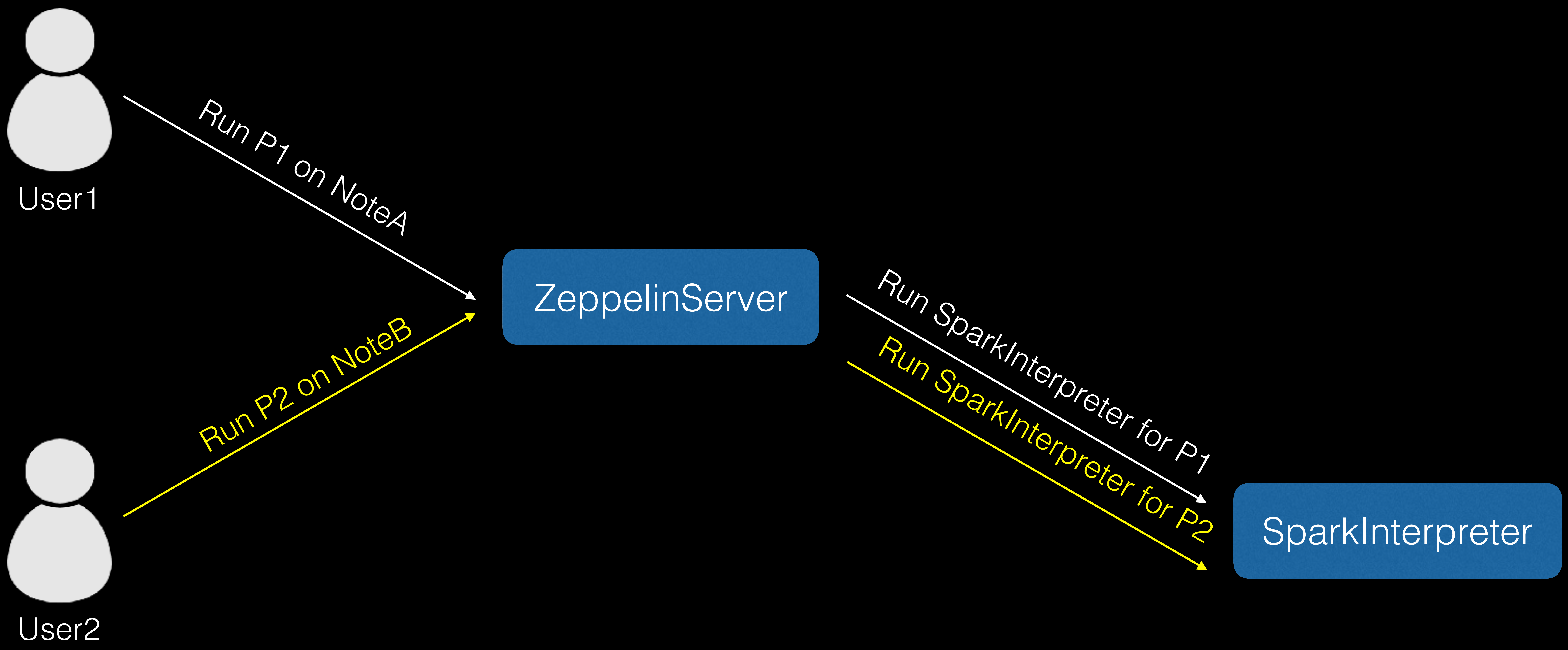
& Enterprise

Multi-tenancy



& Enterprise

Multi-tenancy



& Enterprise

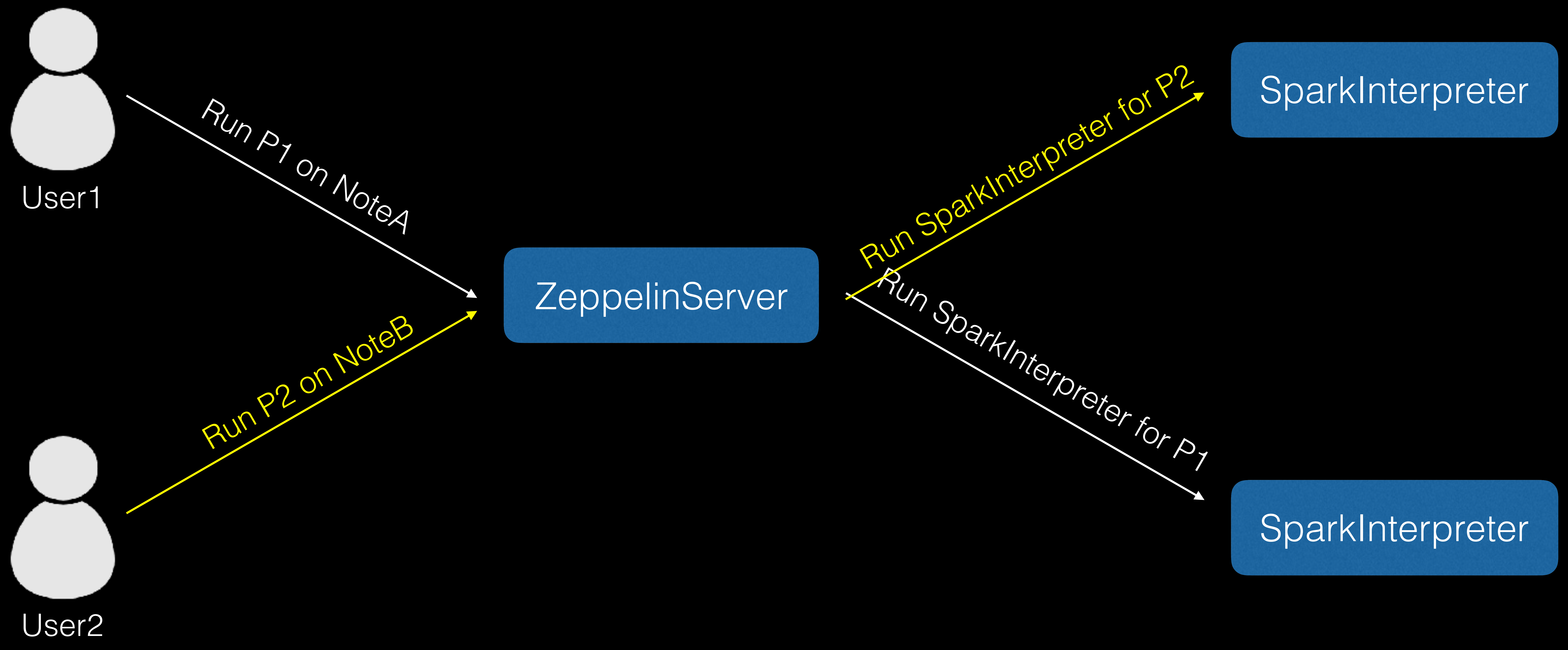
Multi-tenancy

Shared

- Originally implemented
- Pros
  - Simple structure
  - Predictable behavior
- Cons
  - All resources shared
  - Interference among users

