

HADOOP AND SPARK JOIN FORCES AT YAHOO

Andy Feng (afeng@yahoo-inc.com)
Distinguished Architect, Platforms, Yahoo

YAHOO 2012: TODAY MODULE



<http://visualize.yahoo.com/core/#>

YAHOO 2013: PERSONALIZED HOMEPAGE

http://www.yahoo.com

Mobile

YAHOO!

Search

My Yahoo

Hi, Andrew

Mail

Mail

My Yahoo

Finance

Flickr

Games

Messenger

Movies

Music

omg!

Sports

Weather

Autos

News

Shine

Shopping

More Yahoo Sites >

Facebook


Gmail

WSJ

NPR

More Favorites >


Make YAHOO!




7 high-pay careers — no grad school required

The high \$80K median salary of this job is given for working creatively under extreme pressure. Why a bachelor's is enough »


1 - 5 of 50




Suspension in NFL 'bully' case




Hudson's sparkling gown



Scrambled eggs 3 ways



Stars' flirty print dresses



High pay, no grad school

All Stories


News

Local

Entertainment

Sports

More



Rallying for McAuliffe, Obama tears into tea party


ARLINGTON, Va. (AP) — President Barack Obama cast Republican Ken Cuccinelli on Sunday as part of an extreme tea party Republican faction that shut down the government, throwing the political weight of the White House behind Democrat Terry

Associated Press

Google, Apple and other tech giants look to a post-'cookie' era

Google and other online companies are reportedly experimenting with cookie alternatives, in part because smartphones and other mobile gadgets don't support that tracking method.

San Jose Mercury News



Find Your High School Yearbook

View class yearbooks online free. Reminisce & buy a reprint today.

Classmates.com Sponsored

Woman Offering Thanks To Vets Finds It's Not Always Welcome

Trending Now

Watch the show »

1 Ben Roethlisberger

2 Pamela Anderson

3 The Rolling Stones

4 The fox SNL

5 New iPad mini


6 Tatyana McFadden

7 Michelle Pfeiffer

8 Diwali

9 Houston Texans

10 Tropical Storm Sonia



15%

...need I say more?

Get a quote.


GEICO

Ad Feedback

AdChoices


Cupertino

55°F Partly Cloudy




Today

65° 42°



Tomorrow

69° 43°



Tuesday

71° 49°


Quotes

Yahoo Finance

Verizon LTE


9:05 AM

YAHOO!



More than 100 dead in Philippine typhoon

Yahoo News




Microsoft's Tablet Ecosystem Has Just 1 Issue

Motley Fool


Five9 Review – Cloud Contact Center Software

Business 2 Community




The Weekender: cloning cats, revisiting replicants, ...

The Verge




They call it a 'bacon odorant,' not deodorant

Yahoo Shine



Exclusive: Supporters of China's disgraced Bo Xilai ...

Reuters



1,200 feared dead in typhoon-devastated ...

3

YAHOO!

Monday, December 2, 13

YAHOO 2013: PERSONALIZED PROPERTIES

http://finance.yahoo.com

The screenshot shows the desktop version of the Yahoo! Finance website. At the top, there's a search bar and buttons for "Search Finance" and "Search Web". Below the search bar, a list of stock tickers is displayed with their respective percentage changes: YHOO (+3.15%), TWX (+3.47%), AAPL (+1.57%), GOOG (+0.80%), MSFT (+0.75%), FB (-0.06%), CSCO (+1.73%), HPQ (+0.97%), TWC (+2.40%), and INTC (+0.13%). A "Quote Lookup" search bar is also present. The main content area features a large section titled "Asian shares, currencies fall on Fed tapering worries" with a sub-headline "REPLY 25.261 -0.439 (-1.71%)". This section includes a table of Asian market data and a brief article snippet. Below this, there are four smaller news items with images: "Analysis: U.S. retailers tread tight path in shortened holiday...", "Transocean reaches deal with Icahn to resolve proxy battle", "Analysis: As Alabama flames fade, new oil-by-rail questions arise", and "Global Economy: Surprise tactics sweep central banking". The bottom section contains several articles, including "How Nick D'Aloisio Has Changed the Way We Read", "Can Hewlett-Packard Dethrone 3D Systems?", "[video] A New Look for Your iPhone: Curved Screens", and "Wolff: Happiness comes to Rupert Murdoch". The left sidebar includes links to "Finance Home", "My Portfolio", "Markets", "Company News", "Economic News", "Personal Finance", "Yahoo Originals", and "CNBC". At the bottom left, there are advertisements for Ameritrade, shareBuilder, E*TRADE, and Scottrade.

