ECLAIRJS = NODE.JS + APACHE SPARK

David Fallside
IBM

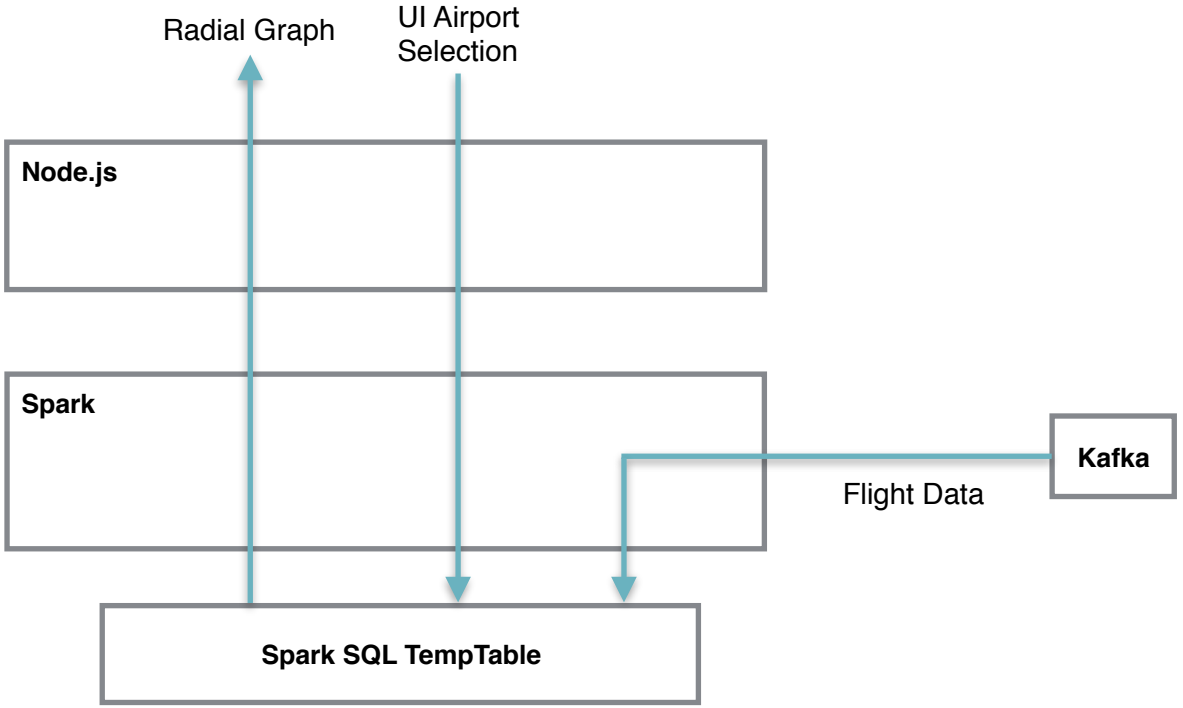
SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO

Why EclairJS?

- Digital business looking to improve customer interactions, capture “perishable” insights (Forrester), use new data sources
- Today’s interactive & user-facing applications often developed in JavaScript using Node.js
  - npm provides largest (Node.js) package repo (www.modulecounts.com)
  - Handles very large numbers of simultaneous requests
  - Compute-intensive workloads handed off to back-end engines
- Apache Spark as a back-end engine
  - Scalable, static & streaming data, Spark SQL, ML analytics, graph engine
  - But no Spark API for Node.js/JavaScript, hence EclairJS
  - So let’s look at an EclairJS application …