& Enterprise

Multi-tenancy



User1

Run P1 on NoteA

User2

Run P2 on NoteB

ZeppelinServer

Run SparkInterpreter for P2

Run SparkInterpreter for P1

SparkInterpreter

SparkInterpreter

& Enterprise

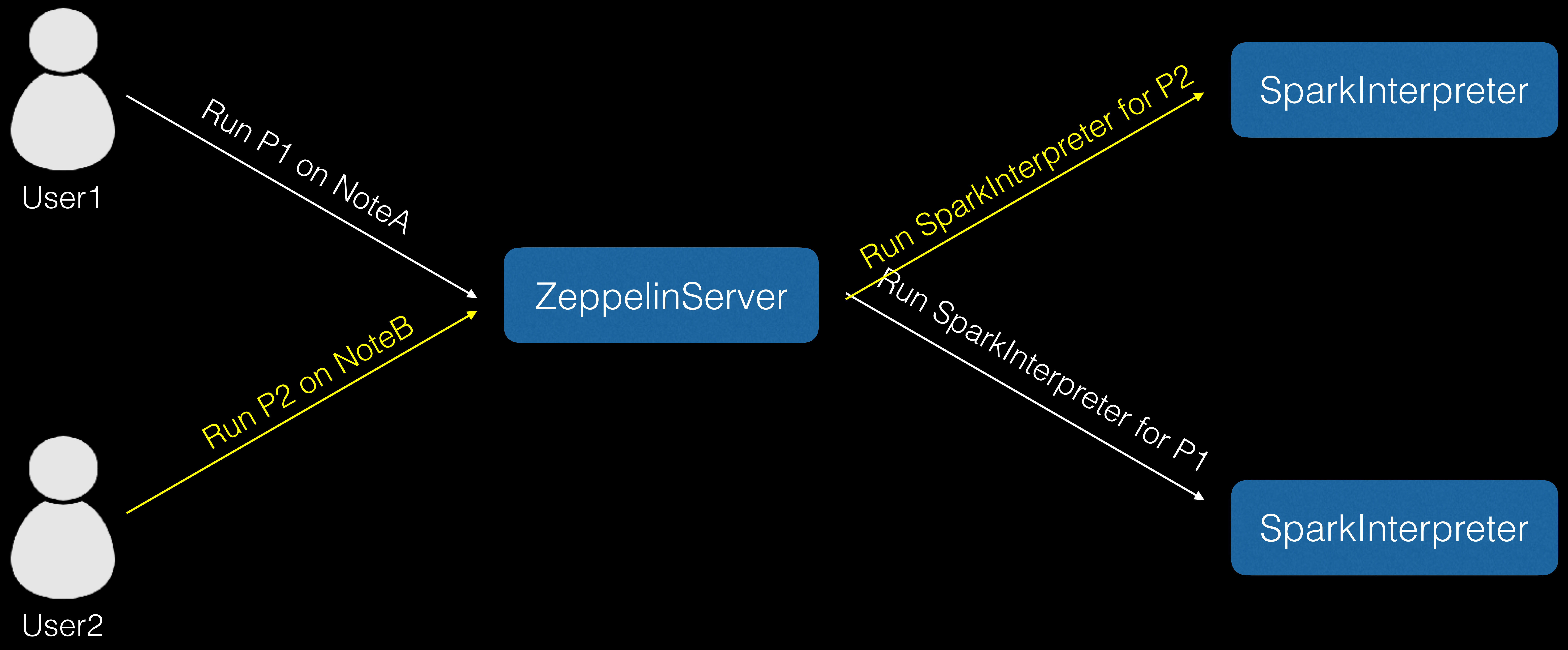
Multi-tenancy

Isolated

- Pros
  - No pending
  - No resources shared
- Cons
  - Lots of memory
  - Inefficiency of using memory
  - Limited by resources

