MOVING MOUNTAINS OF PLAYER DATA

SCALABLE INTERNET SERVICES
UCLA/UCSB - NOV 2016

SEAN MALONEY
RIOT GAMES
@SEAN_SEANNERY

WHO IS THIS GUY?

Lead developer on Riot's ETL and real-time services

FUN FACT:
Was a student in this class 5 years ago
Intern at Appfolio

SEAN MALONEY
BIG DATA ENGINEER

MOVING MOUNTAINS OF DATA

1. INTRODUCTION
2. THE GAME PLATFORM: OUR MAIN DATA SOURCE
3. HOW WE INGEST AND QUERY DATA
4. HOW WE SCALE IN AWS
5. CONCLUSION - SEAN'S PRO TIPS

INTRODUCTION

WHAT IS LEAGUE OF LEGENDS?

YOUR CHAMP

THE TEAM

THE BATTLE GROUND


REPLAY

12 BILLION
GAME RELATED EVENTS

0.5 TRILLION
DATA POINTS

50 TB
STORAGE

DAILY

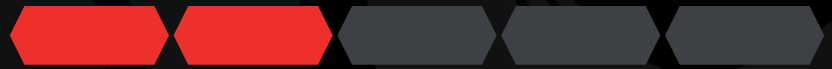
26 PETABYTES
PLAYER DATA

SINCE BETA

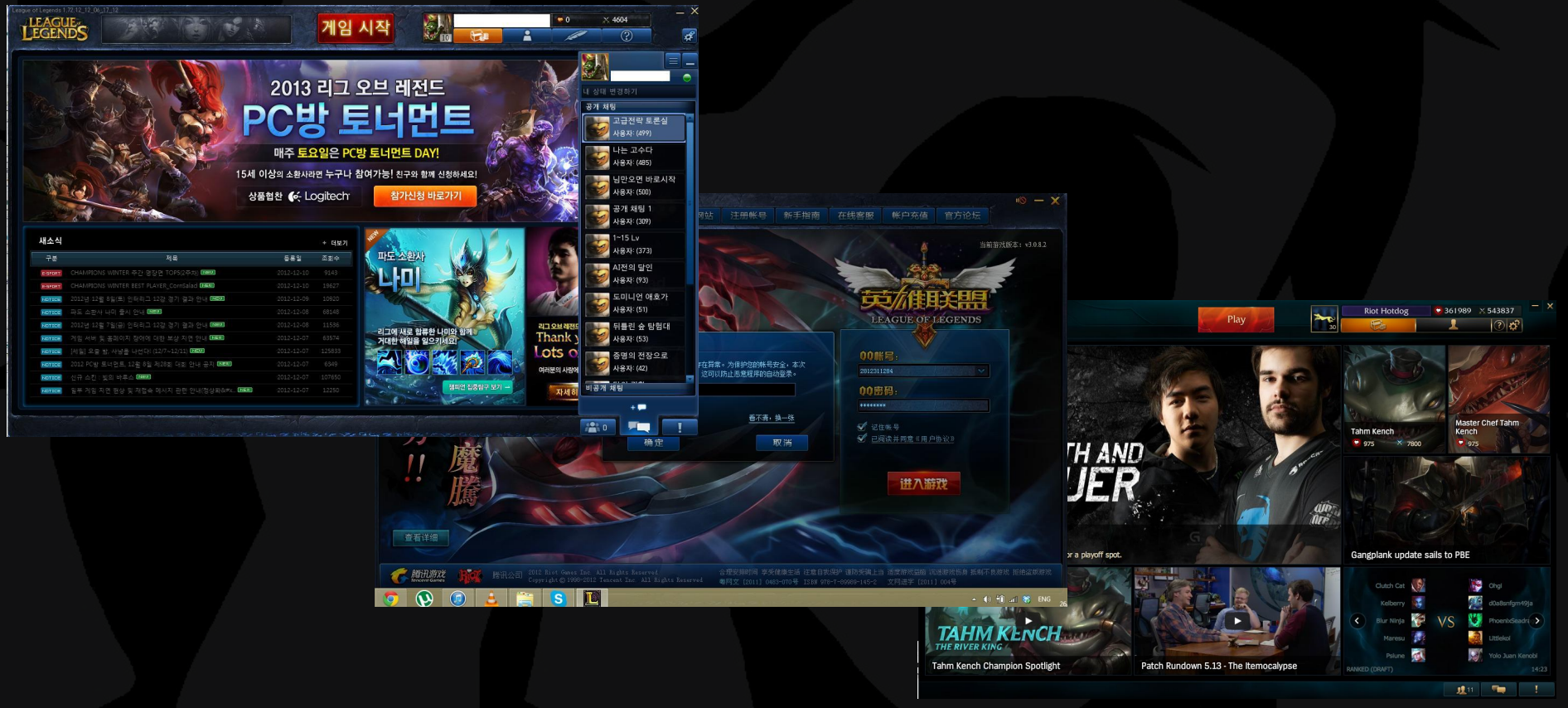
OUR MISSION

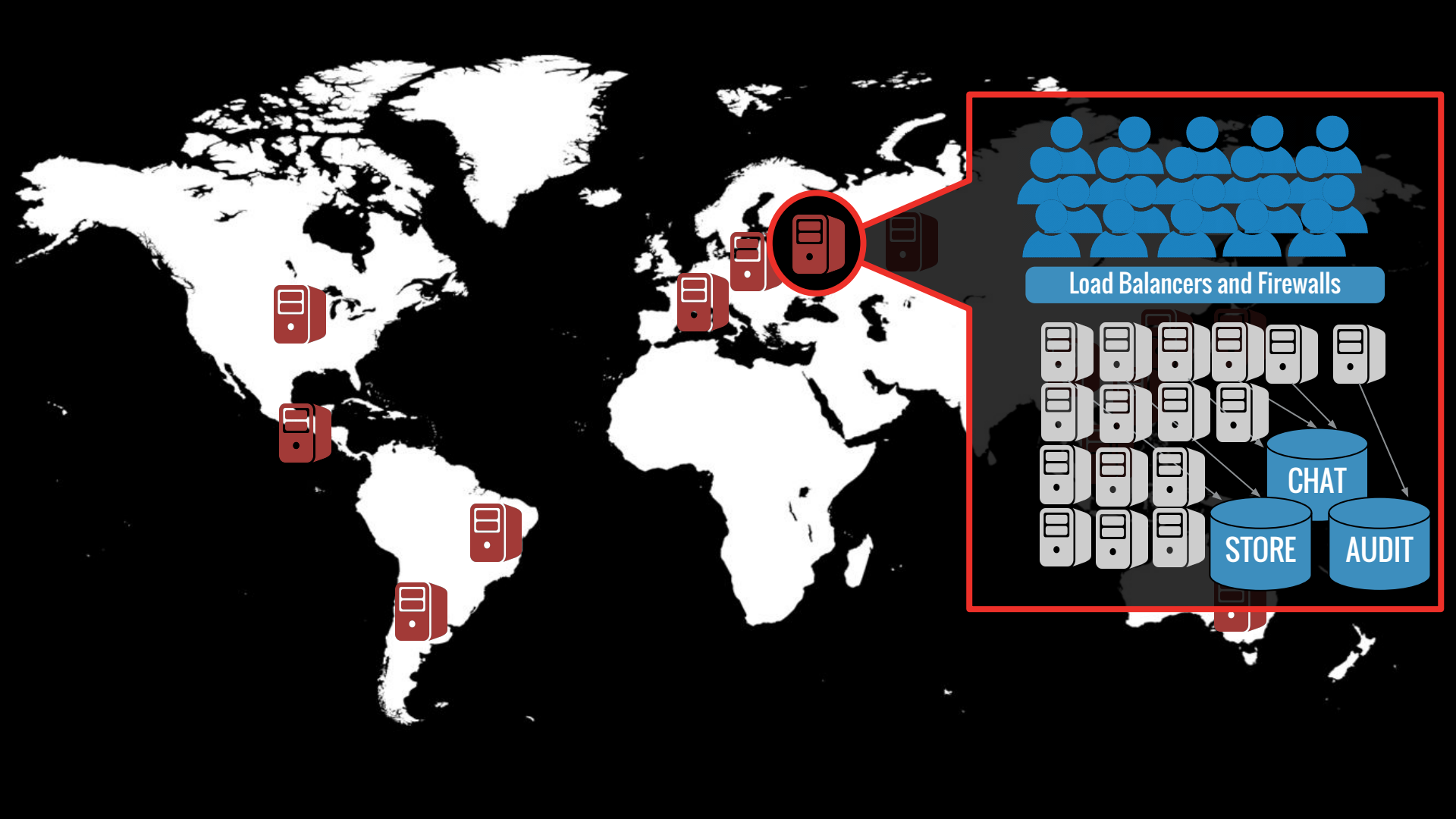
WE ASPIRE TO BE THE MOST PLAYER FOCUSED GAME COMPANY IN THE WORLD

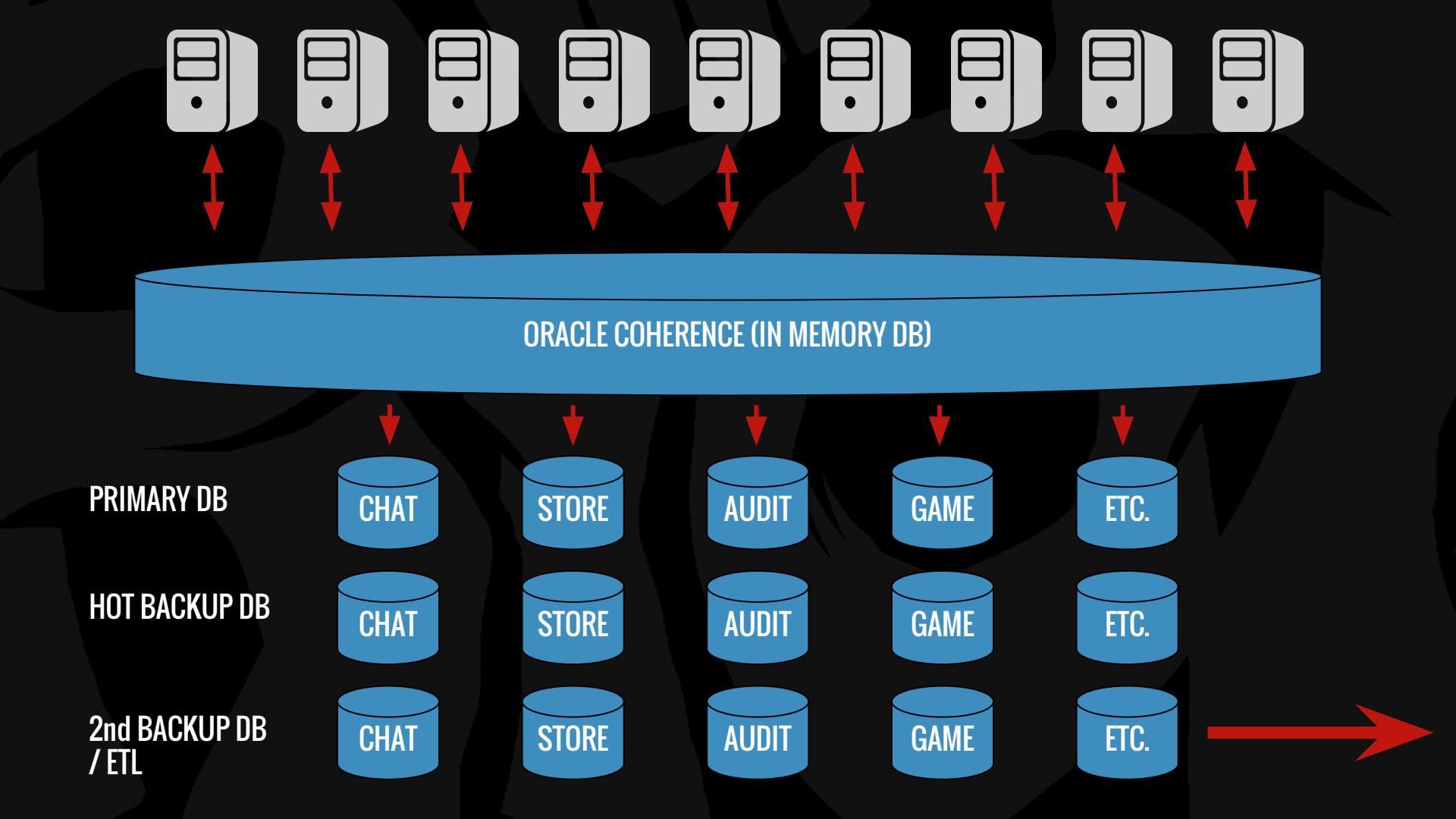
THE GAME PLATFORM

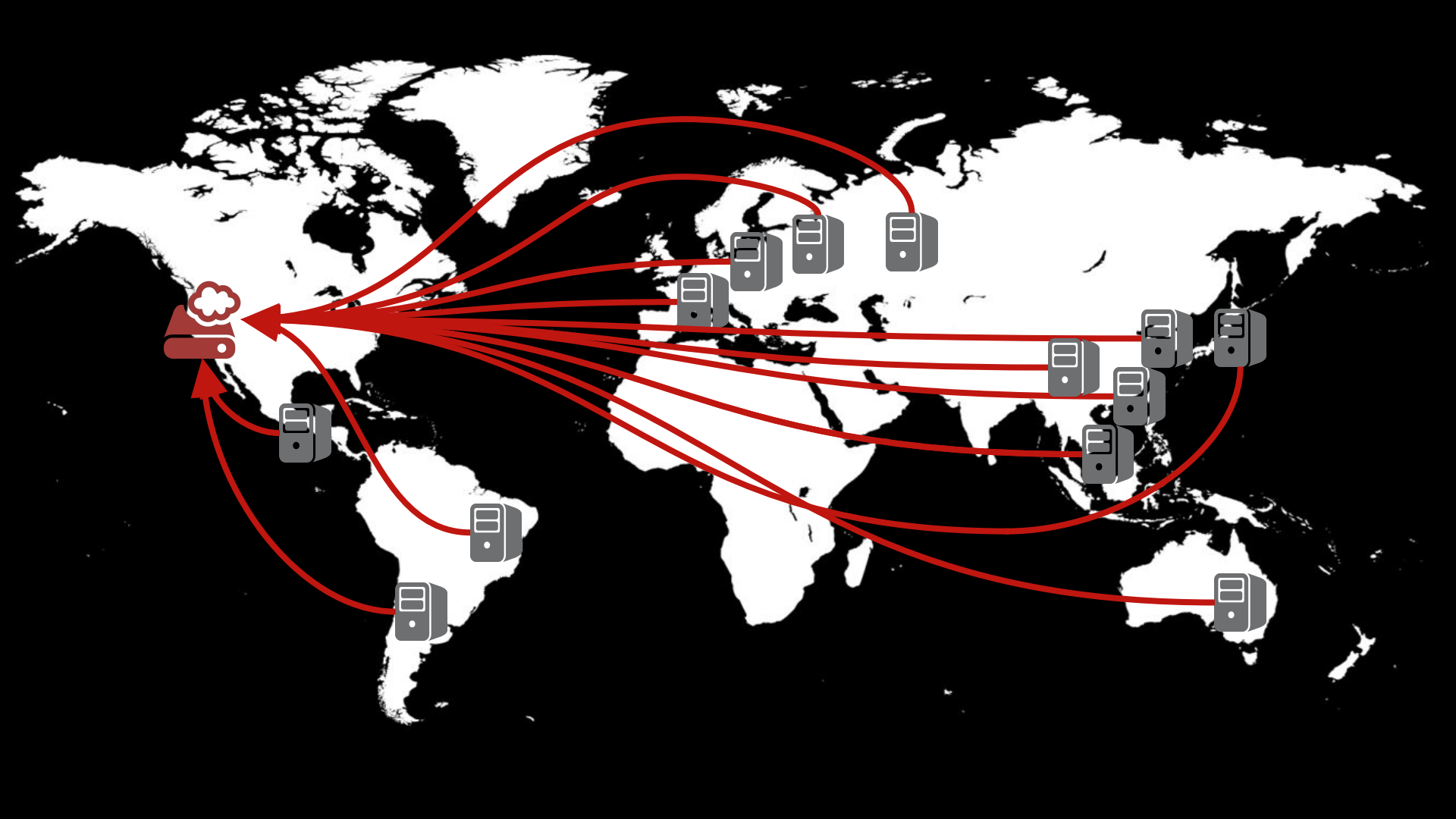


THE CLIENT.


LEAGUE OF LEGENDS
게임 시작
2013 리그 오브 레전드
PC방 토너먼트
매주 토요일은 PC방 토너먼트 DAY!
15세 이상의 소환사라면 누구나 참여가능! 친구와 함께 신청하세요!
상품협찬 Logitech
참가신청 바로가기
세소식
파도 소환사 나미
英雄联盟
LEAGUE OF LEGENDS
进入游戏
Play
Riot Hotdog
Tahm Kench
Master Chef Tahm Kench
Gangplank update sails to PBE
Tahm Kench Champion Spotlight
Patch Rundown 5.13 - The Itemocalypse


