KIXEYE ANALYTICS

HIVE IN KIXEYE ANALYTICS

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

1 INTRO
2
3
4
5
6
7

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

1 INTRO
2
3
4
5
6
7

DAILY STATS

160 M
click events

100G
logs data

1 M
active users

Triple the number
in 2013 Q2

1

2 Requirements

3

4

5

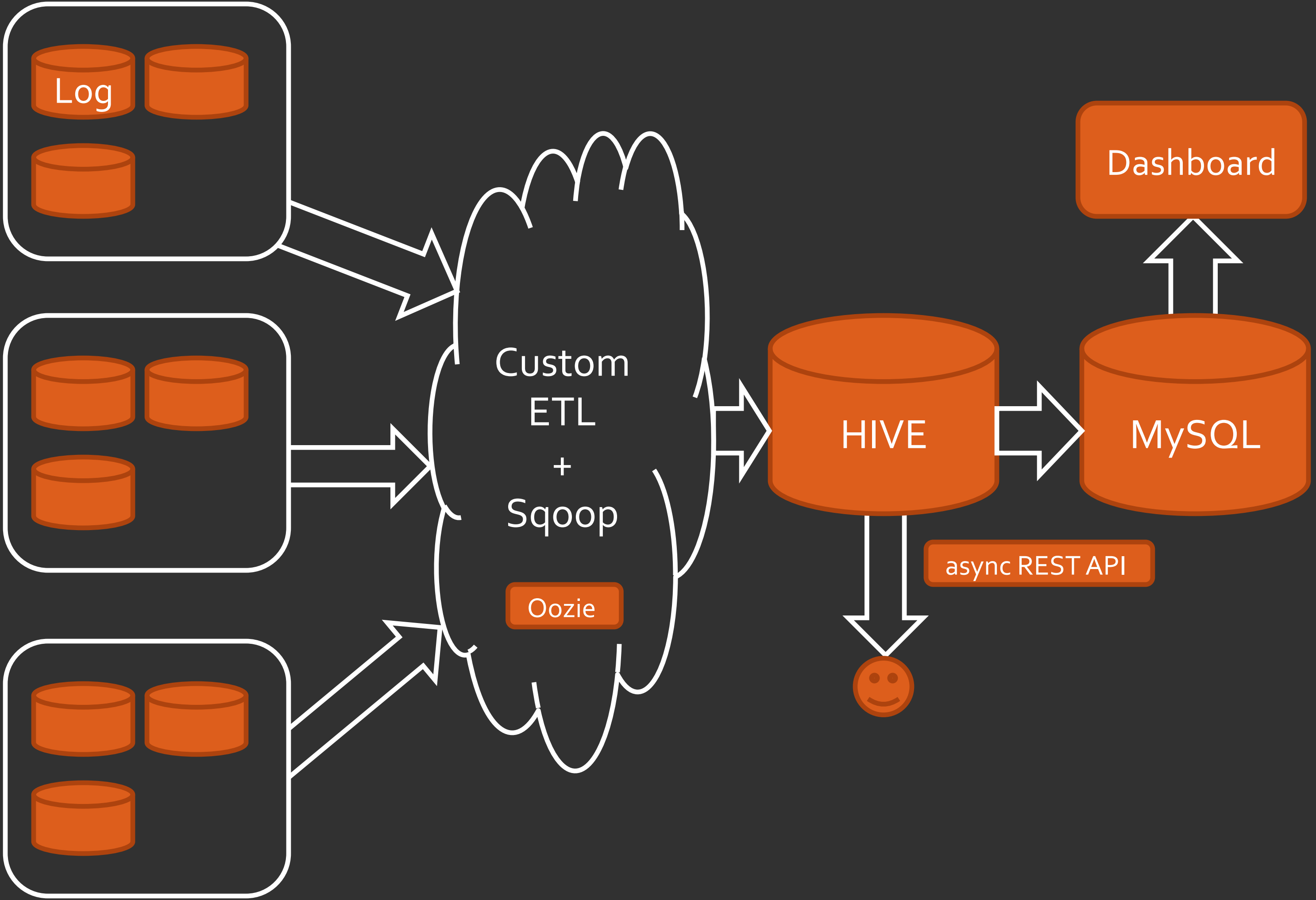
6

7

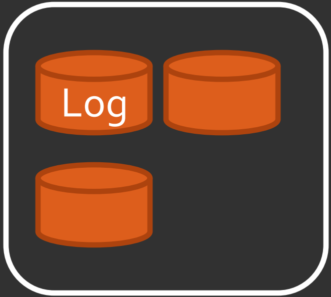
Requirements

- 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



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

Custom ETL
+
Sqoop

Oozie

- 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 for collecting data from legacy ETL system

- All ETL tasks are managed by Oozie

HIVE Tables

HIVE

- 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

1

2

3 Architecture

4

5

6

7

HIVE Tables (Cont’d 1)

