facebook

Facebook's Petabyte Scale Data Warehouse using Hive and Hadoop

Why Another Data Warehousing System?

Data, data and more data

200GB per day in March 2008

12+TB(compressed) raw data per day today

Trends Leading to More Data

Trends Leading to More Data

Free or low cost of user services

Trends Leading to More Data

Free or low cost of user services

Realization that more insights are derived from simple algorithms on more data

Deficiencies of Existing Technologies

Deficiencies of Existing Technologies

Cost of Analysis and Storage on proprietary systems does not support trends towards more data

Deficiencies of Existing Technologies

Cost of Analysis and Storage on proprietary systems does not support trends towards more data

Limited Scalability does not support trends towards more data

Deficiencies of Existing Technologies

Cost of Analysis and Storage on proprietary systems does not support trends towards more data

Limited Scalability does not support trends towards more data

Closed and Proprietary Systems

Lets try Hadoop…

- Pros
  - Superior in availability/scalability/manageability
  - Efficiency not that great, but throw more hardware
  - Partial Availability/resilience/scale more important than ACID
- Cons: Programmability and Metadata
  - Map-reduce hard to program (users know sql/bash/python)
  - Need to publish data in well known schemas
- Solution: HIVE

What is HIVE?

- A system for managing and querying structured data built on top of Hadoop
  - Map-Reduce for execution
  - HDFS for storage
  - Metadata in an RDBMS
- Key Building Principles:
  - SQL as a familiar data warehousing tool
  - Extensibility – Types, Functions, Formats, Scripts
  - Scalability and Performance
  - Interoperability

Why SQL on Hadoop?

```
hive> select key, count(1) from kv1 where key > 100 group by
    key;
```

vs.

$ cat > /tmp/reducer.sh

```
uniq -c | awk '{print $2"\t"$1}'
```

$ cat > /tmp/map.sh

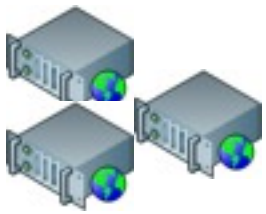
```
awk -F '\001' '{if($1 > 100) print $1}'
```

$ bin/hadoop jar contrib/hadoop-0.19.2-dev-streaming.jar -input /user/hive/warehouse/kv1 -mapper map.sh -file /tmp/reducer.sh -file /tmp/map.sh -reducer reducer.sh -output /tmp/largekey -numReduceTasks 1