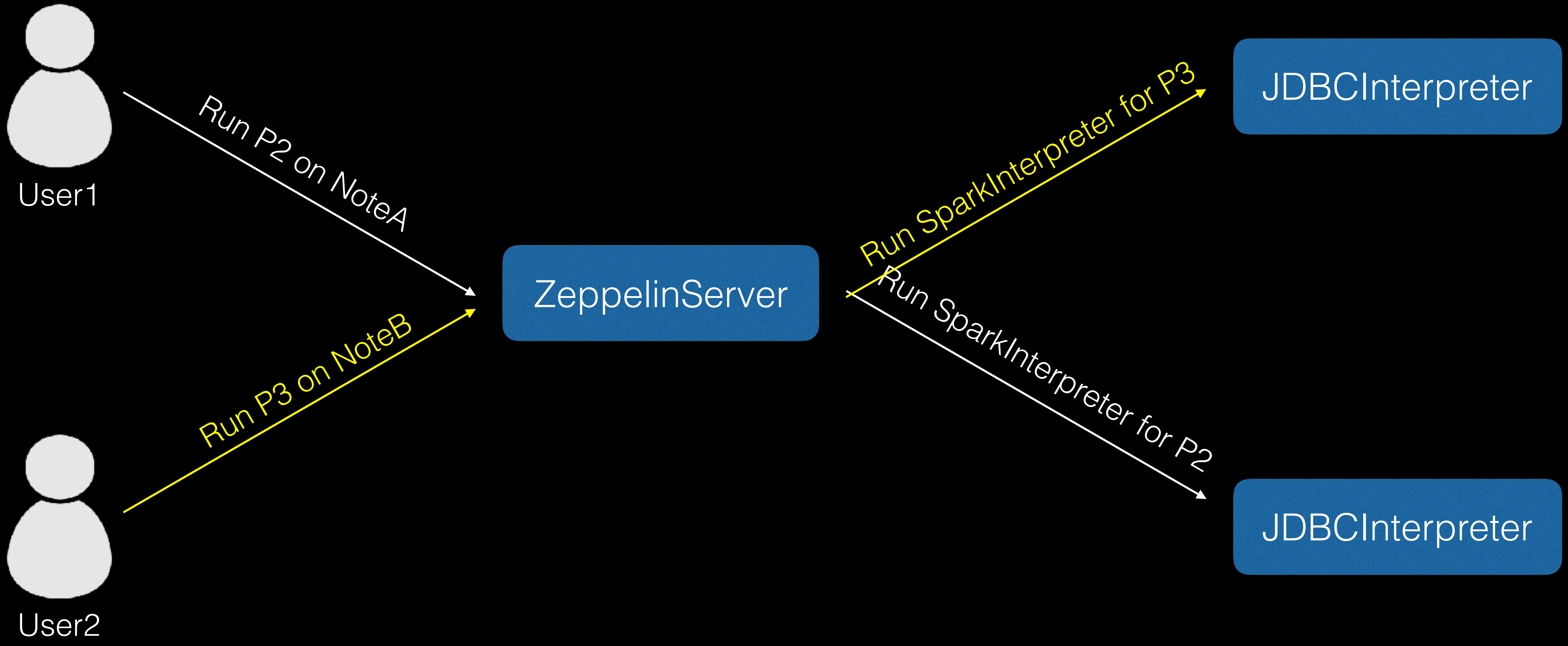
& Enterprise

Multi-tenancy



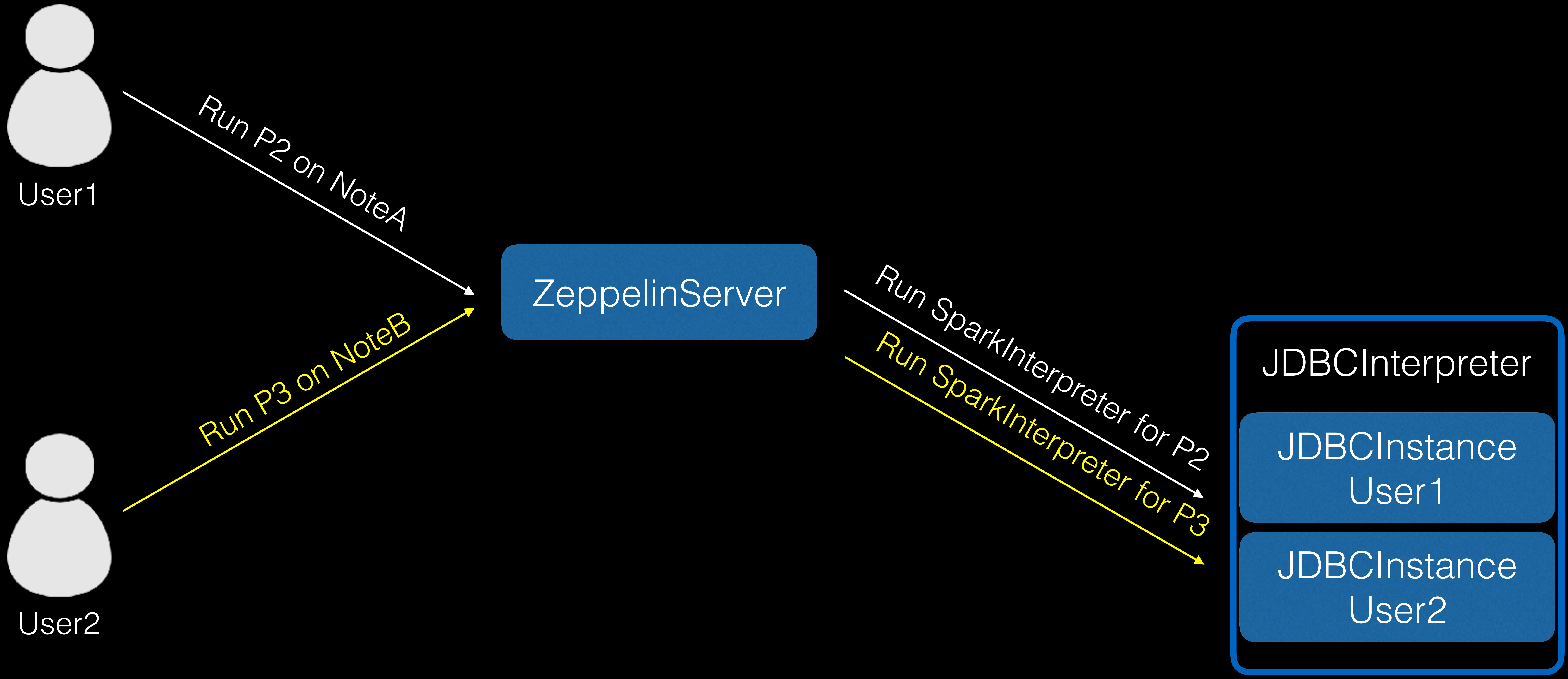
& Enterprise

Multi-tenancy



& Enterprise

Multi-tenancy



& Enterprise

Multi-tenancy

Scoped

- Pros
  - Less memory
  - Some resources Isolated
- Cons
  - Some resources shared
  - Big single process