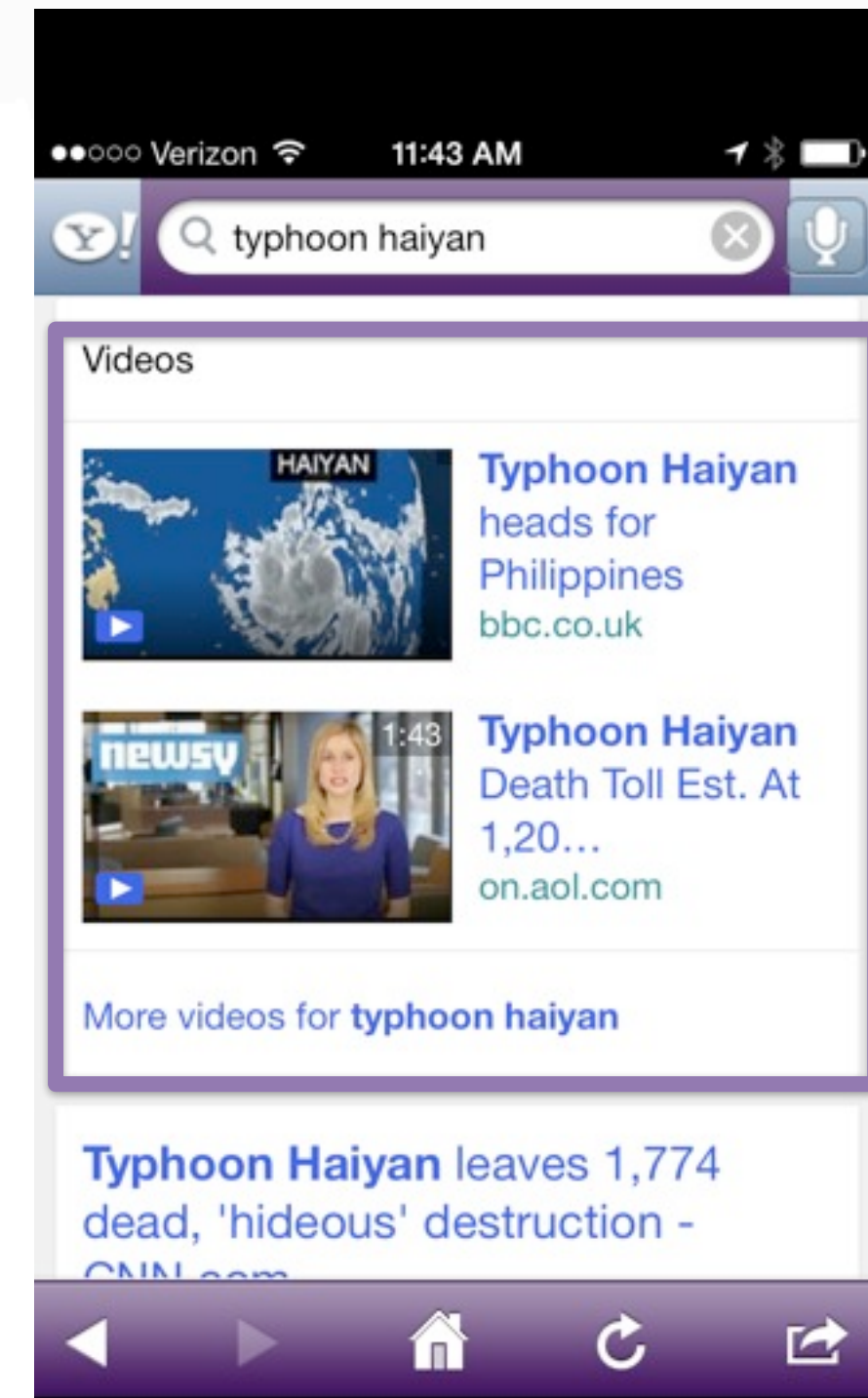
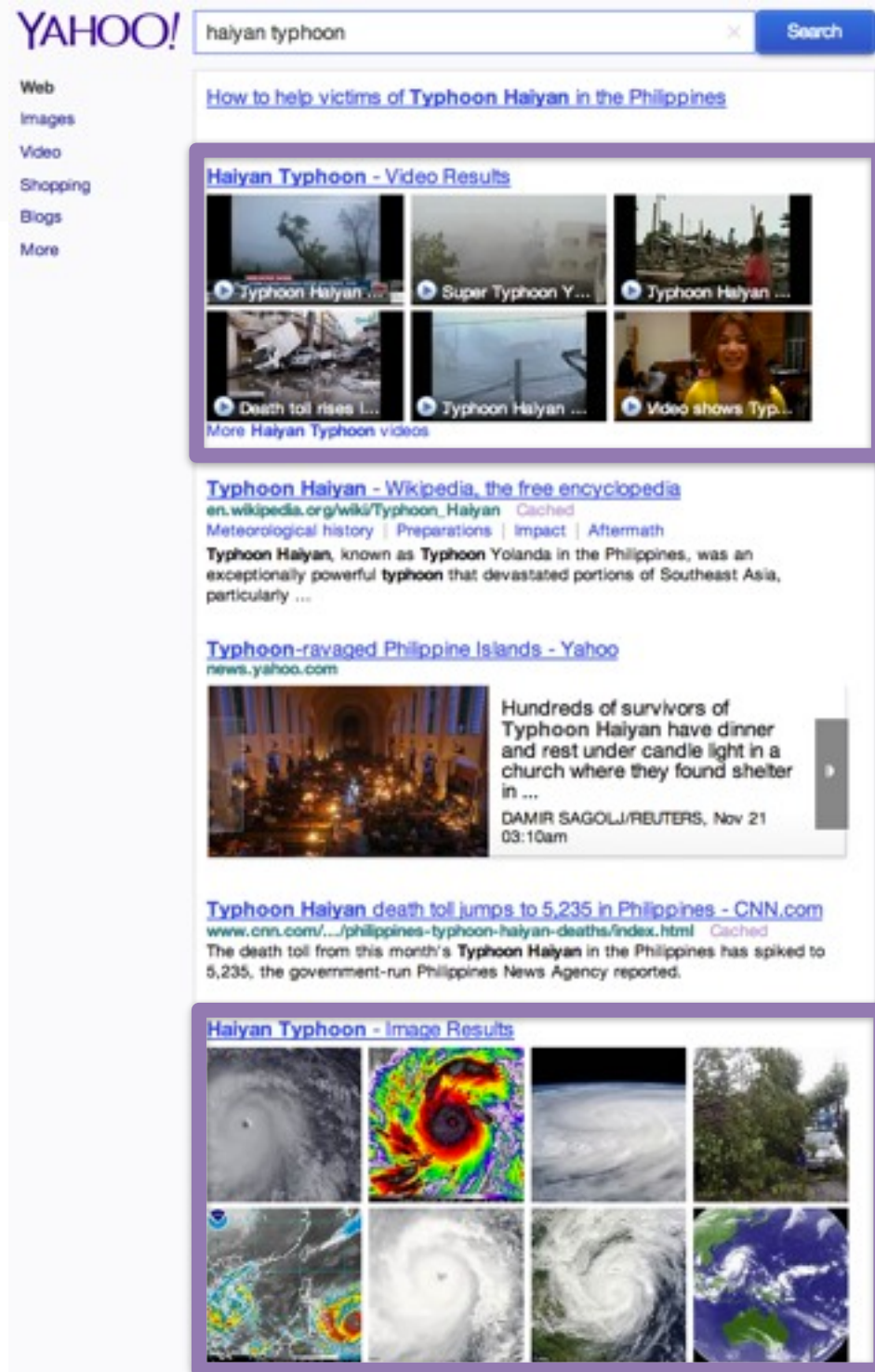
Mobile

