

YAHOO!

Learning from Big Data: Architecture Options

Andy Feng (@afeng76, afeng@apache.org)

Distinguished Architect, Yahoo

Feb. 28, 2014

Who am I?

- Current

- › Distinguished Architect, Yahoo Hadoop Group
- › Committer, Apache Storm
- › Contributor, Apache Spark

- Past

- › Online advertisement
- › Personalization
- › Cloud services
- › NoSQL

YAHOO!

Agenda

1. Machine Learning

- › Use Cases
- › Libraries

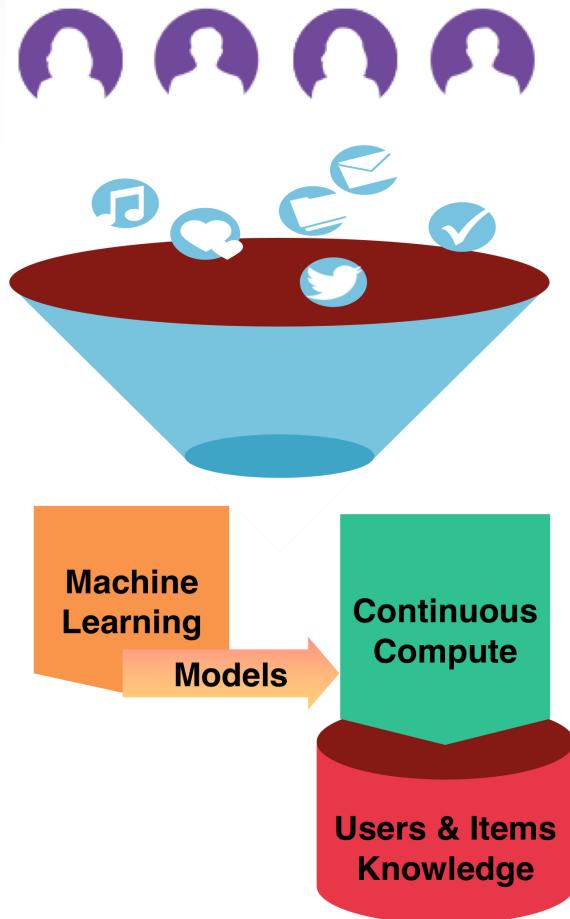
2. Big-Data Technologies

- › Hadoop, Spark and Storm

3. Design Patterns

- › Batch/incremental learning
- › Real-time learning
- › Hybrid learning

Yahoo Machine Learning



- Will a user enjoy a new article?
- What vertical contents are relevant to a user search?
- What products should we recommend a user?
- Is this email a spam?
- Is this ad click fraud?
- Was this photo taken at a beach?
- Find all potential users to be targeted by an ad campaign
- ...

YAHOO!

Use Case 1: Personalized Homepage

<http://www.yahoo.com>

The desktop version of the Yahoo homepage features a purple header with the word "YAHOO!" in white. Below the header is a navigation bar with links to Mail, My Yahoo!, Finance, Flickr, Games, Messenger, Movies, Music, omg!, Sports, Weather, Autos, News, Shine, Shopping, and More Yahoo! Sites. The main content area includes a large image of two men shaking hands, a headline about high-pay careers, and several news thumbnails. At the bottom, there are sections for Google news, a yearbook offer from Classmates.com, and a story about a woman offering thanks to vets.

Mobile

The mobile version of the Yahoo homepage is displayed on an iPhone screen. The top status bar shows signal strength, network (Verizon LTE), time (9:05 AM), and battery level. The main content area is similar to the desktop version, featuring news stories and images. The "YAHOO!" logo is visible at the top right of the screen.

YAHOO!

Use Case 2: Improved WEB Search w/ Vertical Content

<http://search.yahoo.com>

The screenshot shows the Yahoo search interface. The search bar at the top contains the query "typhoon". Below the search bar, the results are displayed under the heading "Typhoon News Results". The first result is a news article from The Christian Science Monitor about Typhoon Haiyan hitting the Philippines. The second result is another news article from CNN about the death toll. The third result is a link to the Wikipedia page on Typhoons. Below these, there's a summary of what a typhoon is, mentioning it's a mature tropical cyclone in the western Pacific. The fourth result is a news article from Yahoo itself about Super Typhoon Haiyan hitting the Philippines. This article includes an aerial photograph showing damaged houses. The fifth result is a link to the official Typhoon website. At the bottom, there's a promotional banner for a Typhoon Fall Tour.

