Apache Spark

Mate Gulyas

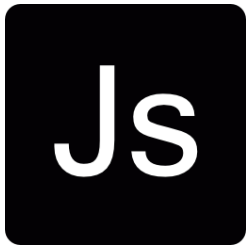
WHY WE DO IT?

36% non-human visitor *	54% average viewability**	What percentage of digital ads reach people?
Clickbots, botnets	Invisible, hidden ads	Transparency in the market

*http://technorati.com/iab-keynote-36-percent-ad-traffic-from-bots-and-threatening-industry/
**http://www.statista.com/statistics/255061/viewability-rates-for-rich-media-ads-worldwide-by-industry/

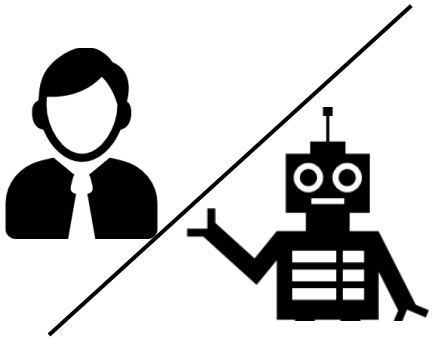
enbrite.ly

WHAT WE DO?



JavaScript

Behaviour analysis



Segmentation

WHAT DO WE NEED?

Distributed data processing

Averge size client	*30 GB / day*
	900 GB / month
20 average size clients	*600 GB / day*
	18 TB / month

WHAT DO WE NEED?

Recurring data transformations

IT'S SO COOL

Interactive / Batch / Streaming / SQL / Graph processing

WHY SPARK?

In-memory

WHY SPARK?

Productive API

WHY SPARK?

Multiple language

WHY SPARK?

Active community

WHY SPARK?

developer
analyst
CFO

friendly

WHY SPARK?

developer
analyst
CFO

friendly

WHY SPARK?