$ bin/hadoop dfs –cat /tmp/largekey/part*

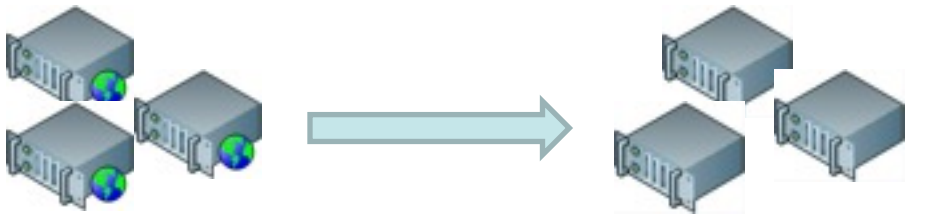
Data Flow Architecture at Facebook

Data Flow Architecture at Facebook

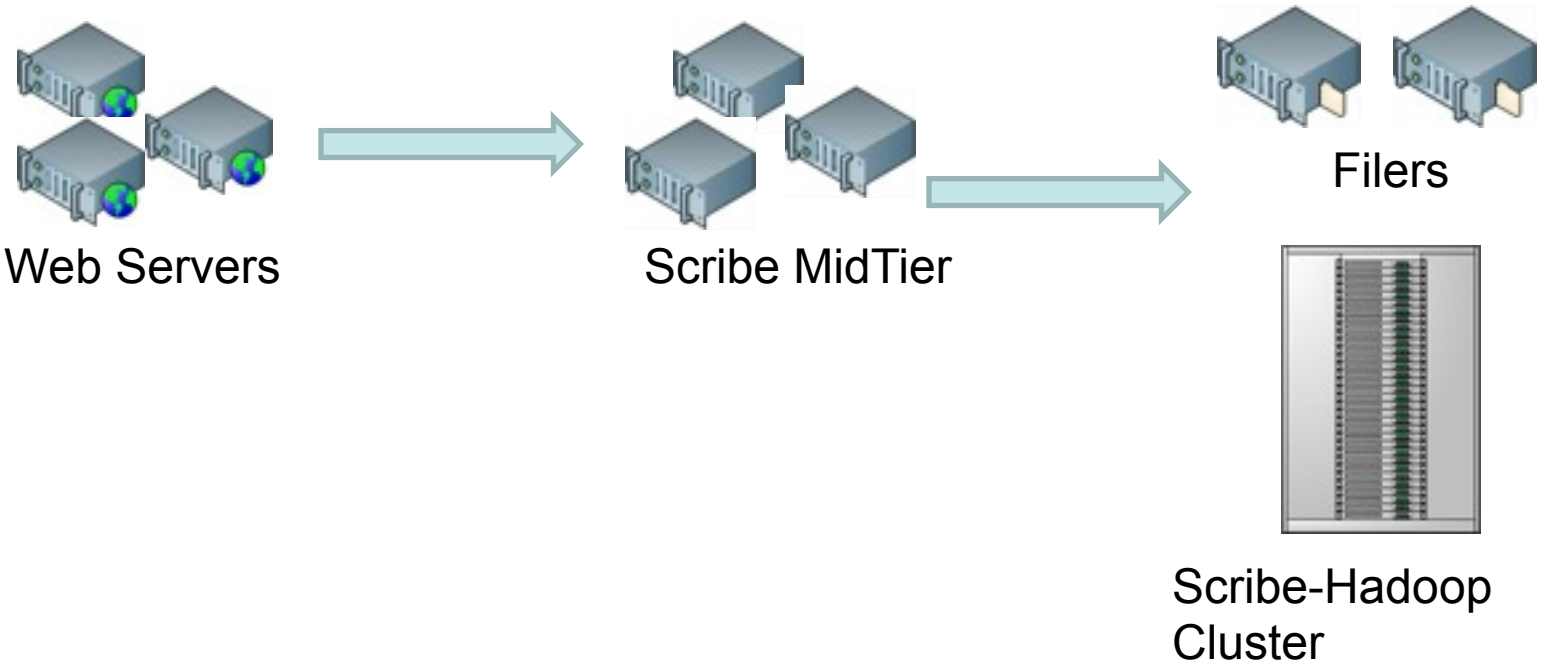


Web Servers

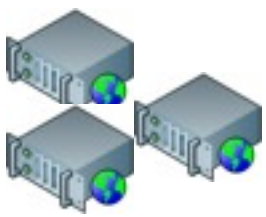
Data Flow Architecture at Facebook



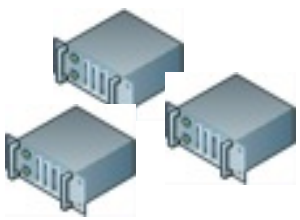
Data Flow Architecture at Facebook



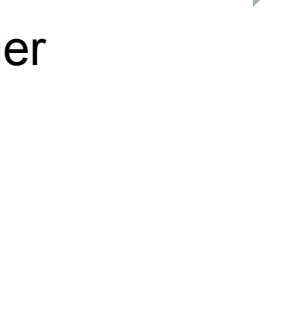
Data Flow Architecture at Facebook



Web Servers

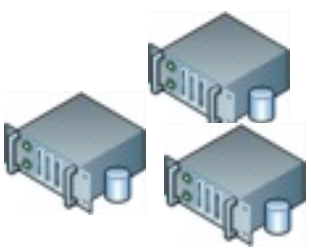


Scribe MidTier



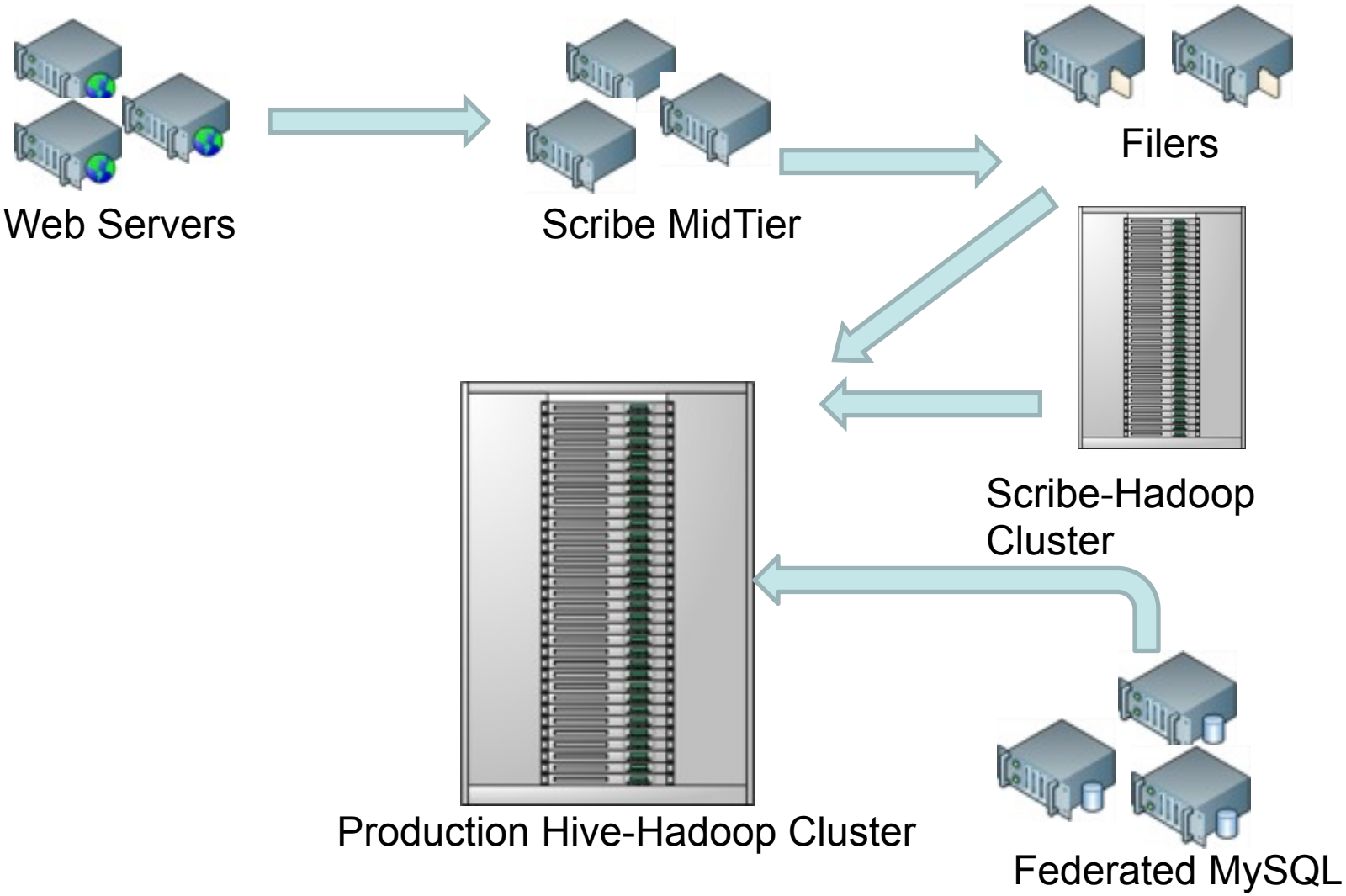
Filers

Scribe-Hadoop Cluster

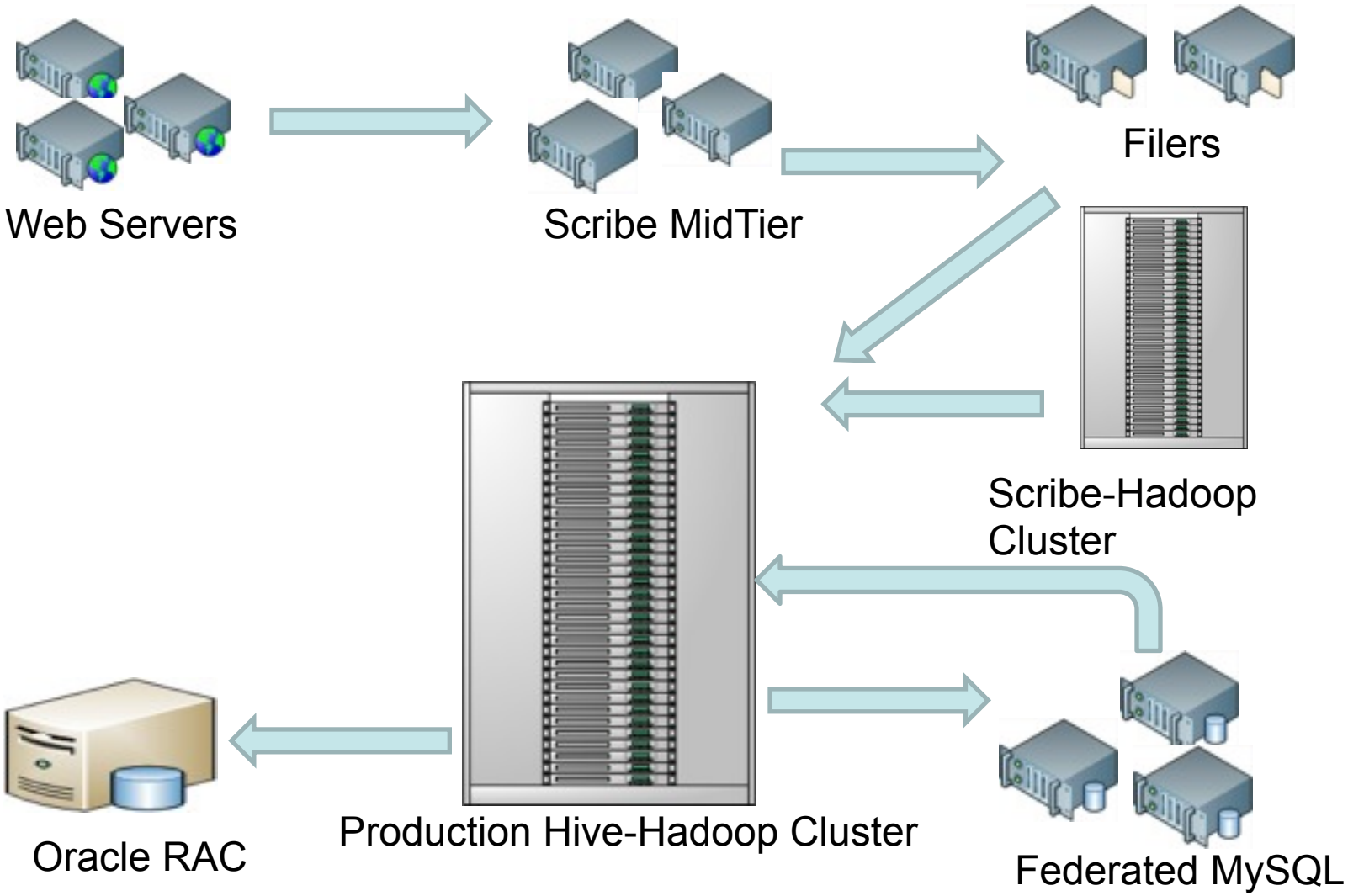


Federated MySQL

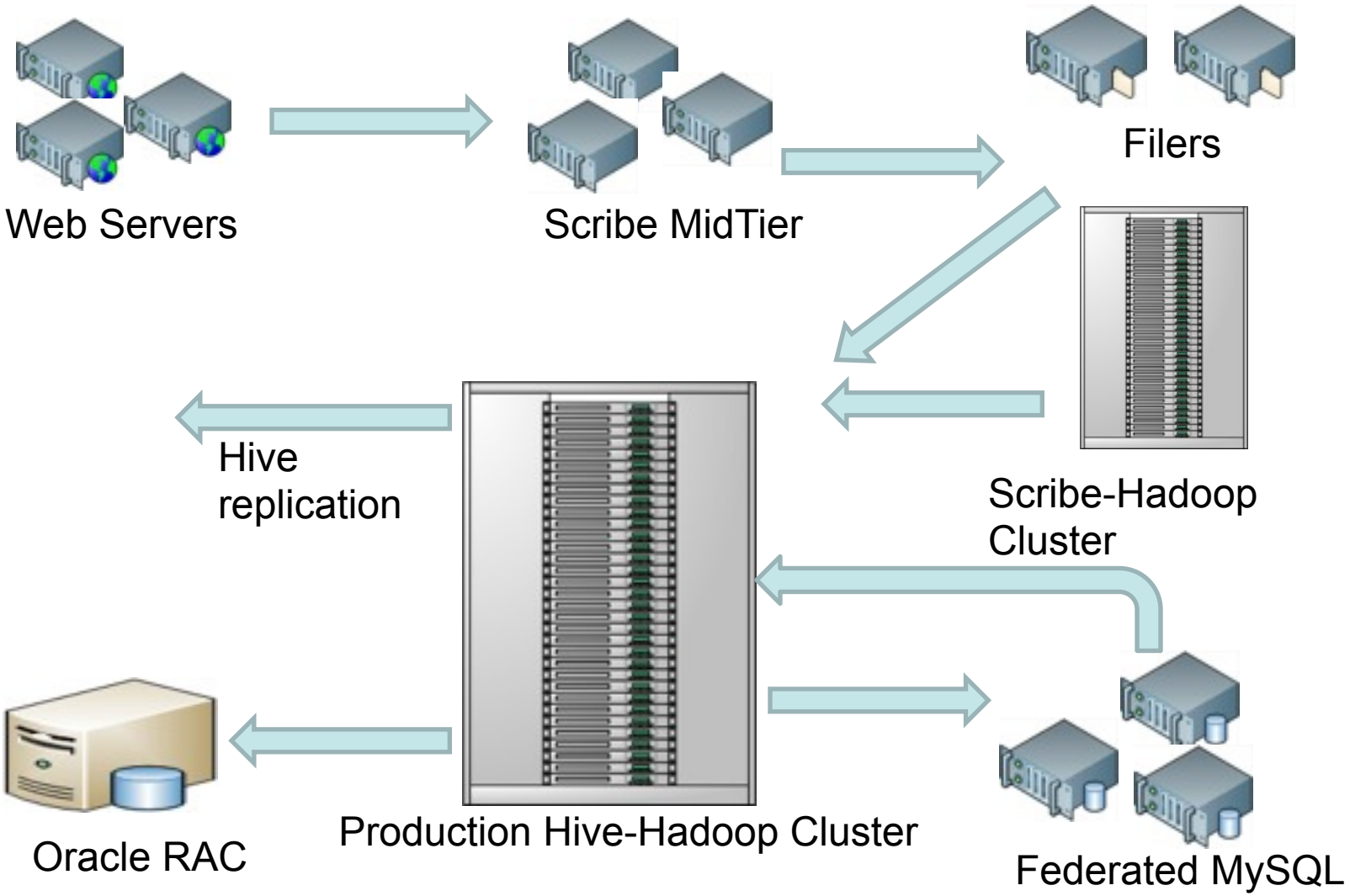
Data Flow Architecture at Facebook



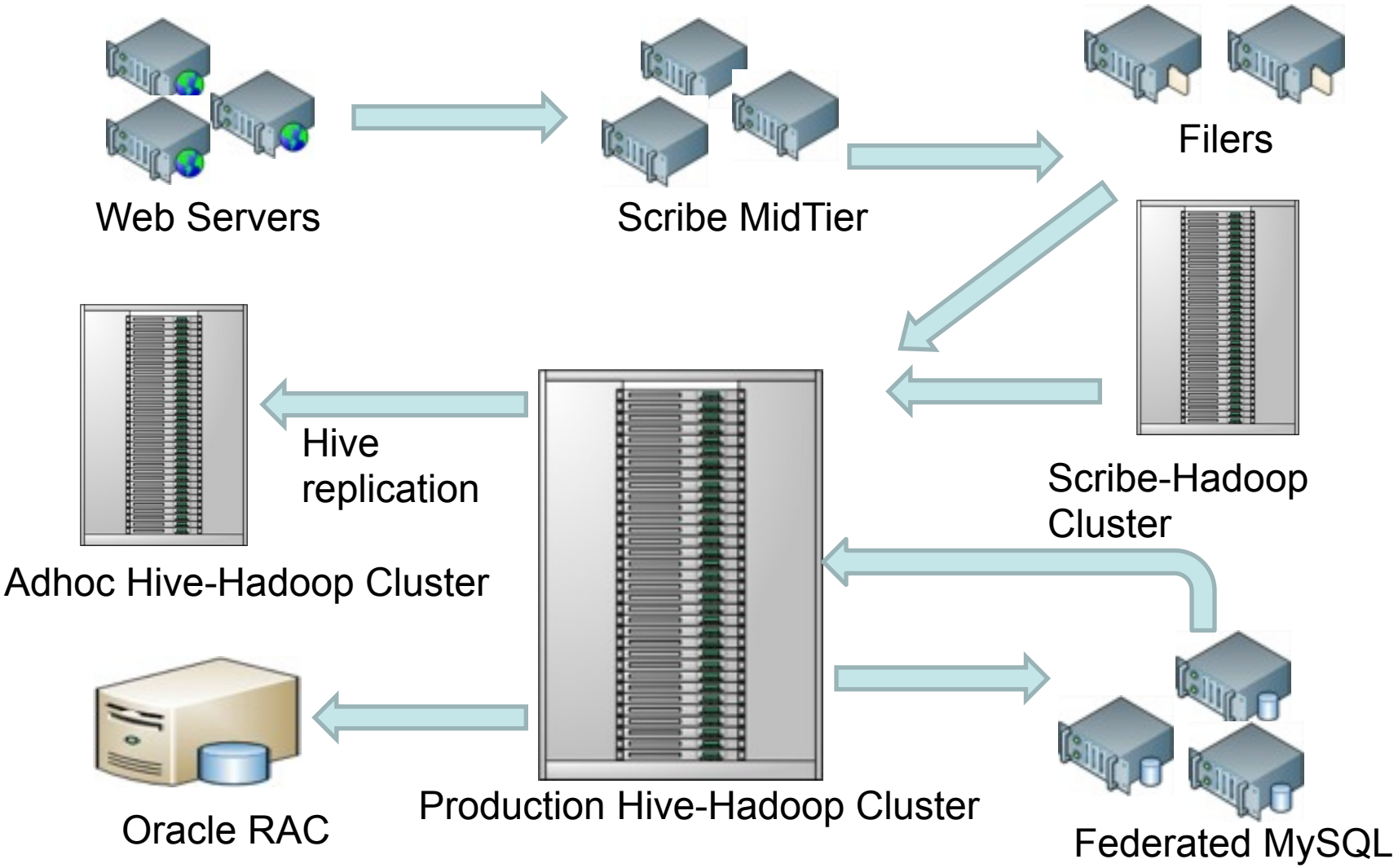
Data Flow Architecture at Facebook



