



Storm-on-YARN: Convergence of Low-Latency and Big-Data

Andrew Feng

Self Introduction

- Current
 - Distinguished Architect, Yahoo! Hadoop Team
 - Core contributor at Storm project
- Past
 - Online advertisement
 - Personalization
 - Serving containers
 - Cloud services
 - NoSQL database
 - Application server

Agenda

- **Business motivation**
- Technical overview
- Open source

Yahoo!: Personalized Web

YAHOO!

Web 

Search

Hi, Andrew

Mail

-  Mail
-  News
-  Finance
-  Sports
-  Movies
-  omg!
-  Shine
-  Autos
-  Shopping
-  Travel
-  Dating
-  Jobs

More Y! Sites 

Make YAHOO!
your homepage



Do a Discount
Double Check®
Get discounts up to
40% on Auto.
Get A Quote



'The worst company in America' (again)

Results of an annual Consumerist poll slam a games company for the second straight year.
[Beat out Bank of America »](#)

1 - 5 of 70 



All Stories News Local Entertainment Sports More 

H-P Introduces Moonshot Server

Hewlett-Packard Co. is all set to woo consumer experience with its latest Moonshot Servers.
Zacks

Tiger Woods Lauds Augusta National's Addition of Female Members

Tiger Woods applauded the addition of Condoleezza Rice and Darla Moore to the list of members at Augusta National Golf Club, home of the Masters Tournament.
at Bloomberg



Render Chiu likes Tiger Woods.

Yahoo Reaches New 52-Week High (YHOO)

Yahoo (Nasdaq:YHOO) hit a new 52-week high Tuesday as it is currently trading at \$23.94, above its previous 52-week high of \$23.90 with 9.6 million shares traded as of 2:41 p.m. ET. Average volume at TheStreet.com



Chris Toomey, Rupesh Chhatrapati and Rajiv Verma like Yahoo!.

The Most-Searched Character On 'Mad Men' Isn't Don Draper

It's not Jessica Paré, either.
Business Insider

4 Tech Stocks Cramer Favors As Markets Threaten To Sell-off

Chris Lau, KAPITALL: For the week ending April 5 2013, Jim Cramer was highly bullish on the companies mentioned. Out [...] at TheStreet.com

Trending Now

- | | |
|---|---|
| 1 Michael J. Fox | 6 Trisha Yearwood weig... |
| 2 Joe Montana son | 7 Jane Fonda not afraid |
| 3 Andy Johns dies | 8 Terry Francona lost |
| 4 2 Navy divers drowned | 9 Tim Tebow plans |
| 5 Kanye West sued | 10 Supersized crabs |

 [Watch the show »](#)


**ZERO TO CARD IN
60 SECONDS™**
Find an offer that's right for you
in under a minute.


[FIND MY CARD >>](#)

Ad Feedback

AdChoices 

Sunnyvale

69°F Fair



Today
73° 53°



Tomorrow
79° 53°



Thursday
71° 52°

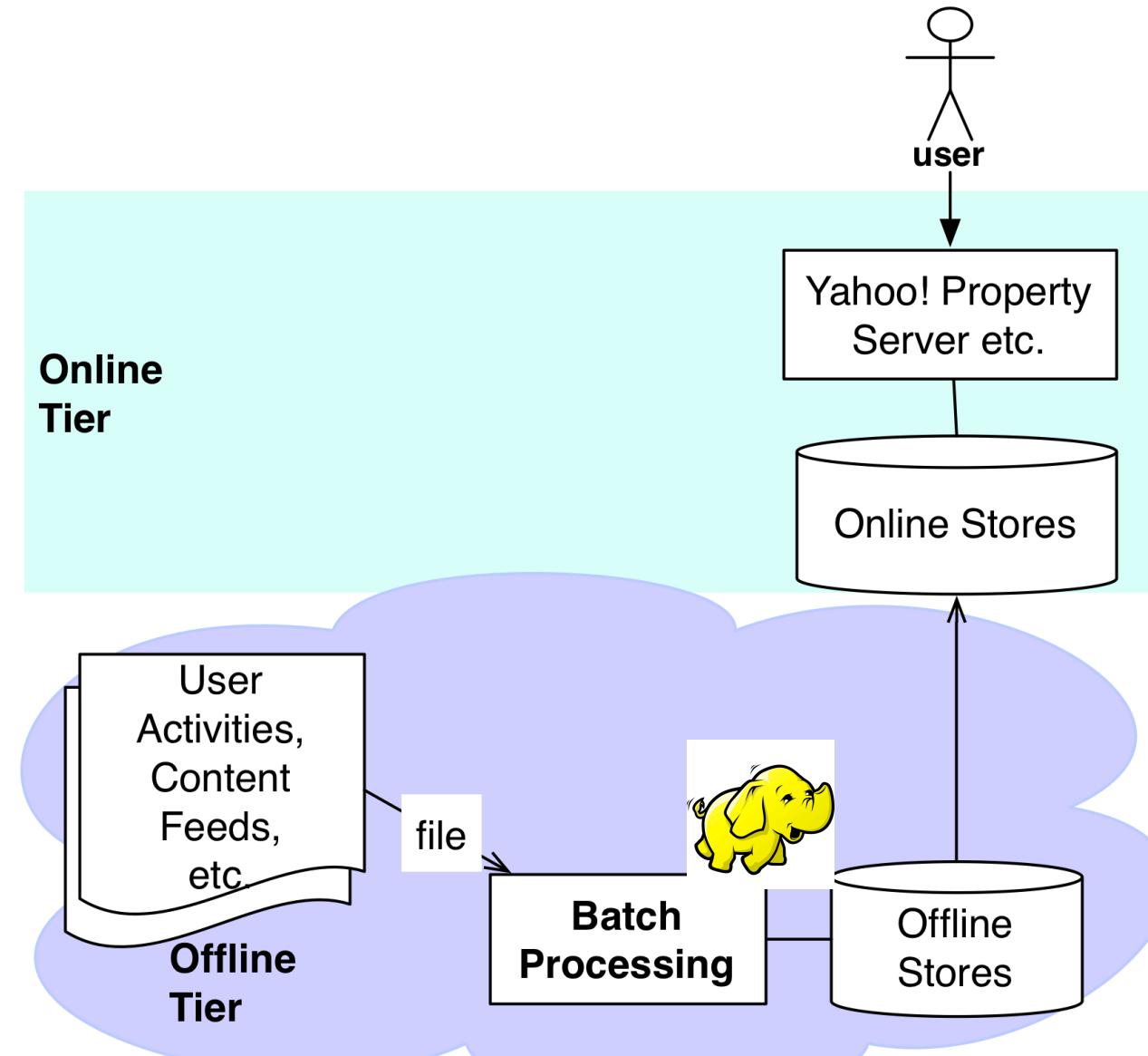
Quotes [Y! Finance](#)

