# Apache Spark and the future of big data applications

Eric Baldeschwieler

# Who is Eric14?

- Big data veteran (since 1996)
- Databricks Tech Advisor
- Twitter handle: @jeric14
- Previously
  - CTO/CEO of Hortonworks
  - Yahoo - VP Hadoop Engineering
  - Inktomi – Web Search

# Last Spark Summit…

"IMO Apache Spark is the most exciting thing happening in big data today"

The potential:

- Spark - the lingua franca for data science
- Spark and Hadoop - great together

- So how are we doing?

# Spark is now bundled with Hadoop

All major Hadoop distributions include Spark






Other big data solutions too!


# Spark is in use for data science

DATABRICKS WorldCup

```
> select sum(score) as total, alpha3 as country from received where team like '%$country%'
group by alpha3 order by total desc
```

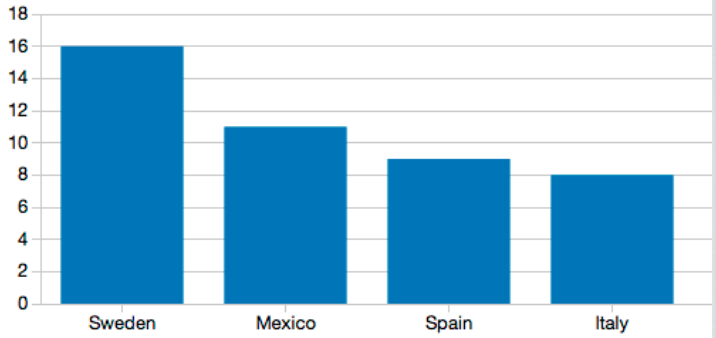
country: Norway


Plot Options...

Command took 0.27s

```
> select sum(score) as total, country from scored where team like '%$co
country order by total desc limit 5
```

country: Brazil



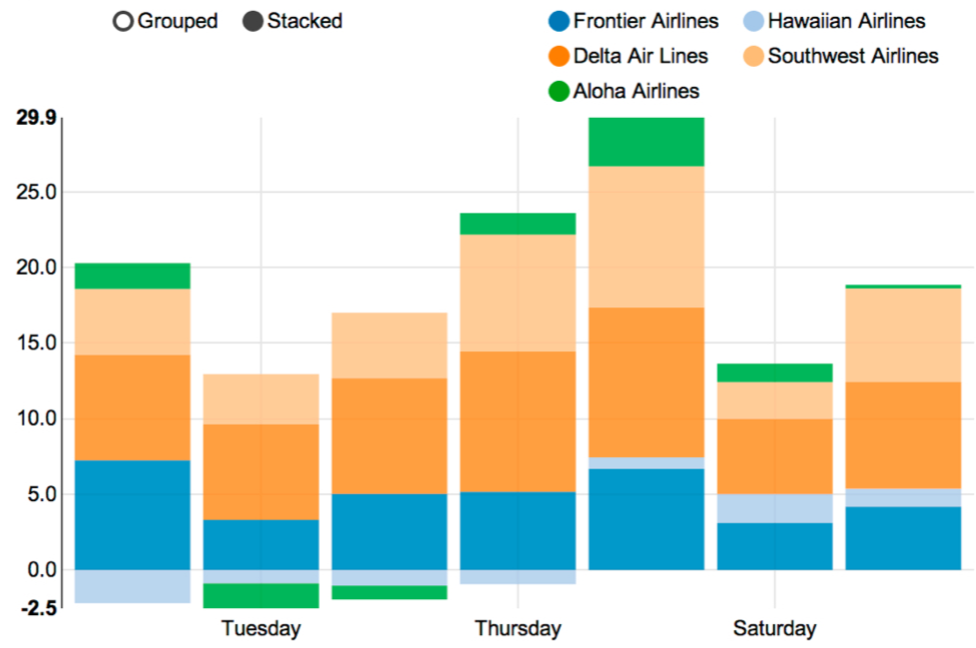

Plot Options...