The screenshot shows the mobile app version of Yahoo! Finance. The top status bar indicates the time is 2:43 PM and the carrier is Verizon. The app header features the "YAHOO! FINANCE" logo and a search icon. Below the header, there's a large image of the New York Stock Exchange building. A "Following" section displays a list of stock tickers with their current prices and percentage changes: FB (47.53, -0.06%), AAPL (520.56, +1.57%), and YHOO (33.12, +3.15%). Below this, a "Market Overview - US" section shows the Dow Jones (15,761.78, +1.08%), S&P 500 (1,770.61, +1.34%), and NASDAQ (3,919.23, +1.60%) indices. The right sidebar contains several news articles, including "Microsoft Patent Chief to Join Shook Hardy: Business of Law", "Lawyers wrapping case in Detroit bankruptcy trial", "Crude Traders Stick With Brent Amid Manipulation Claims", and "UPDATE 2-India's Infosys to pay \$34 million in U.S. visa case".

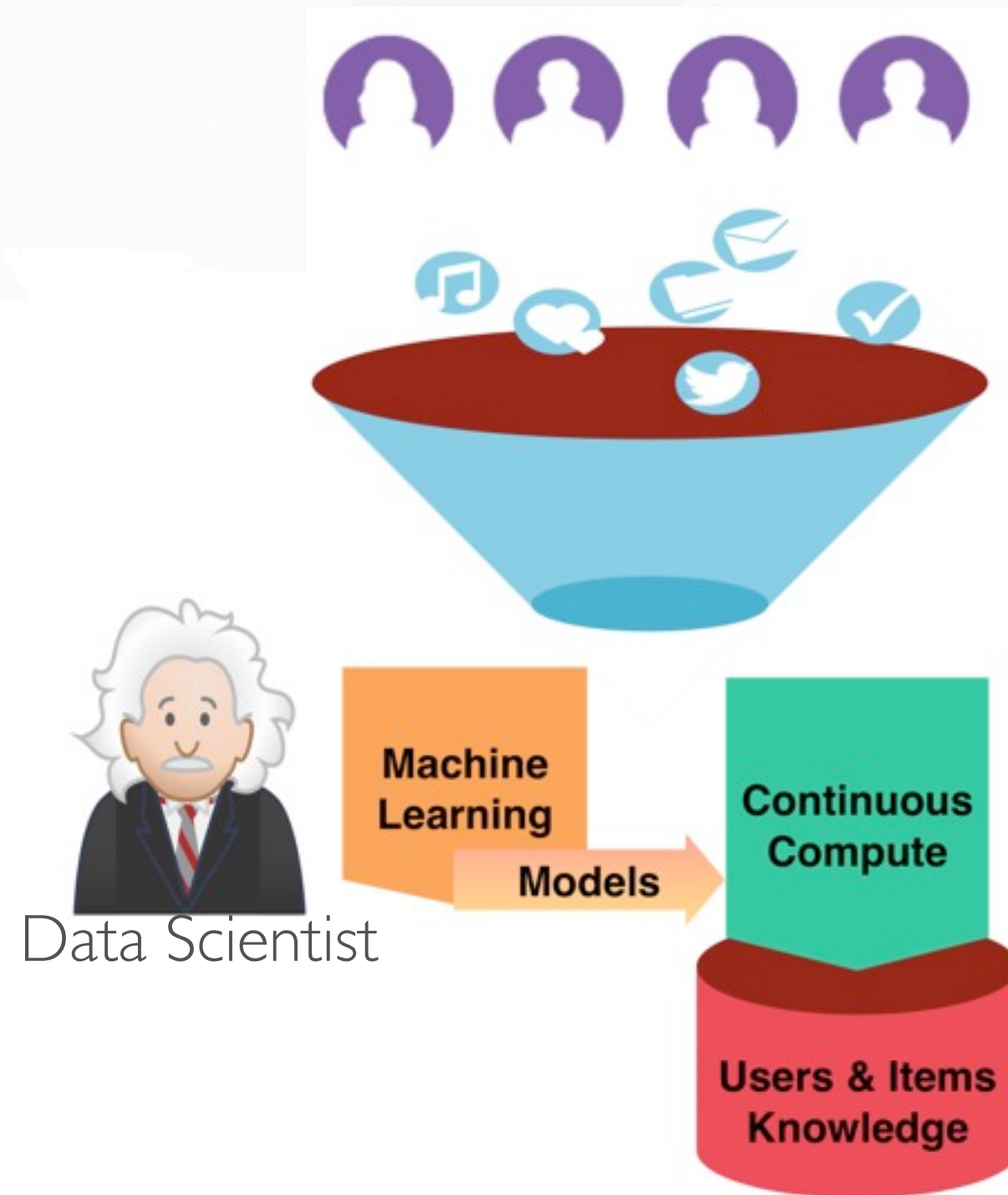
YAHOO 2013: IMPROVED WEB SEARCH W/ VERTICAL CONTENT

http://search.yahoo.com

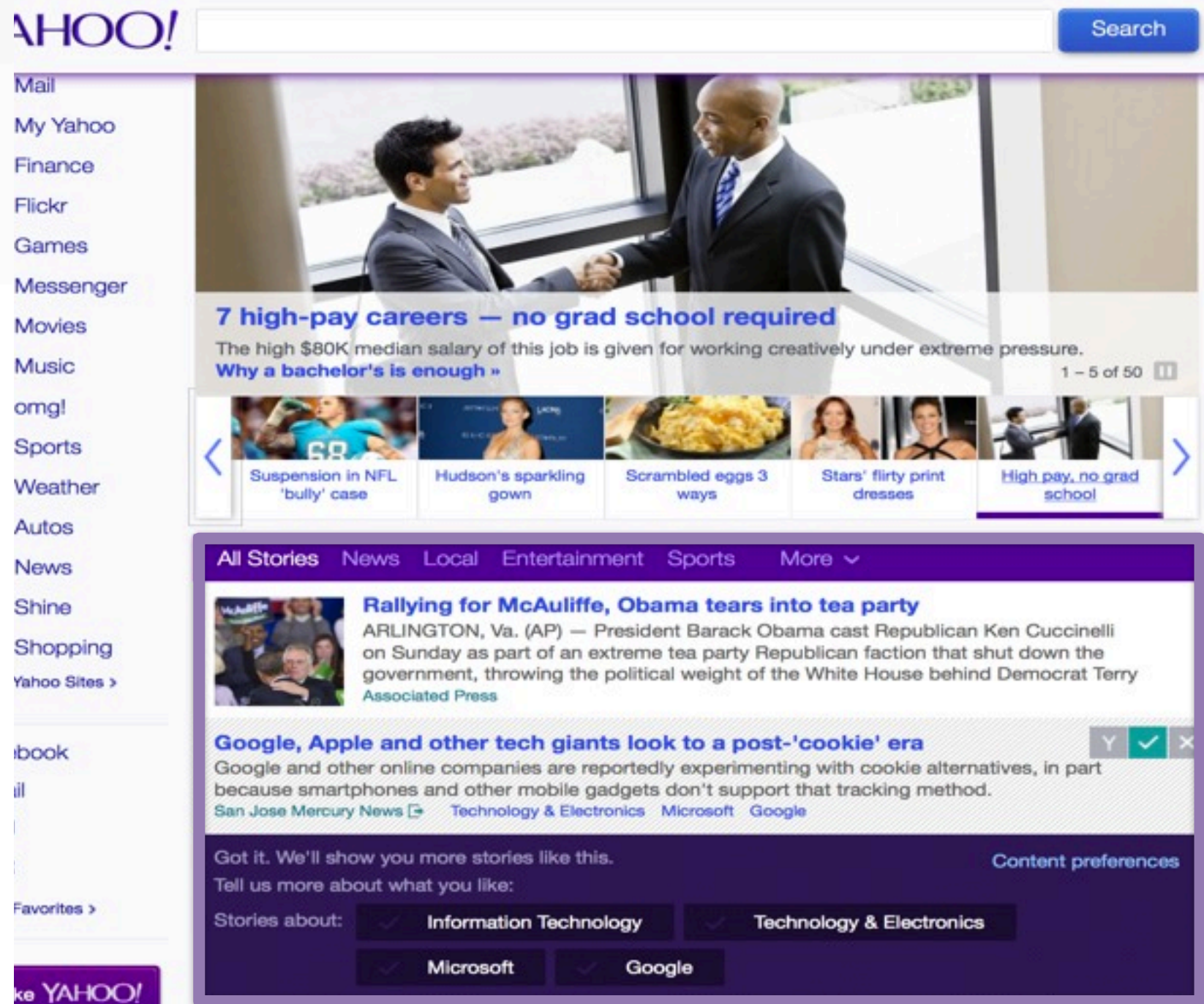
Mobile



DATA SCIENCE AT SCALE



I. CHALLENGE: SCIENCE



- ◆ Single model for all items in homepage stream
 - * Millions of items
 - * 1000's of item/user features
 - Yahoo content categories
 - Wikipedia entity names
 - * Over 800 million users
- ◆ Objective function
 - * Relevance & user engagement
 - * Freshness & popularity
 - * Diversity
- ★ Algorithm exploration
 - * Logistic regression?
 - * Collaborative filtering?
 - * Decision trees?
 - * Hybrid?

