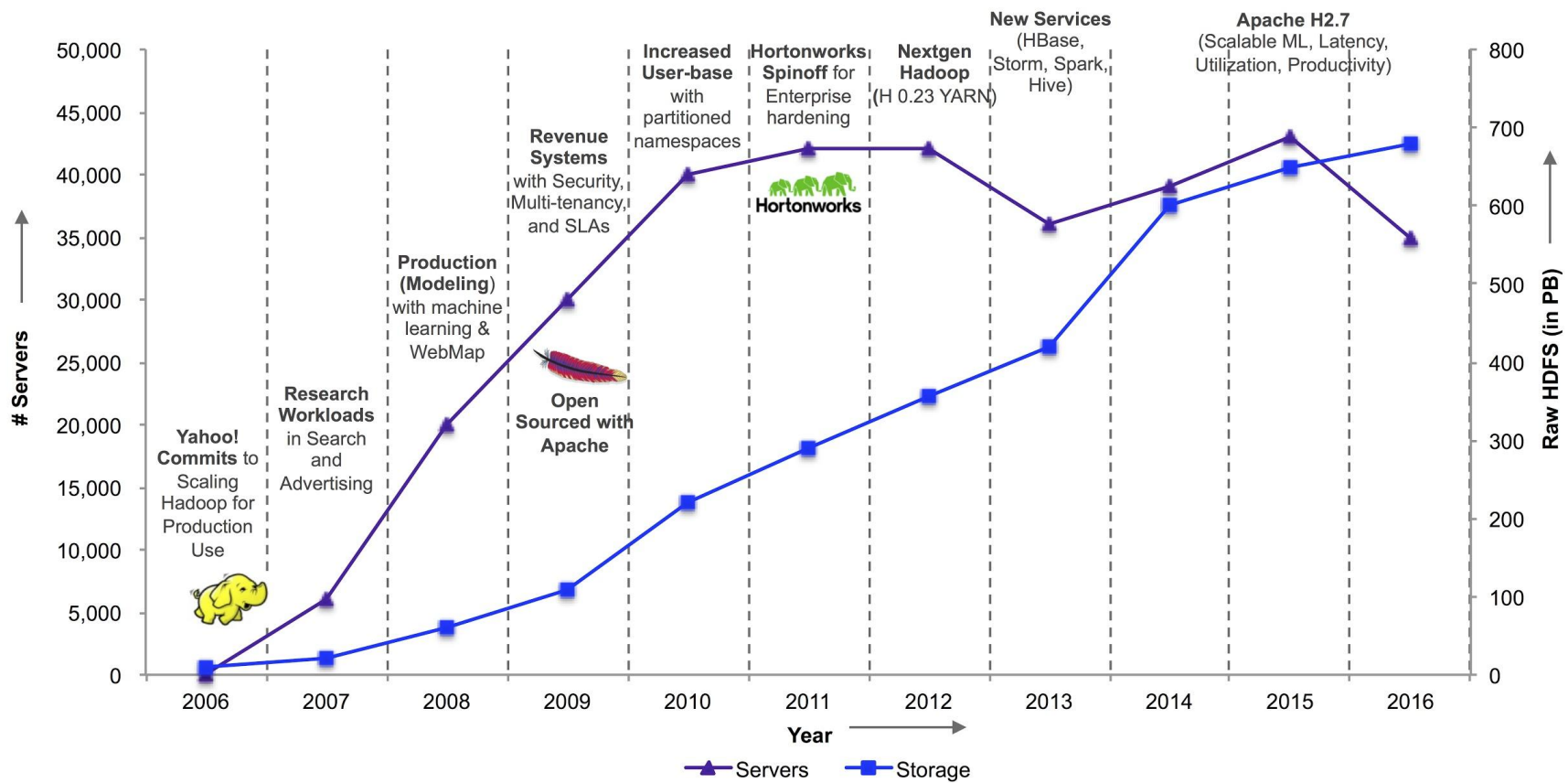




Big-Data Innovation: Hadoop, Real-time, and Machine Learning

Andy Feng
Yahoo

Hadoop at Yahoo



Real-time Processing at Yahoo

Big-Data Applications

Weather App



- **Beauty**
 - › Computational assessed
- **Relevant**
 - › Location
 - › Time
 - › Cloudy
 - › Shower
 - › ...



Yahoo Weather App

YAHOO!

Flickr [flickr.com/cameraroll](https://www.flickr.com/cameraroll)

Camera Roll

Photostream

Albums

Favorites

Groups

Creations

Stats

More ▾

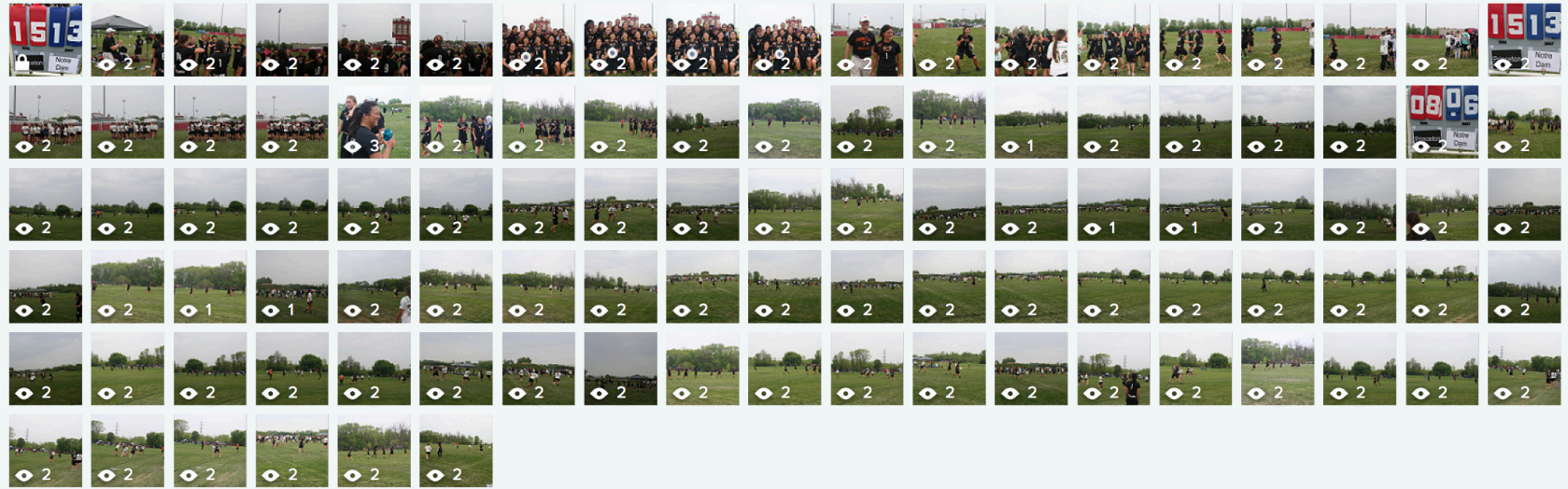
Show info

Date taken ▾

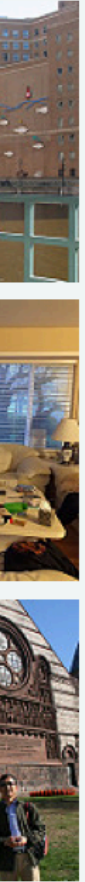
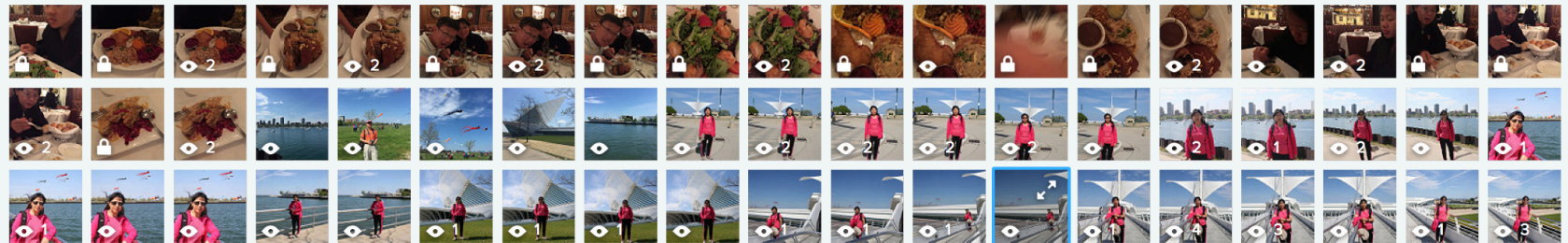
Magic View



May 24, 2015 [Select all](#)



May 23, 2015 [Select all](#)



- Anima
- Archit
- arch
- brid
- buil
- city
- doo
- scul
- tow
- othe
- Food
- Lands
- People
- Plant
- Style

Yahoo Vision Kit

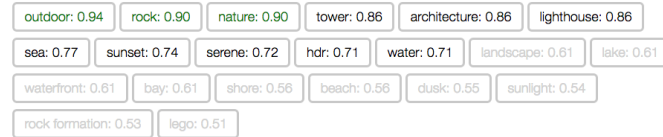


Image Enrichment Results

Dominant Colors



Autotags



Aesthetics



NSFW



Faces

Celebrities

Similarity



Personalized Content

YAHOO!