Mobile

The screenshot shows the Yahoo mobile search interface. The search bar at the top contains the query "typhoon haiyan". Below the search bar, the results are displayed under the heading "Web". The first result is a news article from the Denver Post about a victim of Typhoon Haiyan. The second result is a news article from CBS News about the humanitarian response. The third result is a news article from NBC News about future storms. To the right of the main search results, there's a sidebar titled "Videos" which lists two video thumbnails: one from BBC.co.uk showing a satellite view of Typhoon Haiyan, and another from AOL.com featuring a news anchor. At the bottom, there are navigation icons for a mobile device.

Use Case 3: E-Commerce

The screenshot shows a product page for men's sweatpants on a Yahoo! e-commerce site. The main product is a pair of grey sweatpants with blue stripes on the cuffs. The price is \$799, with a financing option of 6 periods at \$133 per period. The page includes a sidebar with a navigation menu and a 'hot picks' section.

Product Details:

- 早秋長褲新品搶先登場!
- NET - 男素色運動長褲 (共三色)**
- 建議售價 \$799
- \$799** (highlighted in red)
- Yahoo! 內部使用者: 賣貴通報
- 6期 0 利率 每期 133 元起 分期表 >
- 素色款式休閒好穿
- 百搭男裝長褲新品必備精選

Color Options: 022麻灰/03S
[Thumbnail Grid]

Quantity: 数量 [-] 1 [+]

Action Buttons: 放入購物車 (Add to Cart) | 加入追蹤清單 (Add to Wish List)

Product ID: 實場編號 4647308

Related Products: 看此商品的人也看了

- [NET] 男基本款百慕達短褲 (灰卡其)
- [NET] 男合身窄管牛仔褲 (茶褐色)
- [NET] 男基本款百慕達短褲 (灰卡其)
- [NET] 男基本款百慕達短褲 (中灰)
- [NET] 男基本款百慕達短褲 (灰卡其)
- [NET] 男基本款百慕達短褲 (深藍)
- [NET] 男簡約休閒海灘褲 (鐵灰)
- [NET] 春夏男牛仔·任選2件1300元

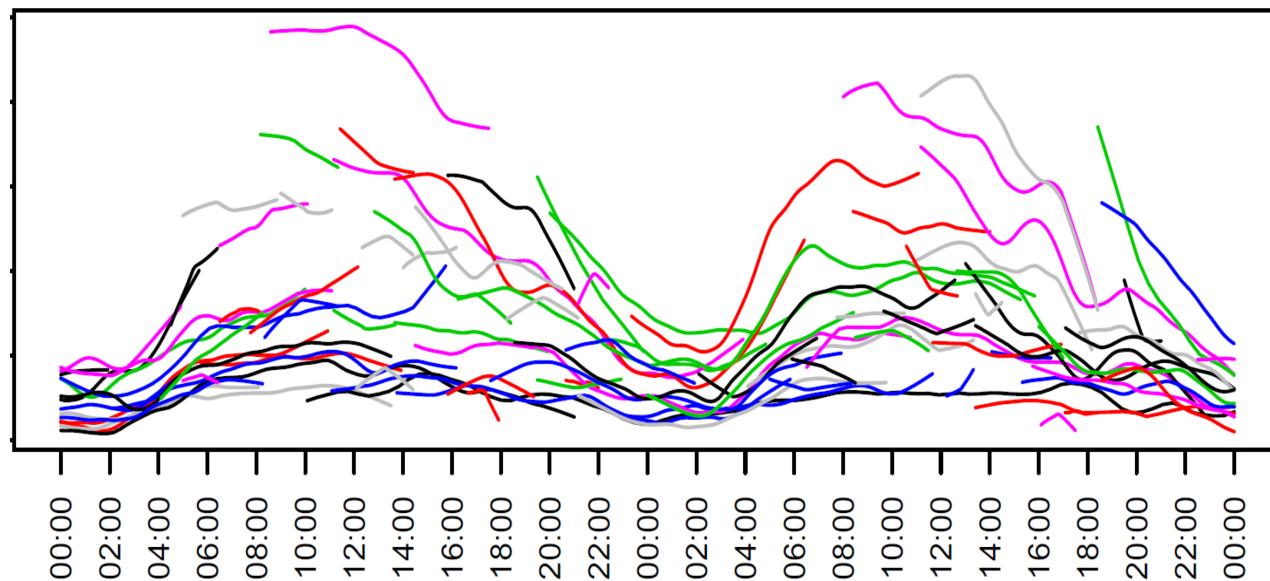
YAHOO!

