EclairJS

A Node.js front end to Apache Spark

Session Outline

- Web applications and big data analytics
- Node.js Overview
- Apache Spark Overview
- Why it's hard to communicate with Apache Spark.
    - Spark Application Overview.
- EclairJS Intro
- Demo
- EclairJS Architecture and Server Requirements.
- Deployment on Bluemix.
- Usage in Notebooks.
- Conclusion and Moving Forward.

Data Analytics and Web Applications

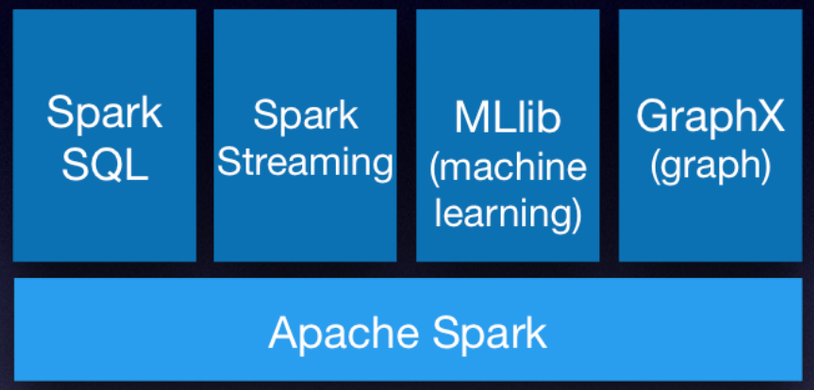
- Increasing need for web applications to be driven by data analytics.
- Companies are generating more and more data (logs, user actions, etc.)
- Node.js is a popular highly scalable web application platform. Large computations are offloaded to back end systems.
- Apache Spark is a popular data analytics platform. Difficult to connect to web application platforms.
- EclairJS combines and connects the two.

Node.js Overview

- Popular server side JavaScript runtime. Used by companies of all sizes and across industries.
- Large number of JEE and .NET users moving to Node.js.
- Event Loop. Non blocking asynchronous IO.
    - Can handle many concurrent connections at once.
    - Scales well in cloud environments.
- Works well at the front end of large web application ecosystems.

Spark Overview

- A distributed compute platform with an API centered around the DataSet abstraction.
    - Immutable in memory data processing.
    - Work distributed over a cluster of worker nodes.
- Runs on the Java Virtual Machine.
- Multiple languages supported.
    - Scala and Java run natively.
    - Python and R proxy into the JVM.