 Search Web

Andrew 2 Corp Mail

- Mail
- News
- Sports
- Finance
- Celebrity
- Shopping
- Movies
- Politics
- Beauty
- Style
- Tech
- TV
- More on Yahoo

Trump's economic plan prompts controversy
A prediction that it would cause a lengthy recession draws a rebuttal from the presumptive GOP nominee. Evidence of bias?

Bernie Sanders Admits for the First Time That 'It Doesn't Appear' He'll Be the Democratic Nominee
Bernie Sanders acknowledged for the first time on Wednesday that he will likely not be the Democratic presidential nominee. "It doesn't appear that I'm

Top 8 Credit Cards For Those With Excellent Credit
Banks are offering insane credit card deals if you have great credit. Take advantage of double cash back, a \$625 intro bonus, 0% interest until 2018.

Even one of the most bullish Tesla analysts is running scared
Morgan Stanley's Adam Jonas, one of the most bullish Tesla Motors (TSLA) analysts, had a change of heart on the stock Thursday. Jonas slashed his

Sia Revealed! Singer Accidentally Shows Her Face During Concert
Sia's attempts at privacy were foiled when a particularly strong gust of wind blew back her trademark hairpiece, giving fans a glimpse of her face for the

Dumping Donald Trump would be worse than keeping him
Conservatives who are understandably appalled by the prospect of a Donald Trump nomination - and therefore understandably intrigued by the notion of a

- Trending Now**
1. Kristen Wiig
 2. Zara Holland
 3. Tiger Woods
 4. Red Robin
 5. Psoriatic Arthritis ...
 6. Flower Delivery
 7. Blake Lively
 8. Calysta Bevier
 9. Buy Auto Tires
 10. Ashley Benson

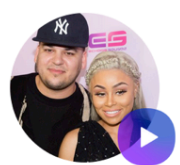
COMPLIMENTARY PREMIER PACKAGE FOR A LIMITED TIME ONLY
This is our standard
VIEW OFFER

Sunnyvale, CA

Today	Fri	Sat	Sun
77° 56°	81° 58°	85° 58°	85° 59°



Yahoo TV
Selena Gomez does 'Carpool Karaoke' on roller coaster

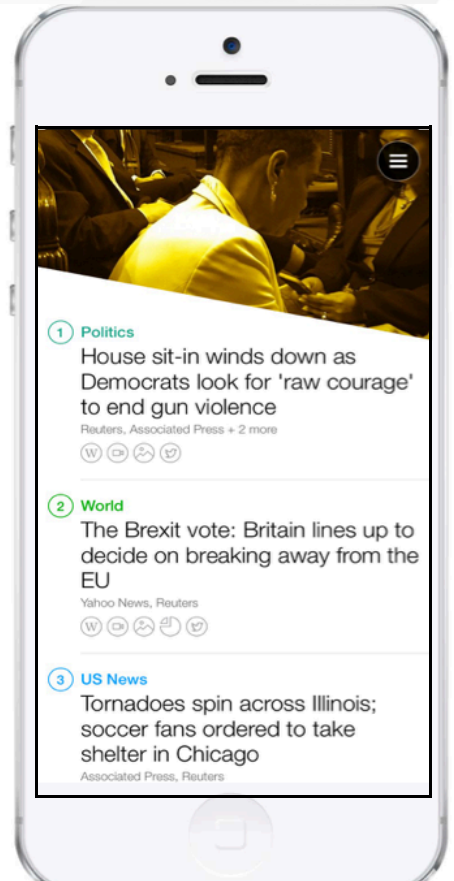


Good Morning America
Blac Chyna gushes about 'teddy bear' Rob Kardashian

Content augmentation

User profiling

Recommendation



YAHOO!

Mail Smart Views

The screenshot displays an email client interface. On the left, a sidebar contains folders: Compose, Inbox (9999+), Drafts (5), Sent, Archive, Spam (240), Trash (9999+), and Smart Views. Under Smart Views, there are sub-categories: Important, Me, Unread, Starred, People, Social, Shopping, Travel, and Finance. Red arrows point to Spam, Smart Views, and Shopping. The main pane shows a list of 97 results, sorted by Date. The list includes messages from KAYAK, HRG North America, Concur Travel, Southwest Airlines, NH Hotel Group, and TripCase. The messages are grouped by date: Last week, Earlier in June, May, February, January, and 2015.

Compose

Messages Photos Documents 97 Results Saved Refine

Archive Move Delete Spam More Sorted by Date

Last week

KAYAK **BOOK Travel Faster: KAYAK - We Search 100s of S** Sponsored

HRG North A... Travel Reservation to BURBA... HRG North America Inbox Jun 21

Concur Travel **Concur Itinerary 07-07-201...** Trip Overview Trip M Inbox Jun 21

Southwest Airli... Flight reservation (9T5J5K) | ... Southwest Airlines, Inbox Jun 21

Earlier in June

NH Hotel Group **NH Collection le ofrece una...** NH Collection le ofr Inbox Jun 9

May

NH Hotel Group ¡Te esperamos en la piscina! NH Rewards Hoteles c Inbox May 31

February

NH Hotel Group **¿Todo listo para tu escapa...** ¿Todo listo para tu Inbox Feb 25

January

HRG North A... Travel Reservation to CHAM... HRG North America Inbox Jan 25

Concur Travel Concur Itinerary 16-02-2016:... Trip Overview Trip M Inbox Jan 25

2015

TripCase **Your Trip 'the Hvatt Redenc...** Dear Andrew. Thank Inbox 12/4/15

Search

YAHOO!

2015 mercedes technology package

Search

Andrew

Corp Mail

Web
Images
Video
News
Local
Answers
Shopping
More

Anytime
Past day
Past week
Past month

Ad related to 2015 mercedes technology package

[Mercedes-Benz® GLA250 - See Offer Details.](#)
www.MBUSA.com/GLA/250

Lease the 2015 GLA250 4MATIC® for \$369 Per Month.

[Buy Area Dealerships](#)

[Build Your Own](#)

[Compare Vehicles](#)

[Coupes](#)

[Current Special Offers](#)

[Sedans](#)

See more ads for: [2015 mercedes technology package](#), [2015 mercedes c class](#)

[2015 Mercedes S-class coupe wraps high-tech, high luxury in ...](#)
www.autonews.com/article/20141017/OEM04/141019879/2015... Cached

The 2015 Mercedes-Benz S-class coupe embodies the automaker's new design philosophy of modern luxury, says Gordon Wagener, the brand's design chief.

[2015 Acura RDX Technology Package 4dr SUV AWD \(3.5L 6cyl 6A ...](#)
cars.findthebest.com > [Motors](#)

Compare Cars: 2015 Acura RDX Technology Package 4dr SUV AWD (3.5L 6cyl 6A) vs Mercedes-Benz GLA-Class GLA45 AMG 4MATIC 4dr SUV AWD (2.0L 4cyl Turbo 7AM). In this side ...

[2015 Mercedes Technology Package - Image Results](#)



[More 2015 Mercedes Technology Package images](#)

[2015 Acura RDX Technology Package 4dr SUV AWD \(3.5L 6cyl 6A ...](#)
cars.findthebest.com > [Motors](#)

Compare Cars: 2015 Acura RDX Technology Package 4dr SUV AWD (3.5L 6cyl 6A) vs Mercedes-Benz GLK-Class GLK350 4MATIC 4dr SUV AWD (3.5L 6cyl 7A). See all the ...

See more ads for:

[2015 mercedes technology package](#)

[2015 mercedes c class](#)

[disney world vacation packages 2015](#)

[mercedes gla 2015](#)

[mercedes benz gla 2015](#)

[disney vacation packages 2015](#)

Ads

[2015 Mercedes-Benz S550](#)
FleetRates.com

New Rebate Now \$699/mo \$82,275
Low Prices Nationwide Delivery

[2015 Mercedes |](#)
Edmunds.com

www.Edmunds.com

★★★★★ 78 reviews

Research, Reviews, Pricing & More.
Edmunds.com® - Ask the Car People!

[Top-Rated SUVs 2015](#)

educatehow.com/Savings

Explore the Best New SUVs for 2015.
Compare Prices & Specs.

Query intention

Page ranking

Ads matching

YAHOO!

