Big Data Architectures@ Facebook

QCon London 2012

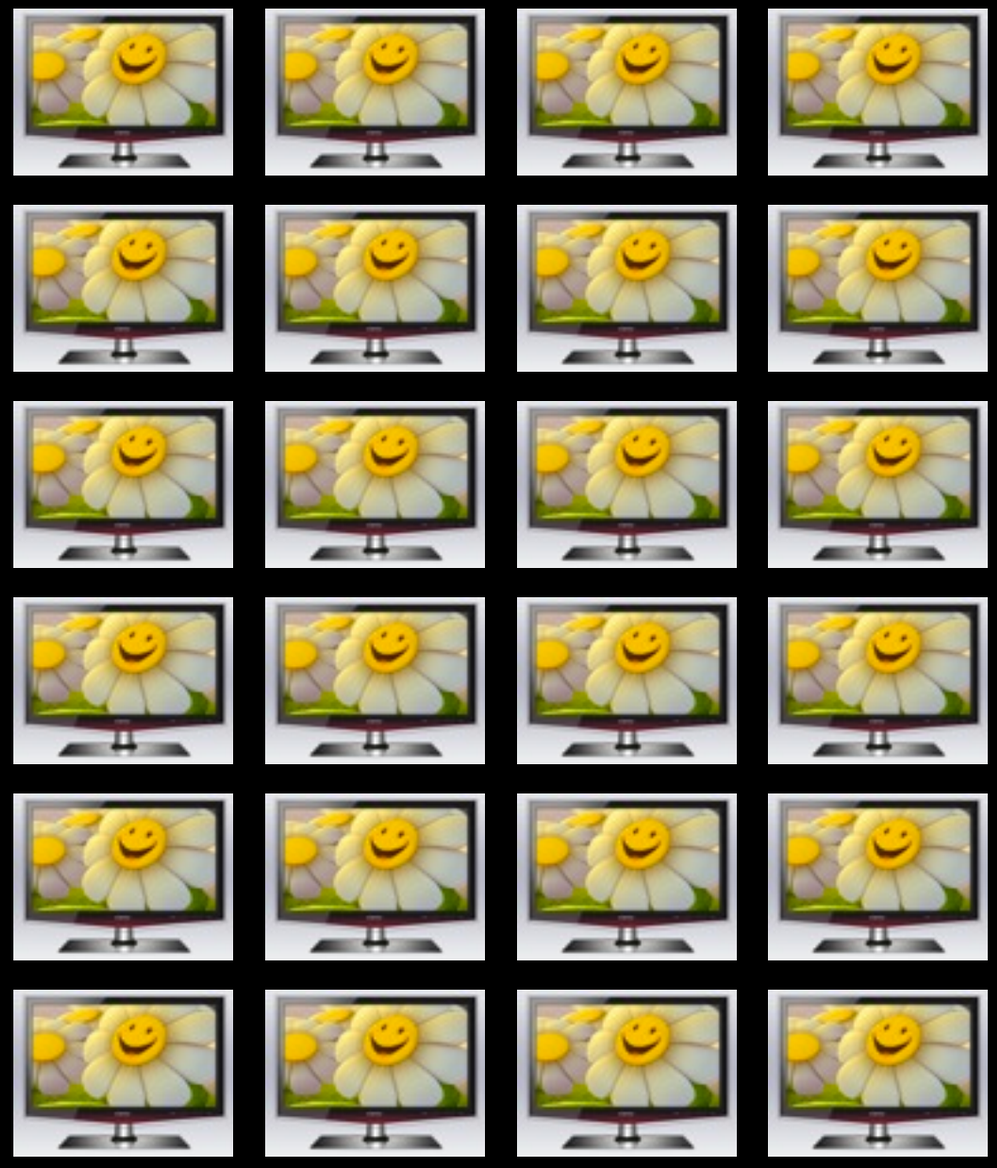
Ashish Thusoo

Outline

- Big Data @ Facebook - Scope & Scale
- Evolution of Big Data Architectures @ FB
  - Past, Present and Future
- Questions

Big Data @ FB: Scale

- 25 PB of compressed data
- equivalent to 300 years of HD-TV video



Big Data @ FB: Scale

- 150 PB of uncompressed data
- equivalent to 3 x the entire written works of mankind from the beginning of recorded history in all languages

Big Data @ FB: Scale

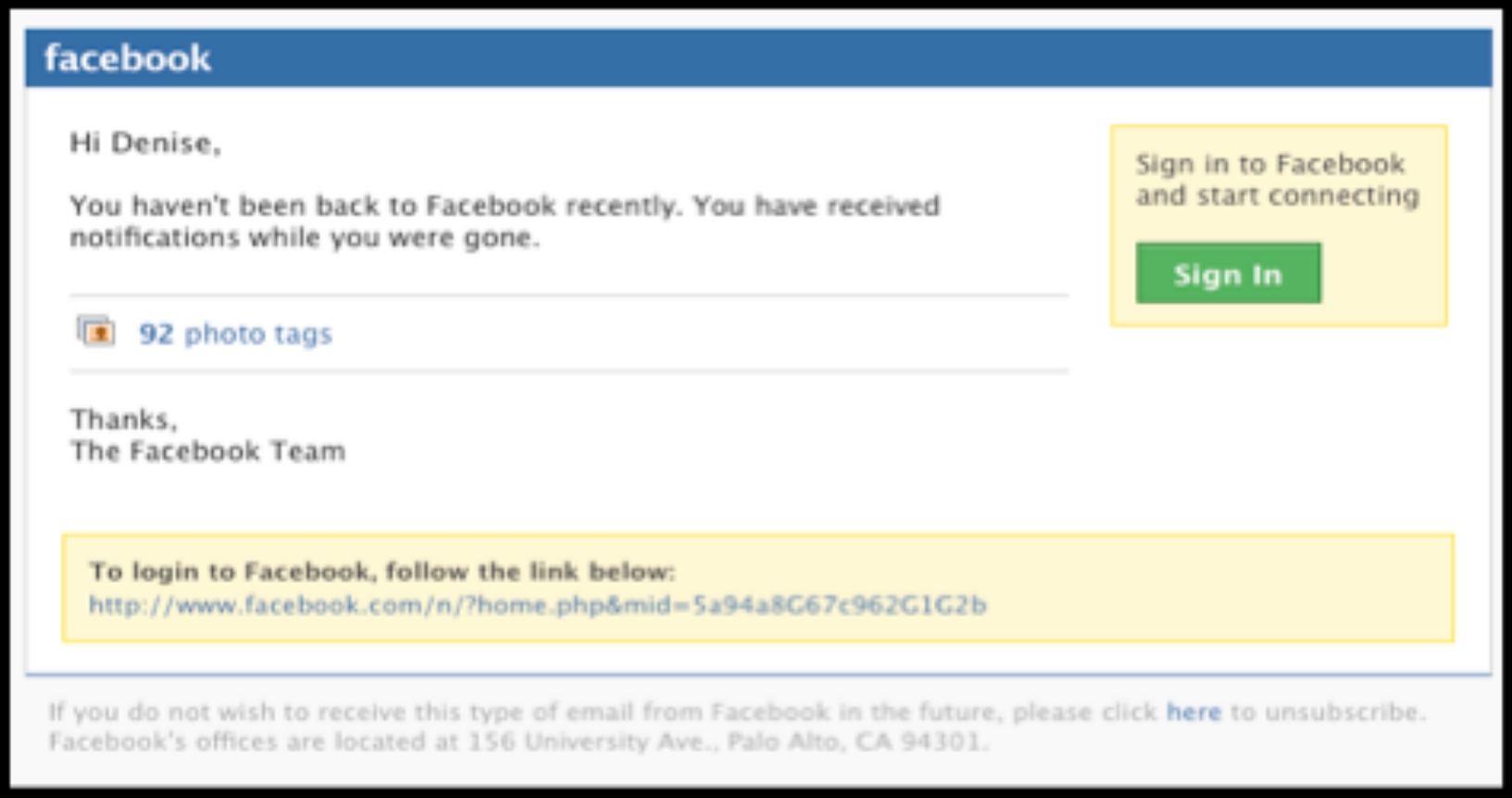
- 400 TB/day (uncompressed) of new data
- That is a lot of disks

Big Data @ FB: Scope

- Simple reporting
- Model generation
- Adhoc analysis + data science
- Index generation
- Many many others...