Challenge: Scale

- 150 PB data on Yahoo Hadoop clusters
- Science projects need to be executed on Hadoop clusters directly
 - *Avoid latency caused by data movement
 - *Leverage existing 35,000 servers for computation

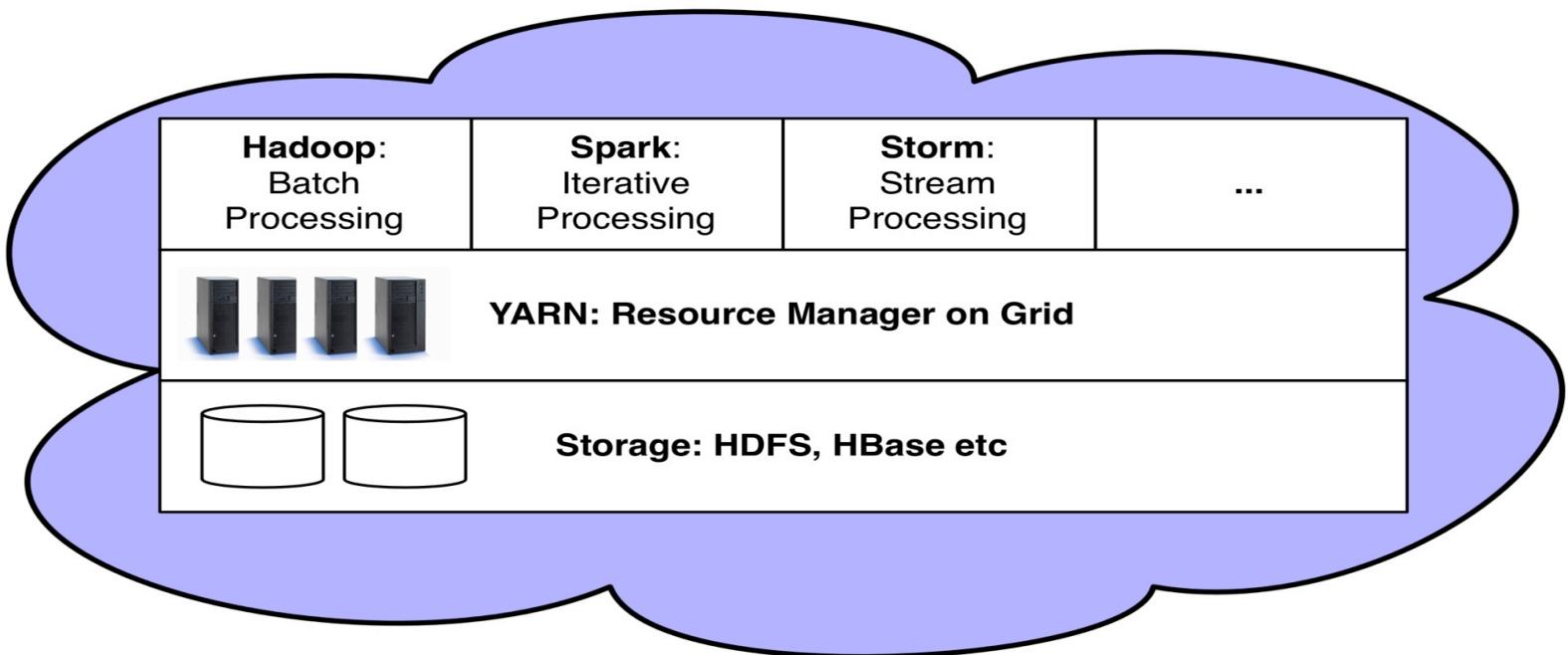
Challenge: Speed

- Ex. Item CTR in Yahoo homepage Today Module
 - * Short Lifetimes
 - * Temporal effect
 - * Break news



*Update user/item knowledge continuously
*Enable item discovery within minutes

Yahoo Big-Data Platform

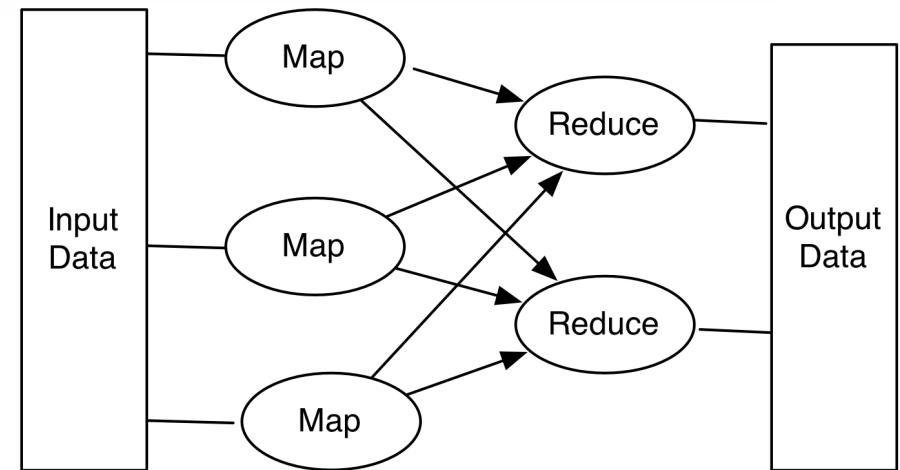


YAHOO!