HIVE

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

- “Clustered by” and “Tablesample”
  - A useful feature for analysts

1
2
3 Architecture
4
5
6
7

HIVE Tables (Cont’d 2)

HIVE

- 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

1
2
3
4 Data Access
5
6
7

Data Access – Pull

- Two data access patterns
- Pull – RESTful service built on top of Beeswax

  async REST API

  - 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

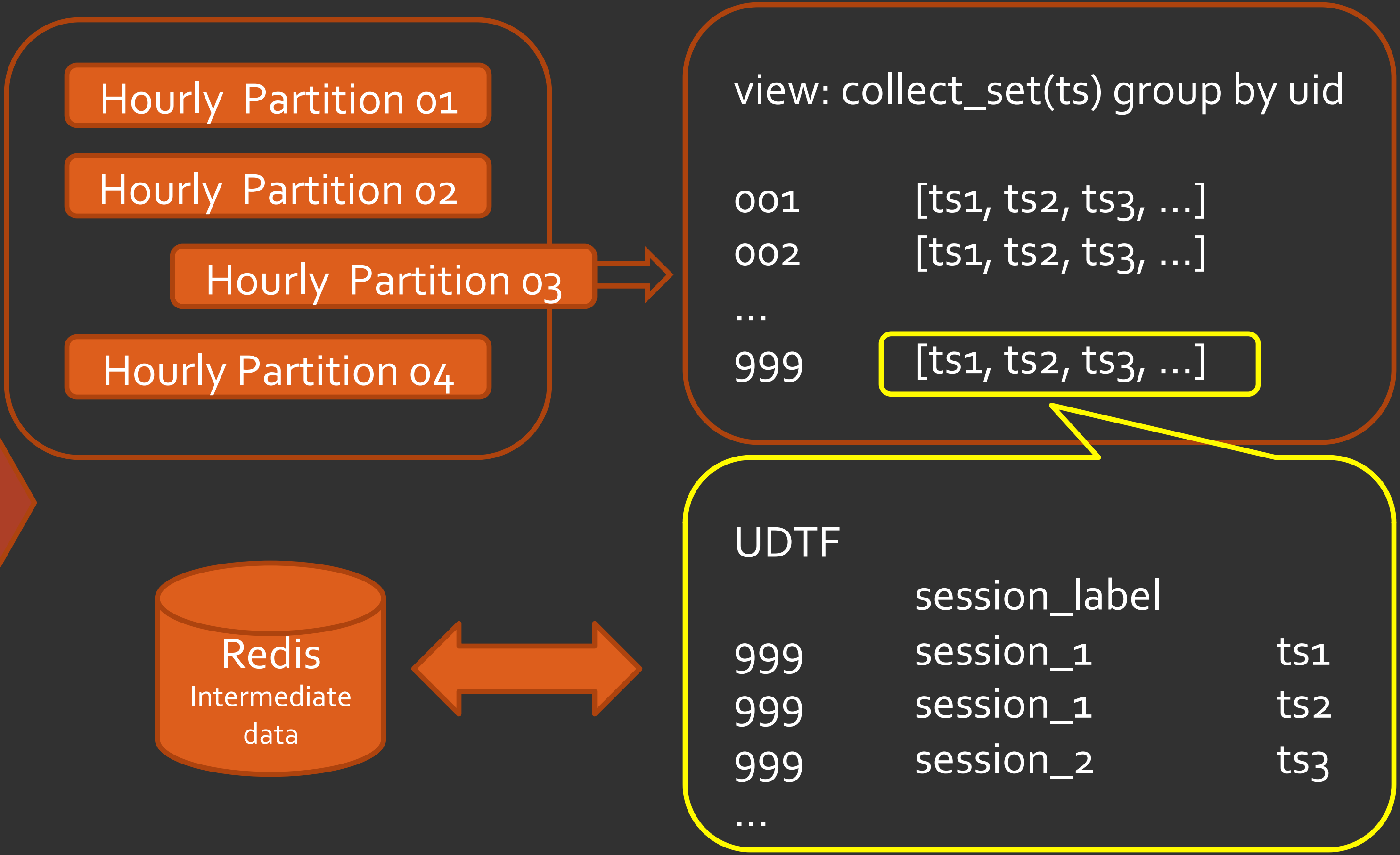
Using Hive UDTF to Generate Session Stats

- 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

2
3
4
5 A Case Study
6
7

Using Hive UDTF to Generate Sessions



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, …]
…
999        [ts1, ts2, ts3, …]
```

UDTF

```
           session_label
999        session_1          ts1
999        session_1          ts2
999        session_2          ts3
…
```

Redis
Intermediate data

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

1

2

3

4

5

6 Final notes

7

Future Work

- Visualization layer
- Integration with Hbase
- Richer UDFs

2
3
4
5
6
7 Questions

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

2
3
4
5
6
7 Final Notes

Q & A

asun@kixeye.com