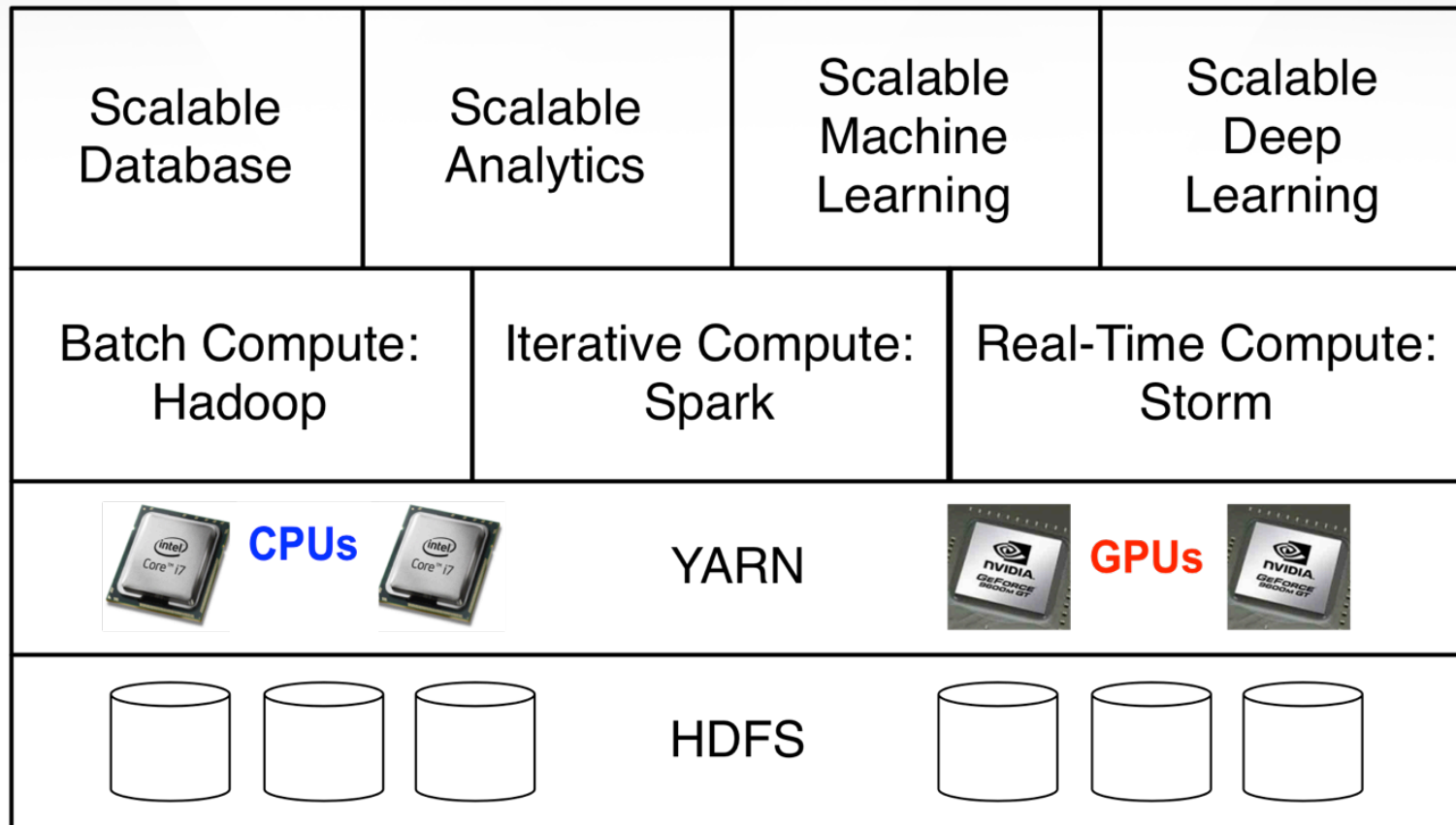
Yahoo Search2Vec: Big Data & Machine Learning

<http://bit.ly/28SLdjU>

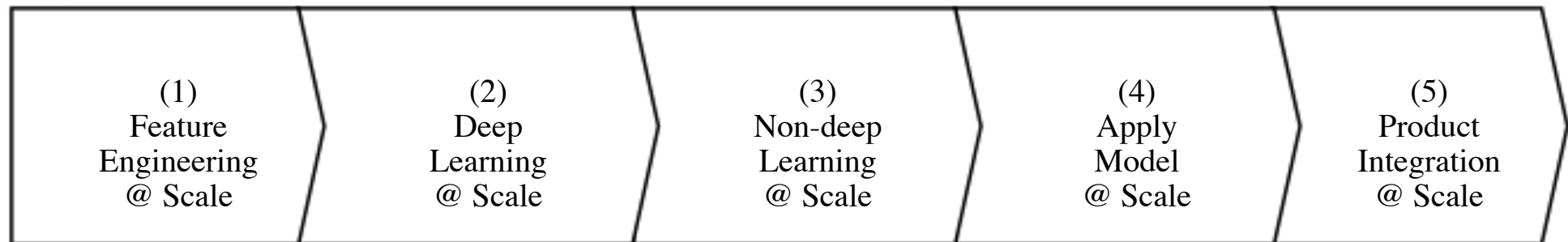
Bucket Tests	Query Coverage	Auction Depth	Revenue per Search
Simple model vs. Baseline	+1.14%	+2.13%	+7.07%
Large model vs. Baseline	+3.61%	+4.57%	+17.12%