Apache Hadoop

<http://wiki.apache.org/hadoop/>

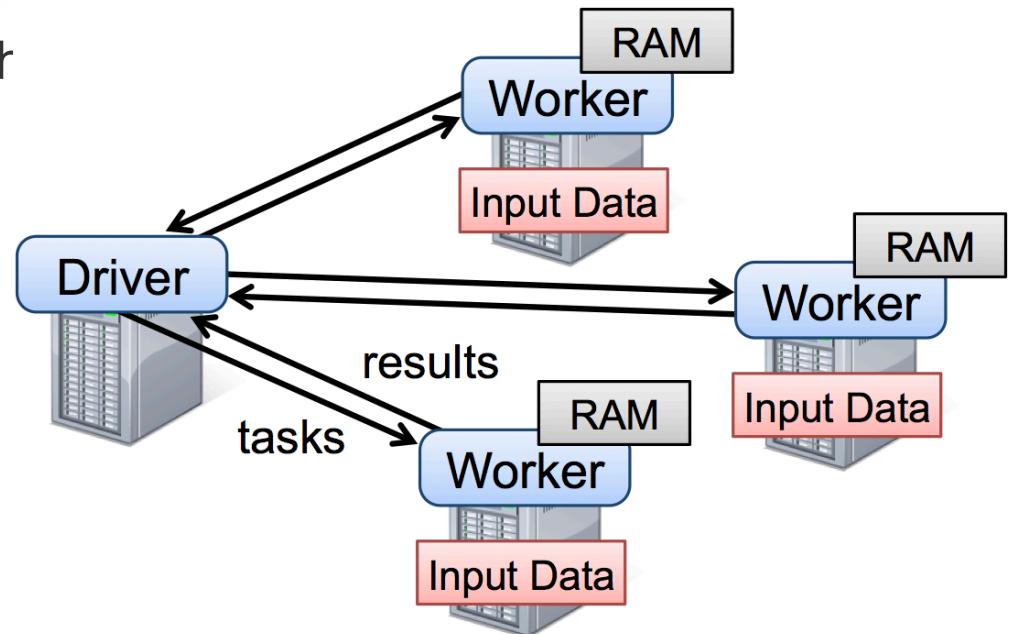
- Popular framework for running applications on large cluster built of commodity hardware
- Implement Map/Reduce and Tez with YARN resource manager
- Provides HDFS to store data on the compute nodes
- Designed for very high throughput and reliability



Apache Spark

<http://spark.incubator.apache.org>

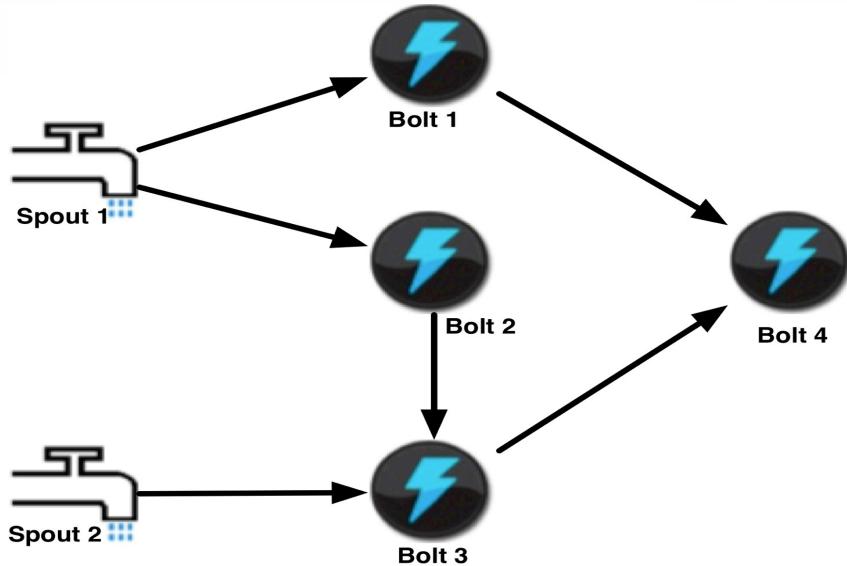
- Fast and expressive cluster computing system compatible with Apache Hadoop
- Support general execution DAGs
 - › Include iterative programming
- Resilient distributed datasets (RDDs)
 - › In-memory storage



YAHOO!