Data Flow Architecture at Facebook



Data Flow Architecture at Facebook



Scribe & Hadoop Clusters @ Facebook

- Used to log data from web servers
- Clusters collocated with the web servers
- Network is the biggest bottleneck
- Typical cluster has about 50 nodes.
- Stats:
  - ~ 25TB/day of raw data logged
  - 99% of the time data is available within 20 seconds

Hadoop & Hive Cluster @ Facebook

- Hadoop/Hive cluster
  - 8400 cores
  - Raw Storage capacity ~ 12.5PB
  - 8 cores + 12 TB per node
  - 32 GB RAM per node
  - Two level network topology
    - 1 Gbit/sec from node to rack switch
    - 4 Gbit/sec to top level rack switch
- 2 clusters
  - One for adhoc users
  - One for strict SLA jobs

Hive & Hadoop Usage @ Facebook

- **Statistics per day:**
  - 12 TB of compressed new data added per day
  - 135TB of compressed data scanned per day
  - 7500+ Hive jobs per day
  - 80K compute hours per day

- **Hive simplifies Hadoop:**
  - New engineers go though a Hive training session
  - ~200 people/month run jobs on Hadoop/Hive
  - Analysts (non-engineers) use Hadoop through Hive
  - Most of jobs are Hive Jobs

Hive & Hadoop Usage @ Facebook

- Types of Applications:
  - Reporting
    - Eg: Daily/Weekly aggregations of impression/click counts
    - Measures of user engagement
    - Microstrategy reports
  - Ad hoc Analysis
    - Eg: how many group admins broken down by state/country
  - Machine Learning (Assembling training data)
    - Ad Optimization
    - Eg: User Engagement as a function of user attributes
  - Many others

More about HIVE

Data Model

	Name	HDFS Directory
Table	pvs	/wh/pvs
Partition	ds = 20090801, ctry = US	/wh/pvs/ds=20090801/ctry=US