Big-Data Innovations

Technology Stack



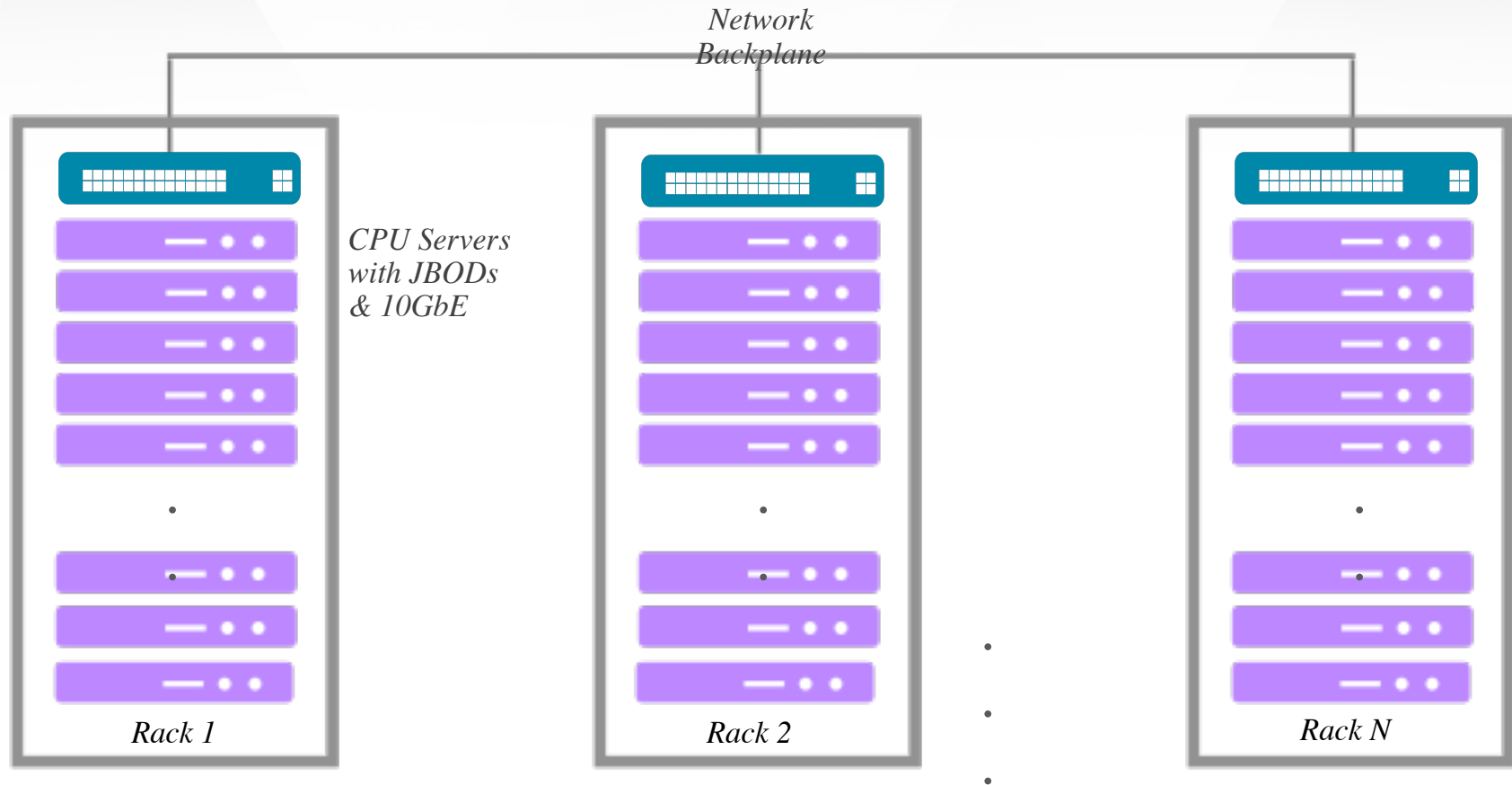
Flickr Pipeline: 10B photos, 8M per day



* <http://bit.ly/1KIDfof> by Pierre Garrigues, Deep Learning Summit 2015

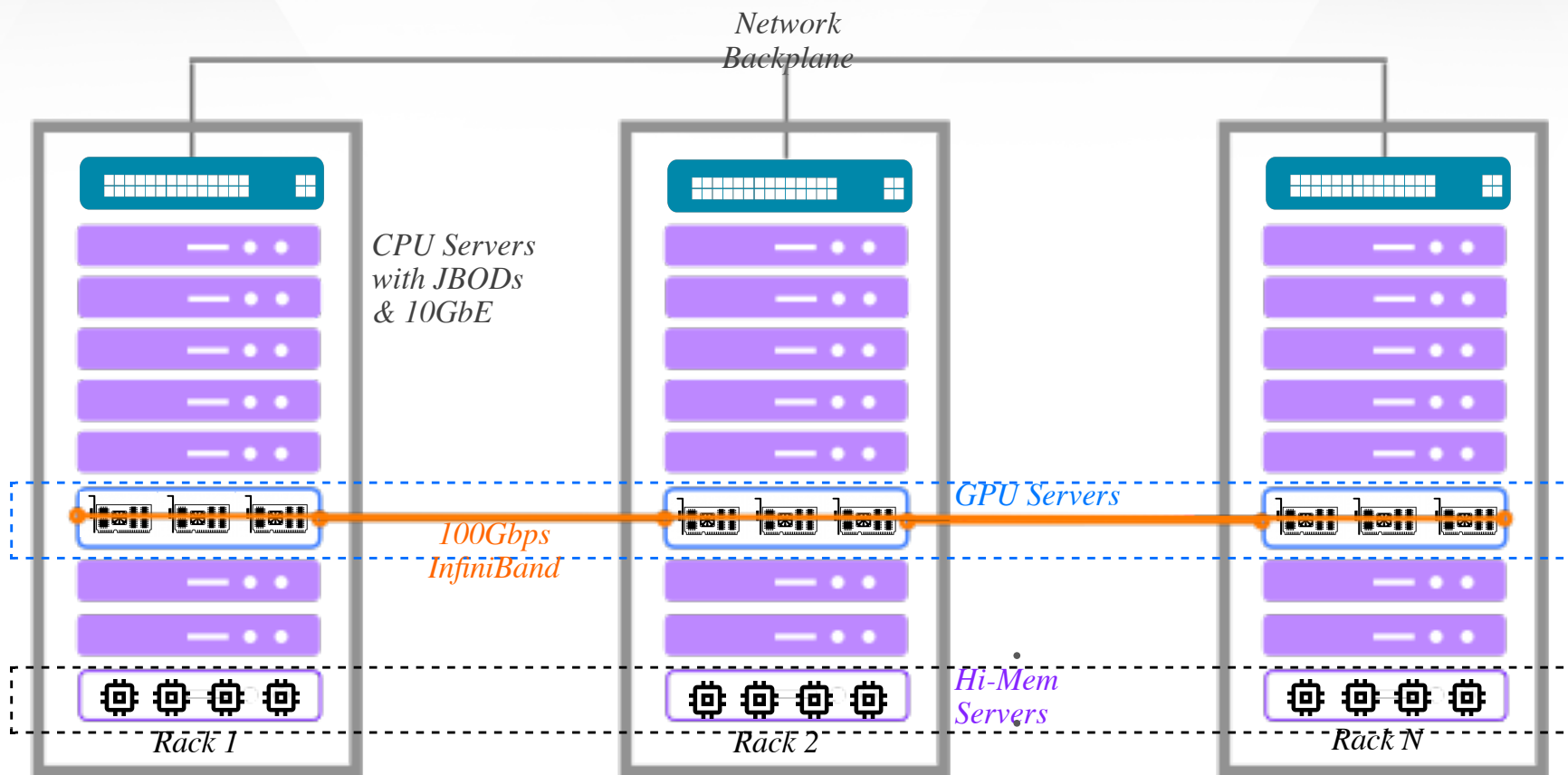
YAHOO!

System Configuration: Homogenous Clusters