Apache Storm

<http://storm.incubator.apache.org>



- “Hadoop for Realtime”

- › a distributed, fault-tolerant, and high-performance realtime computation system that provides strong guarantees on the processing of data

- Emerging standard

- › Initially developed and deployed at BackType in 2011
 - › Open sourced via github.com in Sept. 2011
 - › Apache incubation in Sept. 2013, and Storm 0.9.1-incubating released in Feb 2014
 - › 9 [committers](#), 50 [contributors](#), 12,000 [readers](#)

Hadoop vs. Spark vs. Storm

■ Hadoop

- › Batch computing at data location; Excellent throughput
- › Massively scalable

■ Spark

- › Batch computing at data location; Excellent throughput
- › Enable iterative applications
- › RDDs w/ caching option

■ Storm

- › Realtime computing; Low-latency

Sample ML Libraries

- *Vowpal Wabbit*

- › A fast out-of-core learning system
- › Several optimization algorithms available including sparse gradient descent (GD), BFGS and several loss functions

- *LibSVM*

- › A *popular* library for Support Vector Machines
- › Support vector classification & regression

6 Design Patterns

- Batch/incremental learning
 1. ML in Hadoop Reducers only (Hadoop Streaming)
 2. ML in Hadoop Mappers only (Hadoop Streaming)
 3. Spark Pipe to external ML processes
 4. Spark native ML
- Real-time learning
 5. Storm Bolt w/ ML in JNI
- Hybrid learning
 6. ML in Hadoop/Spark and Storm

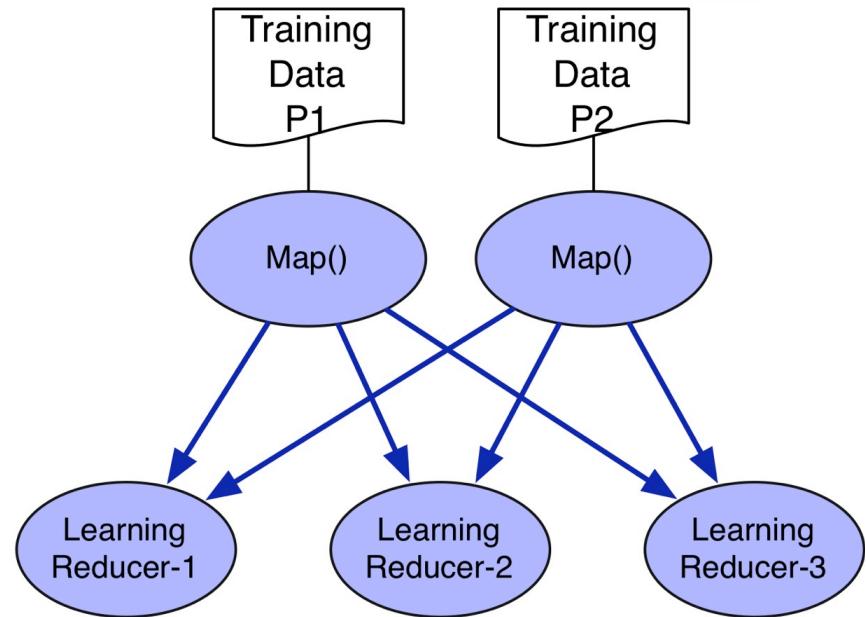
1. ML in Hadoop Reducers

- **Basic Requirements**

- › 100's - 1000's models
- › Training data for each model could be loaded into a single machine

- **Solution: 1 reducer per model**

- › hadoop jar hadoop-streaming.jar
-Dmapreduce.job.reduces=\$num_models
-reducer "vw --passes 20 --cache_file ..."
- › hadoop jar lib/hadoop-streaming.jar
-D mapreduce.job.reduces=\$num_models
-reducer "svm_train_reducer.py ..."