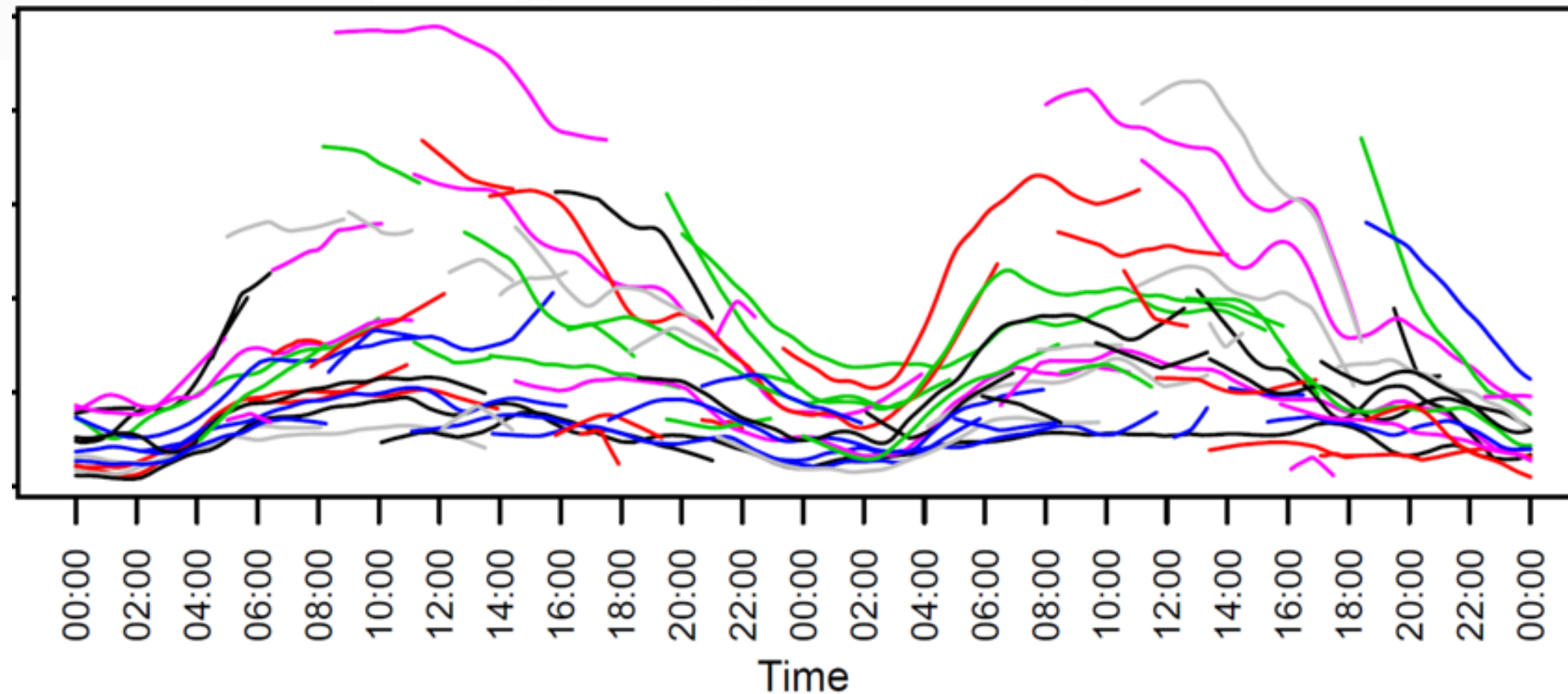
II. CHALLENGE: SPEED

◆ Ex. Item CTR in Yahoo homepage Today Module

* Short Lifetimes

* Temporal effect

* Breaking news

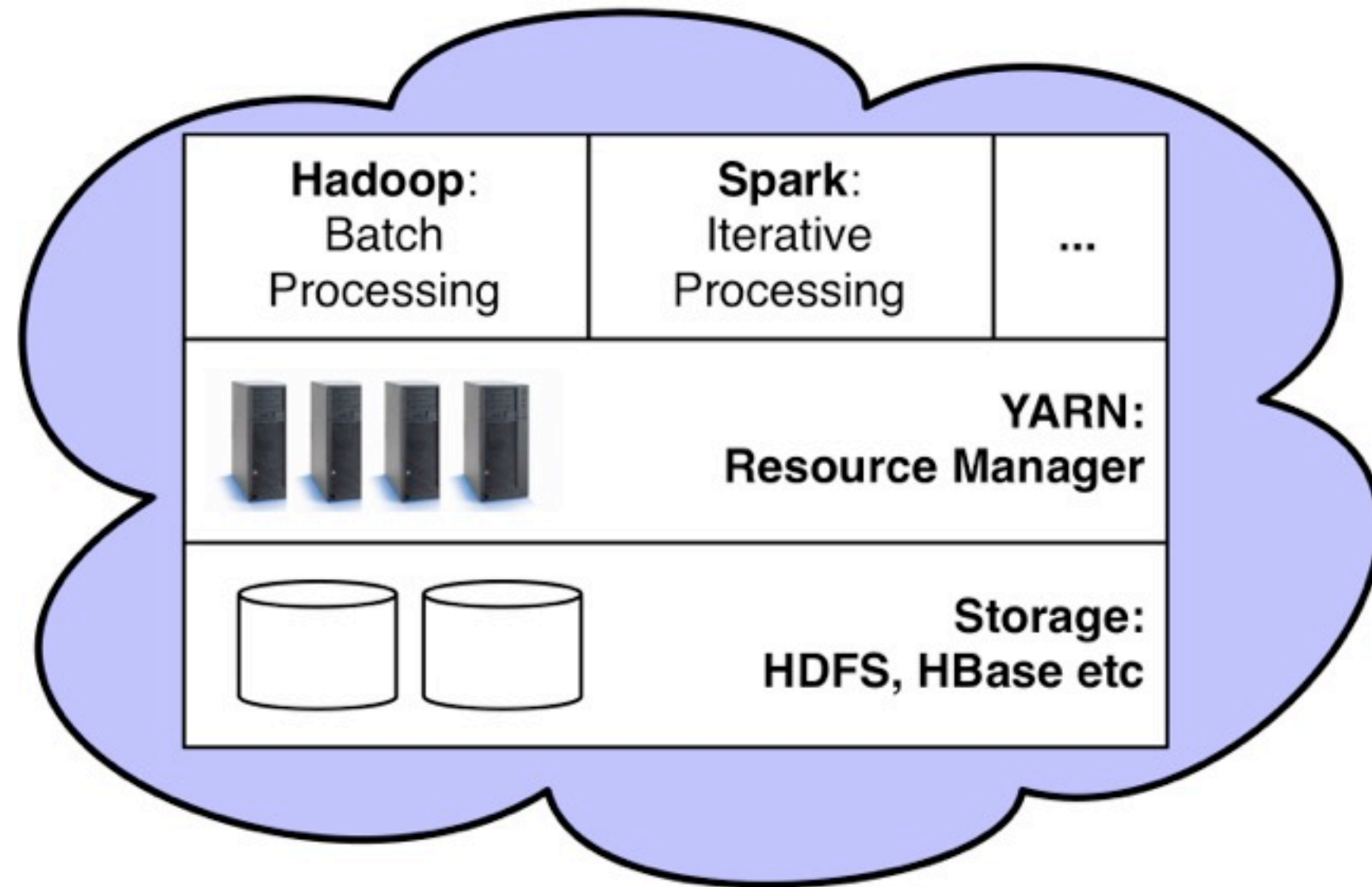


◆ Models should be constructed hourly or faster

III. CHALLENGE: SCALE

- ◆ 150 PB of data on Yahoo Hadoop clusters
 - * Yahoo data scientists need the data for
 - ▶ Model building
 - ▶ BI analytics
 - * Such datasets should be accessed efficiently
 - ▶ avoid latency caused by data movement
- ◆ 35,000 servers in Hadoop cluster
 - * Science projects need to leverage all these servers for computation