S&P 500	1,568.61		0.35%
NASDAQ Composite	3,237.86		0.48%
Dow Jones Indust...	14,673.46		0.41%

 Enter company/ticker

Markets My Portfolio

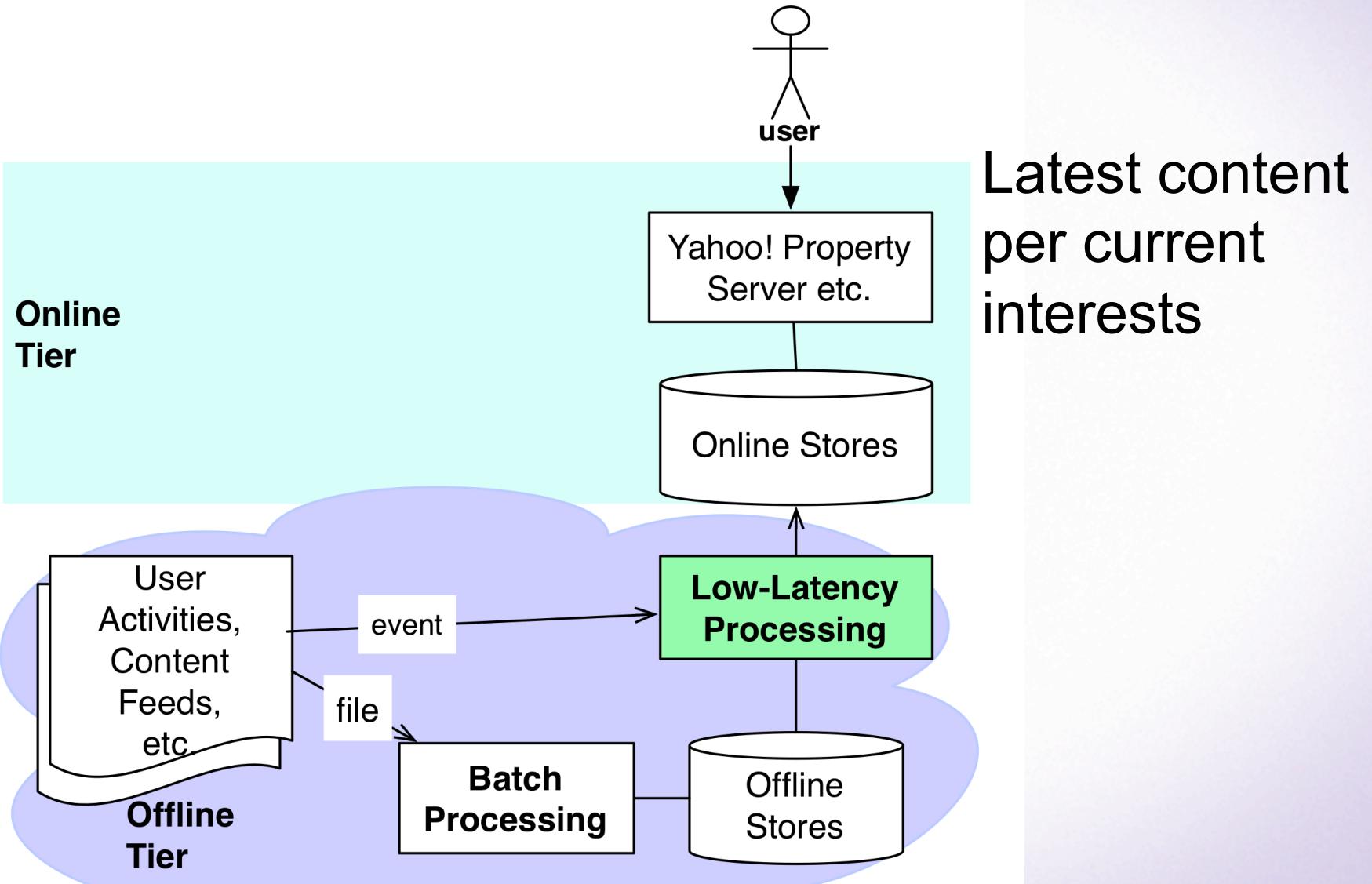
Personalization w/ Hadoop



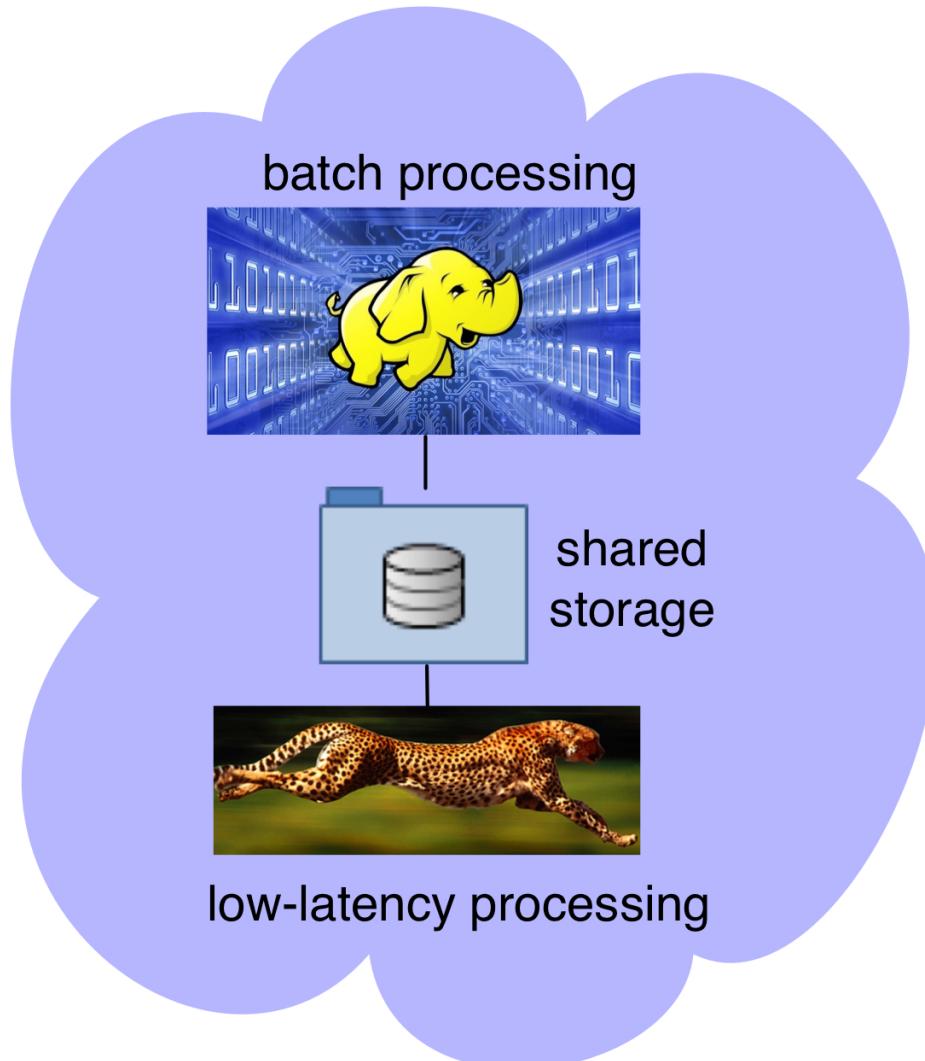