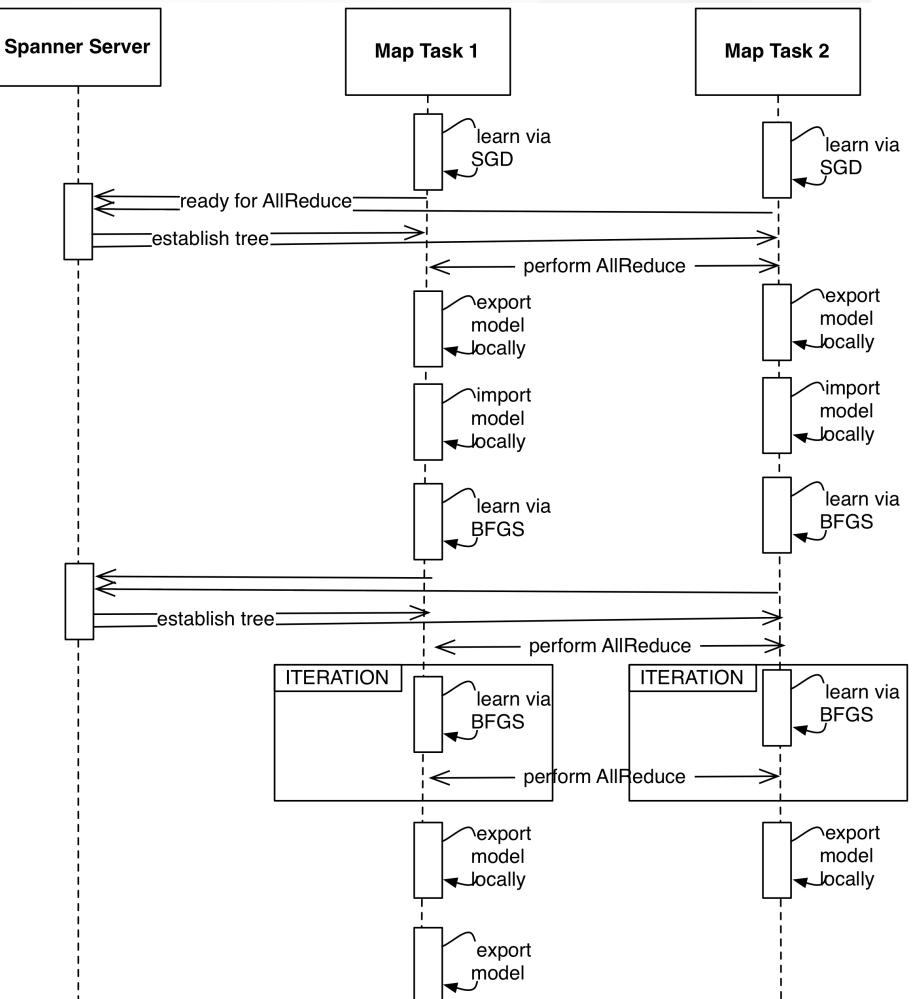
2. ML in Hadoop Mappers

- **Basic Requirements**

- › Small # of models to be trained
- › Training data are too large to be loaded into a single machine

- **Solution: Mappers + MPI AllReduce**

1. spanning_tree
2. hadoop jar hadoop-streaming.jar
-input \$training_data -output \$model_loc
-Dmapreduce.job.maps=\$num_mappers
-mapper "runvw.sh \$model_location
\$span_server \$num_mappers"
-reducer NONE