Command took 0.22s

>

Shift+Enter to run

```
1 bar plot average arrdelay by dayofweek for best 5 uniquecarrier
```




ADATAO
DATA INTELLIGENCE FOR ALL

# Great progress!

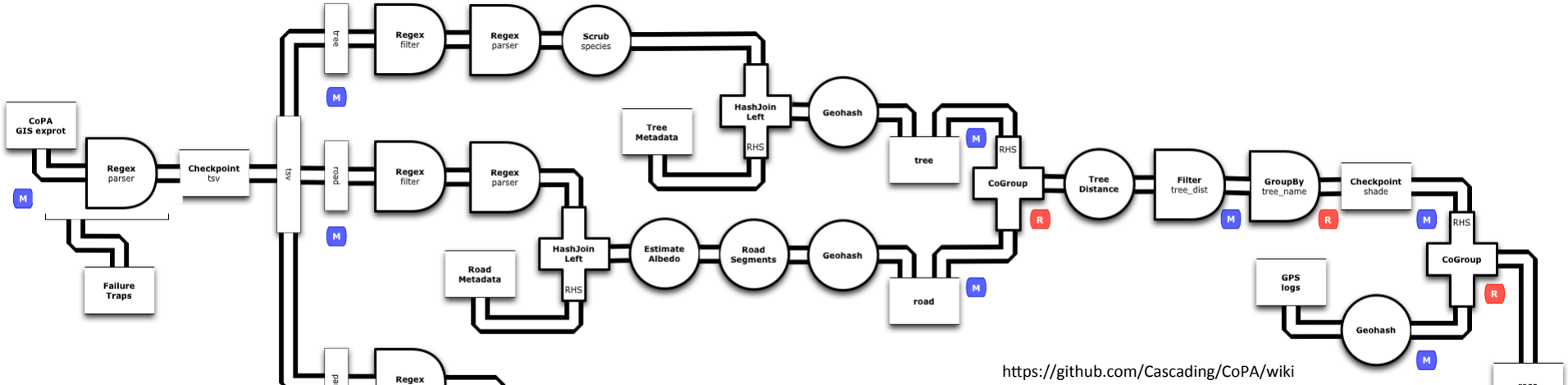
Time to declare victory?

Nope, we're only getting started.

# Making Spark great for Data Science

# Increase focus on ETL

- Science needs data, in the right place and format
- Teams are porting Hadoop ETL languages to Spark
- Better job scheduling tools
- ETL workloads are different – Scale and Throughput
  - Spark 1.0 a big step forward for these workloads
    - 1000 node spark clusters
    - petabyte scale jobs
  - Let's build benchmarks and iterate…