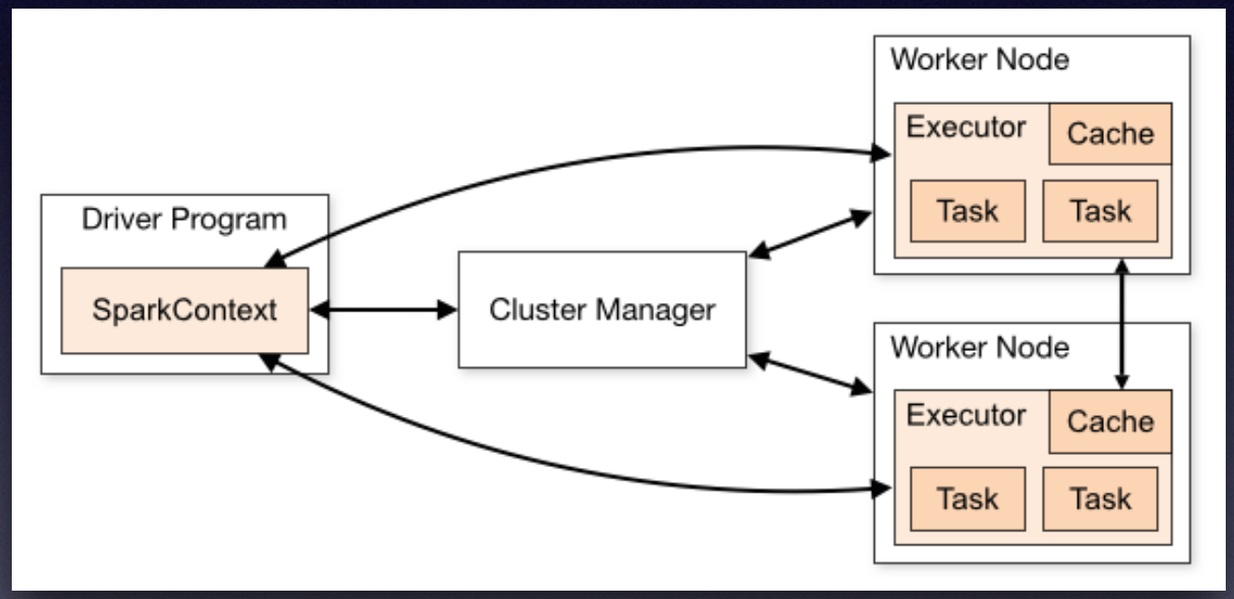


Unified API for SQL, Streaming, ML and Graph.

Why is it hard for web apps to talk to Spark?

Spark Application Overview

- **spark-submit** launches an application called a driver program locally (client mode) or onto the cluster (cluster mode).
- There are no callback mechanisms or rest APIs.
    - Driver programs may put results into a database or message queue.
    - Separate application retrieves data. Never communicates directly with Spark.
- Driver Program needs be on the same LAN as the worker nodes.
- Could possibly capture stdin / stdout from driver app.



EclairJS

- Combines Node's asynchronous IO based web front end with Spark's machine learning and distributed compute capabilities.
- Provides an NPM module (client) for Node.js developers.
    - Implements the Spark API in JavaScript. Including Spark SQL, Streaming, and ML.
- Provides a server environment to proxy API calls into the Spark JVM runtime.
    - Single WebSocket connection from client to server.
    - Docker container for local development.
- Open source project on github
    - https://github.com/EclairJS/eclairjs
    - Apache Licensed

Run the EclairJS Examples

```
# clone the repository to get the examples
git clone https://github.com/eclairjs/eclairjs

cd examples

# install EclairJS node module.
npm install

# start the EclairJS docker container
docker run -d -p 8888:8888 eclairjs/minimal-gateway

# run the simple example
node —harmony simple.js
```

Code Example

simple.js

```
//include the eclairjs module.
var eclairjs = require('eclairjs')

//create an eclairjs client instance.
var spark = new eclairjs();

//Build a spark session.  Will default to local[*] for spark master.
var session = spark.sql.SparkSession.builder()
  .appName("Hello World")
  .getOrCreate()

//Create a simple array dataset.
var ds = session.sparkContext().parallelize([1,2,3,4,5])

//Map over the dataset and add 1 to each element.  collect() returns a
//promise that resolves to the results.

ds.map(function(i) {
  return i+1
}).collect().then(function(r) {
  console.log(r)
})
```