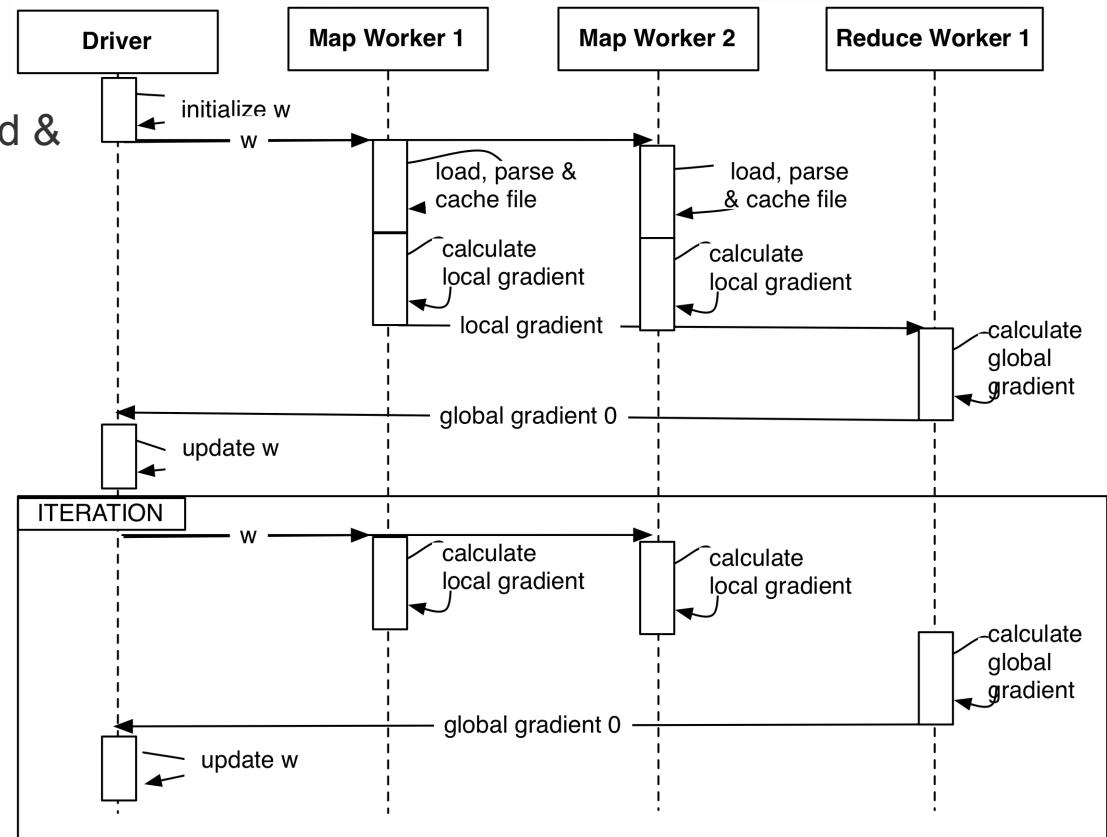
3. Spark Pipe to External ML Processes

▪ Basic Requirements

- › Small # of models built via distributed & iterative computing
- › Hadoop's job launch overhead is significant

▪ Solution: Spark RDD Cache & Pipe

```
val data= ctx.textFile(...).map(...).cache()
var w=data.pipe(GD_cmd).reduce(Avg)
for (i <- 1 to ITERATIONS) {
    val gradient = data.pipe(BFGS_cmd)
        .reduce((a,b) => a+b)
    w -= gradient
}
```



4. Spark Native ML

- **Basic Requirements**

- › Don't need advanced ML features as in VW or LibSVM

- **Sample Solution – Spark or MLLib**

- › Spark
 - Yahoo E-Commerce: 30 LOC Spark program for collaborative filtering
- › MLLib (<http://spark.incubator.apache.org/docs/0.9.0/mllib-guide.html>)
 - Binary classification ... logistic regression; Support Vector Machines
 - Linear regression ... Linear regression; Ridge Regression; Lasso
 - collaborative filtering ... Alternating Least Squares (ALS) matrix factorizaton
 - Clustering ... K-means clustering

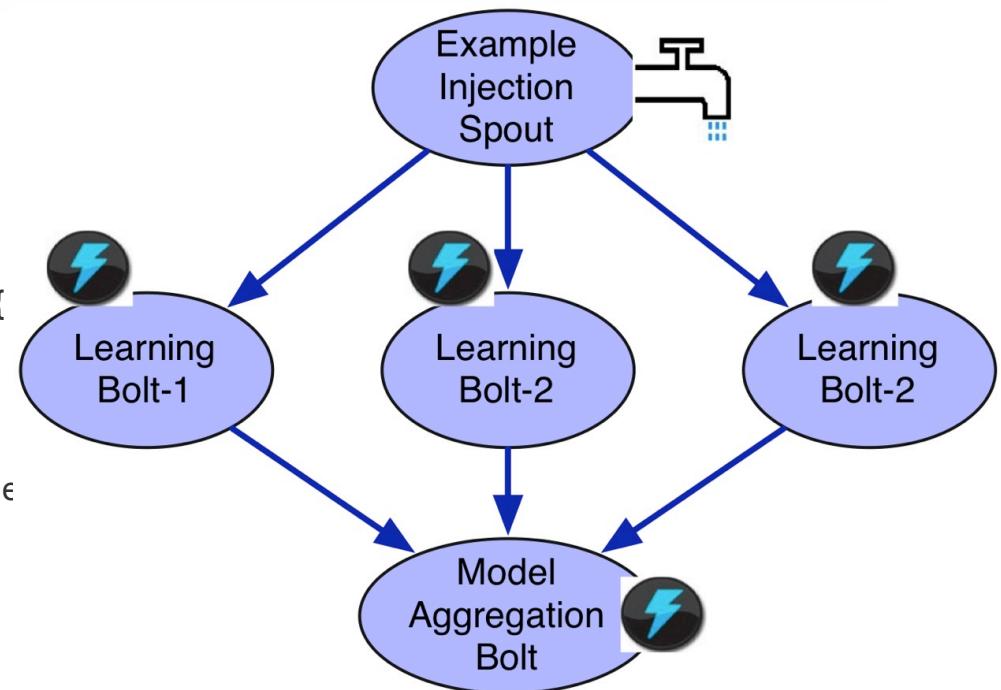
5. Realtime ML in Storm Bolts

- **Basic Requirements**

- › Freshness of ML model is critical

- **Sample Solution**

```
public class TrainingBolt extends BaseBasicBolt {  
    Model model;  
    public void prepare(Map conf, TopologyContext ctx) {  
        System.loadLibrary("VW");  
        model = VW.init(conf);  
    }  
    public void execute(Tuple input, OutputCollector collector)  
    {  
        Instance example = input.getValue(0);  
        model.learn(example);  
        if (Time since last export) collector.emit(model);  
    }  
}
```