https://github.com/Cascading/CoPA/wiki

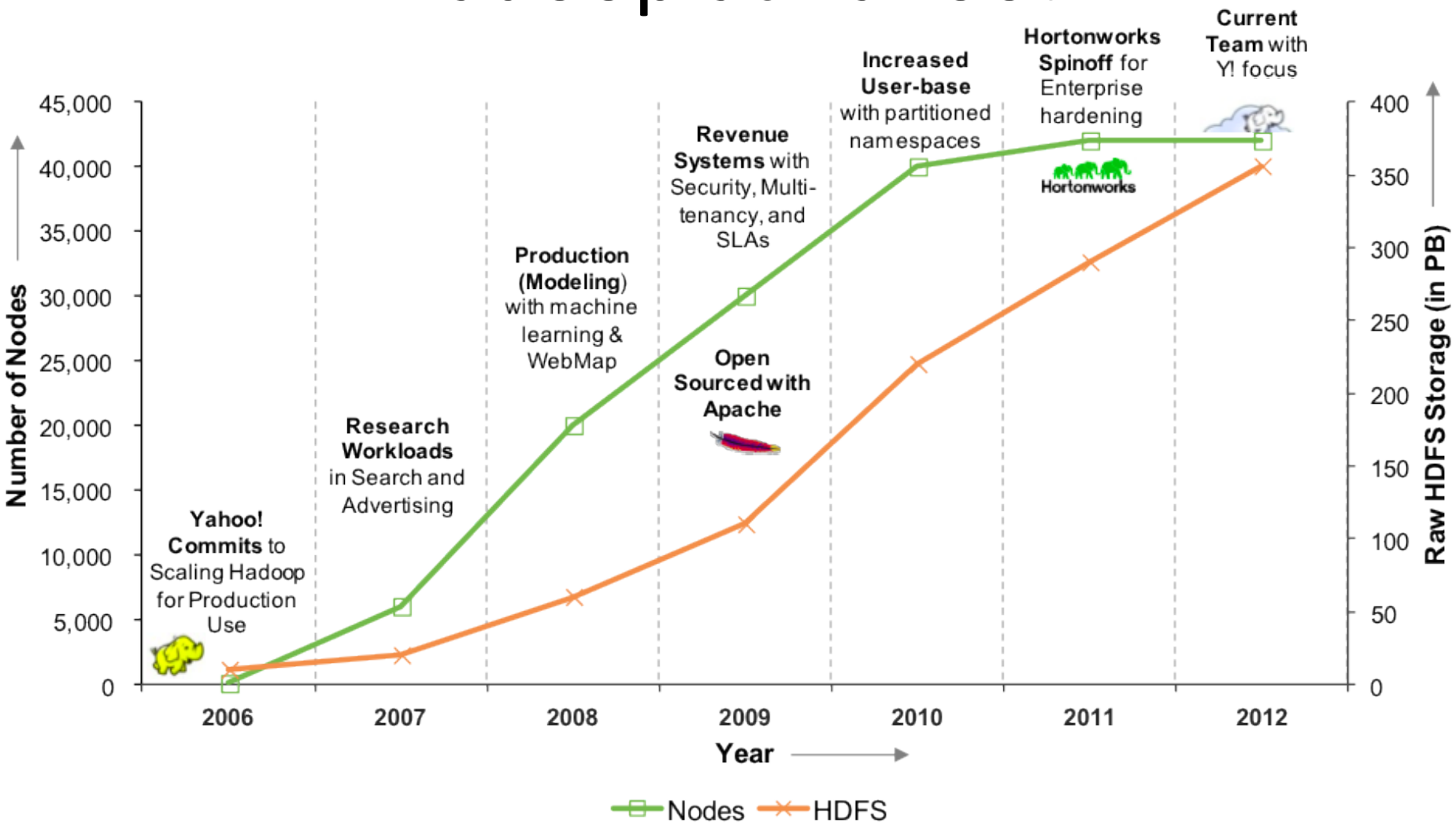
# More stuff

- R bindings!!
- Add features to ease/accelerate code sharing
- SparkSQL needs to be extended to run against more data stores, including object stores
- Deep learning and other Algo support
  - Trade off completeness for speed
  - More communication primitives?
- Developer basics
  - Profiling & debugging, error reporting & logging…

# Data science drives new applications

(some history)