A/B Testing Email #1

A/B Testing Email #2



A/B Testing Email #2 is 3x Better

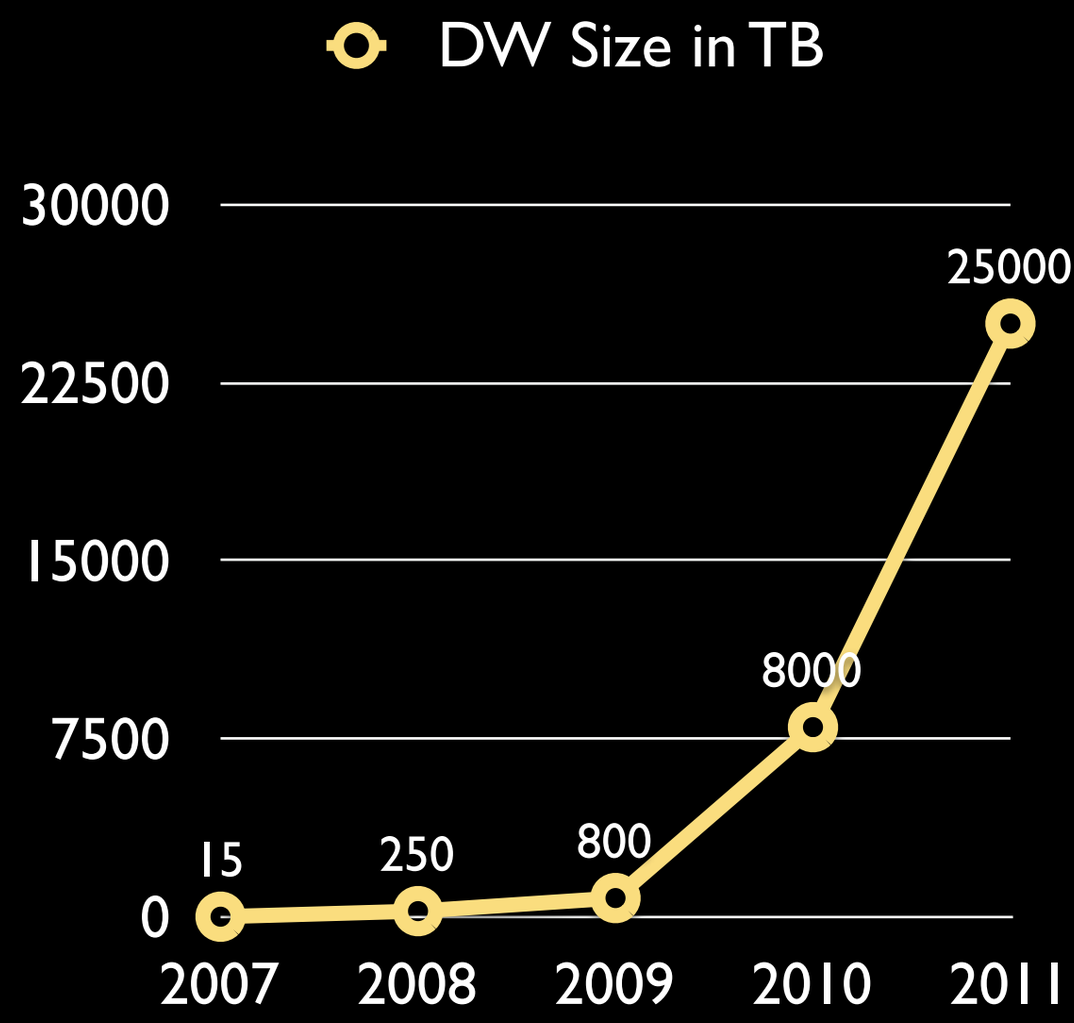
Friend Map

By Paul Butler - https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

Big Data @ FB: Scope

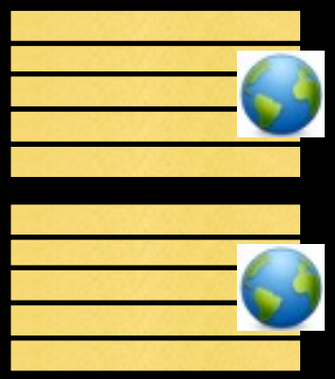
- one new job every second
- ~ 15% of the company uses the clusters

