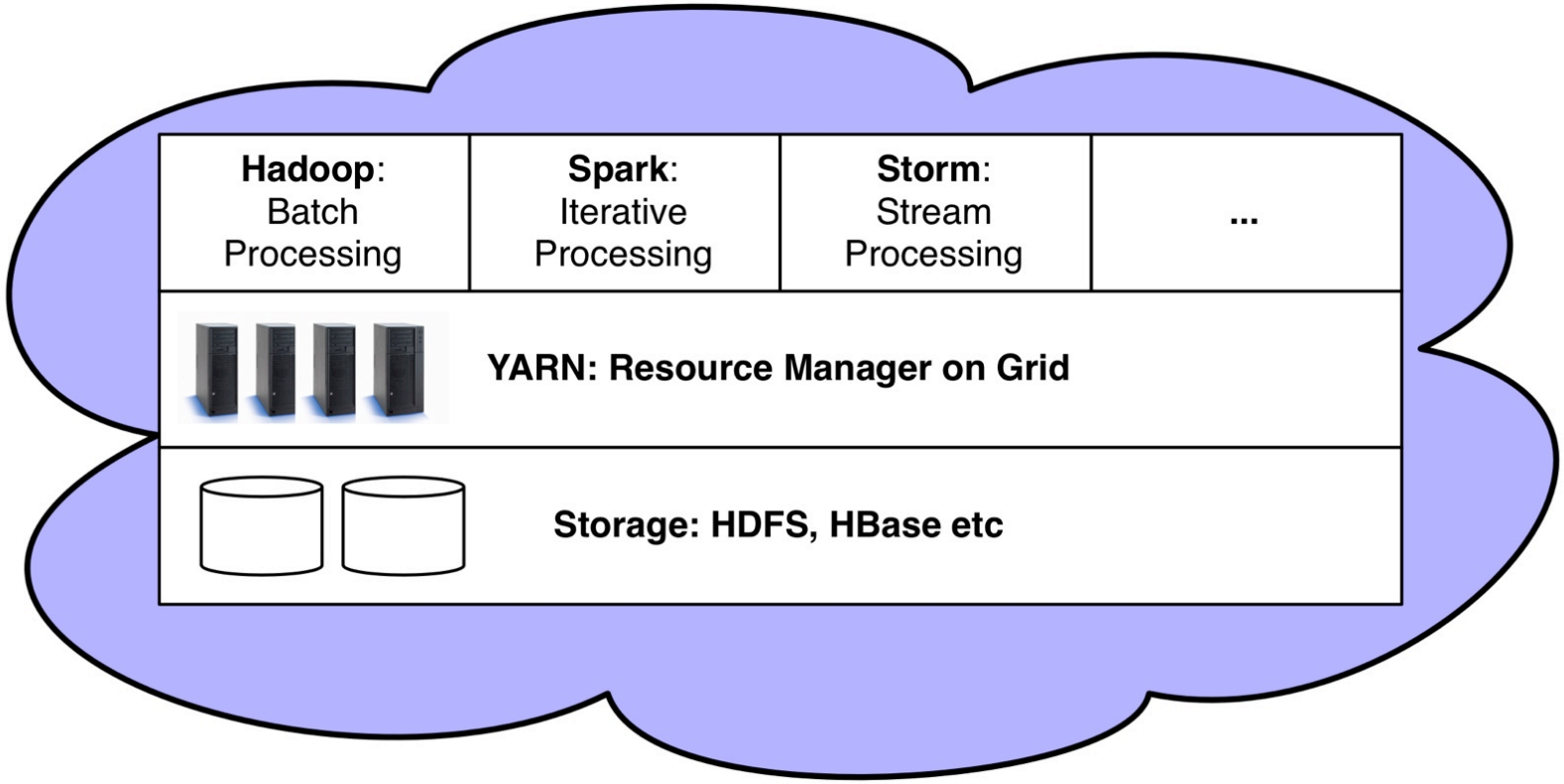
Spark and Hadoop at Yahoo: Brought to you by YARN

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

Personalized Web

Big-Data in Yahoo!