SOLUTION: HADOOP + SPARK

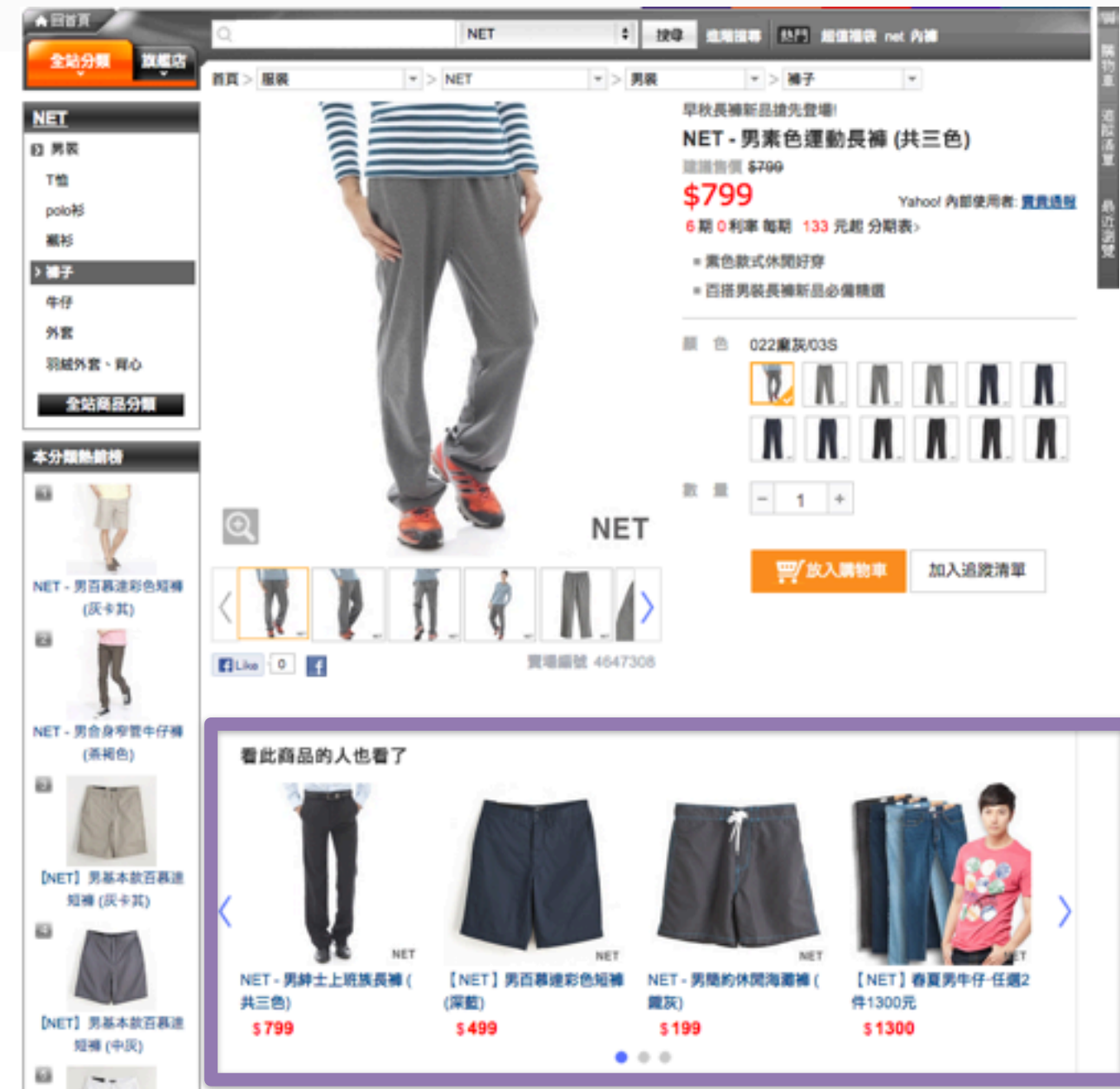


- I. **science** ... Spark API & MLlib ease development of ML algorithms
- II. **speed** ... Spark reduces latency of model training via in-memory RDD etc
- III. **scale** ... YARN brings Hadoop datasets & servers at scientists' fingertips