Evolution: 2007-2011



2007: Traditional EDW

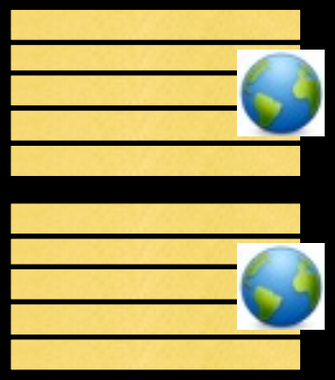
2007: Traditional EDW



Web Clusters

MySQL Clusters

2007: Traditional EDW

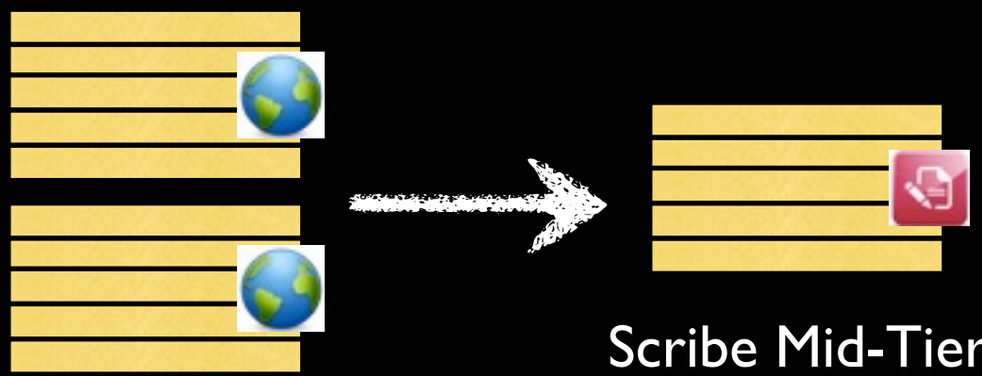


Web Clusters

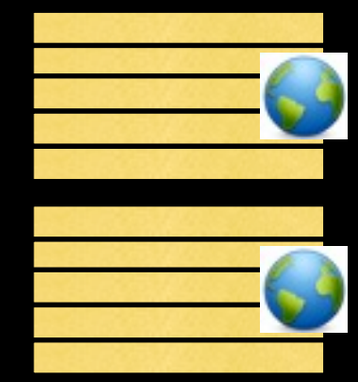
MySQL Clusters

RDBMS Data Warehouse

2007: Traditional EDW



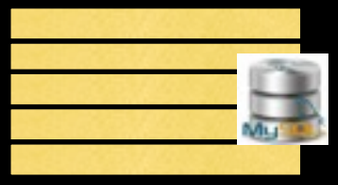
2007: Traditional EDW



Web Clusters

Scribe Mid-Tier

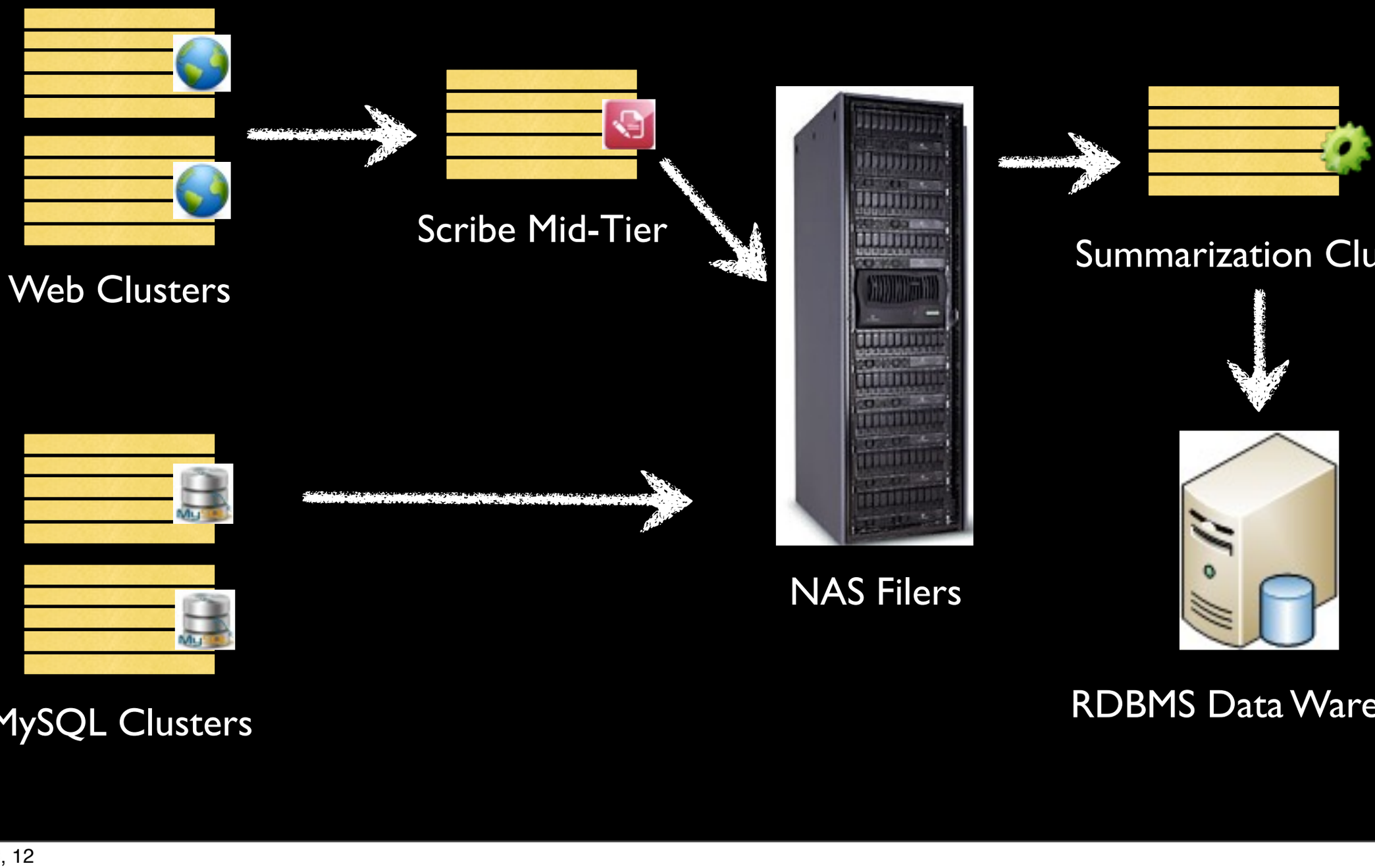
NAS Filers



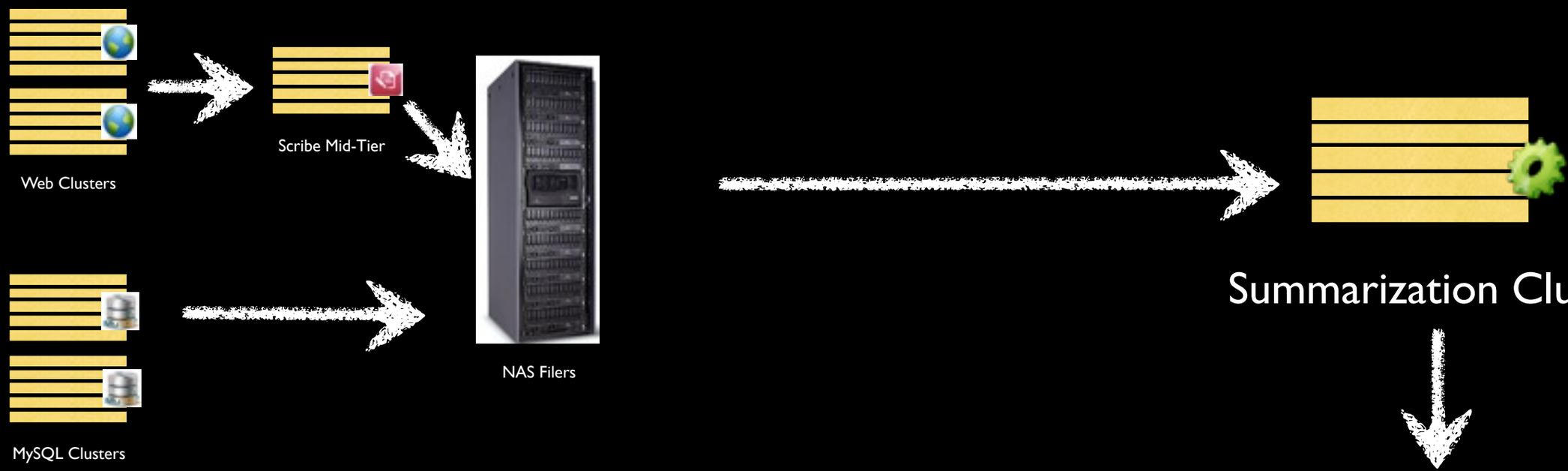
MySQL Clusters

RDBMS Data Warehouse

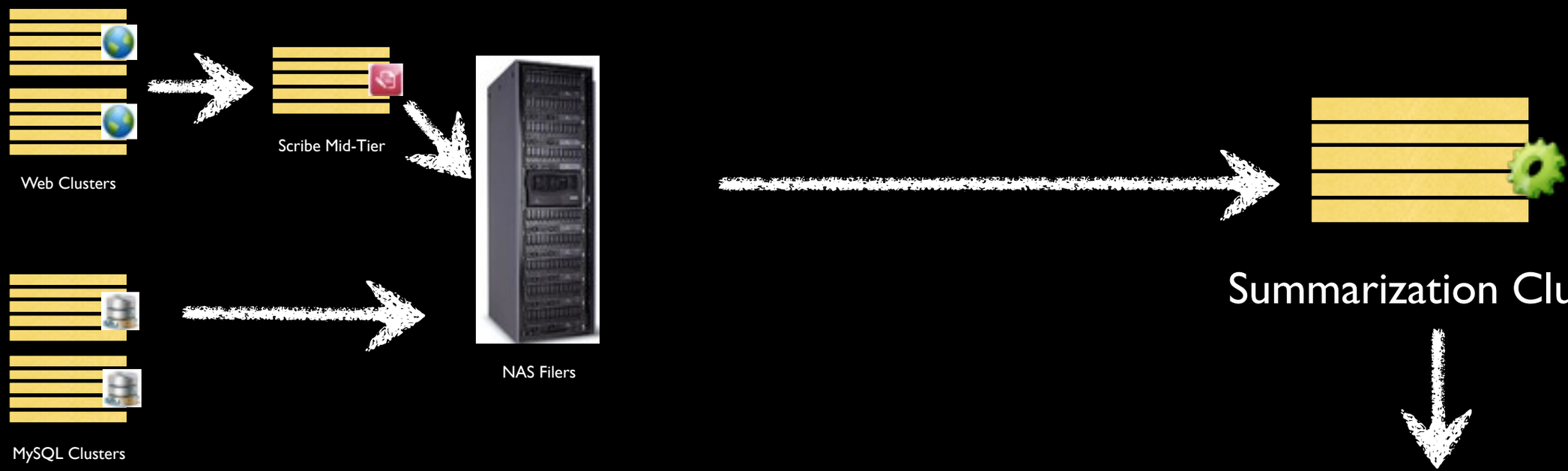
2007: Traditional EDW



2007: Pain Points

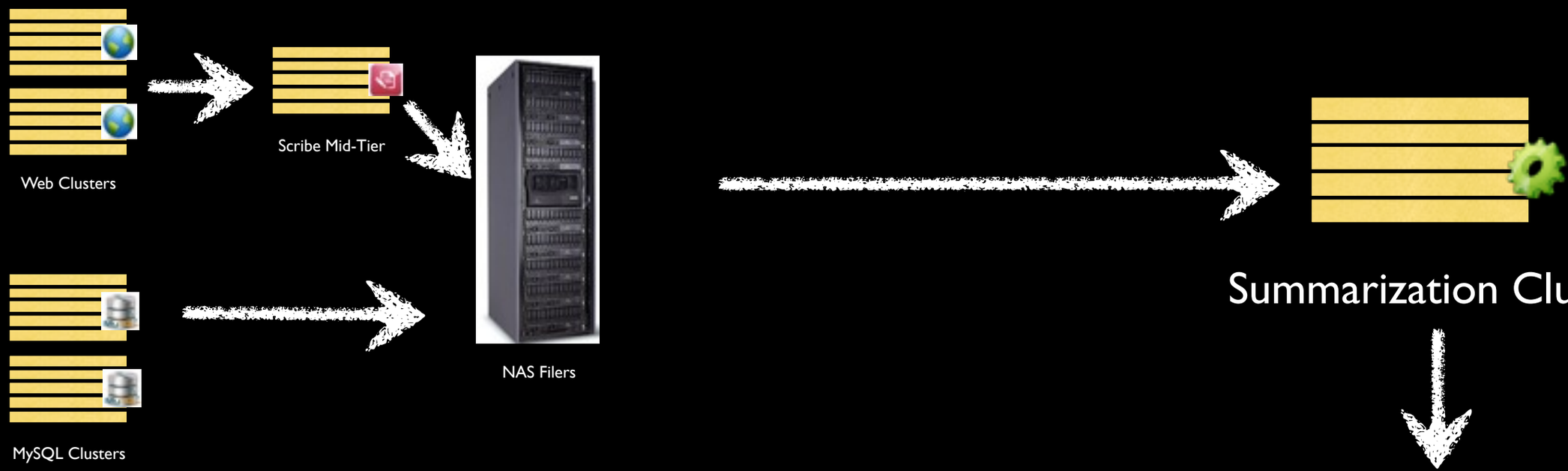


2007: Pain Points



- daily ETL > 24 hours

2007: Pain Points

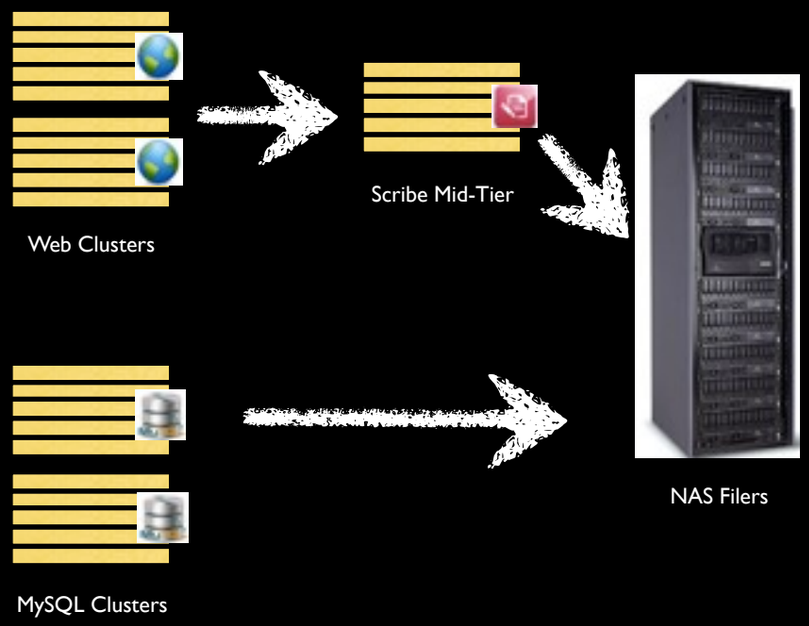


- daily ETL > 24 hours
- Lots of tuning/indexes etc.

2007: Pain Points

- daily ETL > 24 hours
- Lots of tuning/indexes etc.
