

Spark and Hadoop at Yahoo: Brought to you by YARN

Andy Feng
Yahoo! Hadoop
(afeng@yahoo-inc.com)



Personalized Web

YAHOO!

Search



Hi, Andrew



Mai

Show Ad

- Mail
- My Yahoo!
- Finance
- Flickr
- Games
- Messenger
- Movies
- Music
- omg!
- Sports
- Weather
- Autos
- News
- Shine

More Y! Sites >

Facebook

Gmail

WSJ

NPR

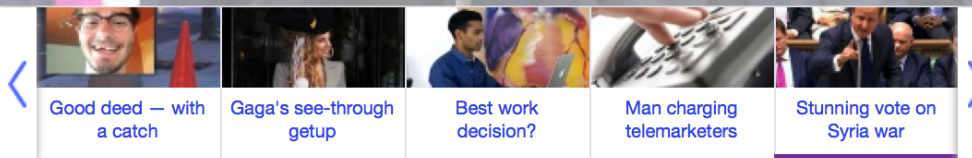
More Favorites >



British prime minister loses Syria war vote

David Cameron says it is clear to him that the British people do not want to see military action. [His response to Parliament »](#)

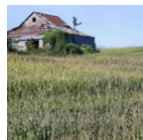
1 - 5 of 95



All Stories News Local Entertainment Sports More ▾

Yosemite Rim Fire: why there's more optimism about taming it

Although the Yosemite Rim Fire is still raging and expanding – now at 193,000 acres – there are reasons for optimism, several officials and fire experts say. But given that Yosemite is a [Christian Science Monitor](#)



Midwest hot, dry spell brings back drought worries

DES MOINES, Iowa (AP) — A growing season that began unusually wet and cold in the Midwest is finishing hot and dry, renewing worries of drought and its impact on crops.

[Associated Press](#)

Hortonworks to seek IPO within two years, CEO says

(Reuters) - Enterprise software start-up Hortonworks plans to seek a public listing in 15 to 24

Trending Now

[Watch the show »](#)

- [Sandra Bullock](#)
- [Military strike Syria](#)
- [Leah Remini](#)
- [Tylenol warnings](#)
- [Scud missiles](#)
- [Colorado marijuana ...](#)
- [Harley-Davidson 11...](#)
- [Holy Grail Jay Z](#)
- [Indian rupee](#)
- [Pressy smartphones](#)

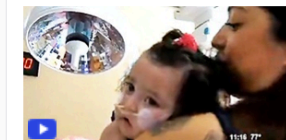
Featured Videos Y! Screen



[90-year-old steam train runs on unusual fuel](#)



[Student sets new McNuggets world record](#)



[Revolutionary device keeps baby alive](#)



[Dramatic waterspout forms off Croatian coast](#)