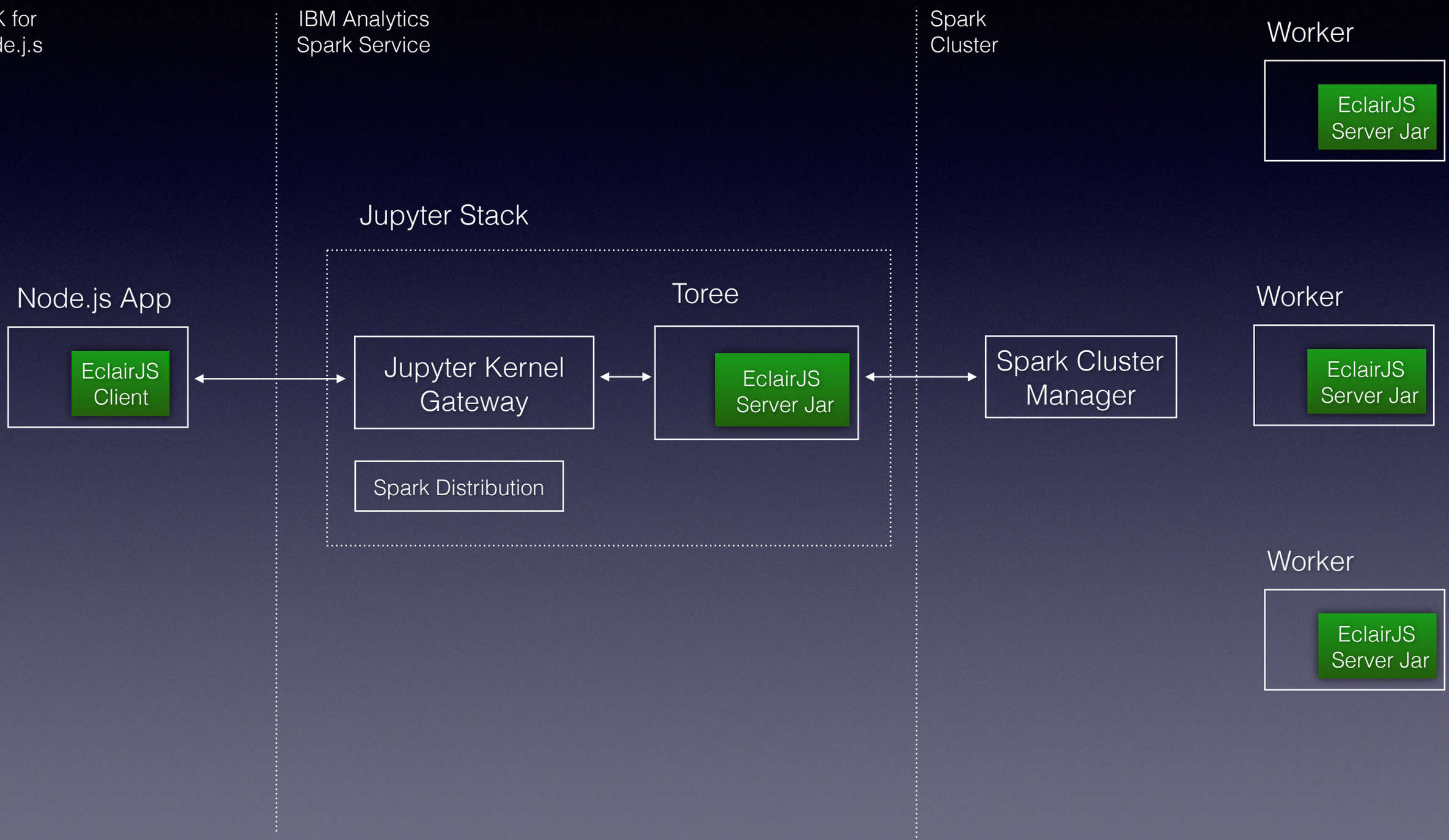
Web Application Demo

https://github.com/EclairJS/eclairjs-examples/tree/master/sales_demo

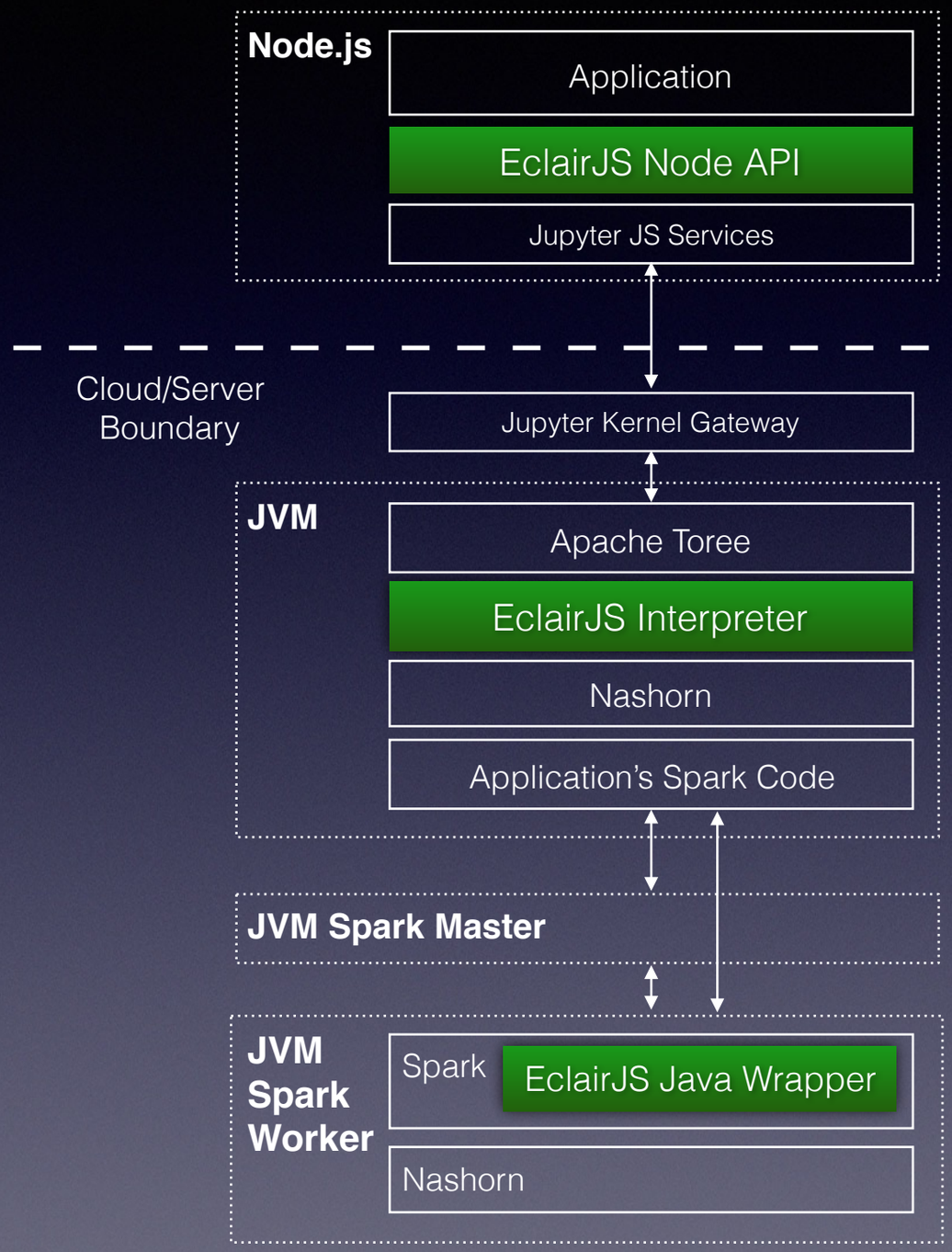
Architecture

- EclairJS utilizes the Jupyter Kernel Gateway environment on the server to execute commands.
  - http://jupyter-kernel-gateway.readthedocs.io/en/latest/
  - Code is executed using the Apache Toree kernel.
    - https://toree.apache.org/
    - EclairJS provides a JavaScript interpreter for Toree.
  - Dataset, Dataframe and RDD lambda functions are executed on the executors using the Java 8 Nashorn JavaScript runtime engine.
- EclairJS client in Node.js makes use of the Jupyter JS Services module to communicate with the kernel gateway.
  - https://github.com/jupyterlab/services

EclairJS Server Components



Continued..




```
var numAs =
logData.filter(function(s) {
   var ret = (s.indexOf("a") > -1);
   return ret;
});
```

EclairJS Node API
Stringify code

```
{code: "var rdd90912 =
rdd72359.filter(function(s) {
   var ret = (s.indexOf("a") > -1);
   return ret;"
}
```

Jupyter JS Services
Resolve variables
Stringify code

```
return this.jvmRdd.filter(
   new org.eclairjs.nashorn
      .JSFunction("function(s) {return (s.indexOf("a")
> -1}")
);
```

Wrap serialized JS functions in Java wrapper.