- Lots of hardware planning

2007: Pain Points



- compute close to storage
(early map/reduce)

Summarization Cluster

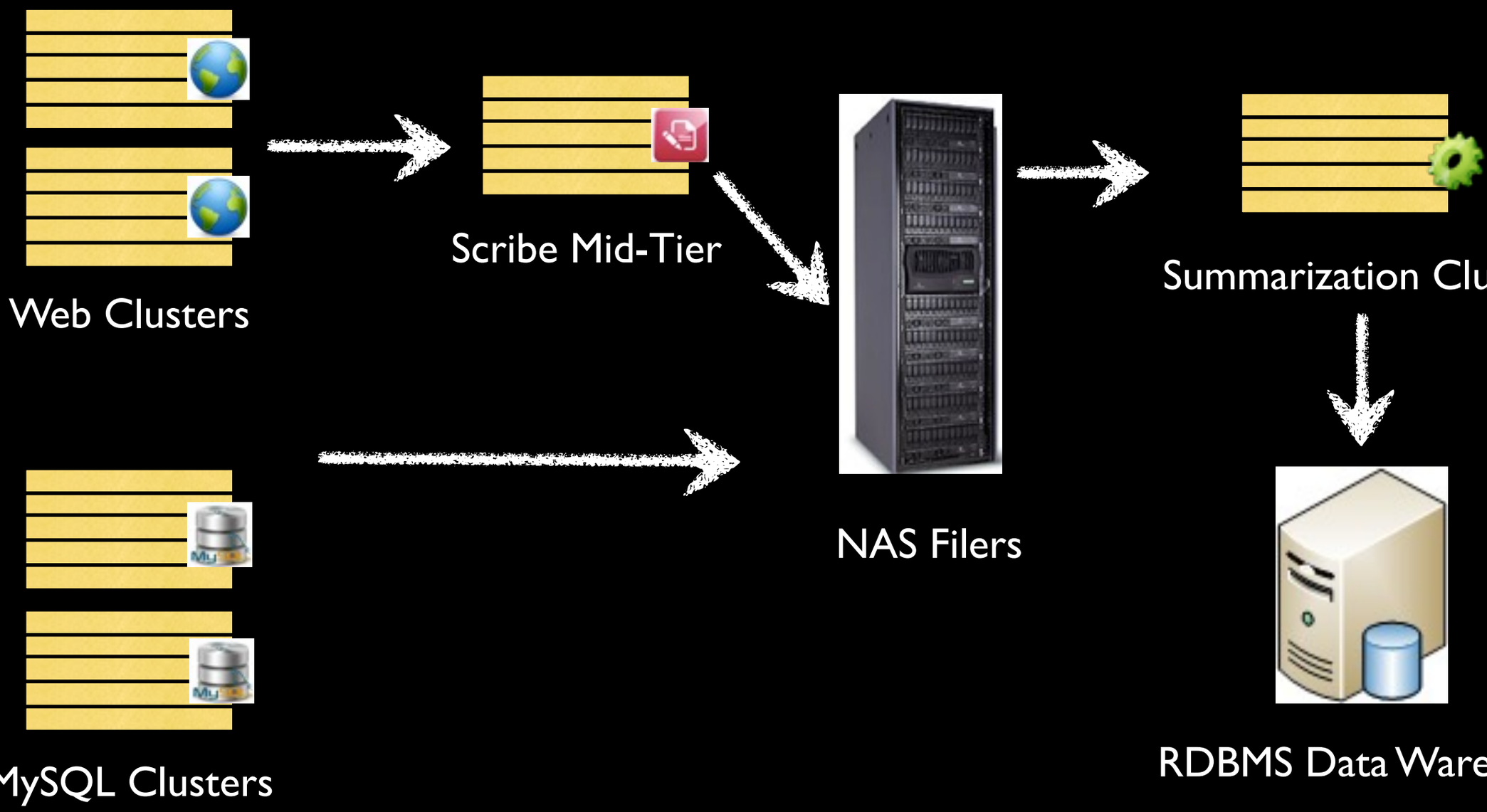
- daily ETL > 24 hours
- Lots of tuning/indexes etc.
- Lots of hardware planning

RDBMS Data Warehouse

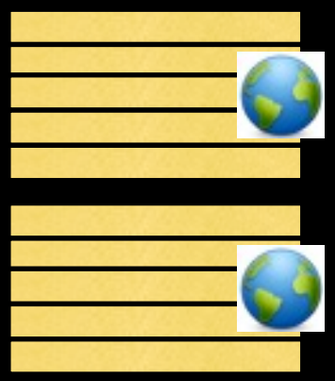
2007: Limitations

- Most use cases were in business metrics - data science, model building etc. not possible
- Only summary data was stored online - details archived away

2008: Move to Hadoop



2008: Move to Hadoop

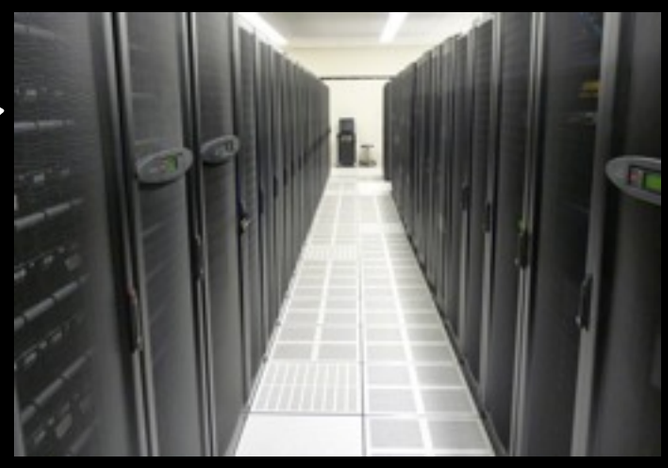


Web Clusters

Scribe Mid-Tier

NAS Filers

Batch copier/ loaders



Hadoop/Hive Data Warehouse

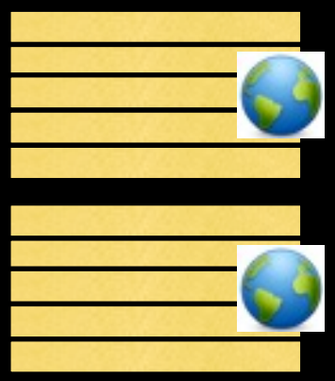
MySQL Clusters

RDBMS Data Mart

2008: Immediate Pros

- Data science at scale became possible
- For the first time all of the instrumented data could be held online
- Use cases expanded