& Enterprise

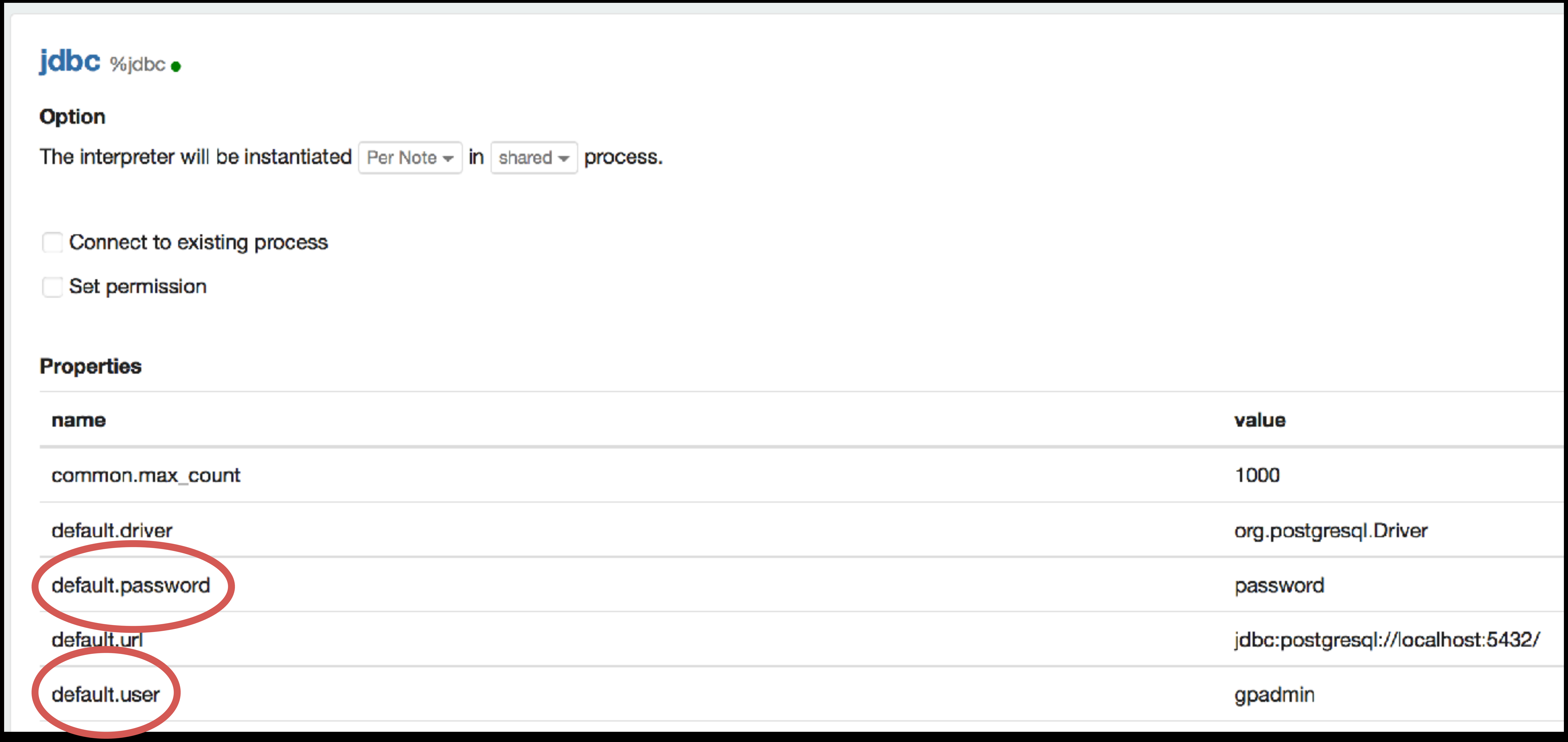
Future

0.7.0

- Enterprise Support
  - Multi users environment
  - **Impersonation on Spark/JDBC interpreter**
  - Job management
- Interpreter
  - Improvement on JDBC/Python interpreter
- Frontend performance improvement
- Pluggable visualization

& Enterprise

Impersonation



jdbc %jdbc ●

Option

The interpreter will be instantiated Per Note ▾ in shared ▾ process.