demo

Program Flow



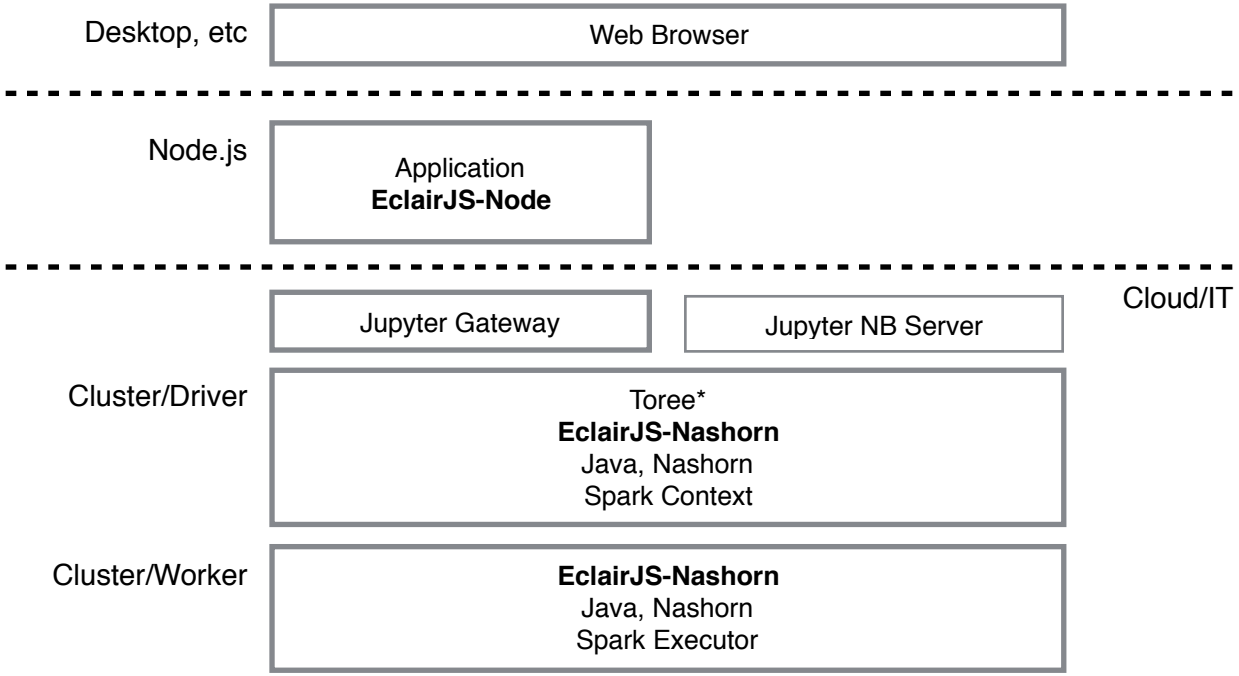
Word Count

```
var spark = require('eclairjs');

var sc = new spark.SparkContext("local[*]", "foo");

var file = __dirname + '/dream.txt';

var rdd = sc.textFile(file);

var rdd2 = rdd.flatMap(function(sentence) {
  return sentence.split(" ");
});

var rdd3 = rdd2.filter(function(word) {
  return word.trim().length > 0;
});

var rdd4 = rdd3.mapToPair(function(word, Tuple) {
  return new Tuple(word.toLowerCase(), 1);
}, [spark.Tuple]);
```

```
var rdd5 = rdd4.reduceByKey(function(value1, value2) {
  return value1 + value2;
});

var rdd6 = rdd5.mapToPair(function(tuple, Tuple) {
  return new Tuple(tuple[1], tuple[0]);
}, [spark.Tuple]);

var rdd7 = rdd6.sortByKey(false);

rdd7.take(10).then(function(val) {
  console.log("Success:", val);
});
```