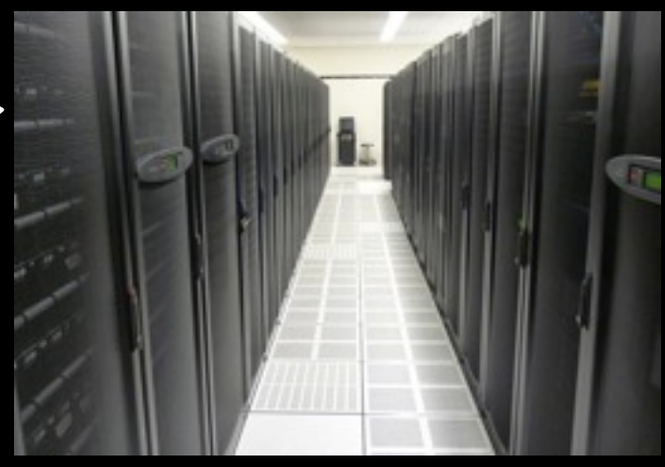
2009: Democratizing Data



Web Clusters

Scribe Mid-Tier

NAS Filers



Hadoop/Hive Data Warehouse

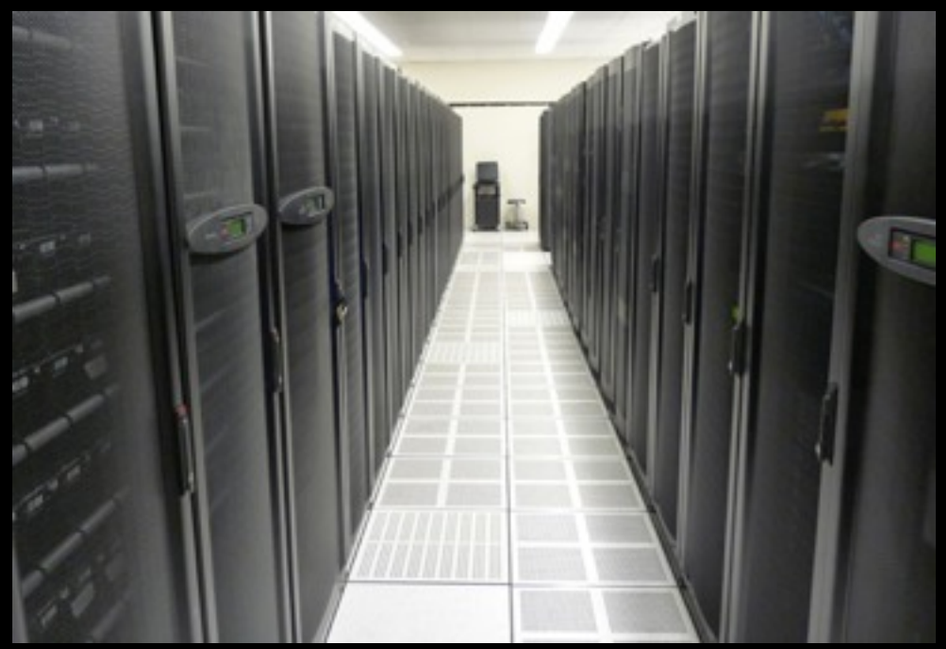
MySQL Clusters

RDBMS Data Mart

2009: Democratizing Data

Databee & Chronos: Data Pipeline Framework

Nectar: instrumentation & schema aware data collection



HiPal: Adhoc Queries + Data Discovery

Hadoop/Hive Data Warehouse

Scrapes: Configuration Driven

2009: Democratizing Data(Nectar)

- Typical Nectar Pipeline
  - Simple schema evolution built in
  - json encoded short term data
  - decomposing json for long term storage

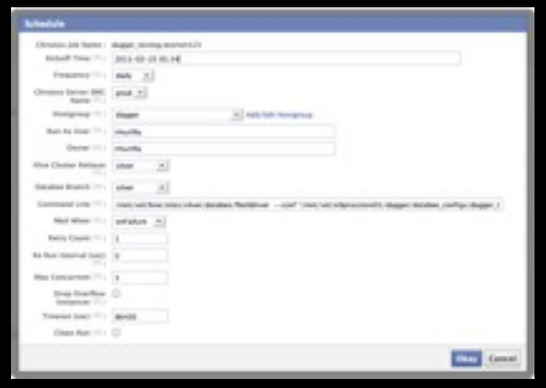
```
// This event has application name 'mobilelog' and app event type
// 'email_mms_upload'.
// NOTE: Make sure you use only one application name per new application.
// Also, app event type should not have any special characters or spaces,
// use underscores instead.  $sampling_rate is the scribe sampling rate and
// has a value between 0 and 100 - sampling is on userid

NectarAppSpecificEvent('mobilelog', 'email_mms_upload', $sampling_rate)

  ->addToOdsKeys(array('k1', 'k2'))        // if you want to add additional
                                           // ODS keys
  ->setODSSamplingRate(1)                  // default is 10000, meaning 1
                                           // in 10000 events is sent to ODS
  ->addToAppSection(array("key" => "val")) // can add different key value
                                           // pairs for different eventtypes
  ->log();                                 // need to explicit log app
                                           // specific events
```

2009: Democratizing Data (Tools)

- HiPal - data discovery and query authoring
- Charting and dashboard generation tools

2009: Democratizing Data (Tools)

- Databee: Workflow language
- Chronos: Scheduling tool



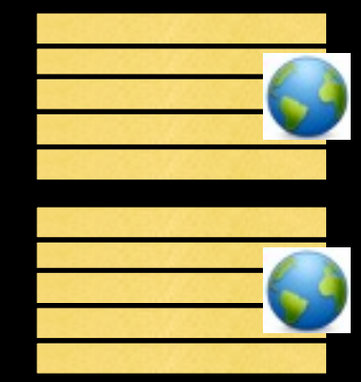
2009: Cons of Democratization

- Isolation to protect against Bad Jobs
- Fair sharing of the cluster - what is a high priority job and how to enforce it

2010: Controlling Chaos

- Isolation
- Reducing operational overhead
- Better resource utilization
- Measurement, ownership, accountability