Search videos

Sunnyvale

76°F Fair

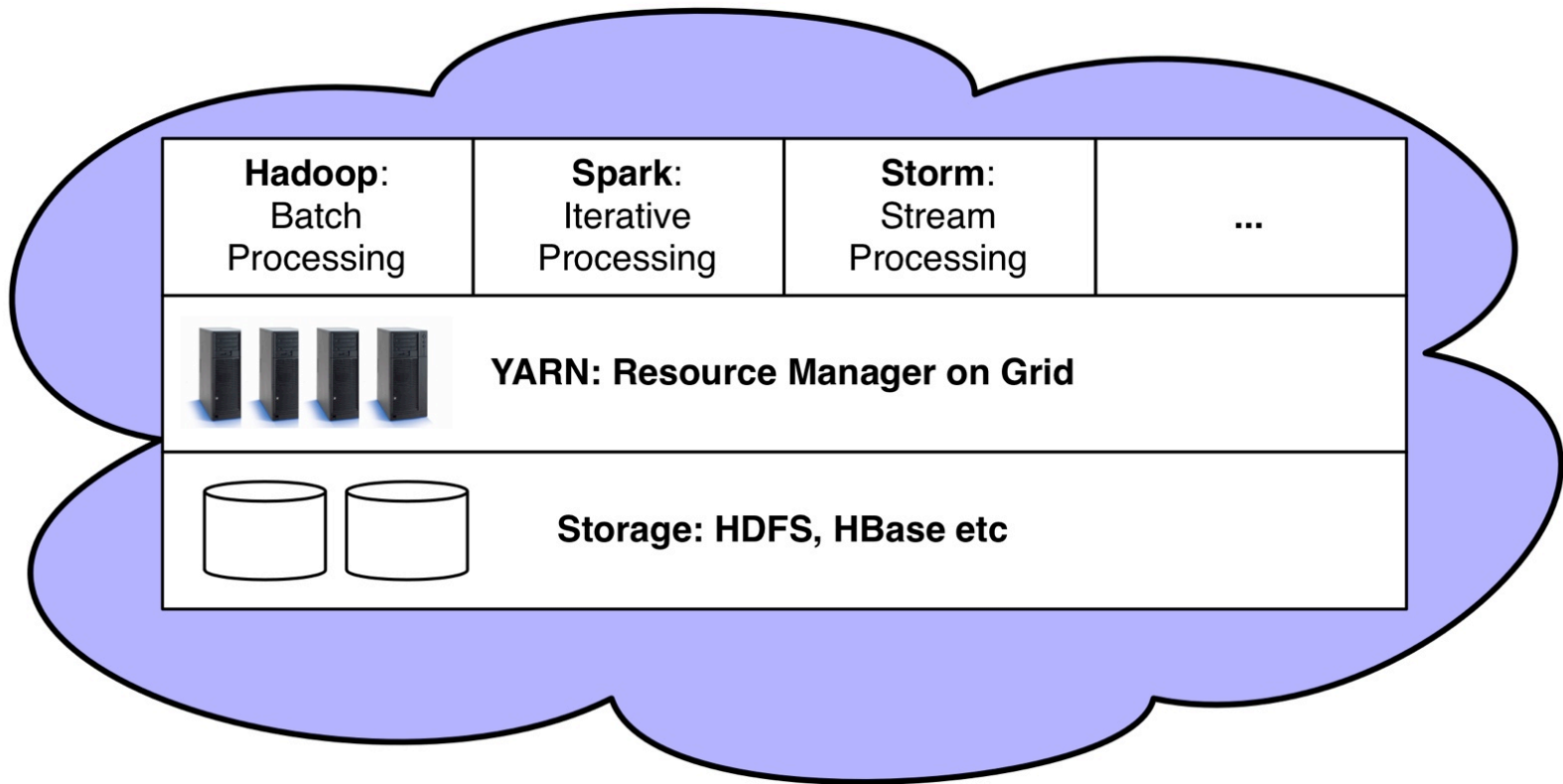


YAHOO!

Big-Data in Yahoo!



Hadoop + Spark: Empowered by YARN



30k+ Yahoo! production nodes on YARN since Q1 2013

YAHOO!