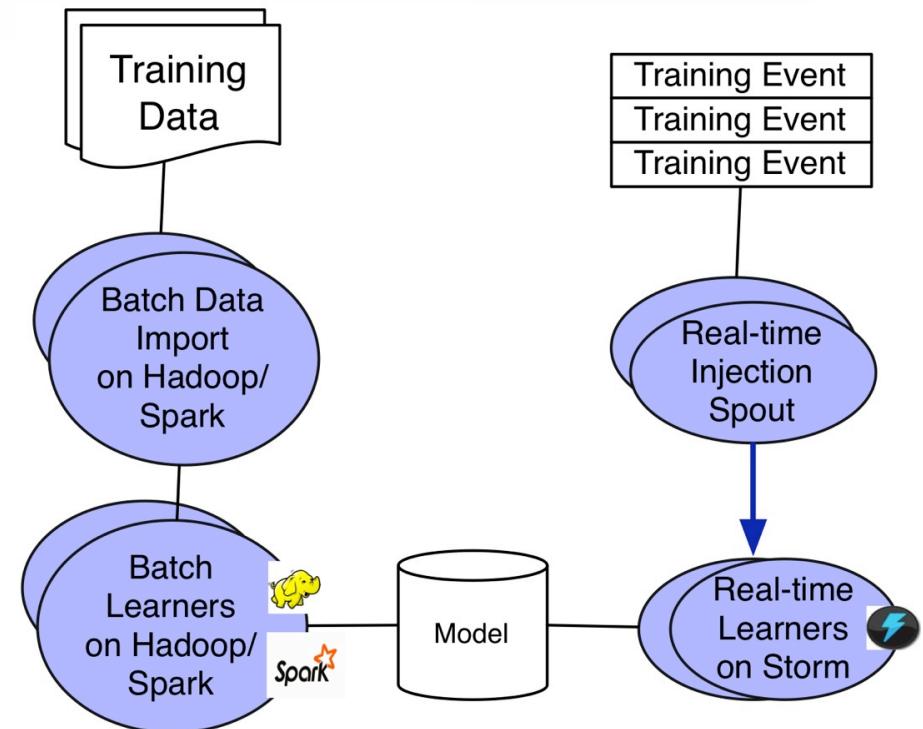
6. Hybrid Learning

- **Basic Requirements**

- › Bootstrap models via batch learning from large datasets
- › Update models via realtime learning from latest events

- **Sample Solution**

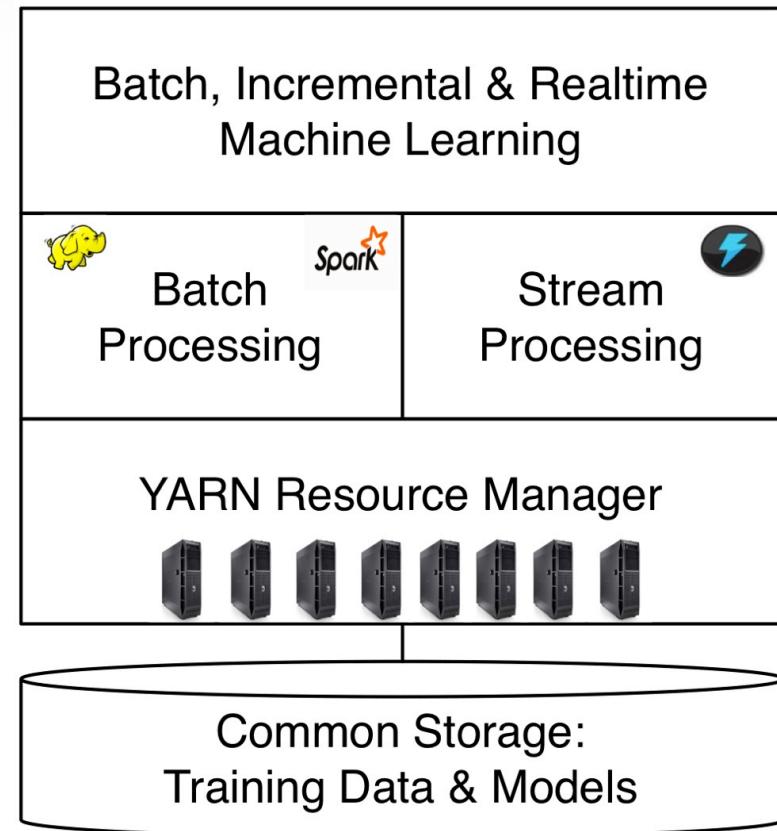
- › ML in Hadoop + Storm
- › ML in Spark + Storm



YAHOO!

Conclusion

- Hadoop, Spark and Storm empower variety of design patterns for machine learning from big-data.
- We expect active research and development of new ML capabilities on these big-data platform.



YAHOO!

We are Hiring!
Email: afeng@yahoo-inc.com
<http://careers.yahoo.com>



YAHOO!