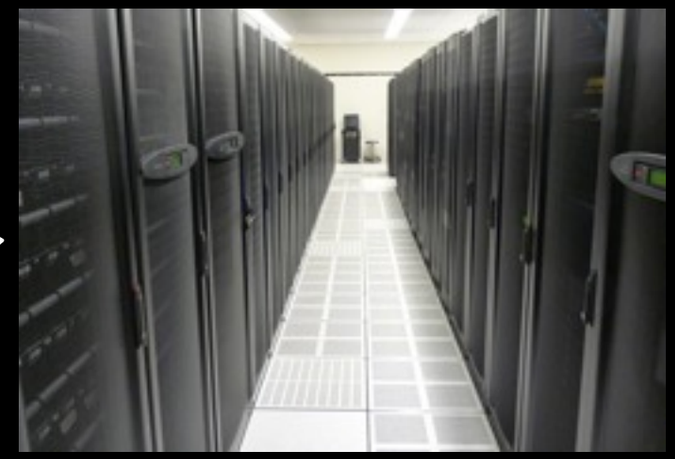
2010: Isolation



Web Clusters

Scribe Mid-Tier

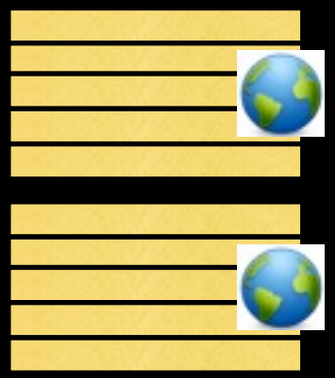
NAS Filers



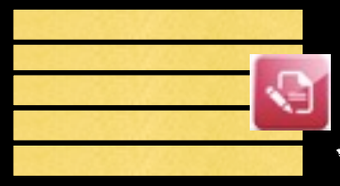
Hadoop/Hive Data Warehouse

MySQL Clusters

2010: Isolation

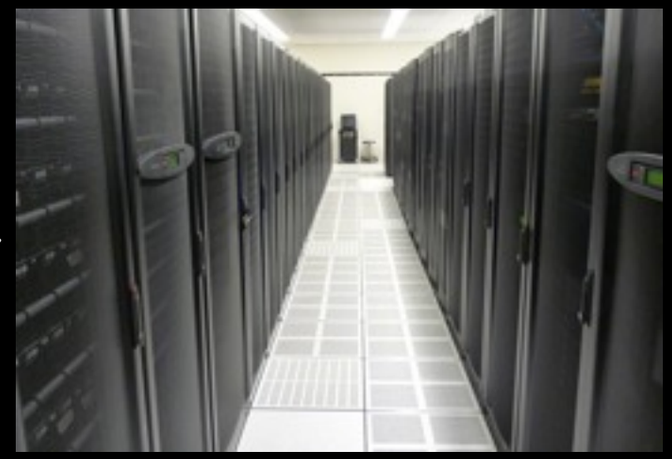


Web Clusters



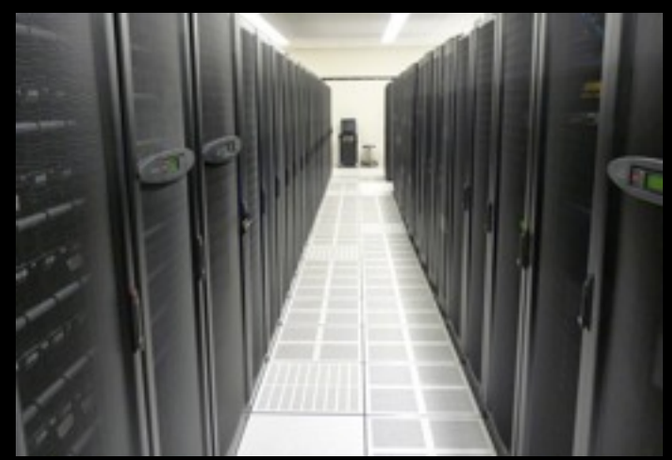
Scribe Mid-Tier

NAS Filers



Platinum Warehouse

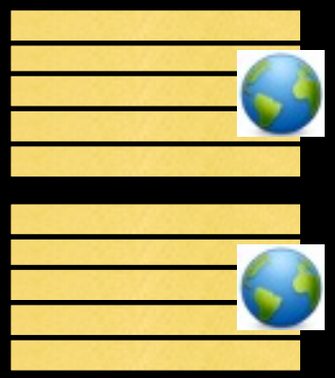
Hive Replication



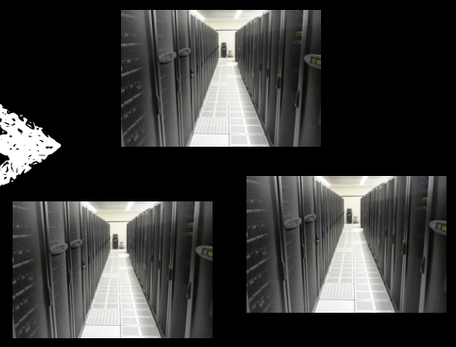
Silver Warehouse

MySQL Clusters

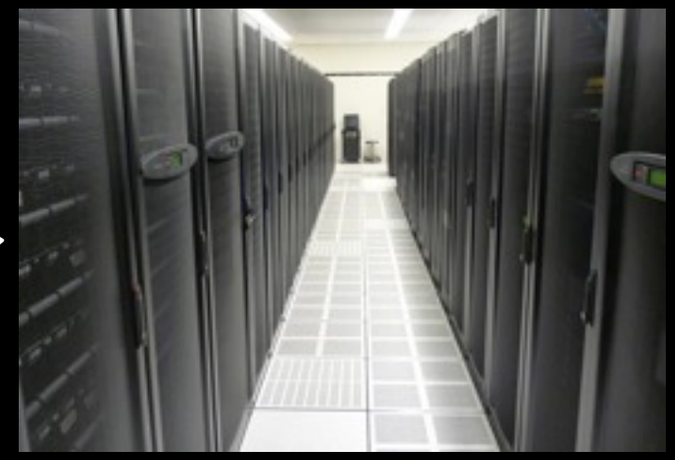
2010: Ops Efficiency



Web Clusters



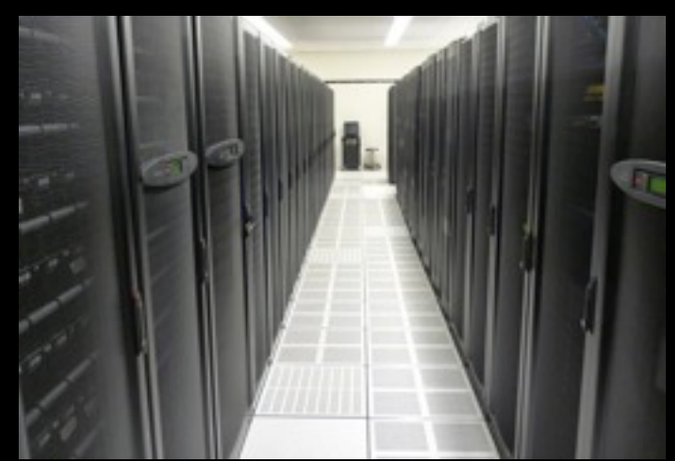
Scribe HDFS



MySQL Clusters

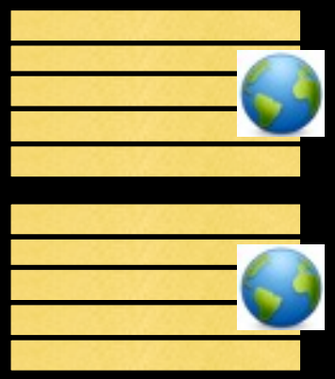
Platinum Warehouse

Hive Replication



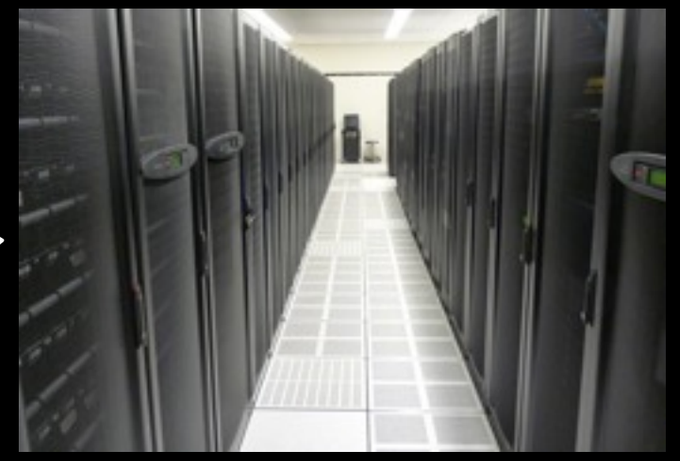
Silver Warehouse

2010: Ops Efficiency



Web Clusters

Scribe HDFS

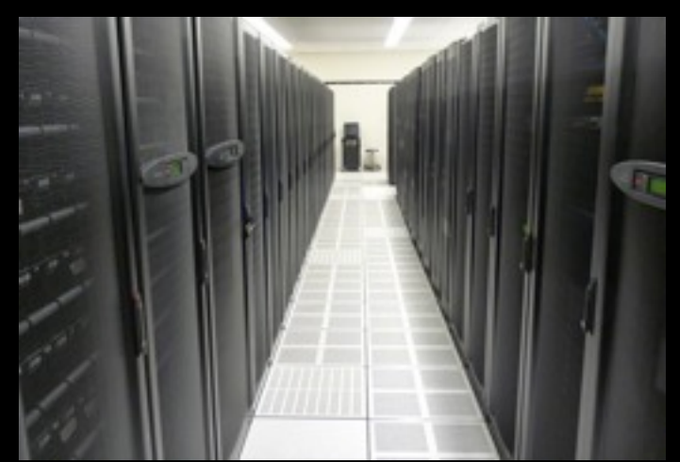