Select relevant content & ads

Understand user & content/ads

Personalization w/ Low-Latency



Big Data + Low Latency: Design Pattern

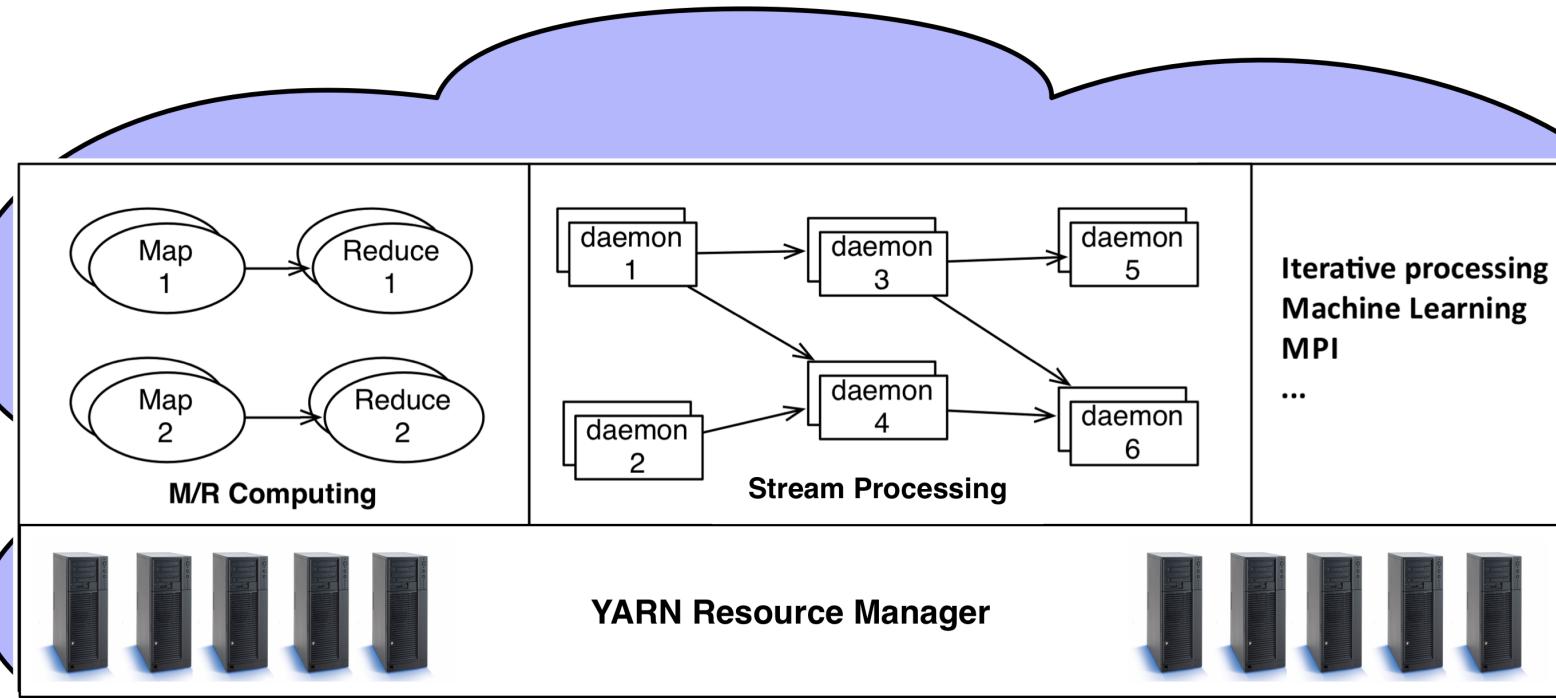


- Personalization
- Ad targeting
- Reporting
- Ad budgeting
- Fraud detection
- Trending topics

