Hadoop Usage at Facebook

Dhruba Borthakur
Facebook Data Infrastructure Team
Data Infrastructure Team

- Cross functional team of 11 members
  - 5 people working in Hive development
  - 2 people on Hadoop development
  - 2 people on Data Pipelines and Oracle Data Mart
  - 1 Production Operations
Hadoop Cluster Hardware

- **Production cluster**
  - 2400 cores, 300 machines, 16GB per machine -- Oct 2008
  - 4800 cores, 600 machines, 16GB/8GB per machine – Nov 2008
  - 4 SATA disks of 1 TB each per machine
  - 2 level network hierarchy, 40 machines per rack

- **Test cluster**
  - 800 cores, 16GB each
HDFS Storage Cluster

- Single HDFS cluster across all cores
  - Running 0.17.2 + patches
  - Over 1.2 PB raw capacity
  - Ingest rate is 2 TB compressed user-data per day
    - 10 TB uncompressed
  - 10 Million files
- NameNode on 64 bit JVM with 20GB heap size
Rate of Growth of Hadoop Data
Map-Reduce Compute Clusters

- 3 static Map-Reduce clusters
  - Running 0.17.2 + patches
  - Main cluster has 2240 cores
    - Serves most users
  - Ads cluster has 80 cores
    - Dedicated for Advertisement related computations
  - Test cluster of 80 cores
    - Test miscellaneous fixes
- Job Tracker(s) on 32 bit JVM with 3GB heap size
A Facebook User has many dimensions
User Data

- 100M+ Active users
- Each user has 100+ dimensions
- 500K new user per day (one San Francisco every 36 hours)
Facebook Friend List
Social Graph Data

- 200+ friends per user (median)
- Interactions among users (100+ types of interactions)
- Interactions among users and application
Data Usage

- Statistics per day:
  - 55TB of compressed data scanned per day
  - 3200+ jobs on production cluster per day
  - 80M compute minutes per day

- Barrier to entry is significantly reduced:
  - All new engineers go through a Hadoop training
  - 50+ engineers have run jobs on Hadoop platform
  - Analysts (non-engineers) starting to use Hadoop through Hive
Typical Job Durations
Cluster Usage Dashboard

- History logs are fed into a mySQL database
- A Dashboard displays cluster usage statistics from the database
- Displays cluster utilization, growth rates of cluster usage, etc
- HADOOP-3708
### Cluster Usage Dashboard

**Jobs**

**Compute Time**

**Map Time**

**Reduce Time**

**Job Durations**

**Map Durations**

**Reduce Durations**

**Jobs by Date**

**Compute Time by Date**

**Task Time by Date**

**I/O by Date**

**HDFS Size**

**HDFS Metadata**

**Largest Hive Tables**

**Largest Home Directories**

**Largest Facebook Project Directories**

---

**Days back:** [ ] **month**  [ ] **year**  [ ] **other**  [ ] **Break down by:** [ ] **job name**  [ ] **status**

---

**Hadoop Job Compute Time (hours) by Job Name**

<table>
<thead>
<tr>
<th>Job Name</th>
<th>Compute Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>extract_used_from_ext</td>
<td>20.95K</td>
</tr>
<tr>
<td>ETL: aggregate asf ad log limps aggregation stage</td>
<td>15.4K</td>
</tr>
<tr>
<td>ETL: aggregate asf ad log limps stage1</td>
<td>14.5K</td>
</tr>
<tr>
<td>ETL: aggregate monthly ad log limps stage2</td>
<td>12.56K</td>
</tr>
<tr>
<td>ETL: aggregate asf ad log limps stage3</td>
<td>11.73K</td>
</tr>
<tr>
<td>ETL: aggregate monthly ad log limps stage4</td>
<td>11.44K</td>
</tr>
<tr>
<td>gen_newsfed_tracking_imp_monthly_stats</td>
<td>10.31K</td>
</tr>
<tr>
<td>ETL: aggregate asf ad log limps stage5</td>
<td>9.11K</td>
</tr>
<tr>
<td>User Profile ETL into Hive</td>
<td>6.45K</td>
</tr>
<tr>
<td>ETL: aggregate weekly ad log limps stage1</td>
<td>5.9K</td>
</tr>
<tr>
<td>gen_newsfed_tracking_imp_weekly_stats</td>
<td>5.85K</td>
</tr>
<tr>
<td>ETL: ad_impressions_unique_monthly_2008-09-26</td>
<td>5.35K</td>
</tr>
<tr>
<td>ETL: aggregate monthly ad log limps aggregation stage</td>
<td>5.34K</td>
</tr>
<tr>
<td>ETL: aggregate monthly ad log limps stage2</td>
<td>5.11K</td>
</tr>
<tr>
<td>Coefficient: partition and load attempt scores</td>
<td>4.67K</td>
</tr>
<tr>
<td>ETL: aggregate asf ad log limps stage2</td>
<td>4.28K</td>
</tr>
<tr>
<td>facebook user plus friends hotset analysis</td>
<td>4.23K</td>
</tr>
<tr>
<td>ETL: aggregate asf ad log limps stage5</td>
<td>4.22K</td>
</tr>
<tr>
<td>Coefficient: partition and load attempt scores</td>
<td>4.01K</td>
</tr>
</tbody>
</table>