Shark Pilot: Advertising Data Analytics

- **Business questions**

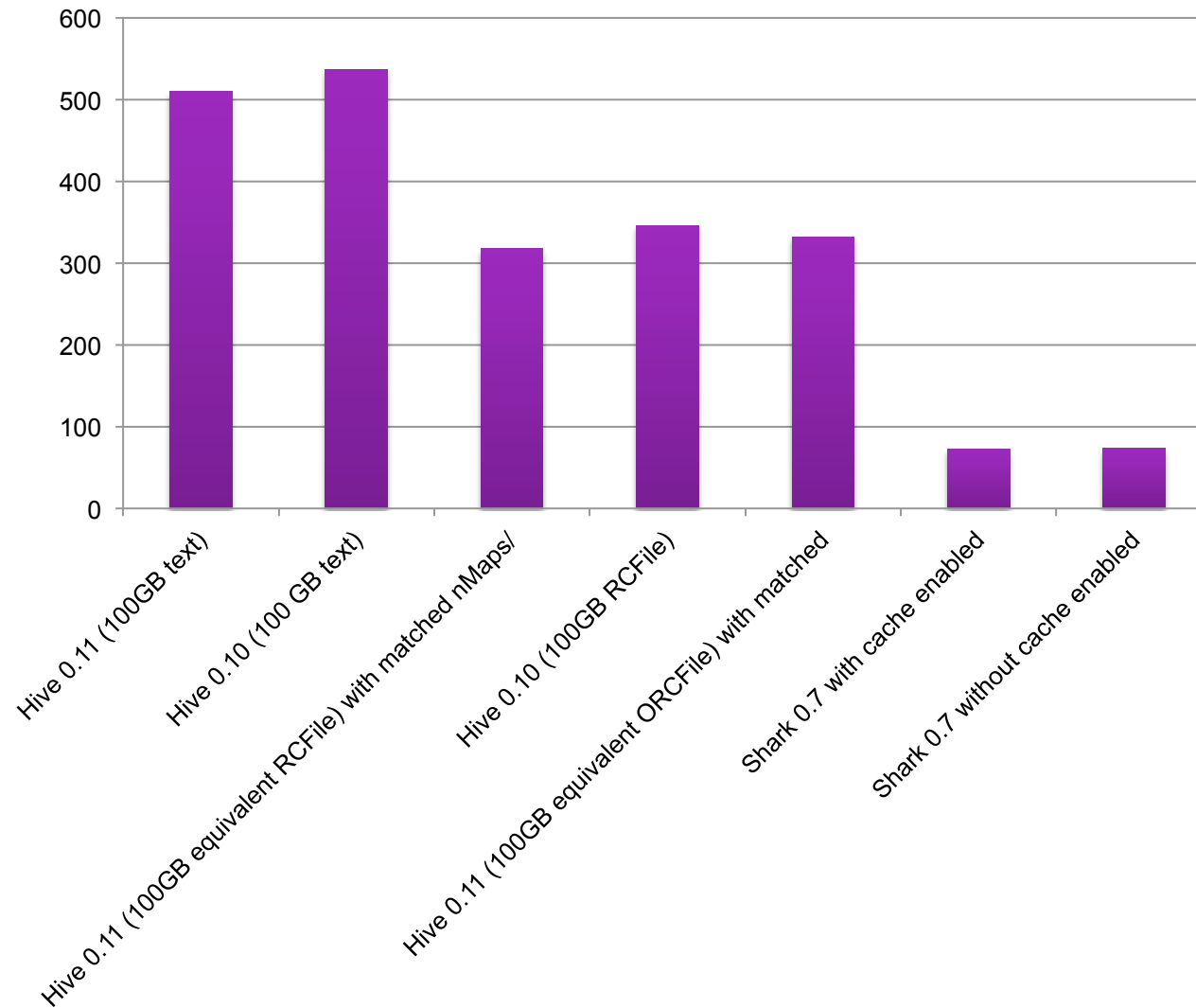
- › Are two sets of audience cohorts similar to each other?
- › What audience segment is most likely to be interested in this ad campaign?
- › In what way was the new front page rollout different than the previous front page as far as audience engagement goes?
- › What are the right metrics to define user engagement?

- **Shark pilot**

- › 50 nodes, each w/ 96GB RAM
 - Currently loaded w/ 3.2 TB sample data in memory
- › Homegrown BI tools for ad-hoc queries
 - Using [Shark Server](#) (contributed to community by Yahoo!)