Hadoop + Spark: Empowered by YARN



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

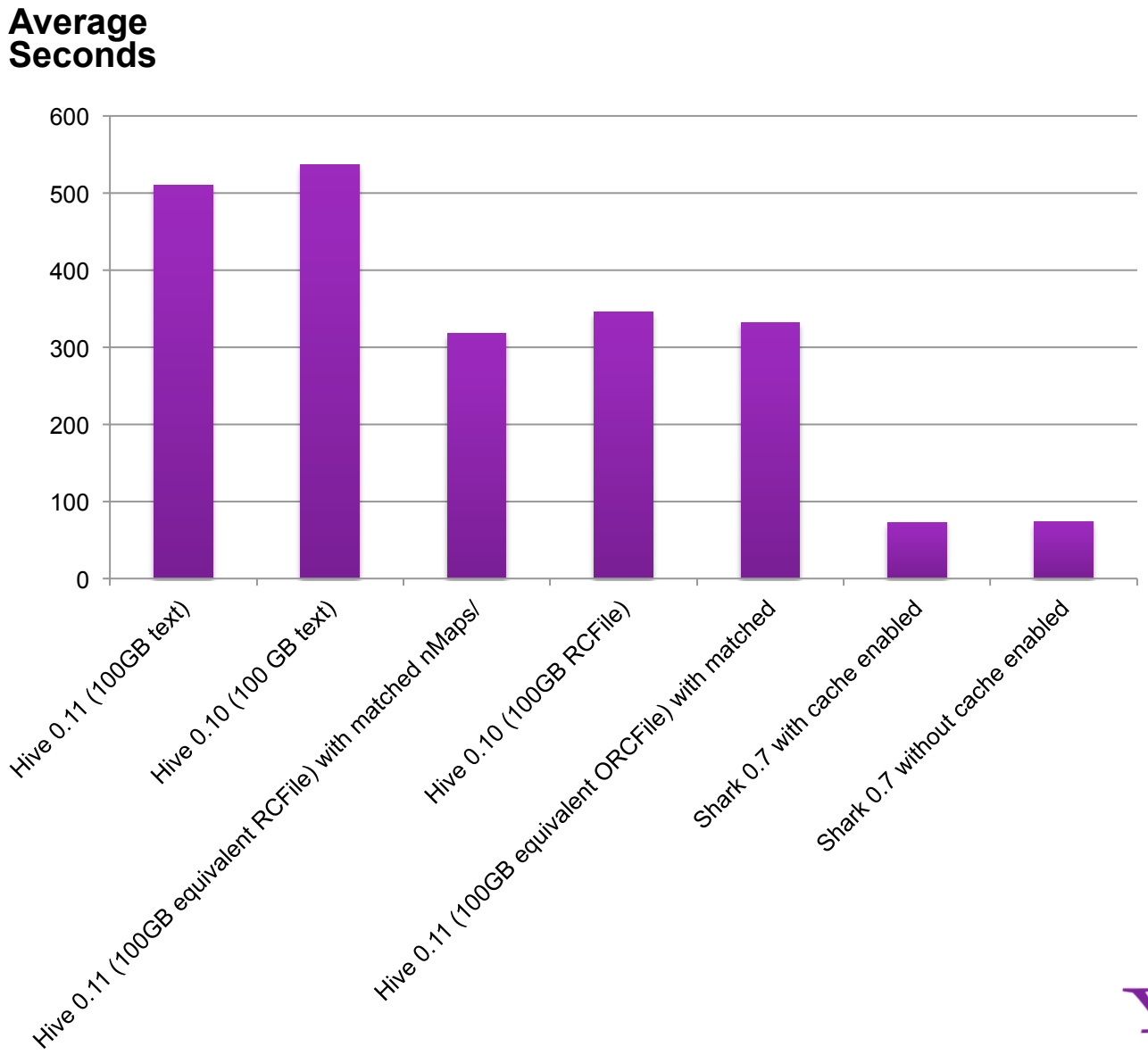
YAHOO!

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!)

YAHOO!