Bucket	user into 32 buckets HDFS file for user hash 0	/wh/pvs/ds=20090801/ctry=US/ part-00000

Hive Query Language

- SQL
  - Sub-queries in from clause
  - Equi-joins (including Outer joins)
  - Multi-table Insert
  - Multi-group-by
  - Embedding Custom Map/Reduce in SQL
- Sampling
- Primitive Types
  - integer types, float, string, boolean
- Nestable Collections
  - array<any-type> and map<primitive-type, any-type>
- User-defined types
  - Structures with attributes which can be of any-type

Optimizations

- Joins try to reduce the number of map/reduce jobs needed.
- Memory efficient joins by streaming largest tables.
- Map Joins
  - User specified small tables stored in hash tables on the mapper
  - No reducer needed
- Map side partial aggregations
  - Hash-based aggregates
  - Serialized key/values in hash tables
  - 90% speed improvement on Query
    - `SELECT count(1) FROM t;`
- Load balancing for data skew

Hive: Open & Extensible

- Different on-disk storage(file) formats
  - Text File, Sequence File, …
- Different serialization formats and data types
  - LazySimpleSerDe, ThriftSerDe …
- User-provided map/reduce scripts
  - In any language, use stdin/stdout to transfer data …
- User-defined Functions
  - Substr, Trim, From_unixtime …
- User-defined Aggregation Functions
  - Sum, Average …
- User-define Table Functions
  - Explode …

Existing File Formats

	TEXTFILE	SEQUENCEFILE	RCFILE
Data type	text only	text/binary	text/binary
Internal Storage order	Row-based	Row-based	Column-based
Compression	File-based	Block-based	Block-based
Splitable*	YES	YES	YES
Splitable* after compression	NO	YES	YES

