HIVE IN KIXEYE ANALYTICS

Aaron Sun, in collaboration with Taehoon Kang, William Greene, Ben Speakmon and Chris Mills
About KIXEYE

- An online gaming company focused on mid-core and hard-core games
  - Founded in 2007
  - Over 400 employees by Feb 2013
  - 5 times longer retention and 20 times higher ARPU

- Analytics Engineering Team
  - Part of the Business Intelligence team
  - 12 team members
INTRO

1 M active users

Triple the number in 2013 Q2

160 M click events

100G logs data

DAILY STATS

BATTLE PIRATES

1 M active users

Triple the number in 2013 Q2
Requirements

• A fault-tolerant and scalable system
• Support standard reports
• Support ad-hoc, exploratory data queries
• Easy to use and manage
• Real-time is nice-to-have, but not necessary
Architecture

Custom ETL + Sqoop

Log

HIVE

MySQL

Dashboard

async REST API

Oozie
Log Collection and Processing

- 4 ~ 6 GB uncompressed logs (player clicks) / hour
- Logs collected by Apache Chukwa
  - Choose over Flume and Scribe for Chukwa’s easy configuration
- Log data cleaned and parsed as Snappy compressed JSON (staging)
  - Choose over Protocol Buffer and Avro for JSON’s simplicity
ETL Component

- Hadoop cluster size
  - 12-core node * 20
  - 240 mappers and 180 reducers

- Run ETL every 30 min
  - Populate RCFile into Hive tables

- Sqoop is used to collect data from legacy ETL system

- All ETL tasks are managed by Oozie

Custom ETL + Sqoop

Oozie
HIVE Tables

- 70+ click types (e.g. “install”, “attack”) are loaded into corresponding tables

- Insertion is done by enabling dynamic partitions
  - “FROM SELECT *** INSERT INTO” is very useful

- Tables are usually partitioned by game and day
  - Some are further partitioned by hour
HIVE Tables (Cont’d 1)

• RCFile with Snappy compression is the data format
  - Excellent performance but expensive “alter table”
  - Evaluated jsonserde and protobuf format with customserde, slow in querying

• “Clustered by” and “Tablesample”
  - A useful feature for analysts
HIVE Tables (Cont’d 2)

- Small files from hourly loading
  - Weekly merge operations
    alter table TBL partition (PART) concatenate;

- Evaluated Hive index on certain fields
  (e.g. level)
  - Improvement is not significant
Data Access – Pull

• Two data access patterns
• Pull – RESTful service built on top of Beeswax
  • Asynchronous and concurrent requests compared to HiveServer1
  • query/status/fetch
  • 100+ queries from the analysts every day

• Fixing bugs and adding features:
  • To support multi-hivedb
  • To support caching, load-balancing, and fail-over
Data Access – Push

• A wrapper library for “hive –f” command
  • Data load
  • Data merge
  • Data migration
  • Metric generation

• Used by ETL engineers
Using Hive UDTF to Generate Session Stats

- Session definition
  - Two consecutive user activities are separated as different sessions if the time interval between them exceeds a time-out threshold (e.g. 30 min)

Requirements:
- Compute incrementally
- Provide as a Hive function
Using Hive UDTF to Generate Sessions

**Redis**
Intermediate data

**Hourly Partition 01**

**Hourly Partition 02**

**Hourly Partition 03**

**Hourly Partition 04**

**view**: `collect_set(ts)` group by `uid`

| 001 | [ts1, ts2, ts3, ...] |
| 002 | [ts1, ts2, ts3, ...] |
| ... | [ts1, ts2, ts3, ...] |
| 999 | [ts1, ts2, ts3, ...] |

**UDTF**

<table>
<thead>
<tr>
<th>session_label</th>
<th>ts1</th>
</tr>
</thead>
<tbody>
<tr>
<td>session_1</td>
<td>ts2</td>
</tr>
<tr>
<td>session_2</td>
<td>ts3</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Lessons Learned

• Analysts are greedy
  • Scan full set of data and ignore partitions
  • Non-optimized joins

• RCFile is a double-edged sword
  • Sqoop does not support RCFile
  • Inflexible schema

• Automate, automate, and automate
  • Constantly-changing ETL requirements
  • New metrics on new features
Future Work

- Visualization layer
- Integration with Hbase
- Richer UDFs
We are hiring!

• Our audacious goals:
  • Build a world-class data and analytics team
  • Deliver high-quality, real-time player behavior intelligence

• Join us to build the “game-changing” analytics system
  • http://www.kixeye.com/#/en/jobs
Q & A

asun@kixeye.com