Spark Operator

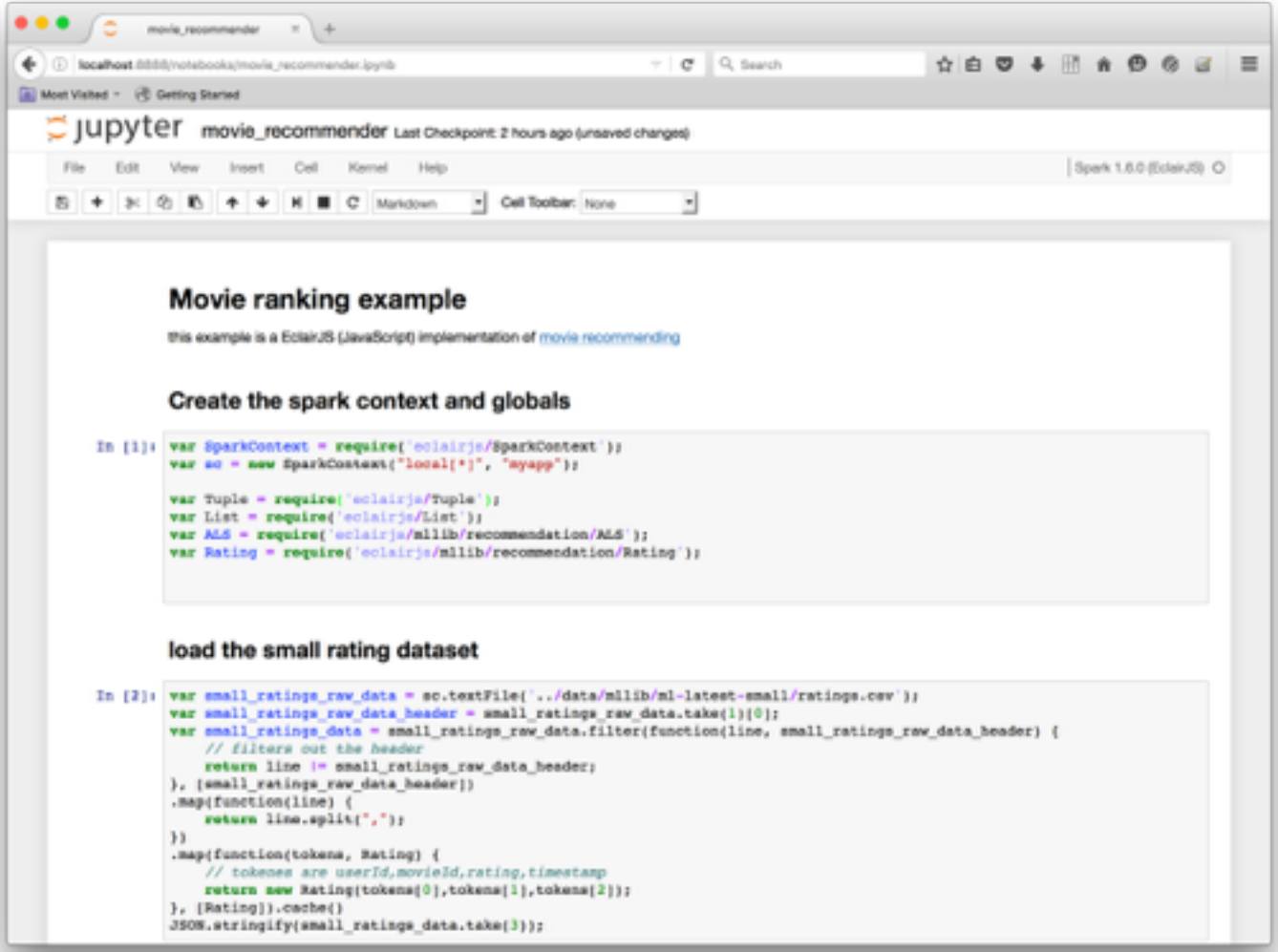
EclairJS Stack



*Toree in Apache Incubator

Notebooks

- Notebooks designed for (data) scientists, widely used for data cleaning and transformation, numerical simulation, statistical modeling, etc
  - Appear in browser as cells, may contain live code, visualizations, formatted text, widgets, etc
  - Jupyter notebooks have pluggable kernel architecture to enable different languages (jupyter.org)
- EclairJS provides JavaScript kernel so data engineers and web developers can try-out code and work with data in notebooks



In Closing

- EclairJS for web application development in Node.js and JavaScript
  - For Data Engineers with JavaScript in Notebooks
- Project under active development in Github, eclairjs.org
  - Examples, documentation, getting-started, etc
  - EclairJS Node and EclairJS Nashorn
  - Open source, Apache v2 license
- Looking for collaborators!

THANK YOU.

eclairjs.org

fallside at us.ibm.com

SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO