Running Spark in Production

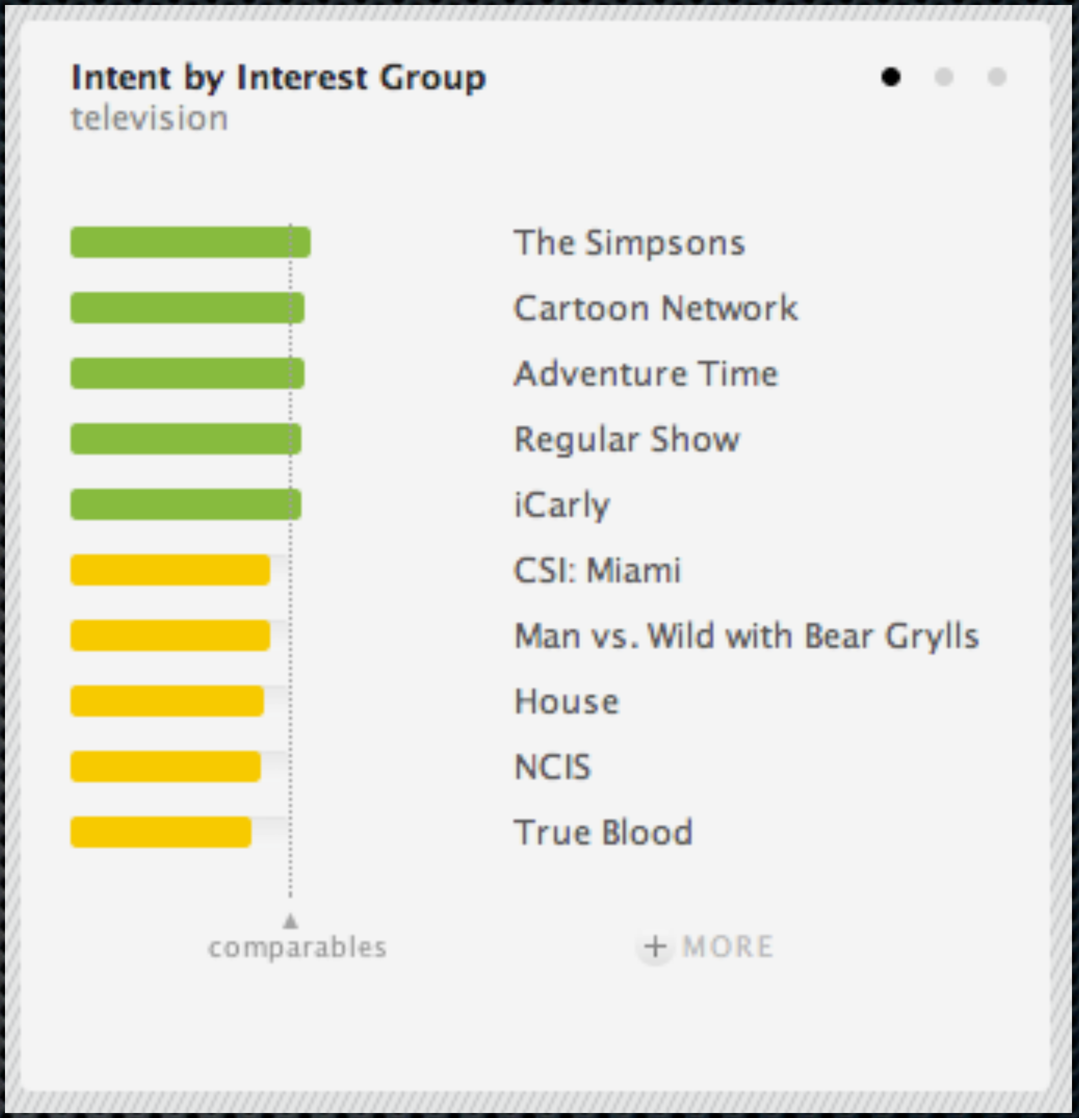
Erich Nachbar
Quantifind

Agenda

- Quantifind?
- Hadoop? No, sir!
- Choices: SLA, MTBF & Money
- Are we there yet?
- Shoot!

Quantifind?

- **Products in several verticals**
  - Movie, Gaming, ...
- **Input Sources**
  - Product Reviews,
  - Comments / Tweets, ...
- **Predicts using (un-)/structure data**
  - Box Office Opening $, Customer Satisfaction, ...
- **Computes the Intentful Audience and its demographics**



Yes, we are hiring!

jobs@quantifind.com

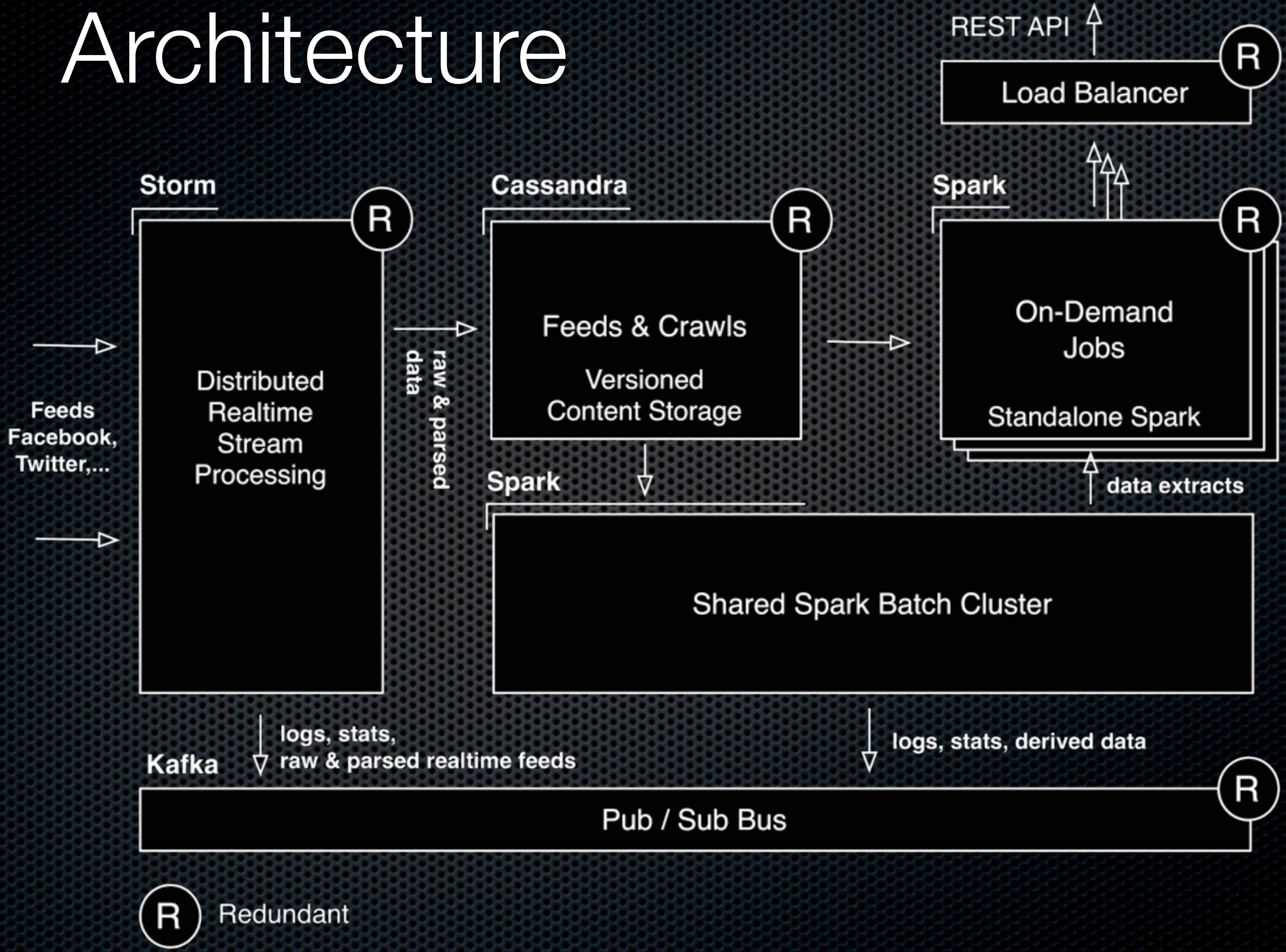
Hadoop? No sir!

- **Development Velocity**
  - Quicker iteration cycles than Hadoop
  - Concise Scala code (10x smaller than Pig UDFs)
  - Excellent for embedding (unlike Pig)
- **Runtime**
  - Orders of magnitude faster for cacheable data sets
  - Forced MapReduce disk spills kill iterative perf.
  - Jobs latency is short. Enables new product features.

SLA, MTBF & Money

- **Architecture Considerations**
  - What is noticeable? - “get out of bed”
    - UI not available
  - What is not (immediately)? - “I have a few hrs”
    - UI Data stale
  - What is irreplaceable? - “Oh, no! Bieber tweets...”
    - Streamed data

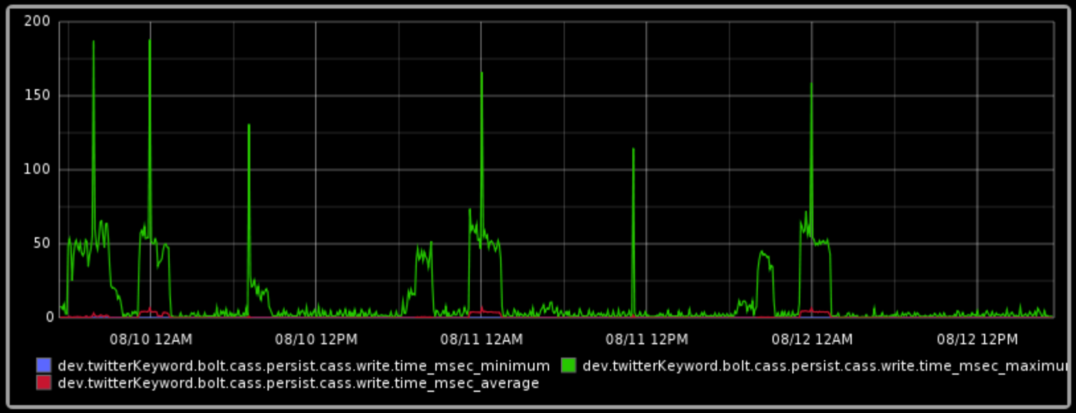
Architecture



Are we there yet?

- **Invest in**
  - External Health Pings
    Like: UptimeRobot (free)
  - Measuring everything (counts, timings,...)
    Like: Twitter Ostrich with Kafka & Graphite
  - Log centrally
    Like: Graylog2, Logstash
  - Automatic Service Restarts
    Like: Supervisord

```
Stats.time("cass.save") {
   cass.save(key, result)
}
```



What’s next?

- **Streaming & Batch Support in a Single System**
    - “Because coding it twice is lame”
        - Spark Streaming
        - Storm Trident
- **RAM Grids**
    - Core i7 RAM = 20GB/s, 20 ns
Fast SSD drive = 0.5 GB/s, 100,000 ns
        - Spark
        - ???

Questions?

Thank you!

erich@quantifind.com

Appendix