*** Splitable: Capable of splitting the file so that a single huge file can be processed by multiple mappers in parallel.**

Map/Reduce Scripts Examples

- add file page_url_to_id.py;
- add file my_python_session_cutter.py;
- FROM

```
  (MAP uhash, page_url, unix_time
     USING 'page_url_to_id.py'
     AS (uhash, page_id, unix_time)
   FROM mylog
   DISTRIBUTE BY uhash
   SORT BY uhash, unix_time) mylog2
REDUCE uhash, page_id, unix_time
  USING 'my_python_session_cutter.py'
  AS (uhash, session_info);
```

UDF Example

- add jar build/ql/test/test-udfs.jar;
- CREATE TEMPORARY FUNCTION testlength AS 'org.apache.hadoop.hive.ql.udf.UDFTestLength';
- SELECT testlength(page_url) FROM mylog;
- DROP TEMPORARY FUNCTION testlength;

- UDFTestLength.java:

```
package org.apache.hadoop.hive.ql.udf;
public class UDFTestLength extends UDF {
  public Integer evaluate(String s) {
    if (s == null) {
      return null;
    }
    return s.length();
  }
}
```

Comparison of UDF/UDAF/UDTF v.s. M/R scripts

	UDF/UDAF/UDTF	M/R scripts
language	Java	any language
data format	in-memory objects	serialized streams
1/1 input/output	supported via UDF	supported
n/1 input/output	supported via UDAF	supported
1/n input/output	supported via UDTF	supported
Speed	faster	slower

Interoperability: Interfaces

- JDBC
  - Enables integration with JDBC based SQL clients
- ODBC
  - Enables integration with Microstrategy
- Thrift
  - Enables writing cross language clients
  - Main form of integration with php based Web UI