□ Connect to existing process

□ Set permission

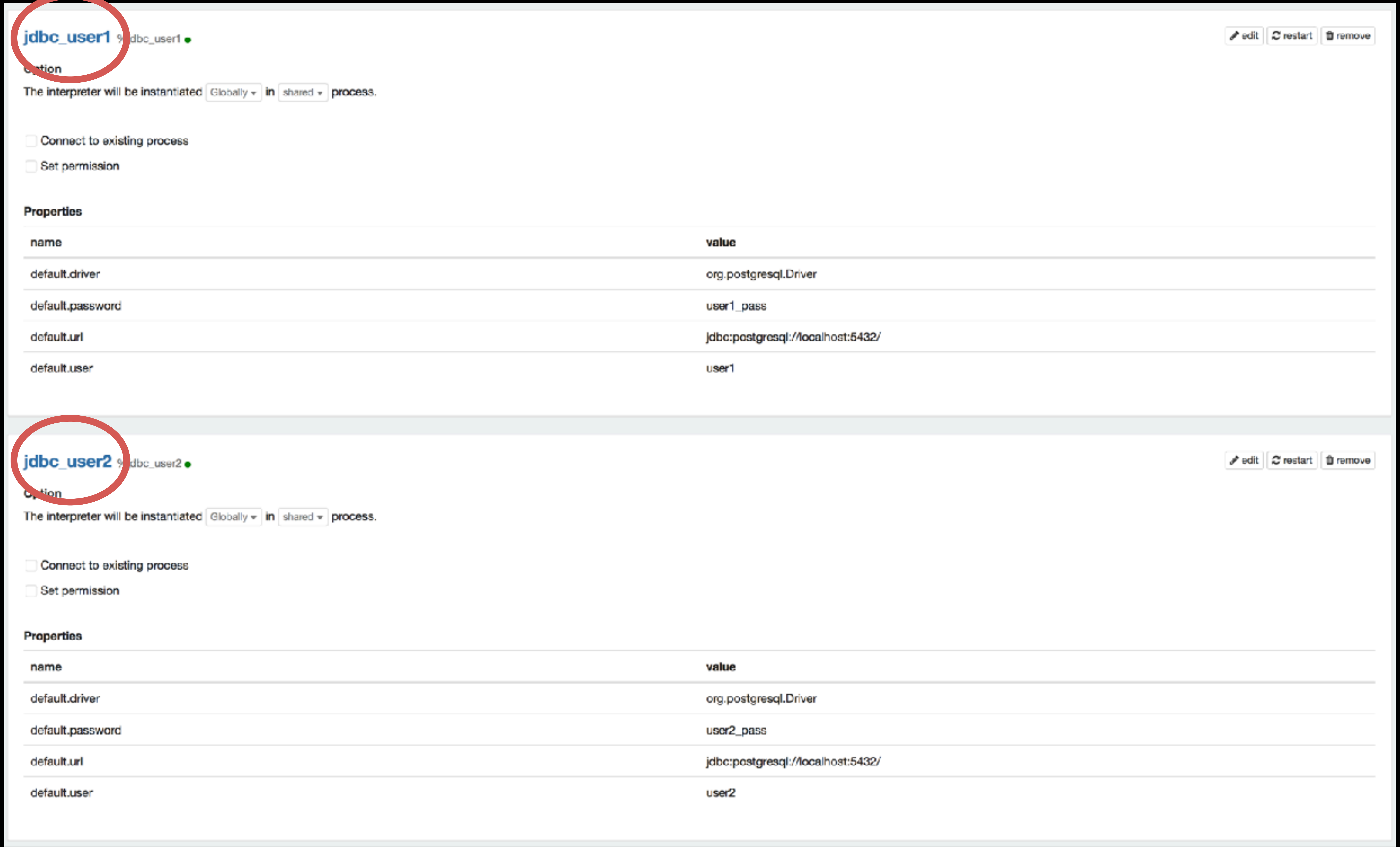
Properties

name	value
common.max_count	1000
default.driver	org.postgresql.Driver
default.password	password
default.url	jdbc:postgresql://localhost:5432/
default.user	gpadmin

What if all users use different credentials?

& Enterprise

Impersonation



jdbc_user1 %jdbc_user1 ●

Option

The interpreter will be instantiated Globally ▾ in shared ▾ process.

□ Connect to existing process

□ Set permission

Properties

name	value
default.driver	org.postgresql.Driver
default.password	user1_pass
default.url	jdbc:postgresql://localhost:5432/
default.user	user1

jdbc_user2 %jdbc_user2 ●

Option

The interpreter will be instantiated Globally ▾ in shared ▾ process.

□ Connect to existing process

□ Set permission

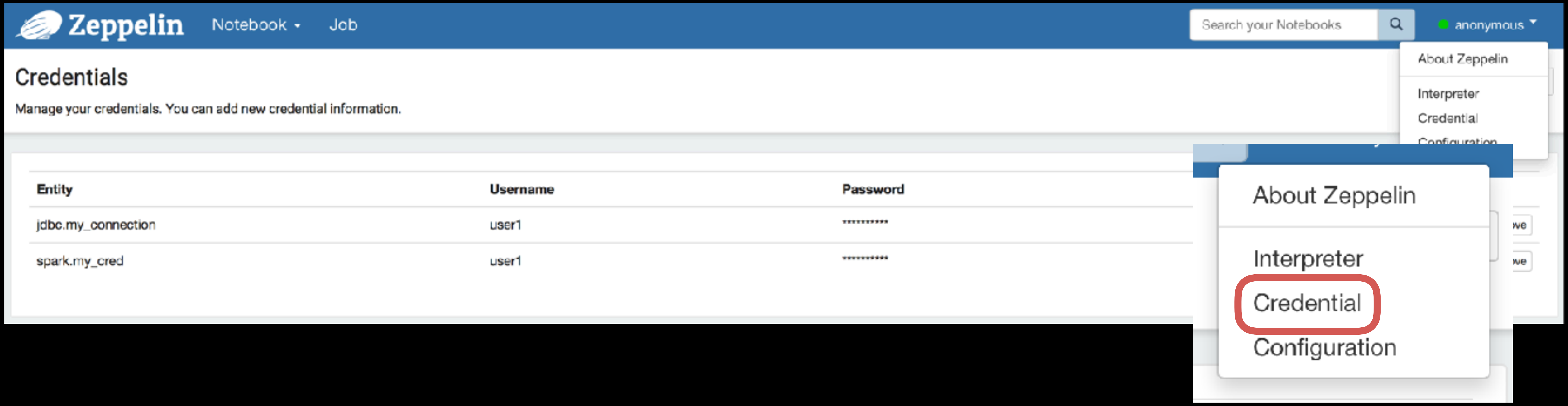
Properties

name	value
default.driver	org.postgresql.Driver
default.password	user2_pass
default.url	jdbc:postgresql://localhost:5432/
default.user	user2