MySQL Clusters

near real time
data consumers

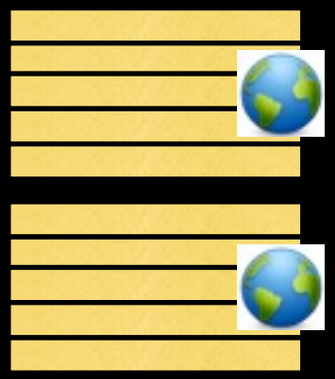
Platinum Warehouse

Hive Replication



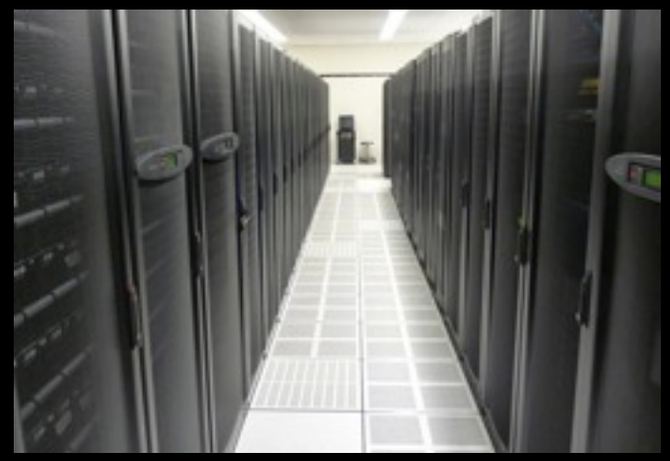
Silver Warehouse

2010: Ops Efficiency



Web Clusters

Scribe HDFS



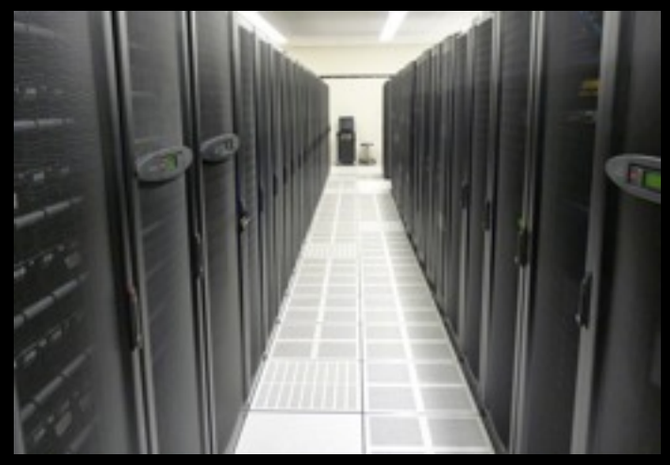
MySQL Clusters

ptail:
parallel tail
on hdfs

near real time
data consumers

Platinum Warehouse

Hive Replication

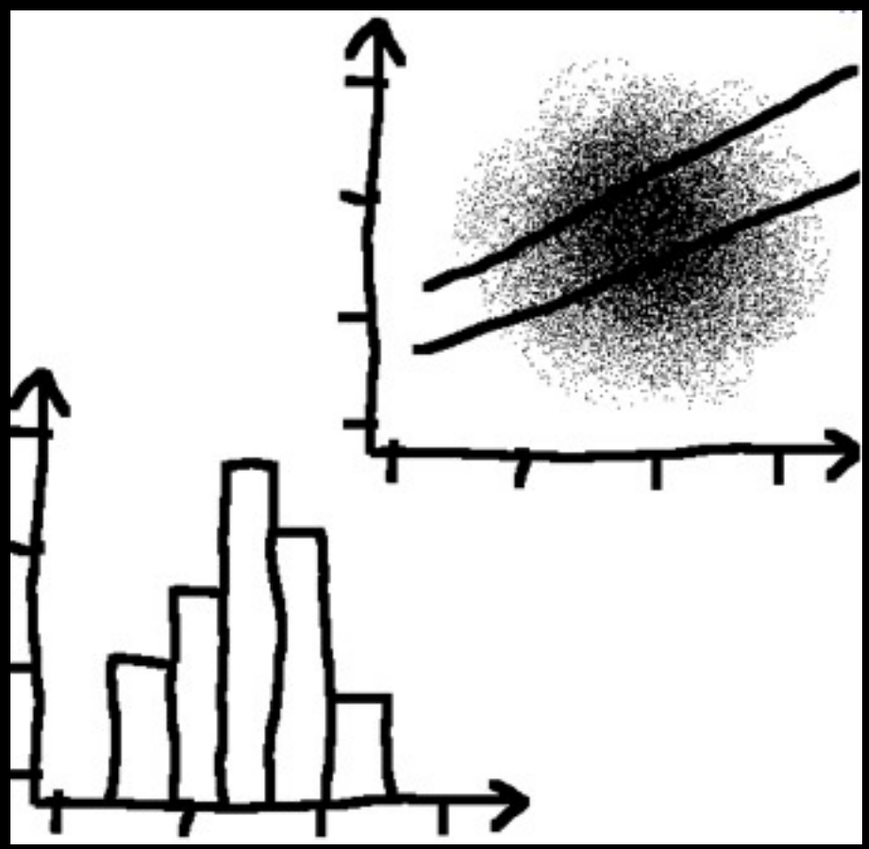


Silver Warehouse

2010: Resource Utilization (Disk)

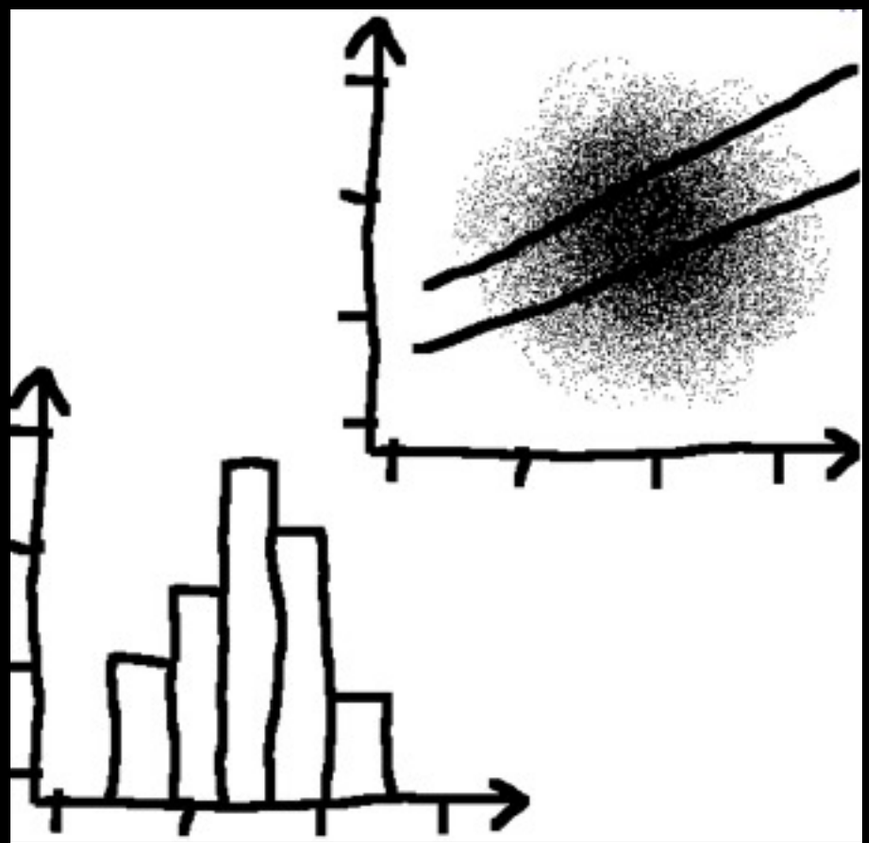
- HDFS-RAID: from 3 replicas to 2.2 replicas

- RCFile: Row columnar format for compressing Hive tables



2010: Resource Utilization (CPU)

- Continuous copier/ loaders
- Incremental scrapes
- Hive optimizations to save CPU



2010: Monitoring(SLAs)

- Per job statistics rolled up to owner/group/team
- Expected time of arrival vs Actual time of arrival of data
- Simple data quality metrics