Shark Perf: TCP-H Benchmark

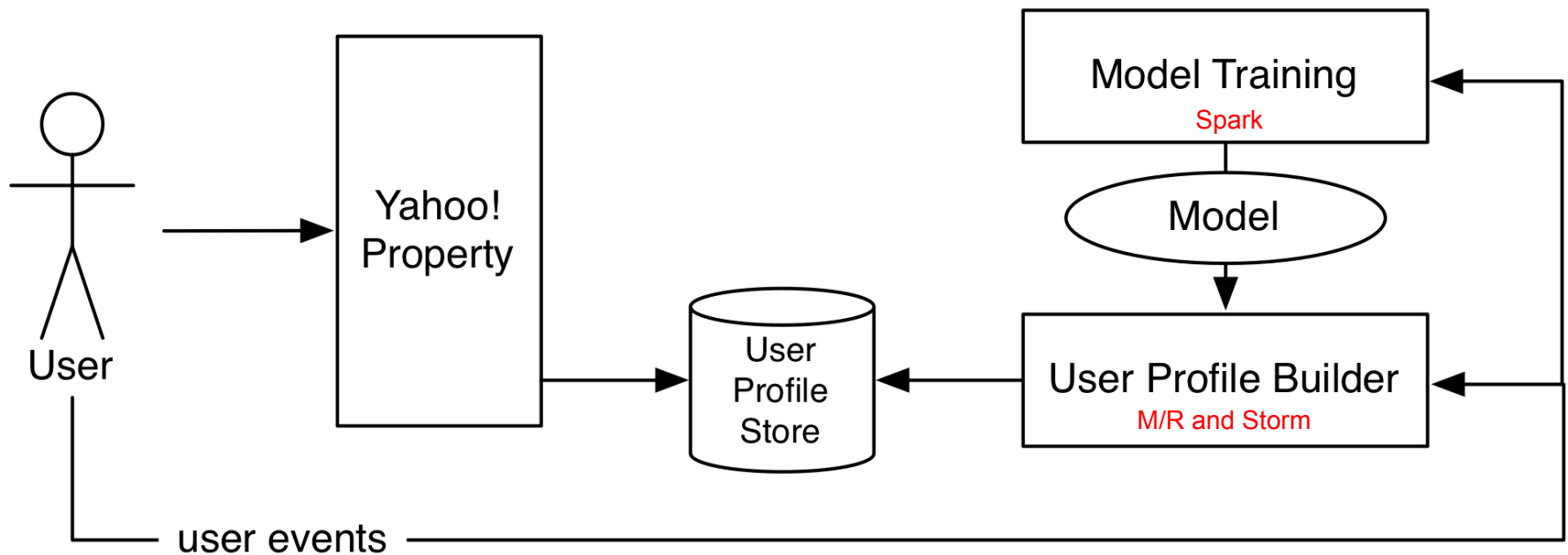
Average
Seconds



Spark Pilot: Model Training Pipeline

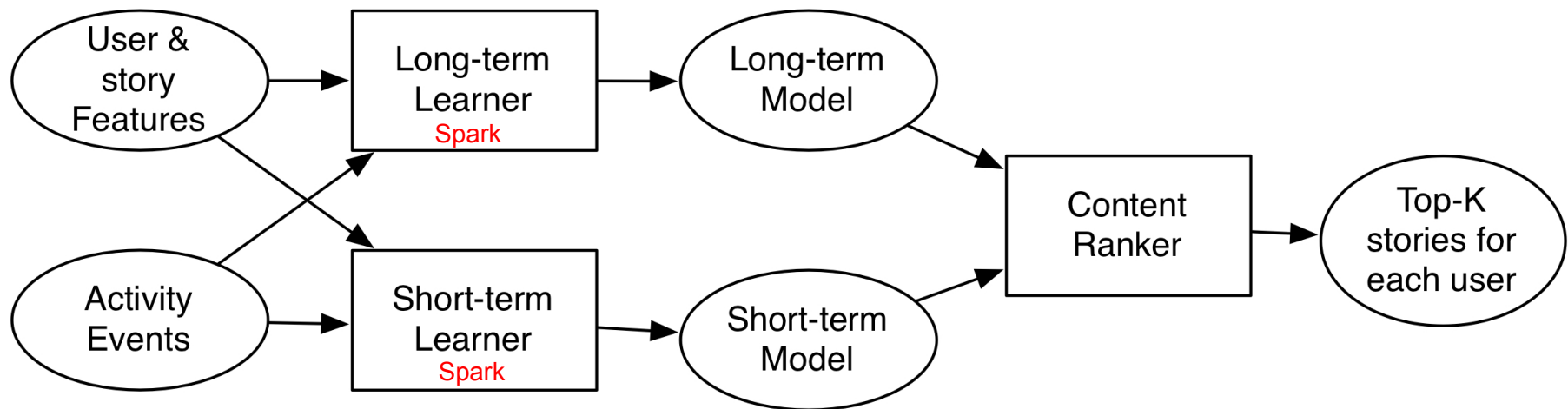
- A DAG of M/R jobs in Hadoop Streaming
 - › Feature extraction
 - › Train models
 - › Score and analyze models
- Initial Spark prototype
 - › 3x speedup on feature extraction
- Production launch
 - › Apply Spark against complete pipeline
 - › Spark on 80 node cluster
 - Thanks to the enhanced UI and metrics in Spark 0.8