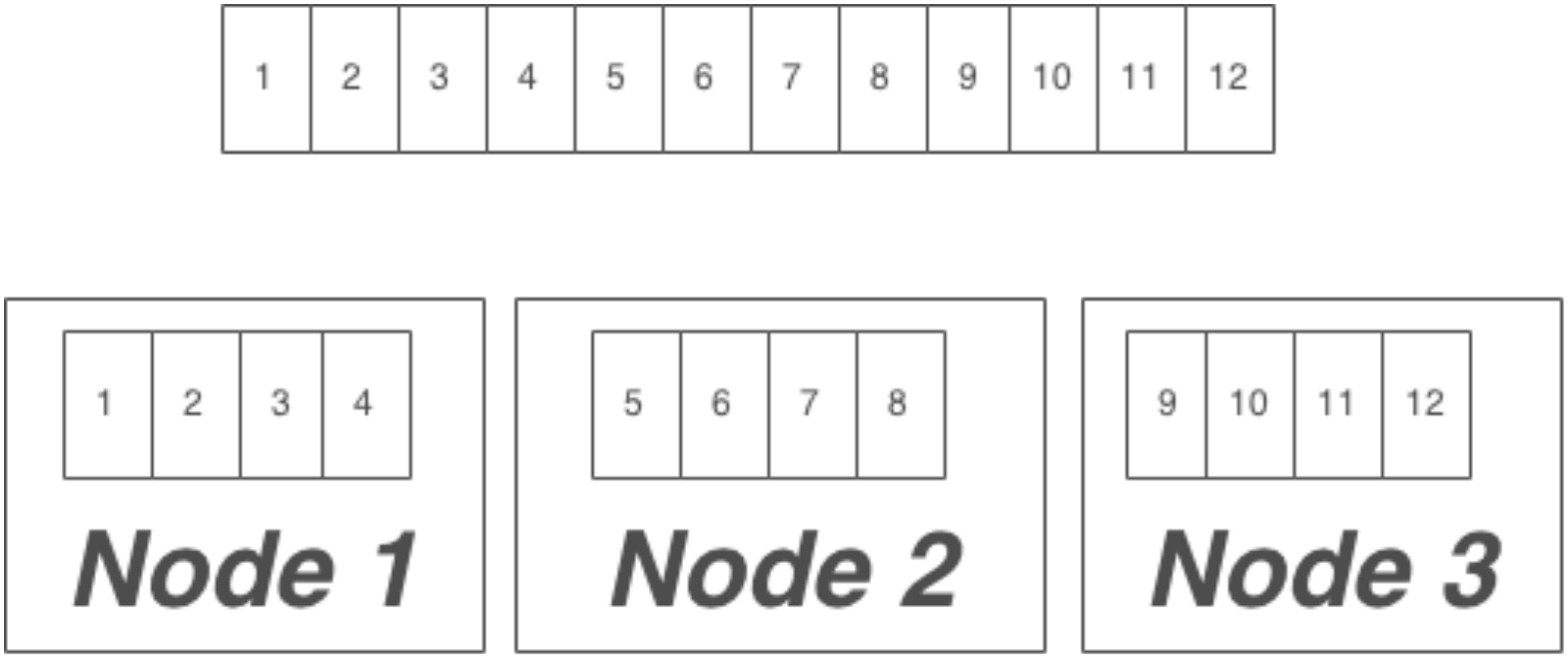
developer
analyst
CFO

friendly

ONE THING TO REMEMBER

Resilient Distributed Dataset (RDD)

RDD



IT RUN'S ON

Mesos

YARN

Standalone

AWS EC2

IT GET'S DATA FROM?

Amazon S3, HDFS, Cassandra, Hive, Hbase, Tachyon, Local Filesystem, ODBC databases, etc...

THE OLD WAY

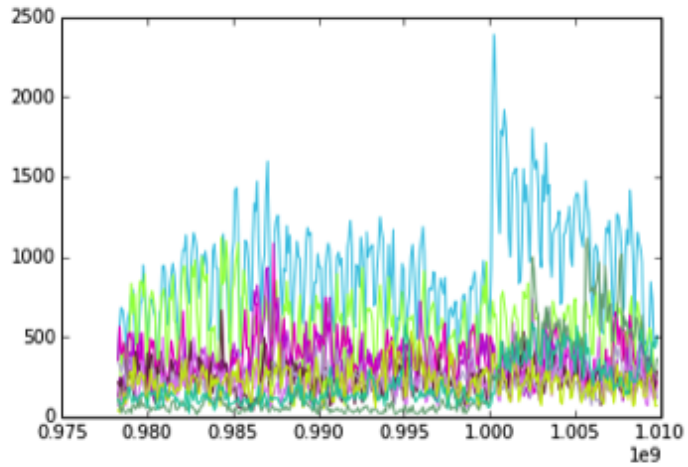
Batch processing

THE NEW WAY

Interactive analytics

Now we're ready to plot. Each object in `country_series` has the information we need to plot a single line.

In [27]:
```
random_color = lambda: '#%02x%02x%02x' % tuple(np.random.randint(0,256,3))
fig = plt.figure()
ax = fig.add_subplot(111)
for (country, times, events) in country_series.takeOrdered(10, lambda x: -sum(x[2])):
    t = ax.plot(times, events, lw=1, c=random_color())
```



What's the big spike for the blue line above?

In [28]:
```
country_day_counts.reduce(lambda x, y: max(x, y, key=lambda z: z[1]))
```

Out[28]:
```
((u'USA', u'20010912'), 2387)
```

Looks like it was the day after September 11th.