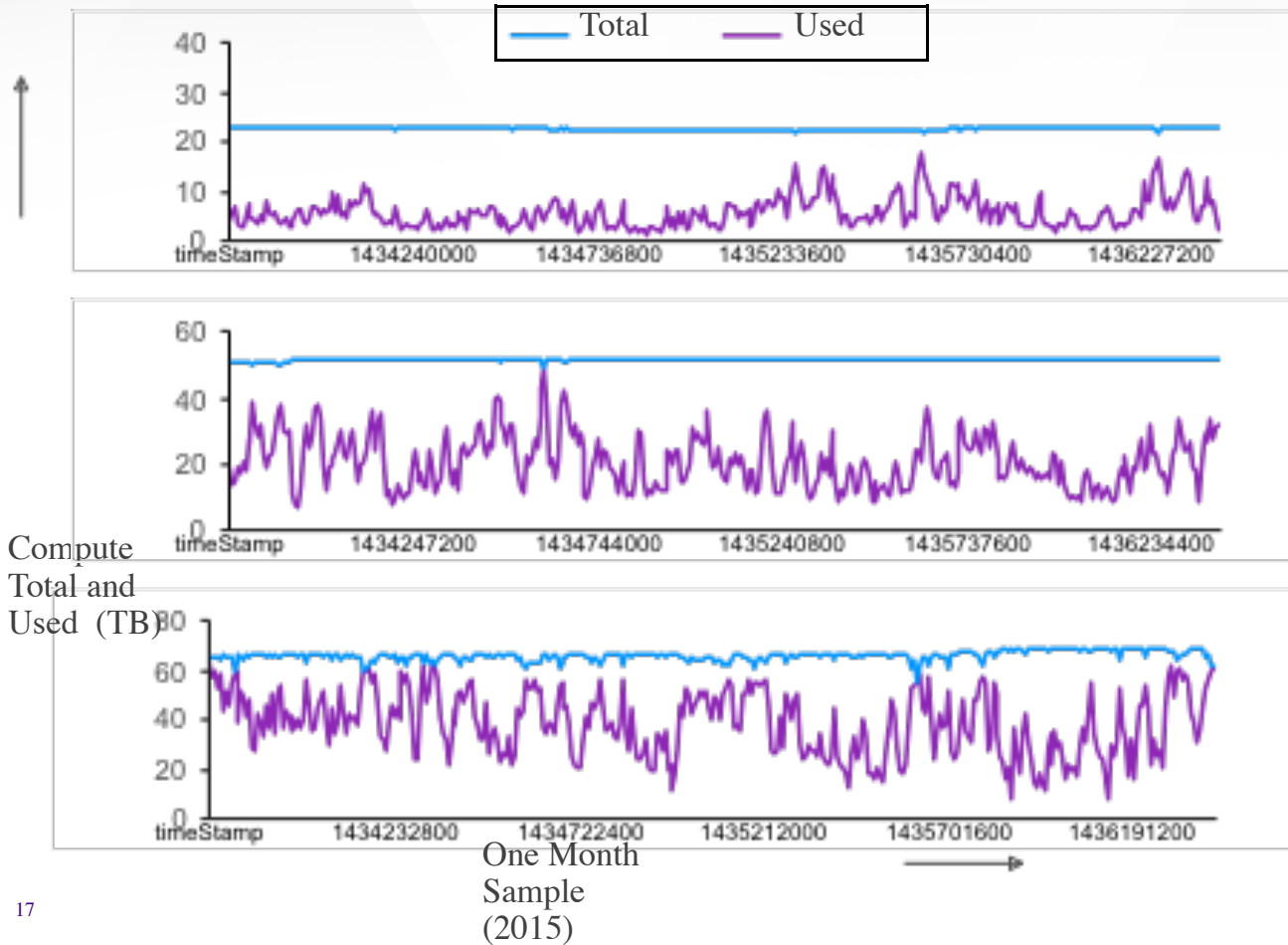


System Configuration: Heterogeneous Clusters

<http://bit.ly/295RNbU>



Resource Allocation: Early Days



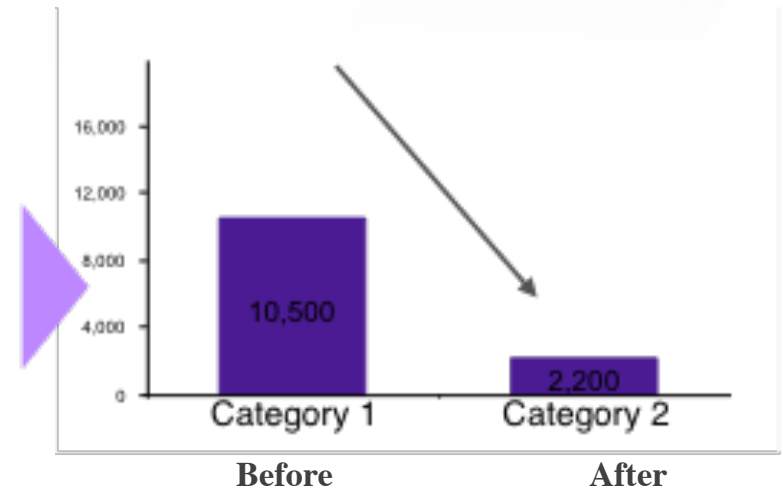
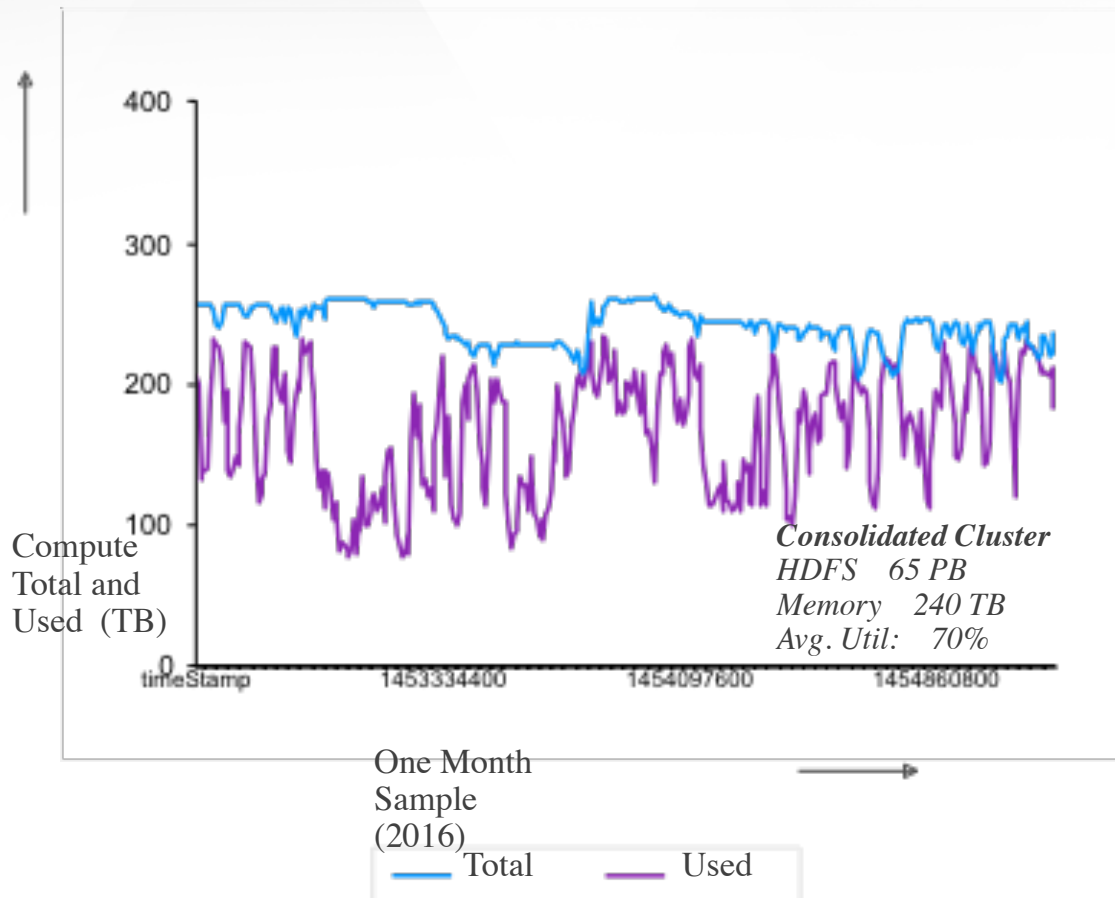
Cluster 1 (2,000 servers)
 HDFS 12 PB
 Memory 23 TB
 Avg. Util: 26%