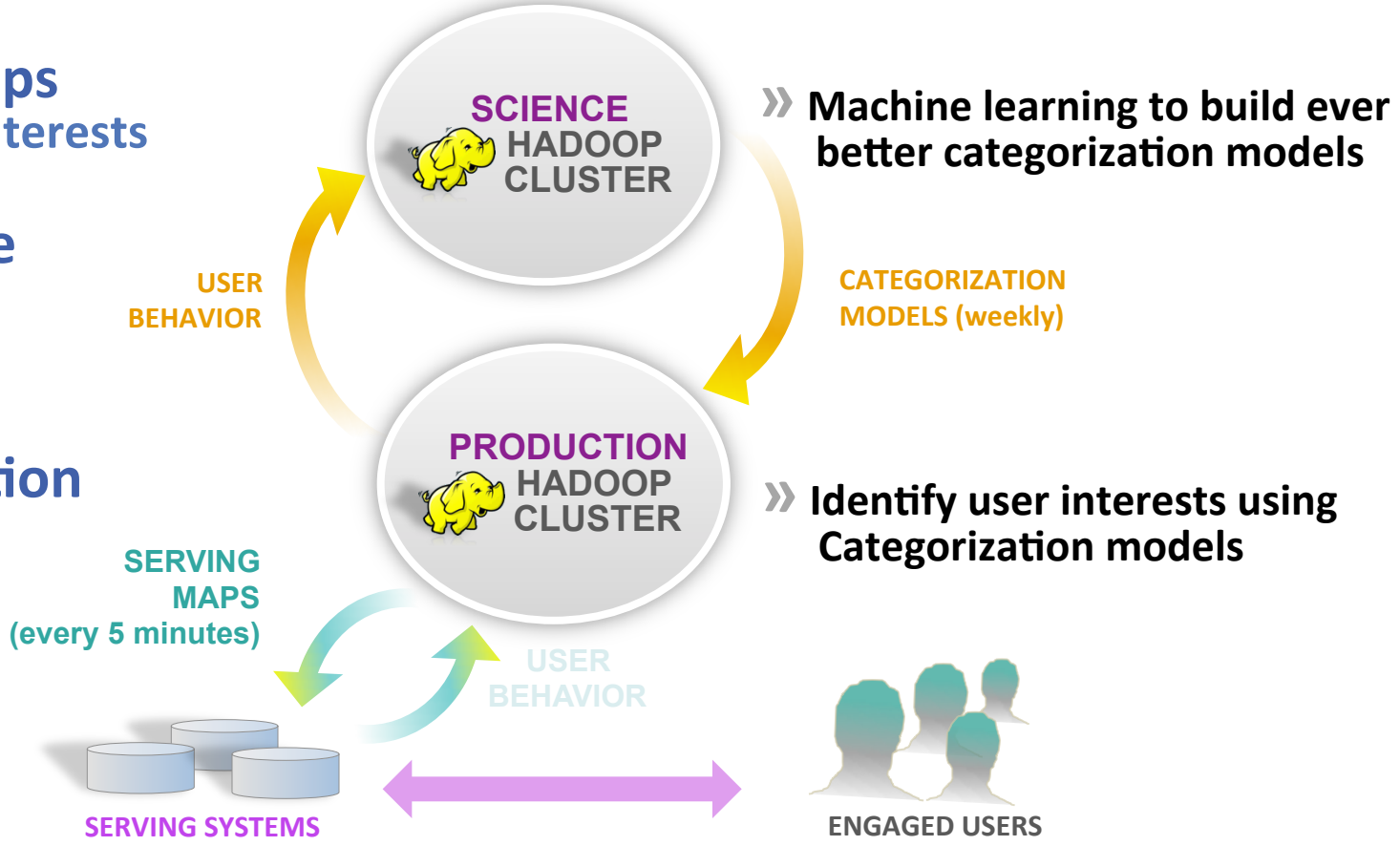
# Hadoop at Yahoo!




Source: http://developer.yahoo.com/blogs/ydn/posts/2013/02/hadoop-at-yahoo-more-than-ever-before/

# CASE STUDY
# YAHOO! HOMEPAGE

- Serving Maps
  - Users - Interests
- Five Minute Production
- Weekly Categorization models