Shark Perf: TCP-H Benchmark

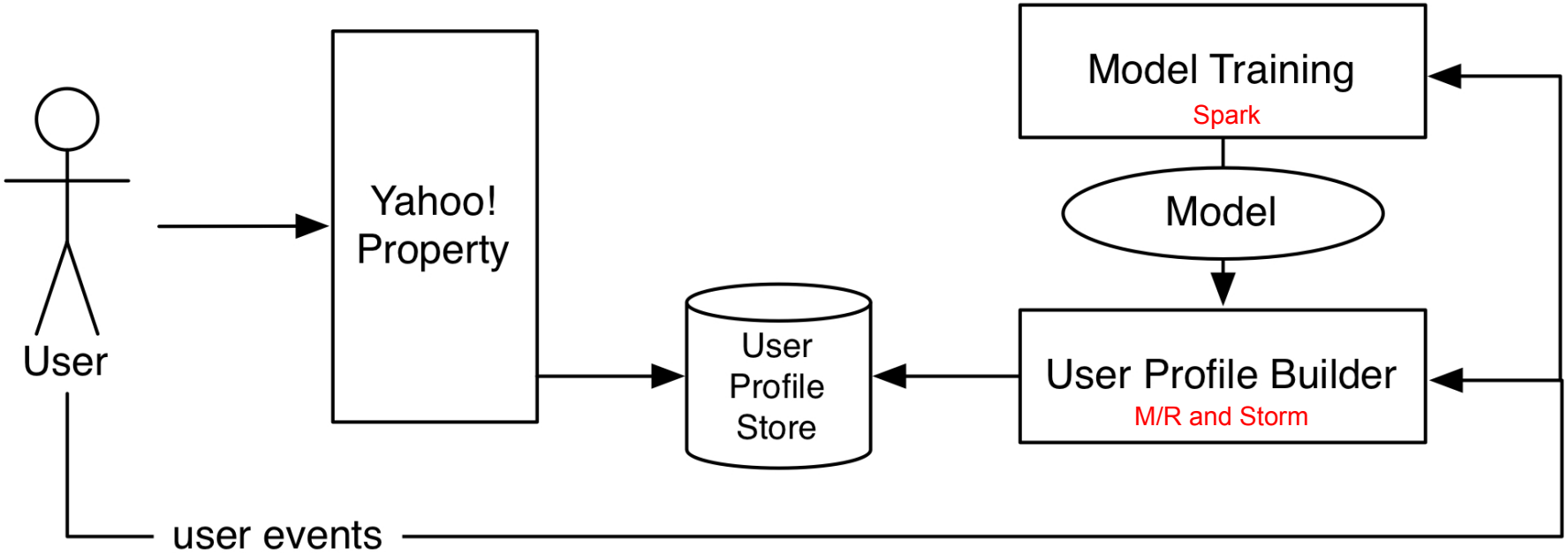


YAHOO!