Cluster 2 (3,100 servers)
 HDFS 21 PB
 Memory 52 TB
 Avg. Util: 40%

Cluster 3 (5,400 servers)
 HDFS 36 PB
 Memory 70 TB
 Avg. Util: 59%

Resource Allocation: 40% decrease in TCO

<http://bit.ly/295RNbU>

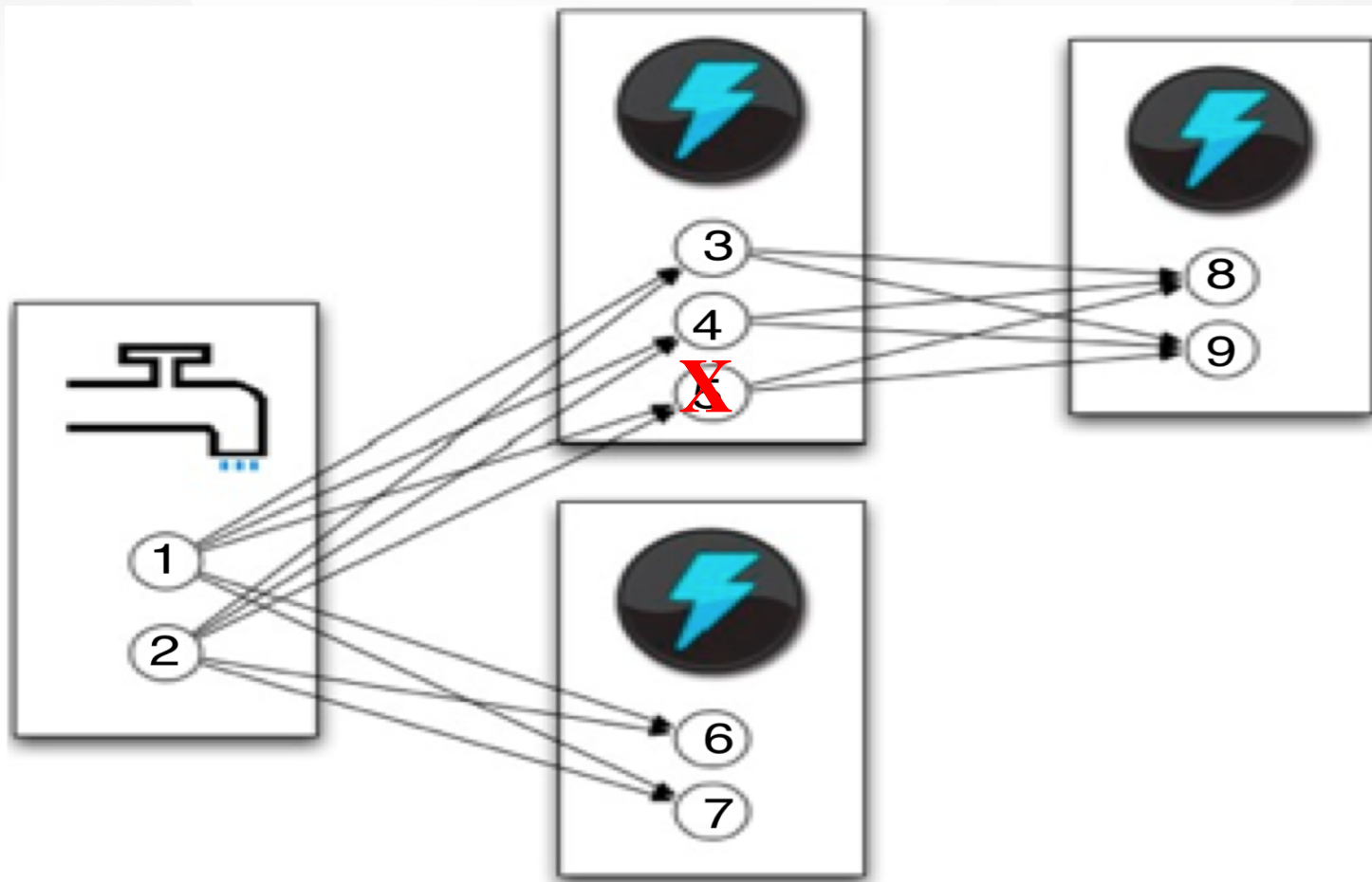


65% increase in compute capacity

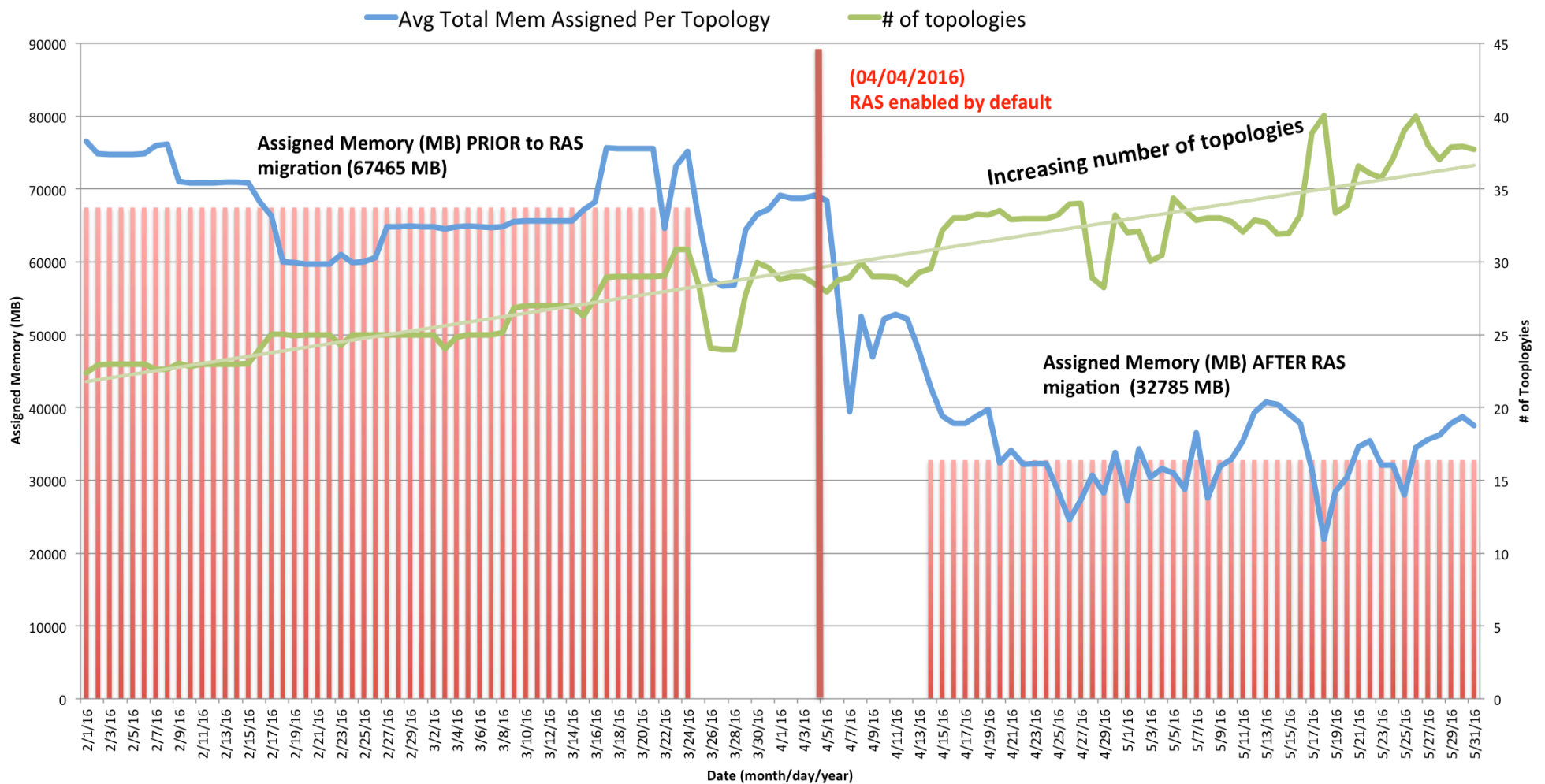
50% increase in avg. utilization

Apache Storm: Real-time Processing

<https://storm.apache.org>

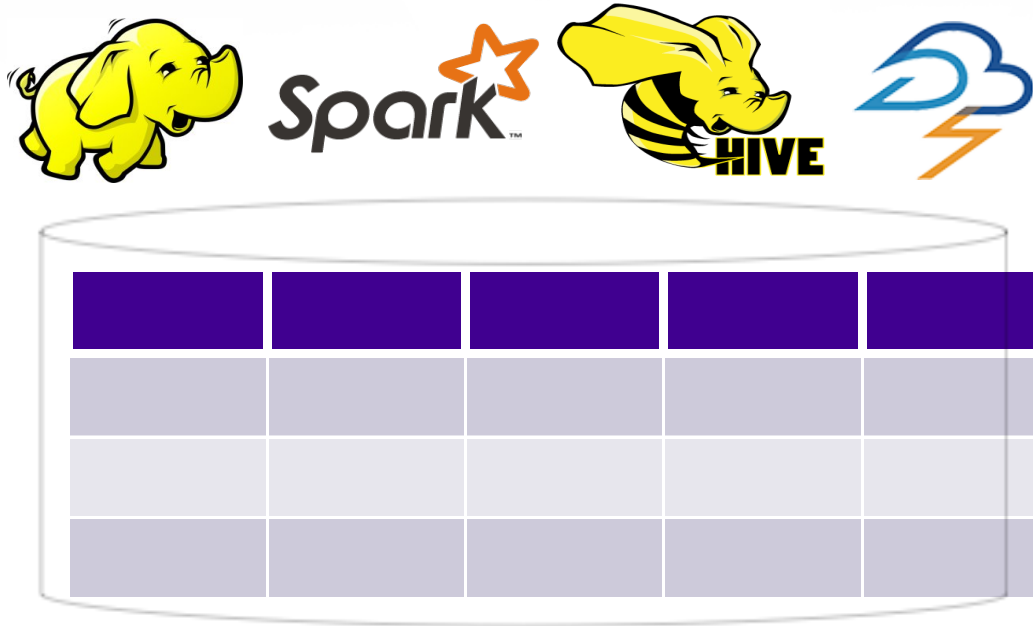


Resource Aware Scheduling: 100% increase utilization



Apache HBase: NoSQL Database for Hadoop

<https://hbase.apache.org>

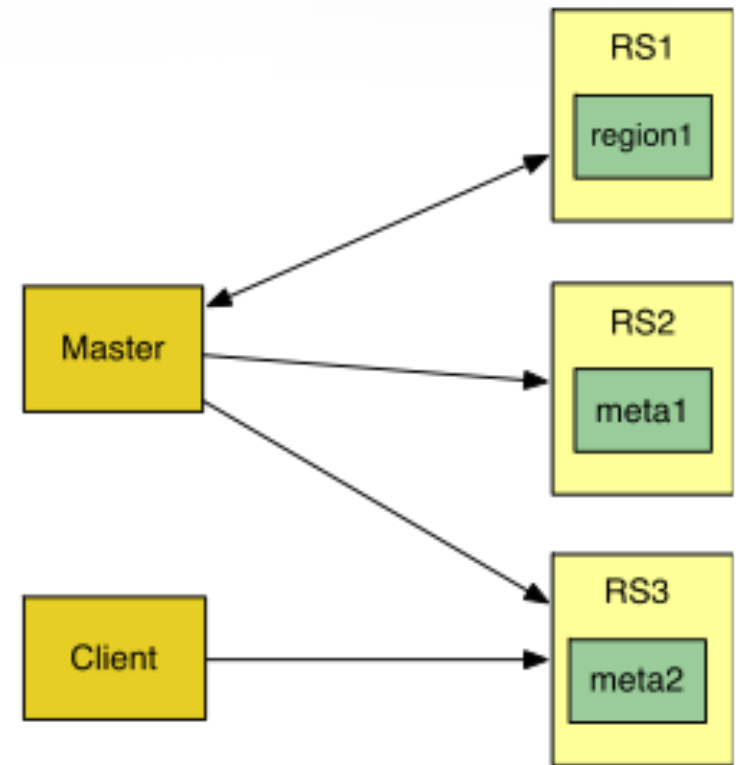
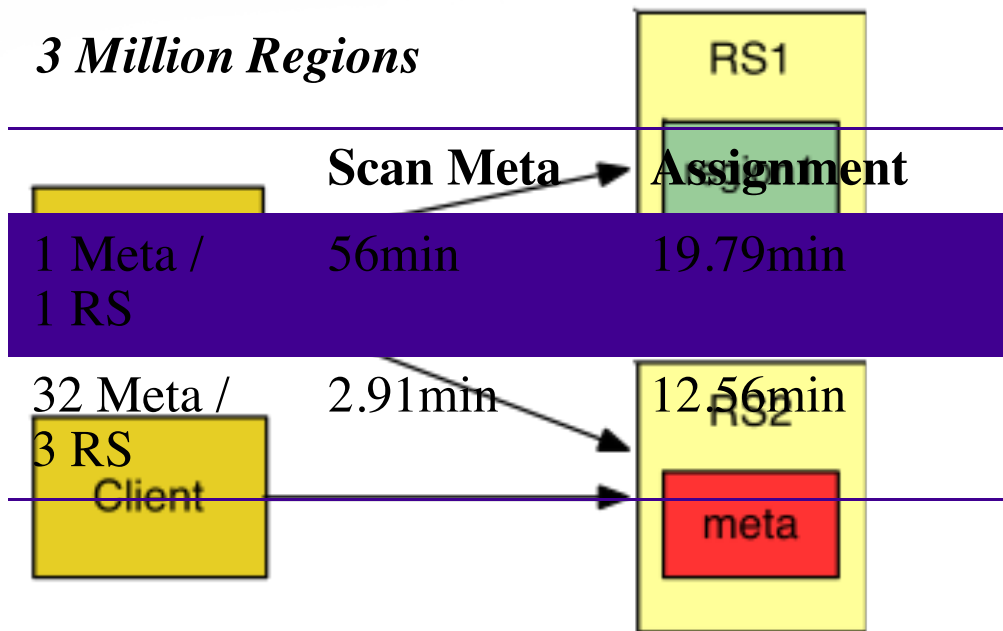


- Several million regions (partitions of tables)
- 800 region servers

Support Millions of Regions

<http://bit.ly/29h1LFW>

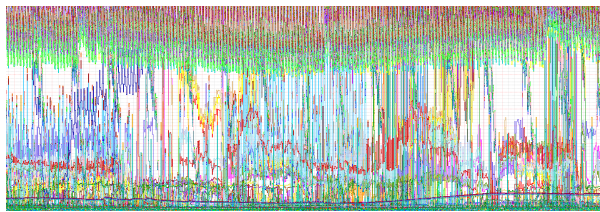
3 Million Regions



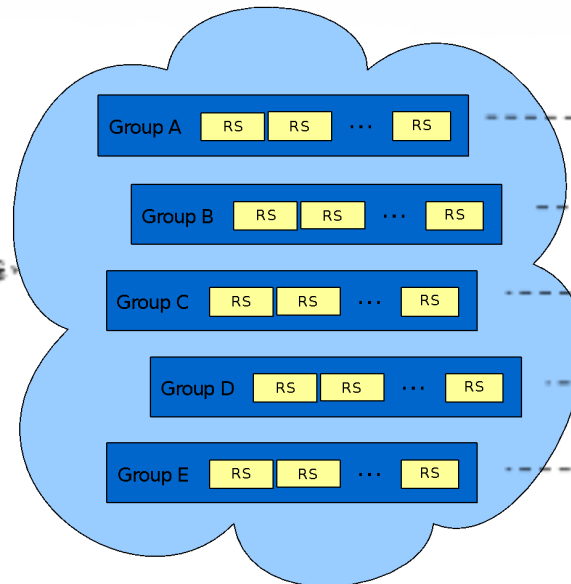
YAHOO!