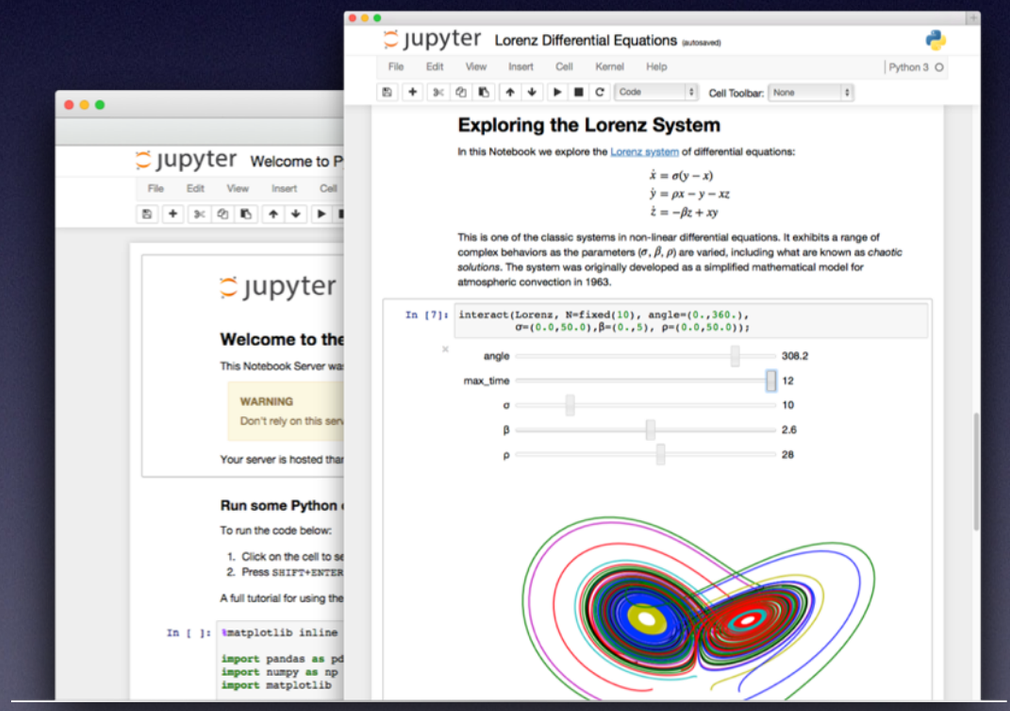
EclairJS executes javascript function in Nashorn.

Deployment on Bluemix

- Node.js applications can be deployed using the SDK for Node.js
  - https://console.ng.bluemix.net/catalog/starters/sdk-for-nodejs/
    - Developers simply include “eclairjs” as a module dependency in their Node.js application.
- Coming Soon. The EclairJS server can be deployed using IBM Analytics for Apache Spark.
  - https://console.ng.bluemix.net/catalog/services/apache-spark
  - Stay tuned for future blog posts and demos..

Notebooks

- Designed for (data) scientists, used widely for data cleaning and transformation, numerical simulation, statistical modeling, etc.
- Notebooks appear in browser, consist of cells that may contain live code, visualizations, formatted text, widgets.
- Often used as reports after analyses completed.
- The EclairJS server uses Jupyter so you can use notebooks with it.



In Conclusion

- EclairJS brings the power of Apache Spark to Node.js web application development.
- Clone our repository and try out some of the examples.
  - https://github.com/eclairjs/eclairjs
    - Developer contributions are welcome under ICLA
- Join our mailing list.
  - https://groups.google.com/d/forum/eclairjs
- Join our Slack Channel
  - http://eclairjs.slack.com/