Agenda

- Business motivation
- **Technical overview**
- Open source

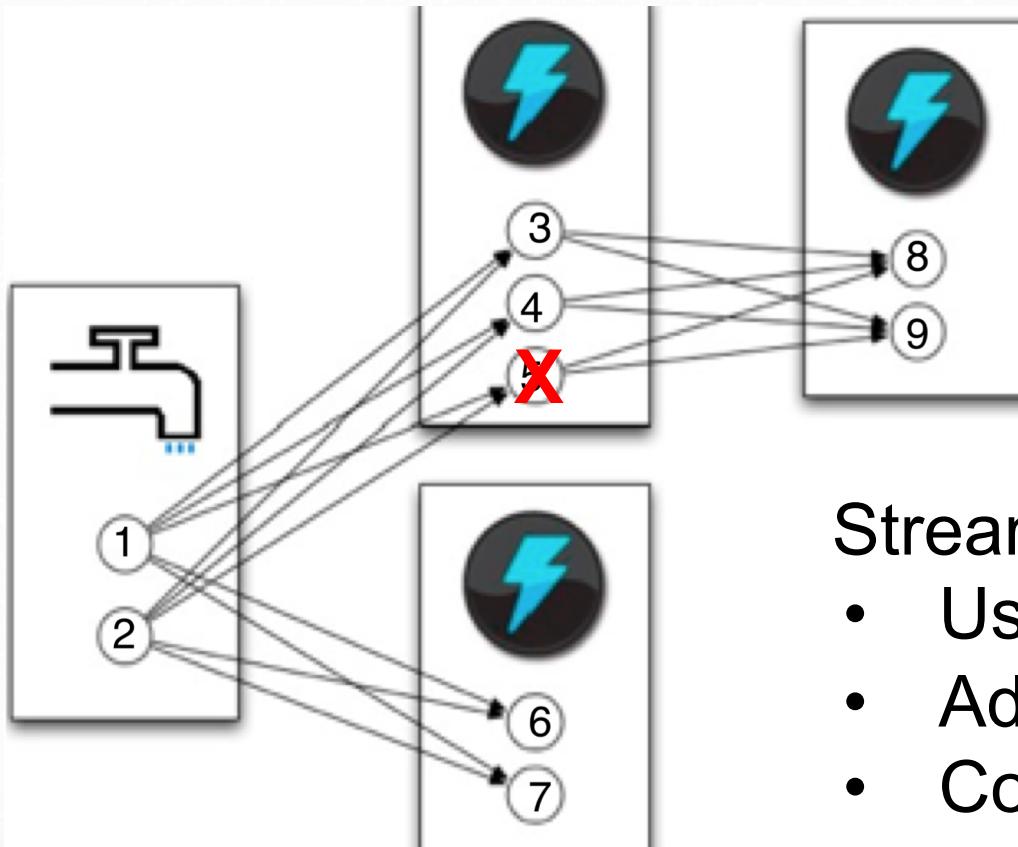
Hadoop YARN: MapReduce & Beyond



- Yahoo! deployed YARN into 30k+ nodes in production.
- YARN Apps ... MapReduce, Storm, etc.