& Enterprise

Impersonation

Credentials



Credentials

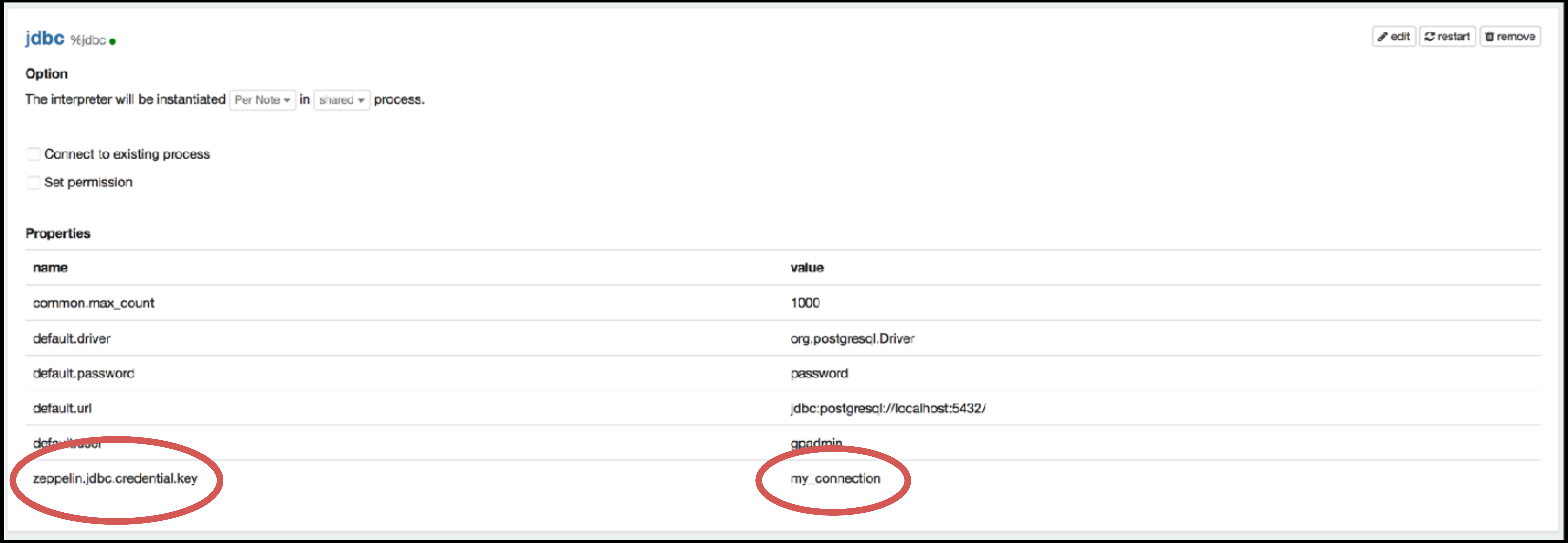
Manage your credentials. You can add new credential information.

Entity	Username	Password
jdbc.my_connection	user1	**********
spark.my_cred	user1	**********

- Already merged by Twitter at Mar. 2016
- Never used in any interpreter

& Enterprise

Impersonation



jdbc %jdbc ●

Option

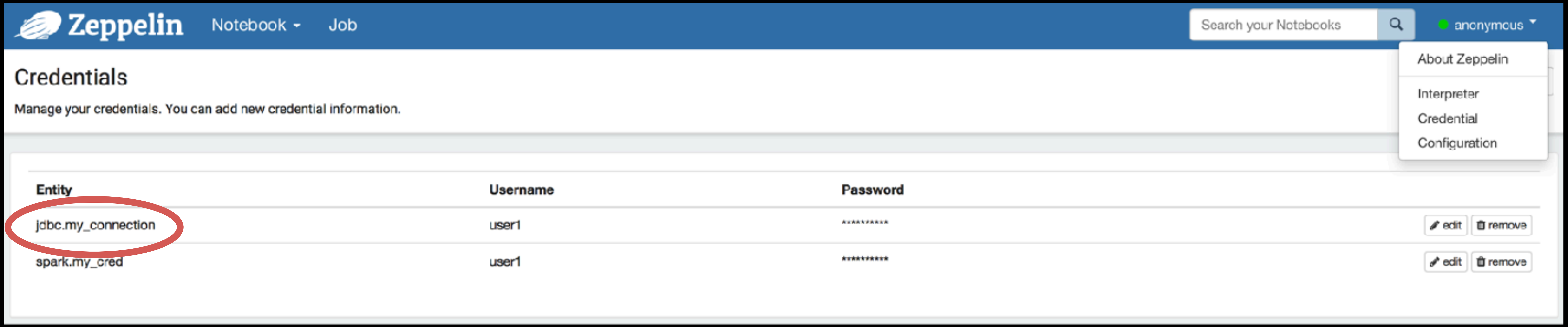
The interpreter will be instantiated Per Note in shared process.

□ Connect to existing process

□ Set permission

Properties

name	value
common.max_count	1000
default.driver	org.postgresql.Driver
default.password	password
default.url	jdbc:postgresql://localhost:5432/
default.user	gpadmin
zeppelin.jdbc.credential.key	my_connection



Credentials

Manage your credentials. You can add new credential information.