PILOT PROJECT: E-COMMERCE

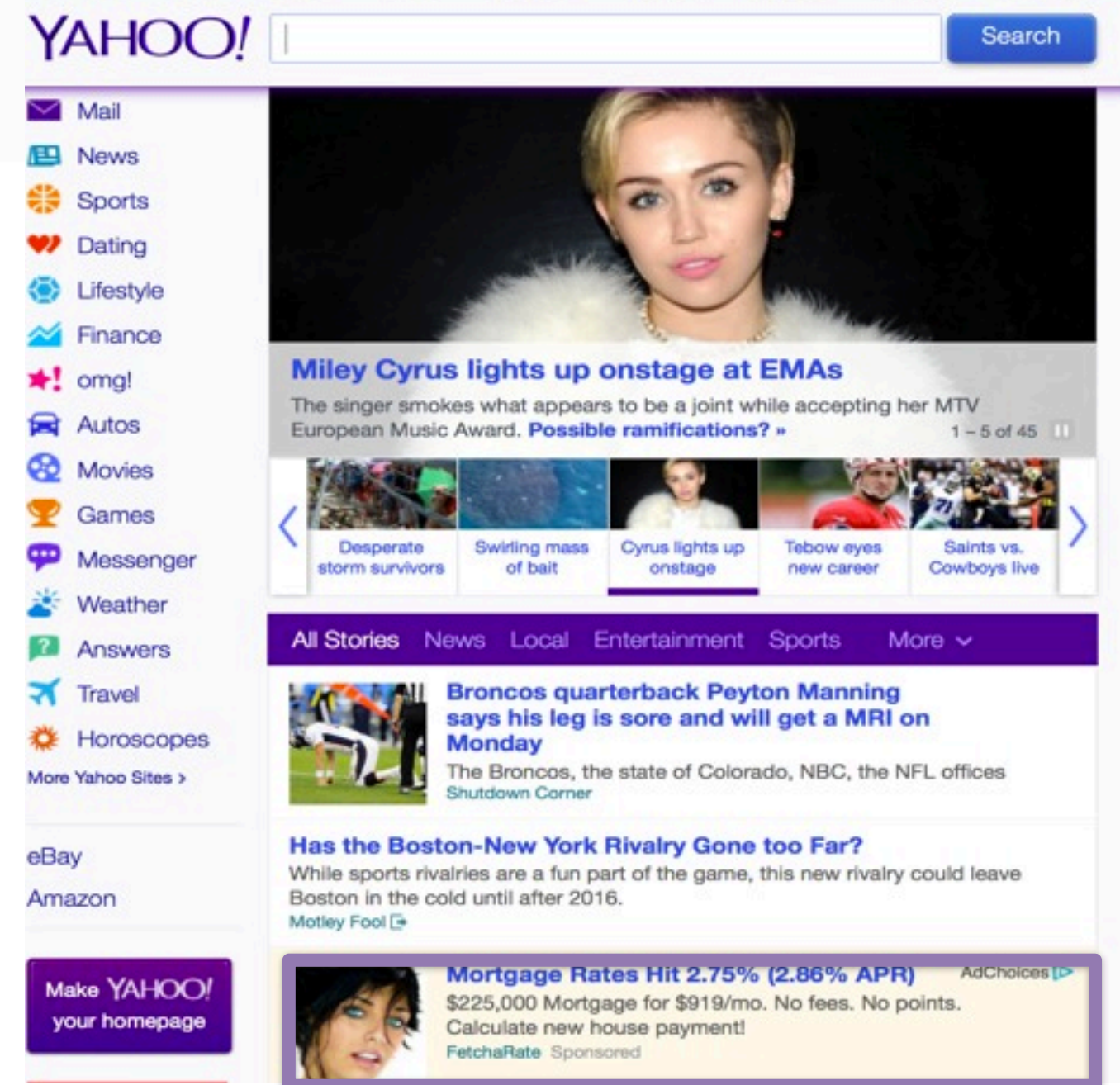
Yahoo Taiwan Shopping & Auction

- ◆ Collaborative filtering algorithms for
 - * Viewed-also-viewed
 - * Bought-also-bought
 - * Bought-after-viewed
- ◆ 30 LOC in Spark/Scala
 - * 14 min. on 10 servers
 - Hadoop-based algorithm: 106 min.



PILOT PROJECT: STREAM ADS

- ◆ A logistic regression algorithm
 - * 120 LOC in Spark/Scala
 - ▶ Alternative: Vowpal Wabbit ... Difficult to extend
 - * 30 min. on model creation for 100M samples and 13K features with 30 iterations
- ◆ “I used Spark-on-Yarn package today, works great.” Amit (July 26, 2013 2:51 PM)
 - * Initial algorithm was launched within 2 hours after Spark-YARN package announcement
 - * Compare: Several weeks on system setup and data movement



SUMMARY

- ◆ Spark plays an important role in machine learning at Yahoo
 - * Hadoop continues to be the core of our big-data platform
- ◆ Yahoo is excited about your continued contribution to Apache Spark
 - * 4 committers
 - * ex. Spark-on-Yarn, Shark, security, scalability, operability etc.