Storm: Distributed Stream Processing

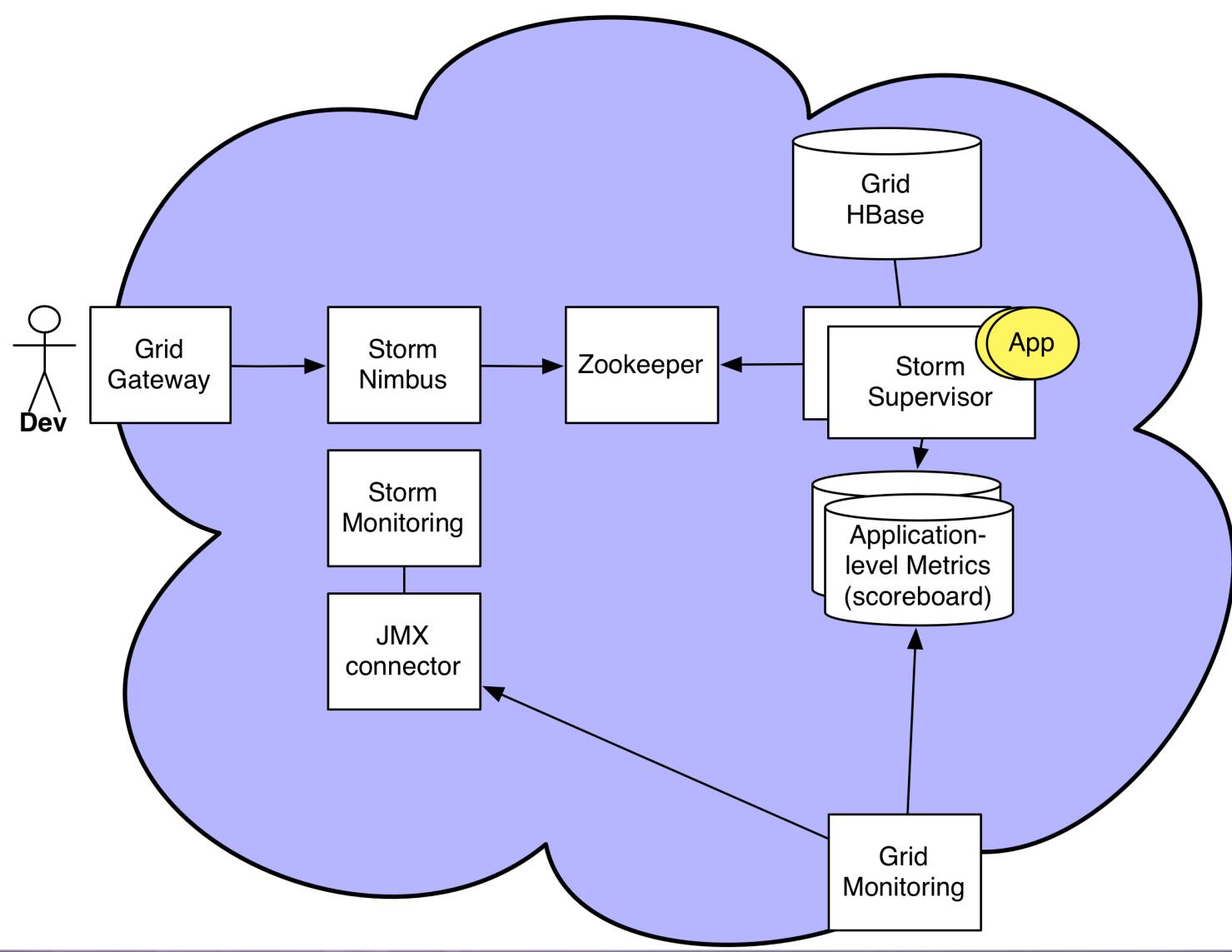
<https://github.com/nathanmarz/storm>



Streams

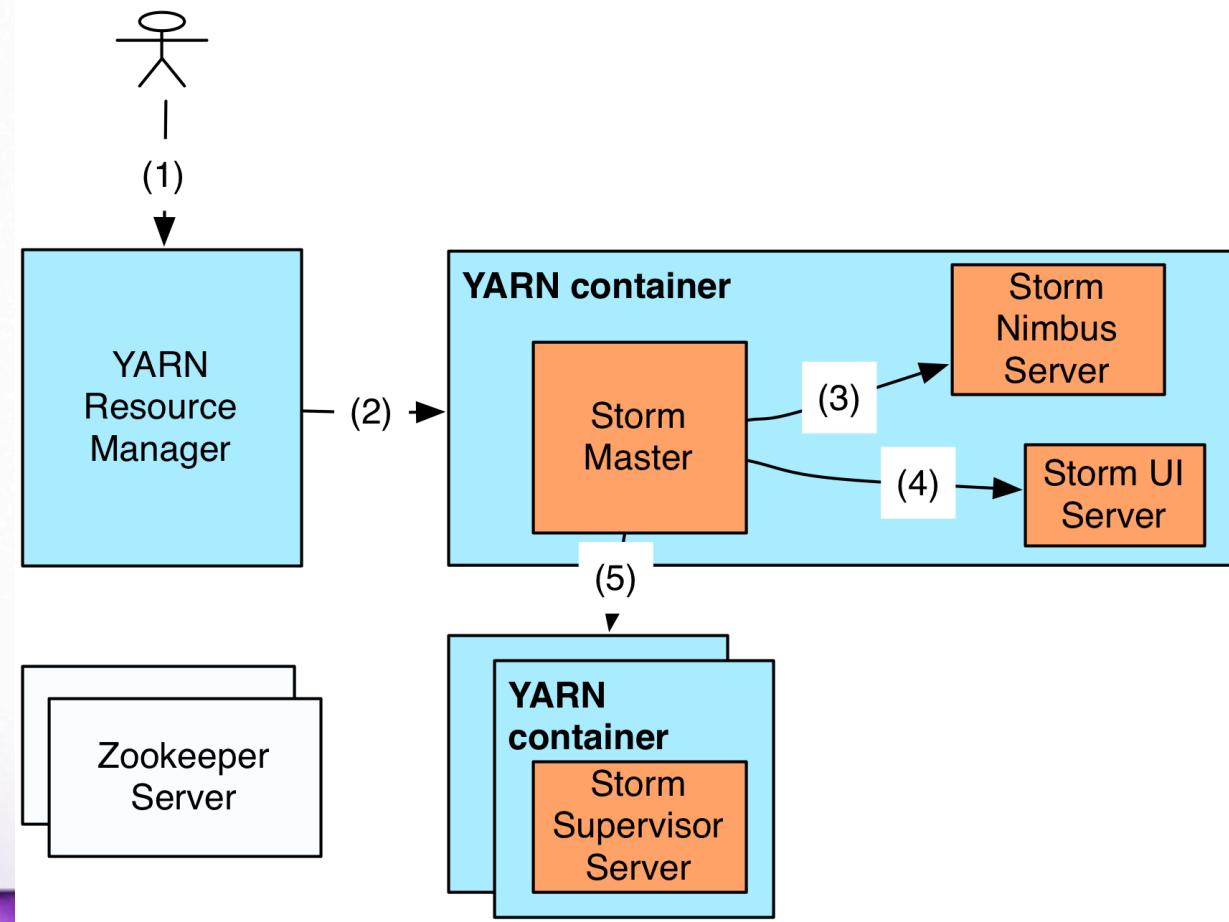
- User activities
- Ad beacons
- Content feeds
- Social feeds
- ...