Entity	Username	Password	
jdbc.my_connection	user1	**********	edit remove
spark.my_cred	user1	**********	edit remove

& Enterprise

Impersonation

- JDBC
  - Set user and password in properties
  - https://issues.apache.org/jira/browse/ZEPPELIN-1567
- Spark
  - Adopt ugi.doAs()
  - https://issues.apache.org/jira/browse/ZEPPELIN-1572

& Enterprise

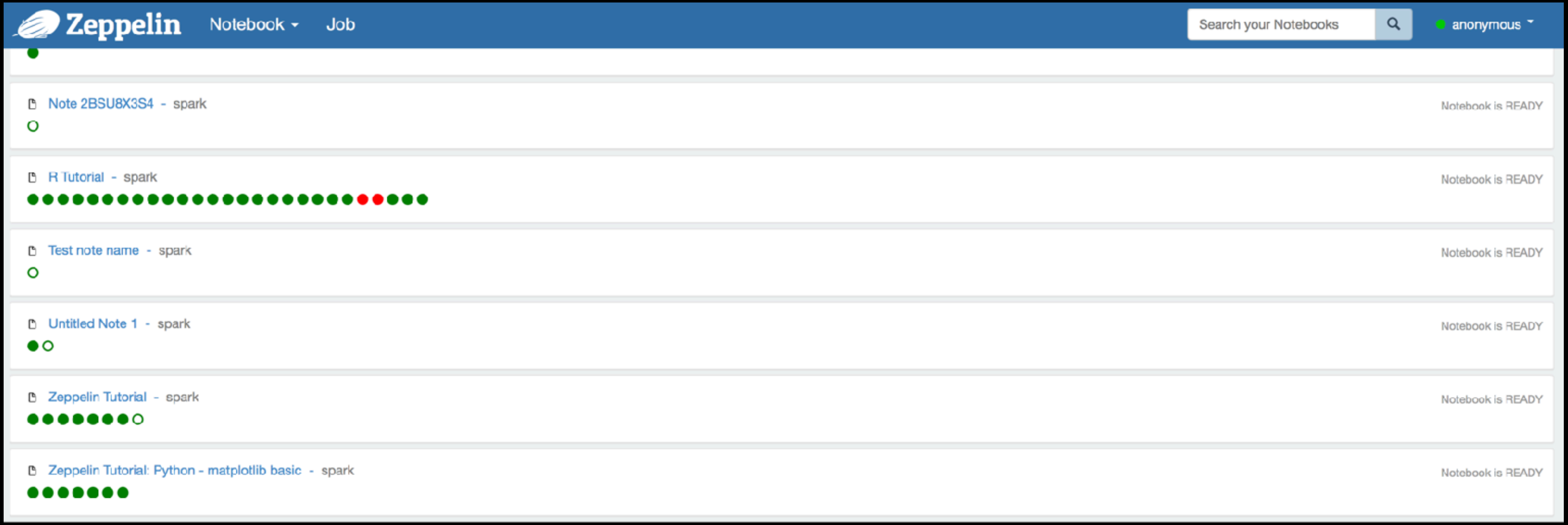
Future

0.7.0

- Enterprise Support
  - Multi users environment
  - Impersonation on Spark/JDBC interpreter
  - **Job management**
- Interpreter
  - Improvement on JDBC/Python interpreter
- Frontend performance improvement
- Pluggable visualization

& Enterprise

Job mgmt



●

Note 2BSU8X3S4 - spark
○
Notebook is READY

R Tutorial - spark
●●●●●●●●●●●●●●●●●●●●●●●●●●●
Notebook is READY

Test note name - spark
○
Notebook is READY

Untitled Note 1 - spark
●○
Notebook is READY

Zeppelin Tutorial - spark
●●●●●●●○
Notebook is READY

Zeppelin Tutorial: Python - matplotlib basic - spark
●●●●●●●
Notebook is READY

& Enterprise

Future

0.7.0

- Enterprise Support
  - Multi users environment
  - Impersonation on Spark/JDBC interpreter
  - Job management
- Interpreter
  - **Improvement on JDBC/Python interpreter**
- Frontend performance improvement
- Pluggable visualization

& Enterprise

Interpreters

- JDBC
  - Connection pool
  - Stabilization for BI
- Python
  - Matplot library
  - Support on python user

& Enterprise

Future

0.7.0

- Enterprise Support
  - Multi users environment
  - Impersonation on Spark/JDBC interpreter
  - Job management
- Interpreter
  - Improvement on JDBC/Python interpreter
- **Frontend performance improvement**
- **Pluggable visualization**

& Enterprise

Frontend

- Frontend
  - Fine-grained broadcast of WebSocket
  - Betterment of rendering DOM
- Pluggable visualization
  - He lium

Demo

Zeppelin

Join the community

Homepage

http://zeppelin.apache.org/

Mailing list

users@zeppelin.apache.org
dev@zeppelin.apache.org

Issue tracker

https://issues.apache.org/jira/browse/ZEPPELIN

Github repository

http://github.com/apache/zeppelin

Thank you

Jongyoul Lee
jongyoul@nflabs.com
https://twitter.com/madeng