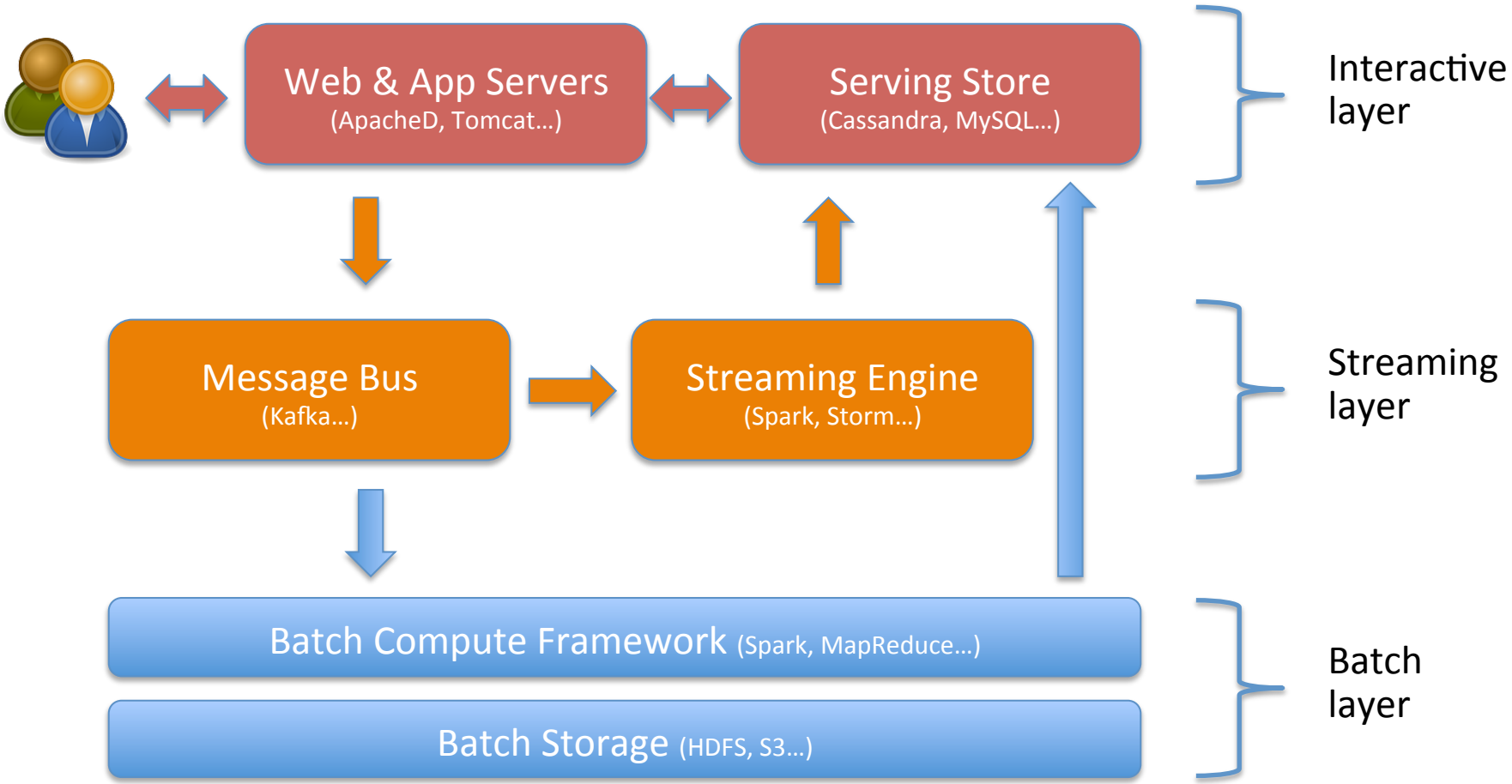

**Build customized home pages with latest data (thousands / second)**





12

# Big data application model

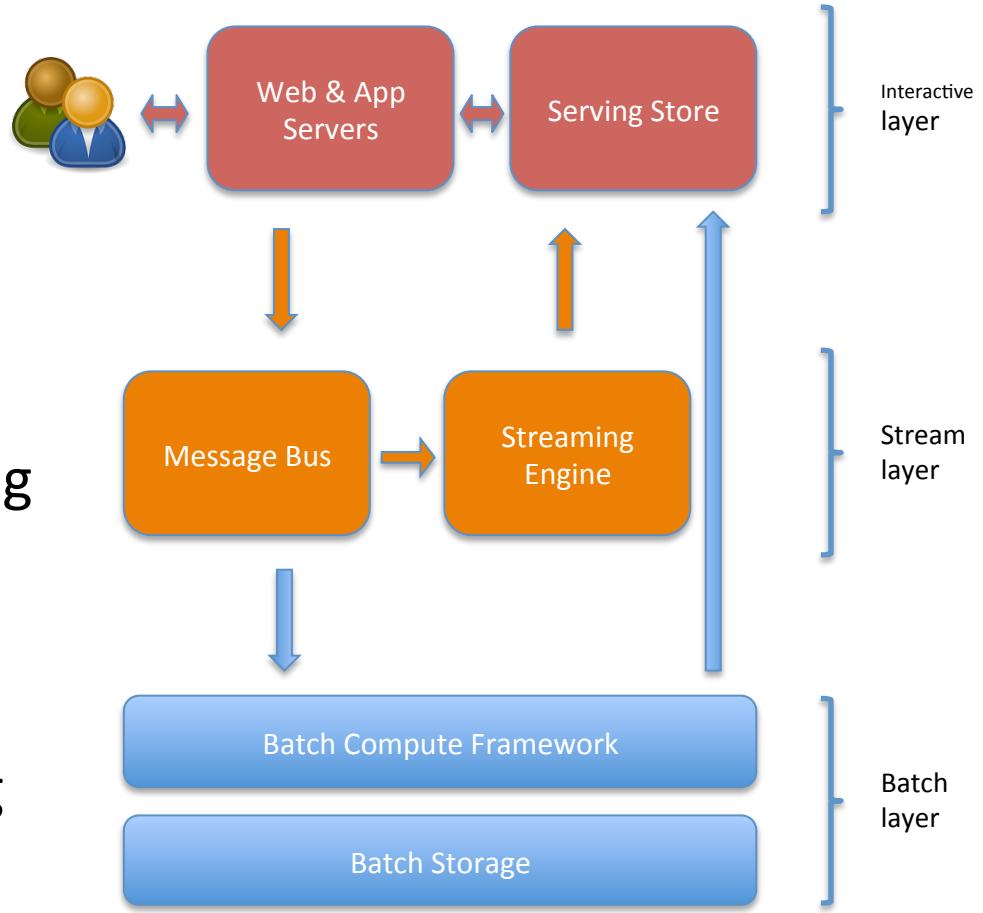


# Spark applications today

# 3rd party applications

# Classes of applications

- Custom Solutions
  - Internal apps
- Enterprise data tooling
  - ETL, BI/Query
- Data science tooling
  - Analytics & ML
  - Collaboration & Reporting
- Vertical specific applications
  - financial, healthcare, IC
  - marketing, retail, gaming



# Why so much activity?

- Spark's speed allows compelling interactivity
- Interactive API eases development
- Spark runs well in many environments
  - Cloud, Hadoop/YARN, Cassandra, …
- Broad open source community support











# Improving Spark for Applications

# Build an Open Certification Suite

- Goals of certification for an application
  - Write an app once and run anywhere
  - Apps continue running after platform upgrades
- Value of an open suite to user community
  - Contributors can validate their specific requirements
  - Drives widest possible ecosystem for applications
- Open source community “does the right thing”
  - Backwards compatibility has been a challenge in the big data ecosystem…
  - Sharing code that defines correct compatibility a win

# Tachyon / More storage APIs

- Better multi-tenancy for Spark
  - Keep objects in RAM between Jobs / users
  - Have a system wide view of RAM requirements
- Provide storage portability
  - Spark can run against S3, HDFS, Cassanda…
  - Can we make this transparent to applications
- Better use of tiered storage
  - RAM, SSD and Disk

“IMO @ApacheSpark is the most exciting thing happening in big data today”

Enjoy the show!
-@jeric14