Use Case: Ad Targeting



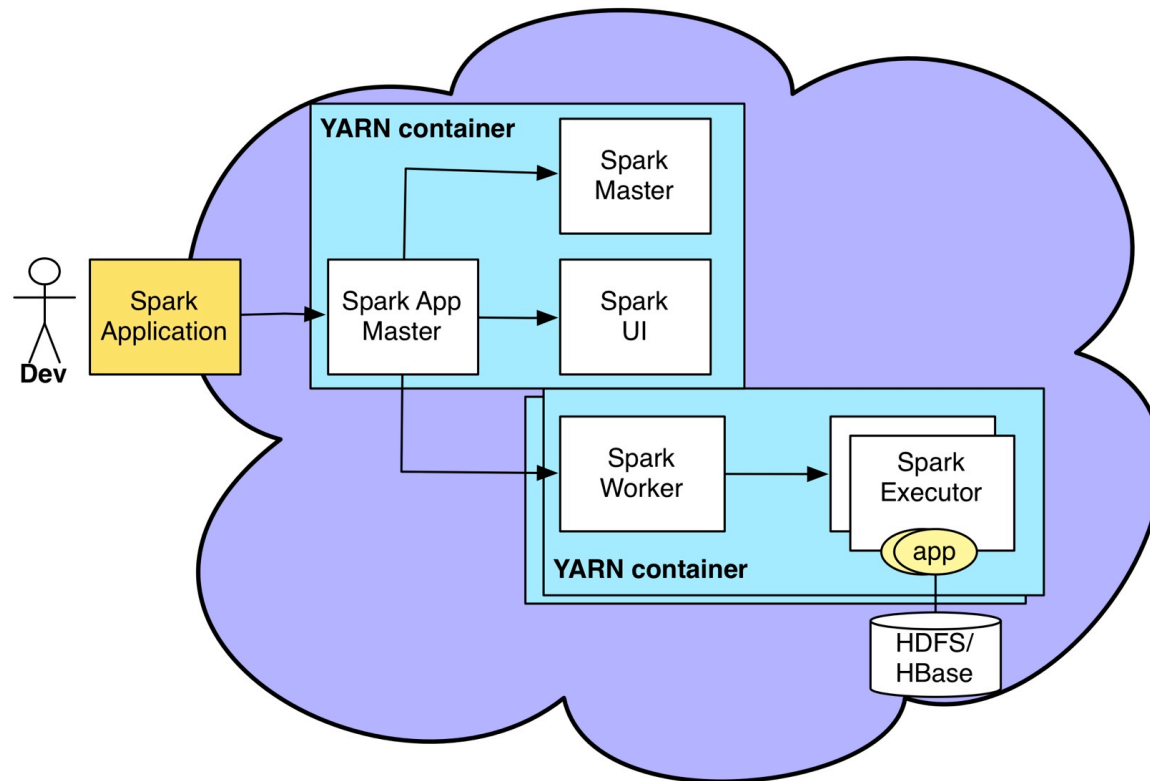
Use Case: Content Recommendation w/ Collaborative Filtering

Input **CF Learning** **Ranking** **Output**



Spark-YARN: Deployment Simplified

```
run spark.deploy.yarn.Client --jar ... --class ... --args ...  
--queue ...--num-workers ... --worker-memory ...
```



Spark-YARN (contributed by Yahoo!) is being adopted by community (ex. Taobao) for production use. You should try it on your Hadoop cluster.

Acknowledgement

- AMPLab team
 - › Outstanding collaboration: Ion, Matei, Reynold, Patrick, Matt, ...
- Yahoo! Hadoop team
 - › Thomas, Bobby, Paul, Rajiv, Mithun, ...
- Yahoo! Lab.
 - › Mridul, Nathan, ...
- Yahoo! data analytics
 - › Supreeth, Ram, Tim, ...
- Yahoo! spark users
 - › Gavin, Jay, Hirakendu, ...

We Are Hiring!

<http://careers.yahoo.com/>