HBase Multi-tenancy & Isolation

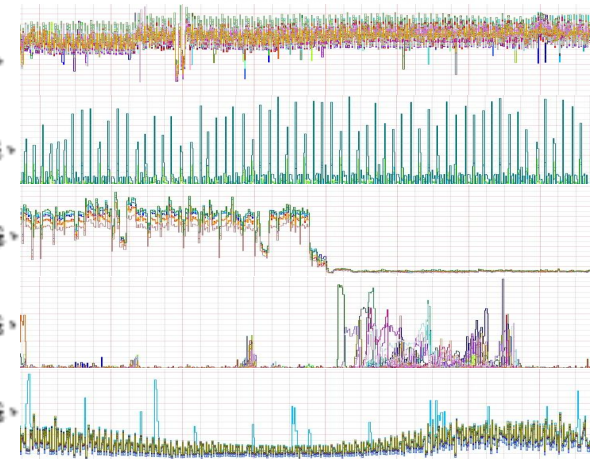
<http://bit.ly/28YK4a7>



Cluster Workload



Region Server Groups



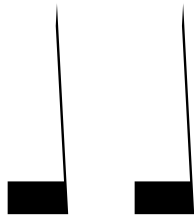
Isolated Group Workloads

Apache Omid: Transactions for NoSQL DB

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



ACID
Transacti
ons



A P A C H E
HBASE

- *Multi-row/multi-table transactions*
- *Snapshot isolation*
- *Lock-free*

Omid: Transactional HBase App

```
TransactionManager tm = HBaseTransactionManager.newBuilder().build();
```

```
Transaction tx = tm.begin();
```

```
Put row1 = new Put(Bytes.toBytes("EXAMPLE_ROW1"));
```

```
row1.add(family, qualifier, Bytes.toBytes("VALUE_1"));
```

```
tt.put(tx, row1);
```

```
Put row2 = new Put(Bytes.toBytes("EXAMPLE_ROW2"));
```

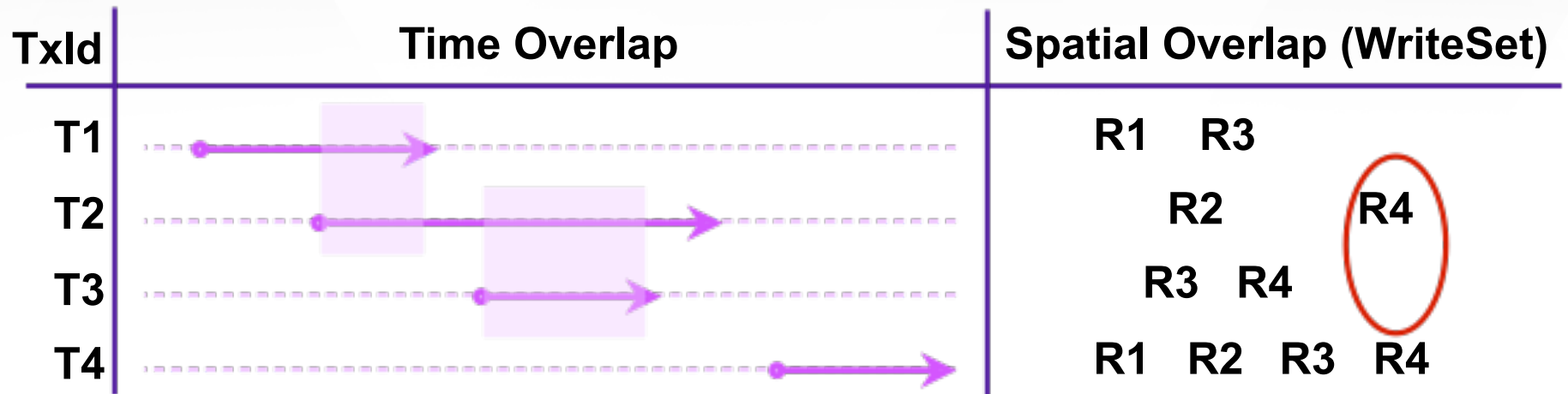
```
row2.add(family, qualifier, Bytes.toBytes("VALUE_2"));
```

```
tt.put(tx, row2);
```

```
tm.commit(tx);
```

Omid: Snapshot Isolation

<http://bit.ly/28YZvQF>



- T2 overlaps in time with T1 & T3, but
- Spatially $T1 \cap T2 = \emptyset$ → T1 and T2 can both commit.
- Spatially $T2 \cap T3 = \{ R4 \}$ → T2 or T3 will abort.

Machine Learning: Search2Vec

<http://bit.ly/28SLdjU>

Input: Training Data

S1: gas_caps gas_cap_replacement_for_car slc_679f037d
gas_door_replacement_for_car slc_466145af1 fuel_door_covers
adid_28540536 slc_348709d7 autozone_auto_parts adid_33183157
auto_zone slc_8dcdab5d slc_58f979b6

S2: hoka_running_shoe_reviews adid_22830711 hoka_shoes_for_bad_feet
hoka_shoes_amazon slc_231gfaw zappos_shoes slc_7c126f71
hoka_walking_shoes

S3: king_tut king_tut_exhibit king_tut_exhibit_seattle_2015 slc_726y6j51
charlies_seattle adid_55774014

Output:

Numeric Vectors

Of queries & ads

Vector("san jose weather") \approx

Vector("weather 95113") \approx

Vector(ad123)

Search2Vec: Cosine similarity b/w ads and queries

adid_243609

ad vector:

0.2 | 1.1 | 7.2 | 0.8 | 3.1



similarity=0.931

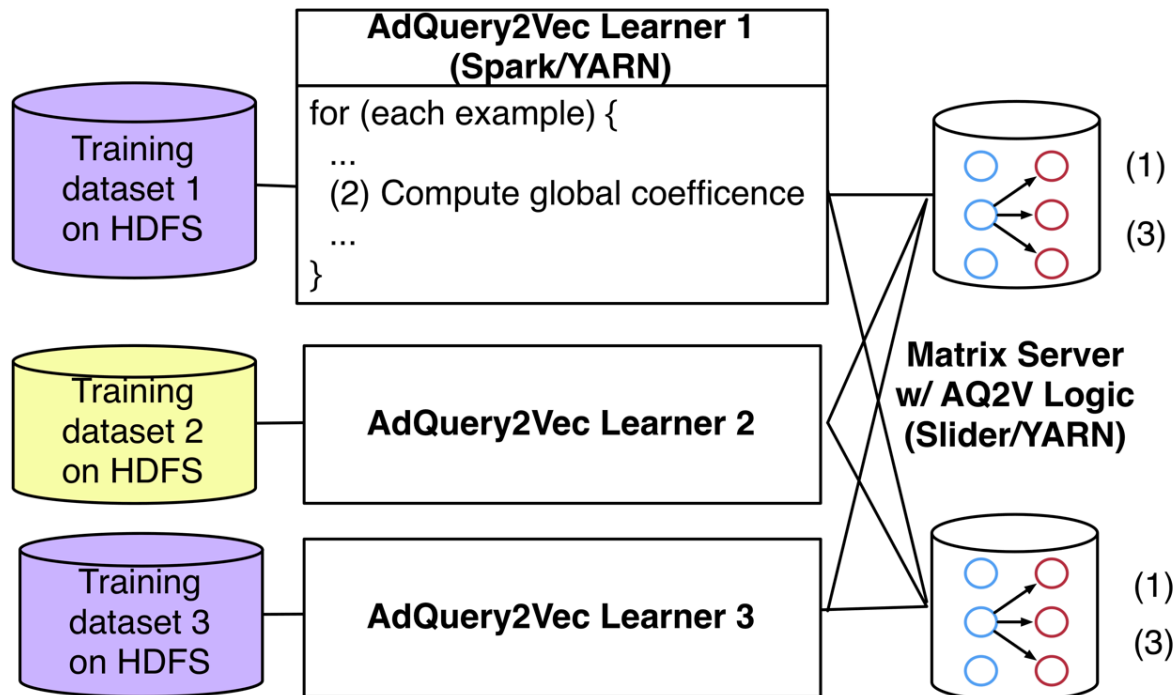
query vector:

0.2 | 1.2 | 6.8 | 0.7 | 3.2

mystery_games

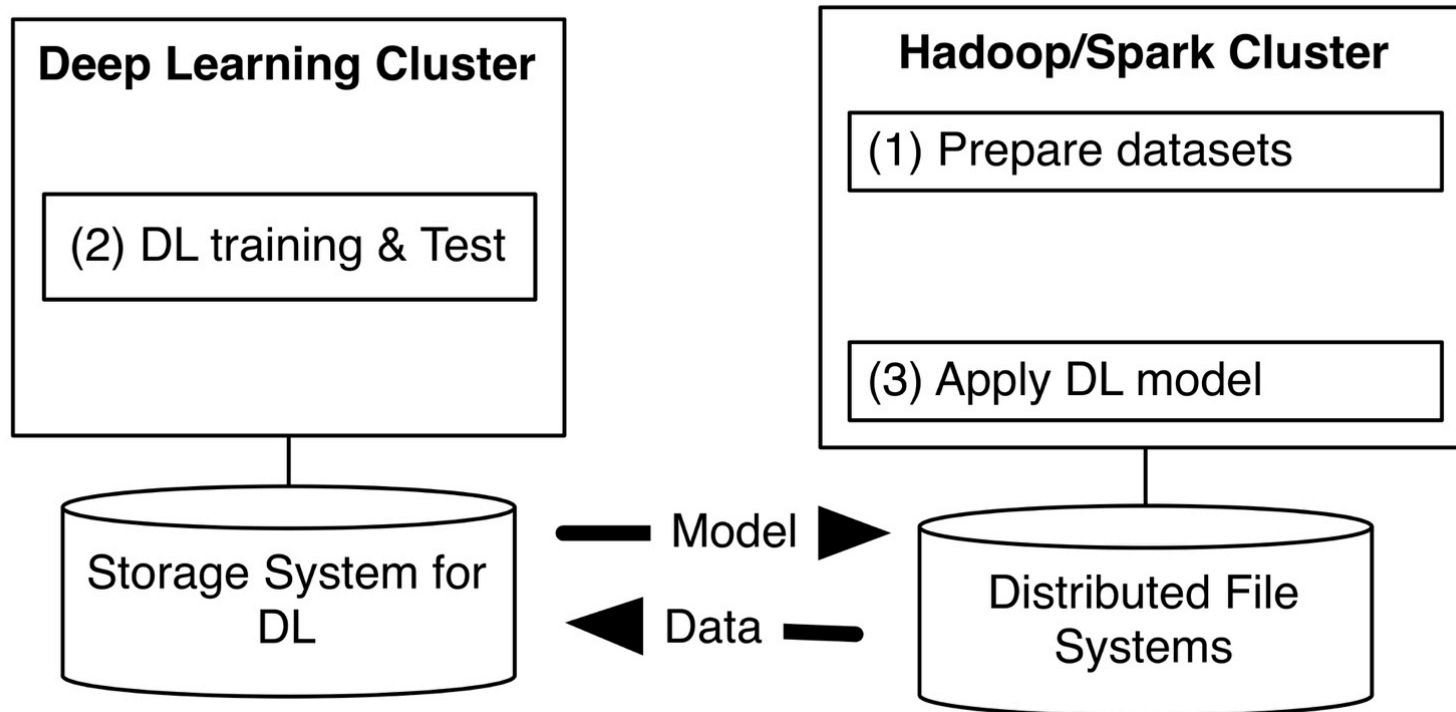
ad metadata	ad id: 243609 bidterm id: 341454 ad title: Host a Fun Murder Mystery Party ad description: Huge selection of fun murder mystery games for all ages, groups. #1 site for instant downloads and boxed sets of exciting murder mystery party games
generated keywords	murder mystery murder mystery games mystery games murder mystery party mystery party free murder mystery games for parties detective games murder mystery game how to host a mystery party for kids murder mystery dinner friends game night murder mystery parties at home murder mystery dinner party ...