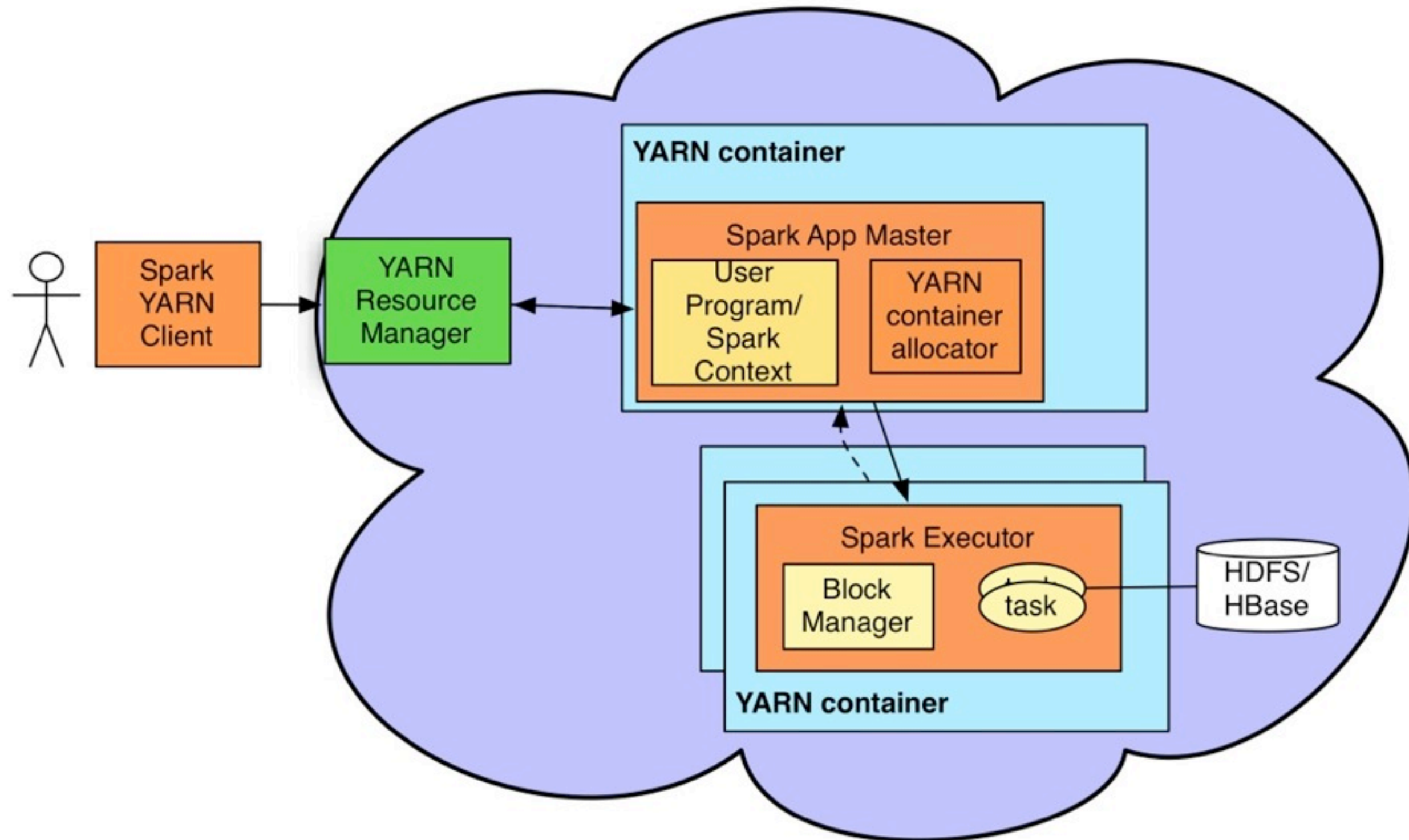
KEY TECHNOLOGY: **SPARK ON YARN**

Thomas Graves (tgraves@yahoo-inc.com)
Spark Committer & Hadoop PMC
Yahoo

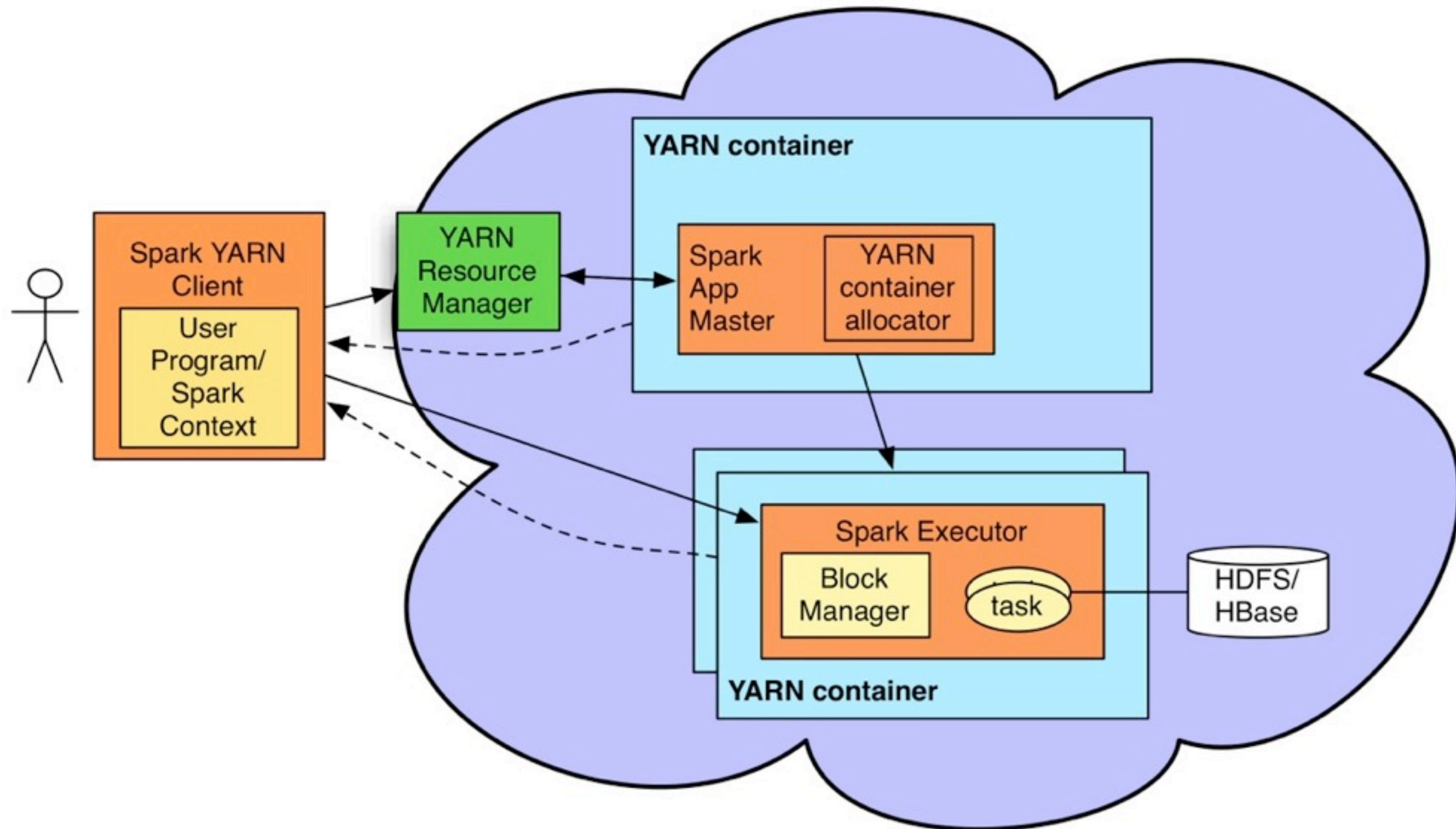
SPARK-ON-YARN: ROADMAP

- ◆ Spark-0.6 & 0.7 - Experimental support of YARN
 - * Hadoop 0.23.X and early Hadoop 2.X releases
- ◆ Spark-0.8.0 - Spark-on-Yarn merged into master Spark branch
 - * Secure HDFS access
 - * Use YARN approved directories
 - * Link Spark UI to YARN UI
- ◆ Spark-0.8.X - Future integration with Hadoop
 - * Add authentication to Spark
 - * Support running spark on YARN from HDFS
 - * Support files/archives in Hadoop distributed cache
- ◆ Spark-0.9.X - Client-mode introduced
 - * Support Spark Shell on YARN
 - * Spark on YARN running on Hadoop 2.2.X

ARCHITECTURE: STANDALONE MODE



ARCHITECTURE: CLIENT MODE



FUTURE DIRECTIONS

- ◆ Support long running jobs
 - * Shark
 - * Spark Streaming
- ◆ Dynamic resource allocation
- ◆ Integrate with Hadoop enhancements
 - * Generic History Server
 - * Preemption, etc.

MORE INFO:

◆ <http://spark.incubator.apache.org/docs/latest/running-on-yarn.html>