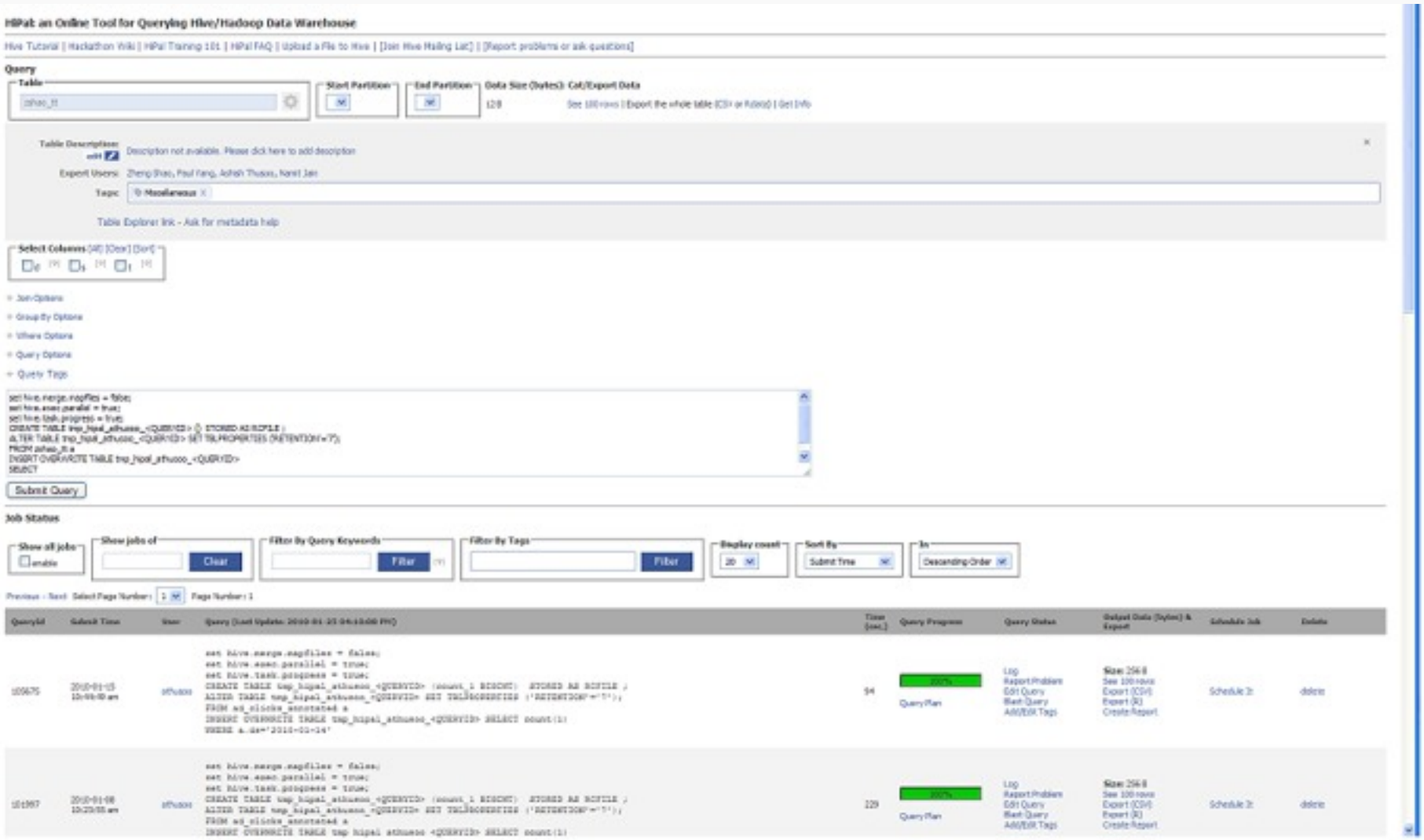
Interoperability: Microstrategy

- Beta integration with version 8
- Free form SQL support
- Periodically pre-compute the cube

Operational Aspects on Adhoc cluster

- Data Discovery
  - coHive
    - Discover tables
    - Talk to expert users of a table
    - Browse table lineage
- Monitoring
  - Resource utilization by individual, project, group
  - SLA monitoring etc.
  - Bad user reports etc.

facebook


HiPal: an Online Tool for Querying Hive/Hadoop Data Warehouse
Hive Tutorial | Hackathon Wiki | HiPal Training 101 | HiPal FAQ | Upload a File to Hive | [Join Hive Mailing List] | [Report problems or ask questions]
Query
Table
Start Partition
End Partition
Data Size (bytes): Cat/Export Data
Table Description:
Description not available. Please click here to add description
Export Users:
Topic
Miscellaneous
Table Explorer link - Ask for metadata help
Select Columns
Join Options
Group By Options
Where Options
Query Options
Query Tags
set hive.merge.mapfiles = false;
set hive.exec.parallel = true;
set hive.task.progress = true;
CREATE TABLE tmp_hipal_athusoo_<QUERYID> () STORED AS RCFILE ;
ALTER TABLE tmp_hipal_athusoo_<QUERYID> SET TBLPROPERTIES ('RETENTION'='7');
FROM ad_clicks_a a
INSERT OVERWRITE TABLE tmp_hipal_athusoo_<QUERYID>
SELECT
Submit Query
Job Status
Show all jobs
Show jobs of
Clear
Filter By Query Keywords
Filter
Filter By Tags
Filter
Display count
Sort By
Submit Time
In
Descending Order
Previous - Next
Page Number: 1
QueryId
Submit Time
User
Query
Time (sec.)
Query Progress
Query Status
Output Data (bytes) & Export
Schedule Job
Delete
set hive.merge.mapfiles = false;
set hive.exec.parallel = true;
set hive.task.progress = true;
CREATE TABLE tmp_hipal_athusoo_<QUERYID> (count_1 BIGINT) STORED AS RCFILE ;
ALTER TABLE tmp_hipal_athusoo_<QUERYID> SET TBLPROPERTIES ('RETENTION'='7');
FROM ad_clicks_a a
INSERT OVERWRITE TABLE tmp_hipal_athusoo_<QUERYID> SELECT count(1)
WHERE a.ds='2010-01-14'
94
100%
Query Plan
Log
Report Problem
Edit Query
Blast Query
Add/Edit Tags
Size: 256 B
See 100 rows
Export (CSV)
Export (R)
Create Report
Schedule It
delete
129
100%
Query Plan
Log
Report Problem
Edit Query
Blast Query
Add/Edit Tags
Size: 256 B
See 100 rows
Export (CSV)
Export (R)
Create Report
Schedule It
delete

Open Source Community

- Released Hive-0.4 on 10/13/2009
- 50 contributors and growing
- 11 committers
  - 3 external to Facebook
- Available as a sub project in Hadoop
  - http://wiki.apache.org/hadoop/Hive (wiki)
  - http://hadoop.apache.org/hive (home page)
  - http://svn.apache.org/repos/asf/hadoop/hive (SVN repo)
  - ##hive (IRC)
  - Works with hadoop-0.17, 0.18, 0.19, 0.20
- Mailing Lists:
  - hive-{user,dev,commits}@hadoop.apache.org

Powered by Hive