Storm Clusters on Hadoop Grid



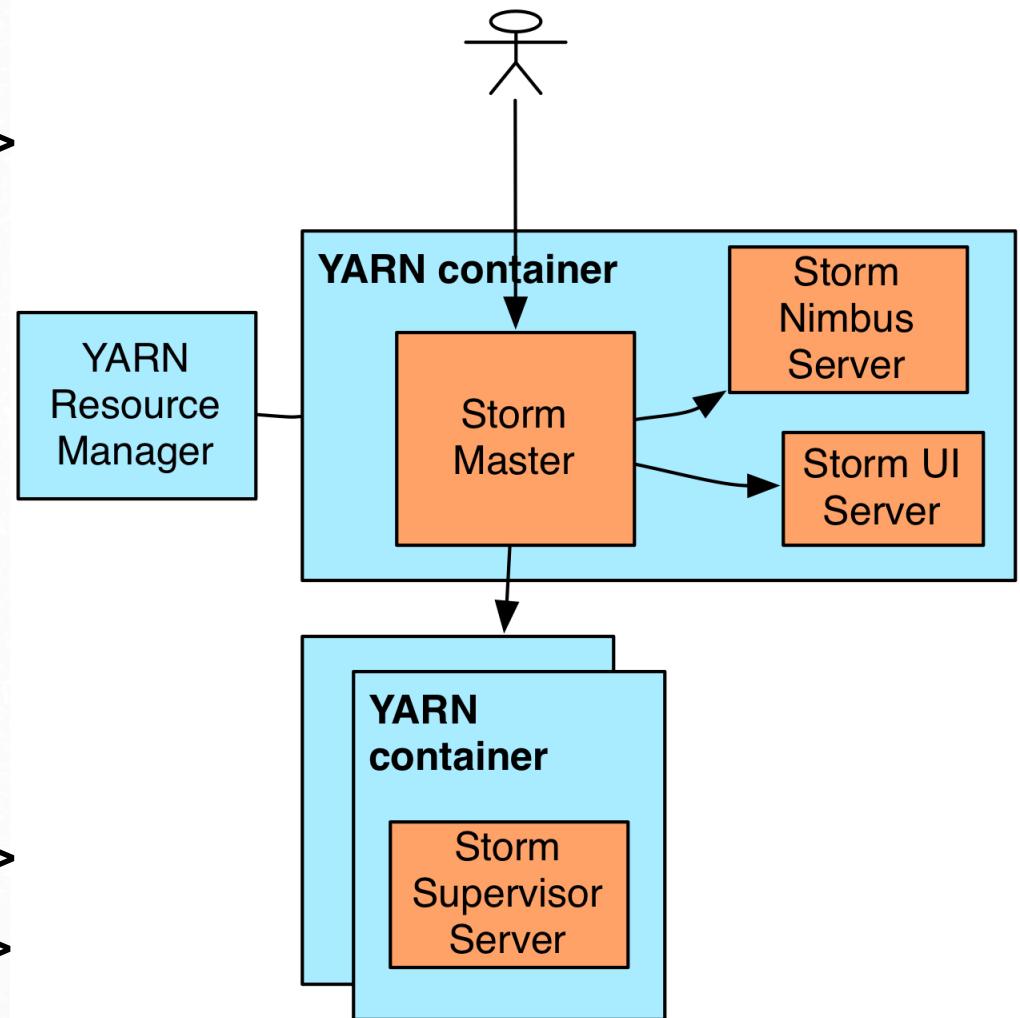
Storm-YARN: Launch Cluster

- `storm-yarn launch <conf>`
 - *Initial # of supervisors*
 - *memory size of allocated container*
- Result: `<appId>` of the newly launched Storm master