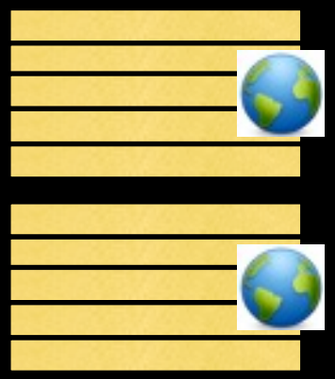
2011: New Requirements

- More real time requirements for aggregations
- Optimizing resource utilization

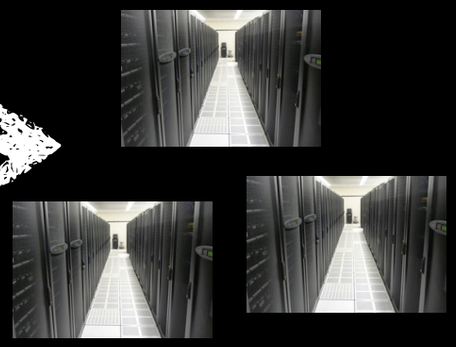
2011: Beyond Hadoop

- Puma for real time analytics
- Peregrine for simple and fast queries

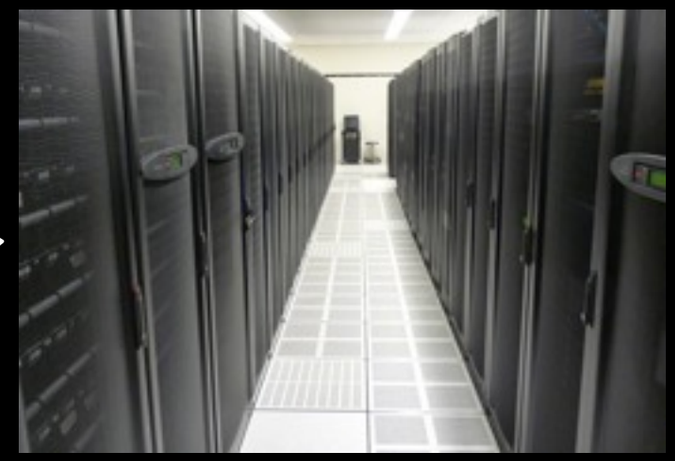
2010: Puma



Web Clusters



Scribe HDFS



ptail:
parallel tail
on hdfs

near real time
data consumers

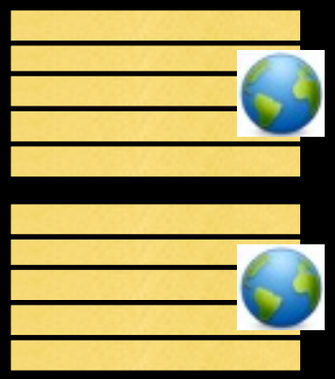
Platinum Warehouse

Hive Replication

MySQL Clusters

Silver Warehouse

2010: Puma

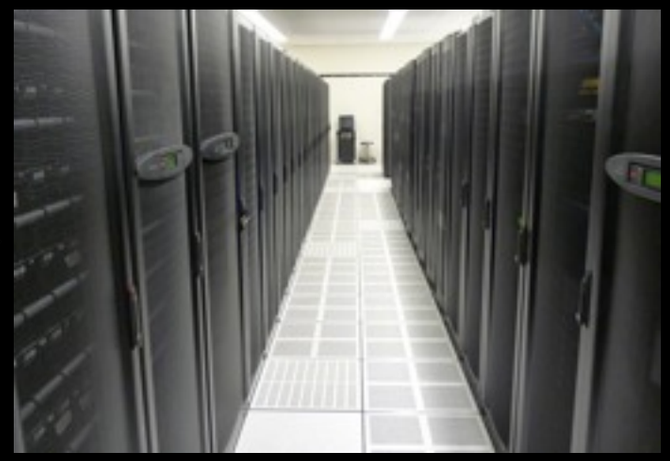


Web Clusters

Scribe HDFS

ptail:
parallel tail
on hdfs

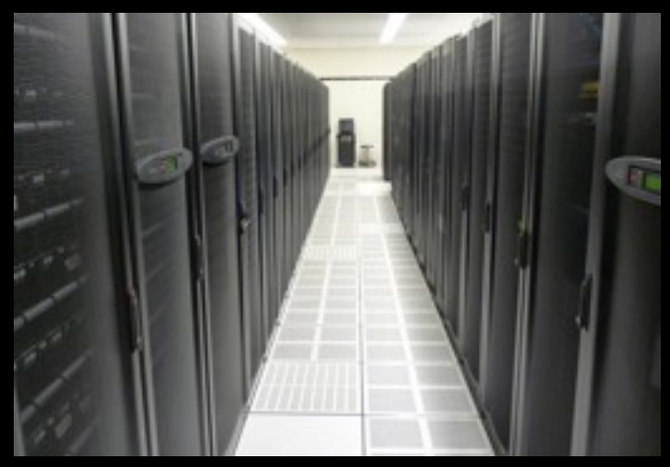
near real time
data consumers



Platinum Warehouse

Hive Replication

MySQL Clusters



Silver Warehouse

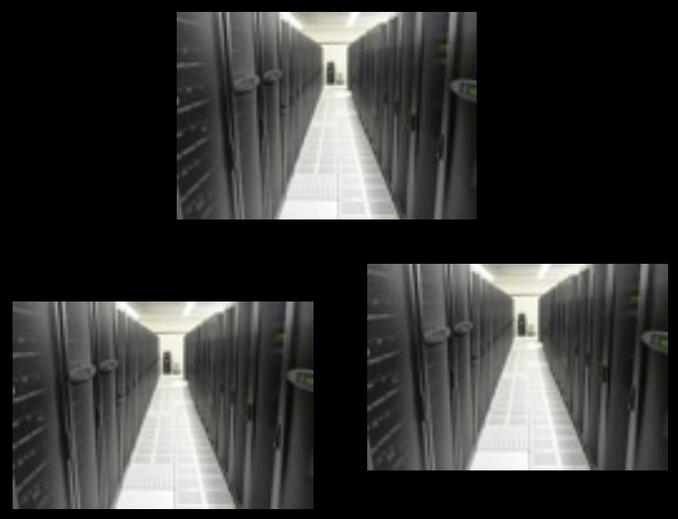
2010: Puma

2010: Puma

Scribe HDFS

ptail: parallel tail on hdfs

2010: Puma



Scribe HDFS

ptail: parallel tail on hdfs

Puma Clusters

2010: Puma

Scribe HDFS

ptail: parallel tail on hdfs

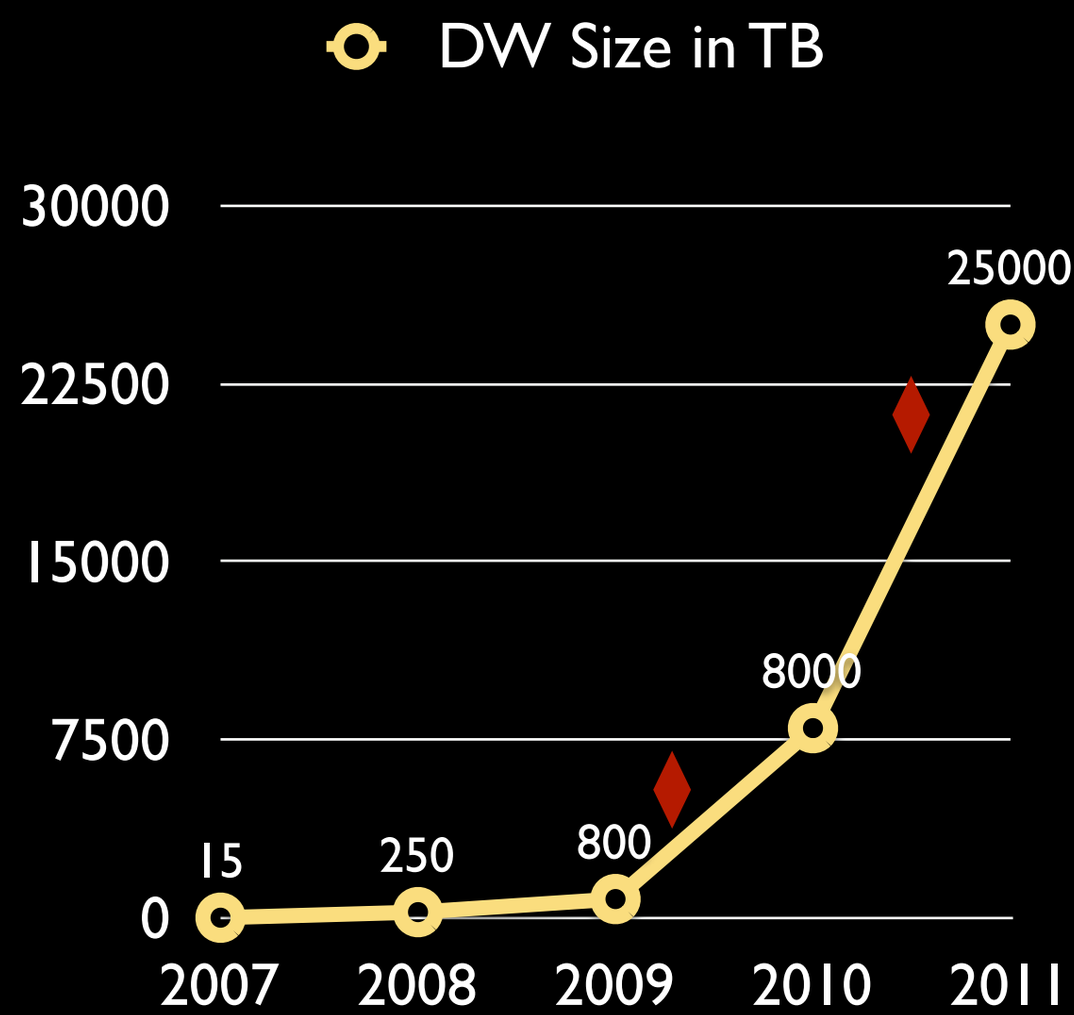
Puma Clusters

Hbase Cluster

Other Challenges Of HyperGrowth

- Moving data centers
- Moving sustainably fast

HyperGrowth - Moving Data Centers



HyperGrowth - Moving Data Centers

- Moved 20 PB of data
- Leverage replication with fast switch
- 2-3 months to accomplish the entire move

Blog Post on FB by Paul Yang: http://www.facebook.com/notes/paul-yang/moving-an-elephant-large-scale-hadoop-data-migration-at-facebook/10150246275318920

Questions

Contact Information:

ashish.thusoo@gmail.com

http://www.linkedin.com/pub/ashish-thusoo/0/5a8/50
https://www.facebook.com/athusoo
https://twitter.com/ashishthusoo