Search2vec: 300M Unique Queries/Ads/Links

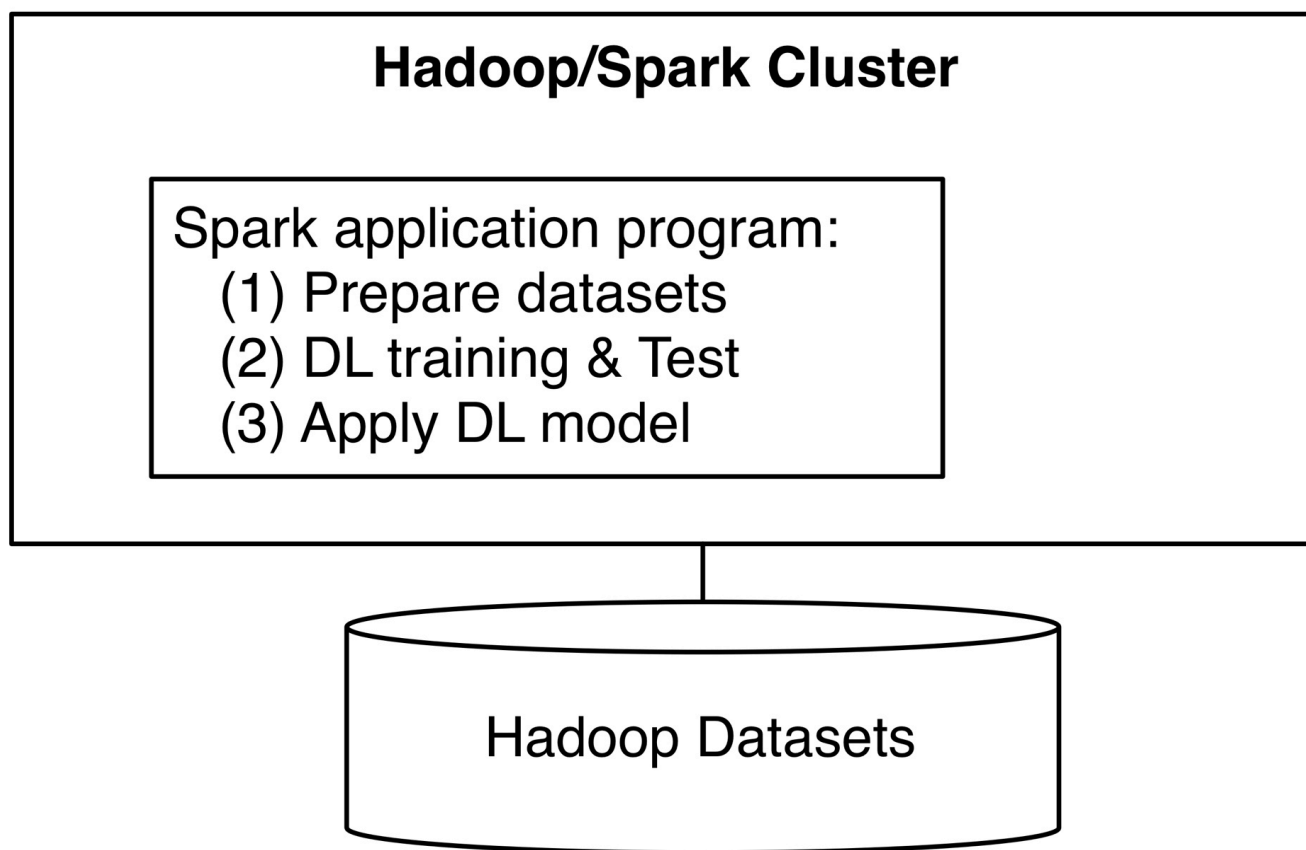


- Computation on servers
 - › (1) Negative sampling
 - › (1) Compute gradient: $X*Y$
 - › (3) Adjust vectors: $Y=aX+Y$
- Daily training enabled
 - › weeks → hours

Deep Learning: Previous Practice



Scalable Deep Learning for Big Data



CaffeOnSpark Open Sourced



- Released in Feb. 2016
 - Apache 2.0 license
- Distributed deep learning
 - GPU or CPU
 - Ethernet or InfiniBand
- Easily deployed on public cloud or private cloud

github.com/yahoo/CaffeOnSpark

CaffeOnSpark: One Notebook <http://bit.ly/1REZ0cN>

File Edit View Insert Cell Kernel Help

Python 2

Logistic Regression using MLlib

```
In [57]: from pyspark.mllib.linalg import Vectors
         from pyspark.mllib.regression import LabeledPoint
         from pyspark.mllib.classification import LogisticRegressionWithLBFGS
```

```
In [58]: data = f.map(lambda row: LabeledPoint(row.label[0], Vectors.dense(row.ip1)))
```

```
In [59]: lr = LogisticRegressionWithLBFGS.train(data, numClasses=10, iterations=10)
```

```
In [60]: predictions = lr.predict(data.map(lambda pt : pt.features))
```

```
In [61]: predictions.take(5)
```

```
Out[61]: [7, 2, 1, 0, 4]
```

```
|00000001| [1.0]| [1.3683326, -0.0, ...] | [2.0906663, 1.048...] | [2.0] |
|00000002| [1.0]| [1.5641443, -0.0, ...] | [-0.773368, 10.61...] | [1.0] |
|00000003| [1.0]| [-0.0, 1.9505613, ...] | [16.46351, -6.917...] | [0.0] |
|00000004| [1.0]| [0.5979191, 0.075...] | [-0.48371825, -2....] | [4.0] |
+-----+-----+-----+-----+-----+
```

```
In [45]: dl_train_source = DataSource(sc).getSource(cfg, True)
```

```
In [46]: cos.train(dl_train_source)
```

Datasets for YOU

Data, Data, Data.

- Machine learning is all about data
 - More data → Better model
- Industrial researchers work on large-scale data
- Academic researchers need data, too ☹️

Yahoo Webscope Datasets

Categories	datatype	# of datasets
L: Language Data	l	25
G: Graph and Social Data	g	8
R: Ratings and Classification Data	r	10
A: Advertising and Market Data	a	4
C: Competition Data	c	3
S: Computing Systems Data	s	4
I: Image Data	i	5

- *Ex. <https://webscope.sandbox.yahoo.com/catalog.php?datatype=r>*
- **Non-commercial use by academics and scientists: 8,000+**

L5 - Query Vectors

- **8M** query vectors trained using our search2vec system
- May serve as a testbed for query rewriting task
 - › IR research
 - › Word and sentence similarity task in NLP research

Method	oAUC	Macro NDCG@5
word2vec	0.817	0.929
search2vec	0.880	0.959

I3 - Flickr Creative Commons

- 100M Flickr photos & videos
 - › Largest public multimedia dataset ever released
- ImageNet ... 14M images
- Metadata
 - › Tags, title, des & geo
 - 68M photos have at least one tag
 - › Comments, favorites, social network data queried via Flickr API



longing by Robert-Crouse Baker

Tags: Ocean Beach, San Francisco, Pacific, Surf, surfing, waves, green, water, man, surfer, January, océan, le surf, vagues, mer, sea
Description: “My soul is full of longing for the secret of the sea,
and the heart of the great ocean sends a thrilling pulse through me.” — Henry Wadsworth
YAHOO!

Summary

- Big-data technologies and continued innovation are critical for Yahoo business.
- Yahoo continues open source contribution ever since our donation of Hadoop to Apache
 - › Apache Spark/Storm/Hbase/Omid, CaffeOnSpark
 - › Webscope datasets

Thanks!

yahoohadoop.tumblr.com

bigdata@yahoo-inc.com