---

Confidential Materials — For Internal Use Only

Hadoop Job Compute Time (hours) by Job Name

---

[Add Comment]
Fair Share Scheduler

- Short jobs finish fast while not starving longer jobs
- A pool has a guaranteed minimum allocation
  - Data loading jobs use a pool with large resource allocation
- Allows limiting the number of concurrent jobs per pool
- No preemption yet
- HADOOP-3746
# Fair Share Scheduler

## Pools

<table>
<thead>
<tr>
<th>Pool</th>
<th>Running Jobs</th>
<th>Min Maps</th>
<th>Min Reduces</th>
<th>Running Maps</th>
<th>Running Reduces</th>
</tr>
</thead>
<tbody>
<tr>
<td>akramer</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>214</td>
<td>0</td>
</tr>
<tr>
<td>dierrante</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hwang</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>itamar</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>jssarma</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>lev</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>prha</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>realtime</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>root</td>
<td>5</td>
<td>1000</td>
<td>1000</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>viyer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>default</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Universal Access to HDFS

- Mount HDFS through fuse (HADOOP-4)
  - An user can use common utilities, e.g. find, ls on HDFS files
  - Goal to have HDFS fuse-mounted on all developer machines
- Use Thriftfs (HADOOP-3754)
  - Access HDFS through Python, PHP
  - Enables custom joins written in Python
HiPal: an Online Tool for Querying Hive/Hadoop Data Warehouse

Query:
- **Table**: u_full
- **Start Partition**: 
- **End Partition**: 
- **Data Size (bytes)**: 734,184,513,313
- **Get 10 rows** | **Export the whole u_full**

**Select Columns** [All] [Clear]
- `userid`
- `base`
- `affiliations`
- `last`
- `friends`
- `ext`
- `groups`
- `fbpages`
- `whs`
- `events`
- `photo_tags`
- `schools`
- `applications`
- `regionid`

**Join Options**
- **Join Options**
- **Group By Options**
- **Where Options**
- **Query Options**

CREATE TABLE tmp_hipal_<QUERYID> (userid STRING, friends STRING, schools STRING);
FROM u_full TABLESAMPLE (BUCKET 1 OUT OF 1024) SELECT a.userid, a.friends, a.schools

**Submit Query**

[Join HiPal User Mailing List] [Report problems or ask questions]

**Job Status**

<table>
<thead>
<tr>
<th>QueryId</th>
<th>Submit Time</th>
<th>User</th>
<th>Query (Last Update: 2008-10-27 12:42:58 AM)</th>
<th>Time (sec)</th>
<th>Query Progress</th>
<th>Latest Hadoop Job Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>3393</td>
<td>2008-10-15 1:48:20</td>
<td>dhnabs</td>
<td>CREATE TABLE tmp_hipal_&lt;QUERYID&gt; (userid STRING, friends STRING, schools STRING) FROM u_full TABLESAMPLE (BUCKET 1 OUT OF 1024) SELECT a.userid, a.friends, a.schools</td>
<td>37</td>
<td>100%</td>
<td>Status</td>
</tr>
</tbody>
</table>

- **Time (sec)**: 37
- **Query Progress**: 100%
- **Latest Hadoop Job Status**: Status
In the Works

- Hierarchical Storage Manager (HADOOP-4058)
  - Migrate data that will be rarely used in future

- Global scheduler
  - Ability to schedule more tasks to be rack-local
  - Ability to process multiple local splits by a single task

- Data Protection
  - HDFS Snapshots (HADOOP-3637)
  - HDFS Symlinks (HADOOP-4044)
Questions?
dhruba@facebook.com
Credits

Suresh Anthony
Zheng Shao
Prasad Chakka
Pete Wyckoff
Namit Jain
Raghu Murthy
Joydeep Sen Sarma
Rama Ramasamy
Matei Zaharia
Ashish Thusoo
Hao Liu