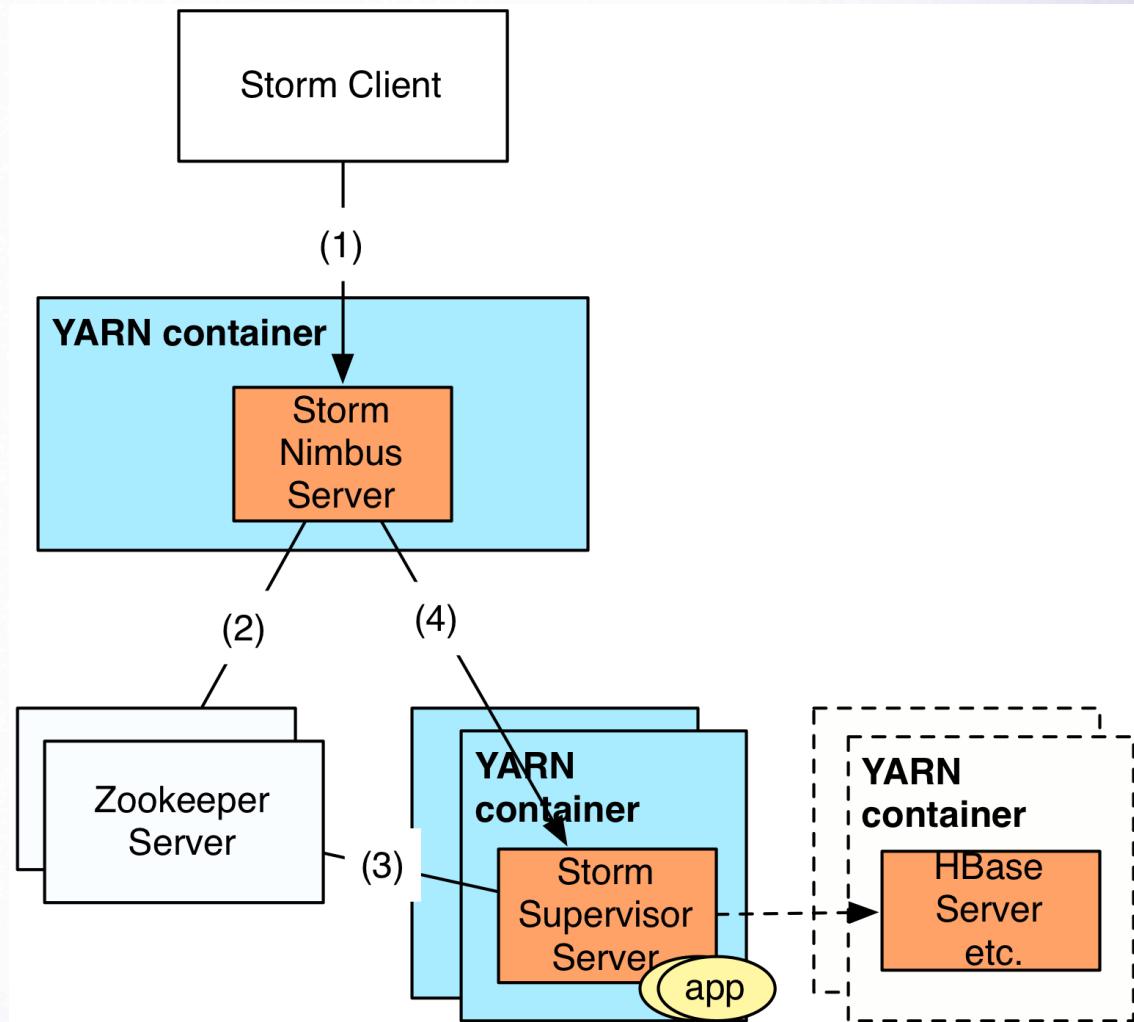
Storm-YARN: Manage Cluster

1. addSupervisors <appID> <count>
2. getStormConfig <appID>
3. setStormConfig <appID>
4. startNimbus <appID>
5. stopNimbus <appID>
6. startUI <appID>
7. stopUI <appID>
8. startSupervisors <appID>
9. stopSupervisors <appID>

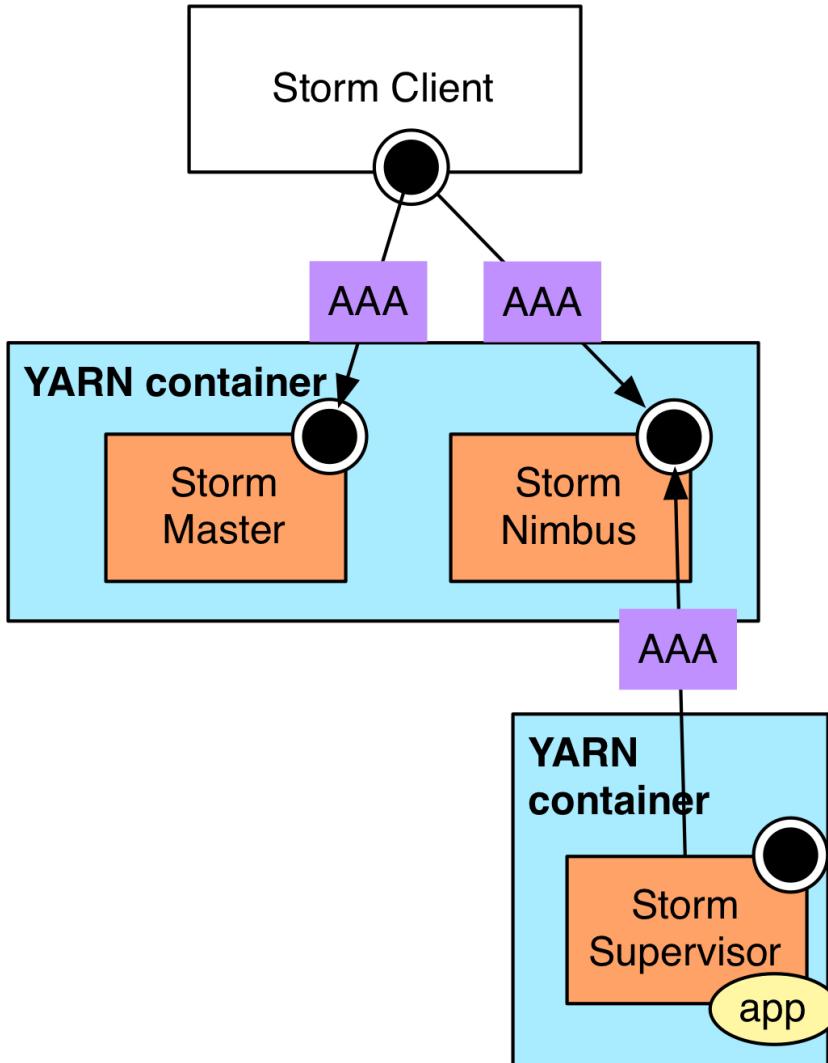


Storm-YARN: Deploy Apps

storm jar <appJar>



Authentication/Authorization/Audit



- Authentication plugins
 - Digest
 - Kerberos (soon)
 - None
 - *Bring your own*
- Authorization plugins
 - Accept all
 - Limited operations only
 - User whitelist
 - *Bring your own*
- Audit
 - Access log

Agenda

- Business motivation
- Technical overview
- **Open source**

Storm-YARN: Open Source

- Code released for early access
 - under the Apache 2.0 License
 - move to apache.org later
- Welcome contribution!
 - Submit proposals
 - Sign Apache style CLA
 - Submit git pull requests

<https://github.com/yahoo/storm-yarn>

The screenshot shows the GitHub repository page for 'storm-yarn'. At the top, there are download options: 'Clone in Mac', 'ZIP', 'HTTP', 'SSH', and 'Git Read-Only'. The URL 'git@github.com:yahoo/storm-yarn.git' is also displayed. Below this is a navigation bar with 'branch: master', 'Files', 'Commits', and 'Branches'. The main area shows a list of files and their details:

File	Last Commit	Description
bin	6 days ago	read-link simplified [anfeng]
lib	11 days ago	almost ready for integration test [anfeng]
src	a day ago	Merge pull request #12 from anfeng/master [revans2]
CLA.pdf	a day ago	Added a copy of the CLA for contributors to sign. [revans2]
LICENSE.txt	a month ago	Initial commit. Works with Hadoop 2.0.4-alpha and storm-0.9.0-wip17. [revans2]
README.md	2 days ago	Nimbus hostname should always be decided by StormMaster [anfeng]
create-tarball.sh	a month ago	Initial commit. Works with Hadoop 2.0.4-alpha and storm-0.9.0-wip17. [revans2]
pom.xml	8 days ago	Integrated test now available [anfeng]

Storm-YARN: mvn test

1. storm-yarn launch

- ./conf/storm.yaml --stormZip lib/storm.zip --appname storm-on-yarn-test --output target/appId.txt

2. storm-yarn getStormConfig

- ./conf/storm.yaml --appId application_1372121842369_0001 --output ./lib/storm/storm.yaml

3. storm jar

- lib/storm-starter-0.0.1-SNAPSHOT.jar
- storm.starter.WordCountTopology
- word-count-topology

4. storm kill

- word-count-topology

5. storm-yarn shutdown

- ./conf/storm.yaml --appId application_1372121842369_0001

Storm UI

Topology summary

Name	Id	Status	Uptime
word-count-topology	word-count-topology-1-1372210293	KILLED	38s

Topology actions

Activate Deactivate Rebalance Kill

Topology stats

Window	Emitted	Transferred	Complete latency (ms)
10m 0s	20200	10860	0.000
3h 0m 0s	20200	10860	0.000
1d 0h 0m 0s	20200	10860	0.000
All time	20200	10860	0.000

Spouts (All time)

Id	Executors	Tasks	Emitted	Transferred	Complete latency (ms)
spout	5	5	1460	1460	0.000

Bolts (All time)

Id	Executors	Tasks	Emitted	Transferred	Capacity (last 10m)	Execute latency (ms)
count	12	12	9340	0	0.005	0.113
split	8	8	9400	9400	0.001	0.056

Storm-YARN: Deployment

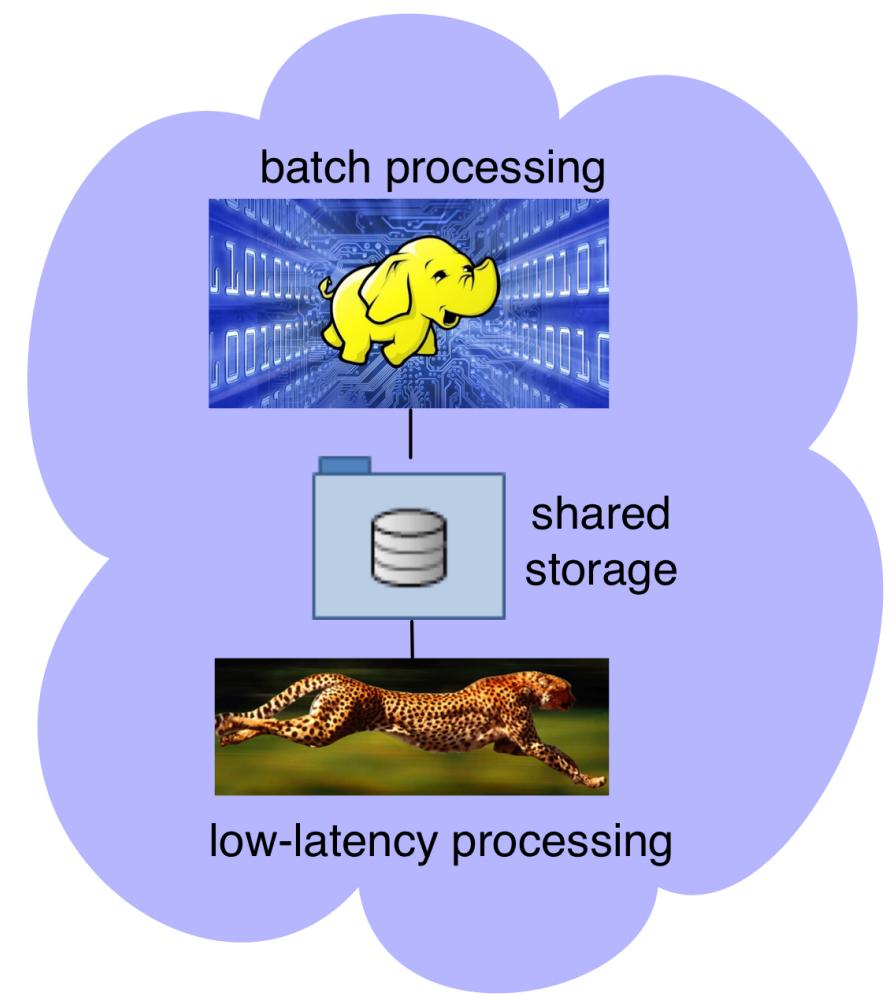
Install Storm S/W

1. hadoop fs –put
storm.zip /lib/storm/
<version>/storm.zip

Apply Storm-YARN

2. storm-yarn launch
→ <appID>
3. storm-yarn
getStormConfig
<appID>
→ <storm.yaml>
4. storm jar <appJar>

Conclusion



- YARN empowers the emergence of big-data & low-latency processing
- Yahoo! open source:
 - Storm-yarn @ [github/yahoo](https://github.com/yahoo)
 - Spark-yarn @ spark-project.org

MORE THAN EVER BEFORE



?

Questions