THE NOT THAT OLD WAY

Spark SQL

SQL WITH JSON

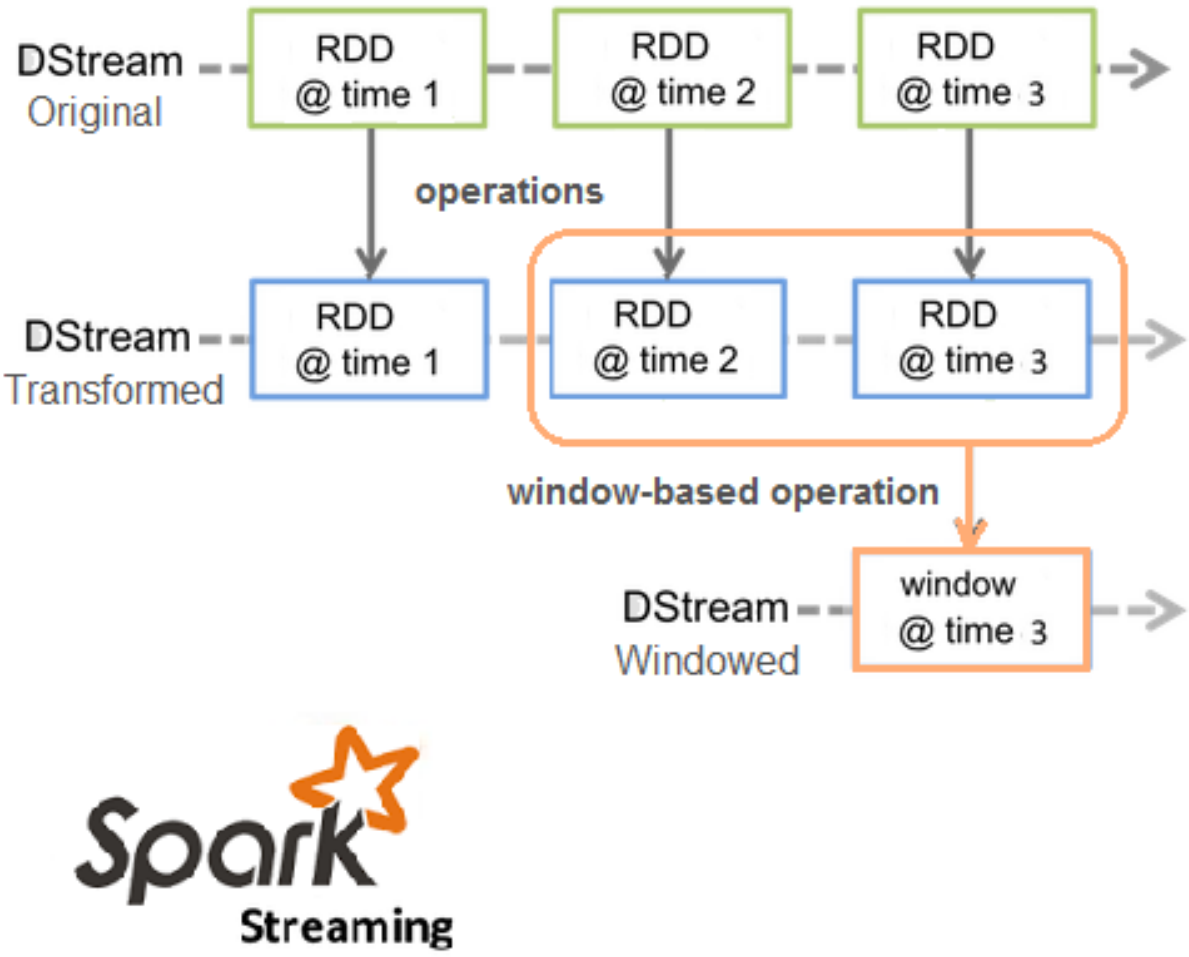
```
{"name": "Mate Gulyas", "twitter": "gulyasm"}
{"name": "John Doe", "email": "jdoe@freemail.com"}
{"name": "Jane Doe", "email": "janedoe@citromail.com"}
```

```
val input = hiveCtx.jsonFile(“example.json”)
input.registerAsTable(“users”)
hiveCtx.sql(“SELECT name, twitter FROM people;”)
```

THE LOW LATENCY WAY

Spark Streaming

DStream



Spark Streaming

THE SKYNET WAY

MLlib

I LOVE GRAPHS

GraphX

THE OTHERS WAY

Third party modules

BUT... WHERE TO GO?

On-premises
AWS
Databricks Cloud

TAKEAWAY I

Spark can provide one platform to cover most of the use-cases in data analytics

TAKEAWAY II

Productive, fast data processing framework that helps you minimize to time business impact.

THANK YOU!

MATE GULYAS
gulyasm@enbrite.ly

@gulyasm
@enbritely

enbrite.ly