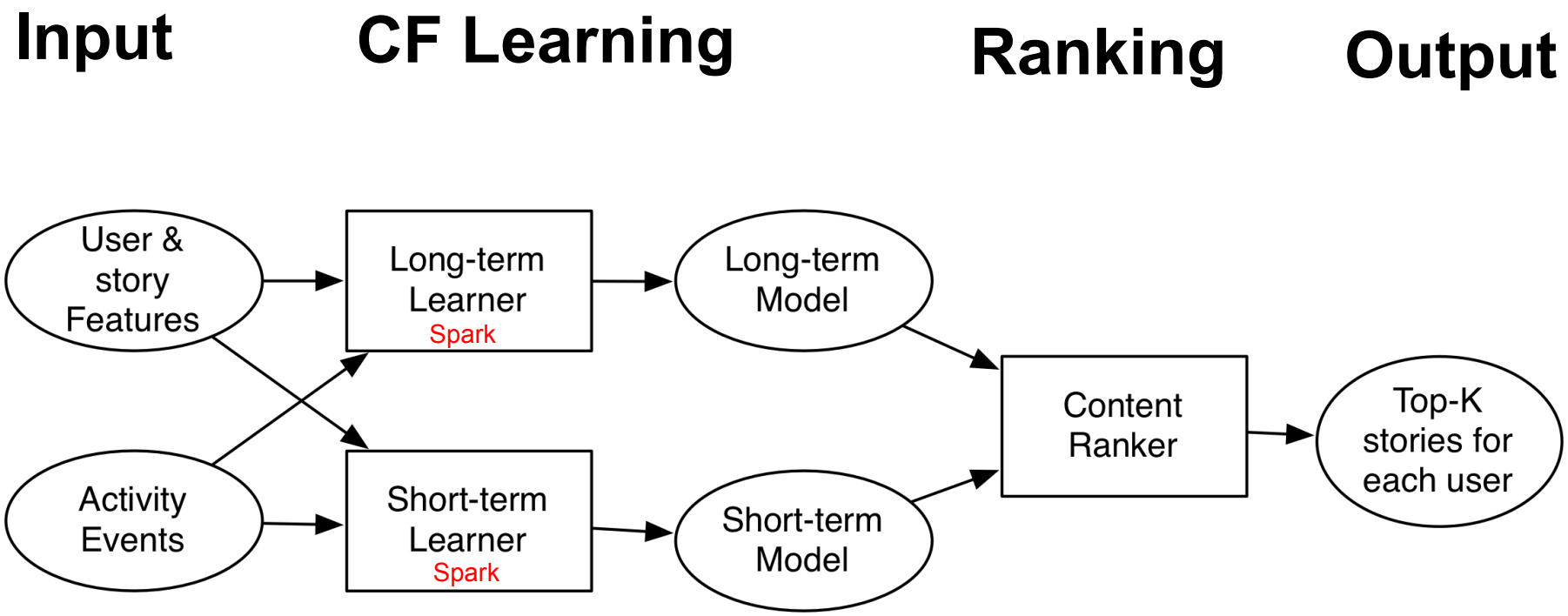
Spark Pilot: Model Training Pipeline

- 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



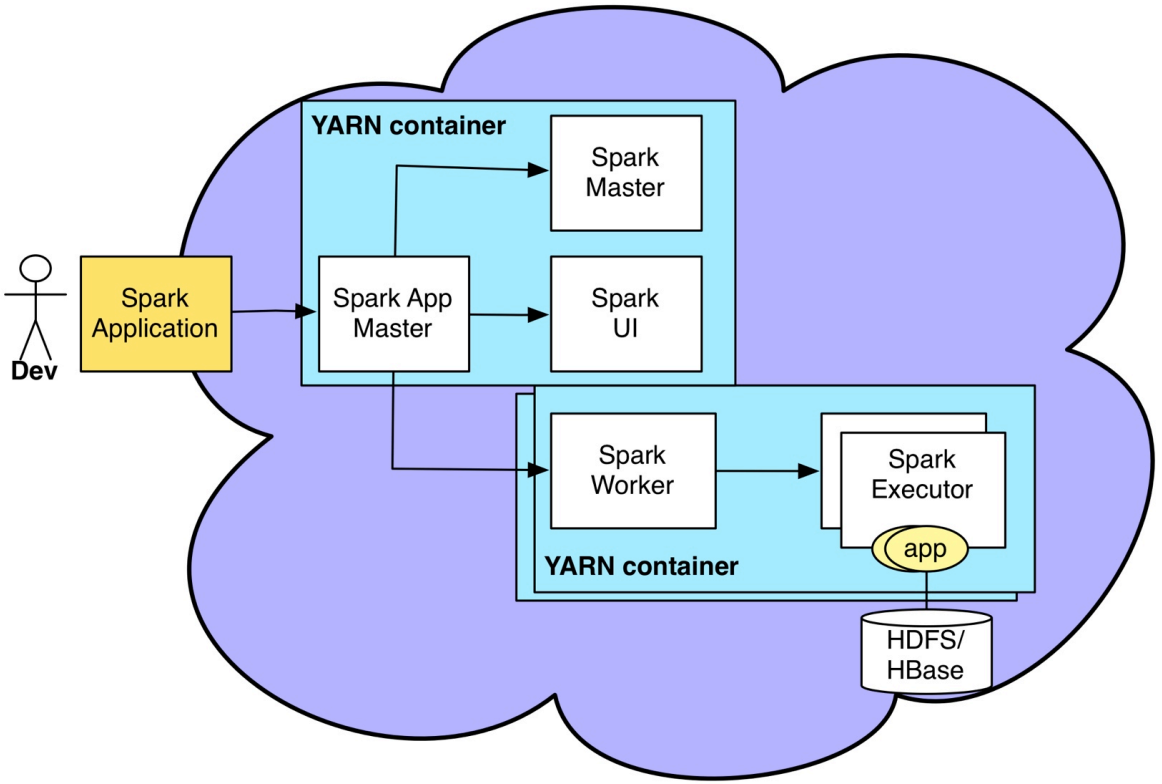
Use Case: Content Recommendation w/ Collaborative Filtering



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.

YAHOO!

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/