Load Balancers and Firewalls
CHAT
STORE
AUDIT


ORACLE COHERENCE (IN MEMORY DB)
PRIMARY DB
HOT BACKUP DB
2nd BACKUP DB / ETL
CHAT
STORE
AUDIT
GAME
ETC.
CHAT
STORE
AUDIT
GAME
ETC.
CHAT
STORE
AUDIT
GAME
ETC.



DATA INGESTION

INGESTION

PULL-BASED / ETL

FuETL
- OLTP game data
- External Data Sources

PUSH-BASED

kafka

- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE

amazon web services™
S3 Simple Storage Service

DATA AUDITING

Wasp

QUERY / VIEWS

AGGREGATE QUERIES

VERTICA
An HP Company

BATCH QUERIES

HIVE

SINGLE-ROW QUERIES

amazon DynamoDB

VIZ. TOOLS

+ableau

platfora
You should know.

INGESTION

PULL-BASED / ETL

FuETL
- OLTP game data
- External Data Sources

PUSH-BASED

kafka

- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE

amazon web services™
S3 Simple Storage Service

DATA AUDITING

Wasp

QUERY / VIEWS

AGGREGATE QUERIES

VERTICA
An HP Company

BATCH QUERIES

HIVE

SINGLE-ROW QUERIES

amazon DynamoDB

VIZ. TOOLS

+ableau

platfora
You should know.

Distributed ETL Software written in Ruby.

Same ETL applied to multiple regions / datacenters

Scales Horizontally

BEST LOGO EVER!

NA

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Korea

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Russia

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

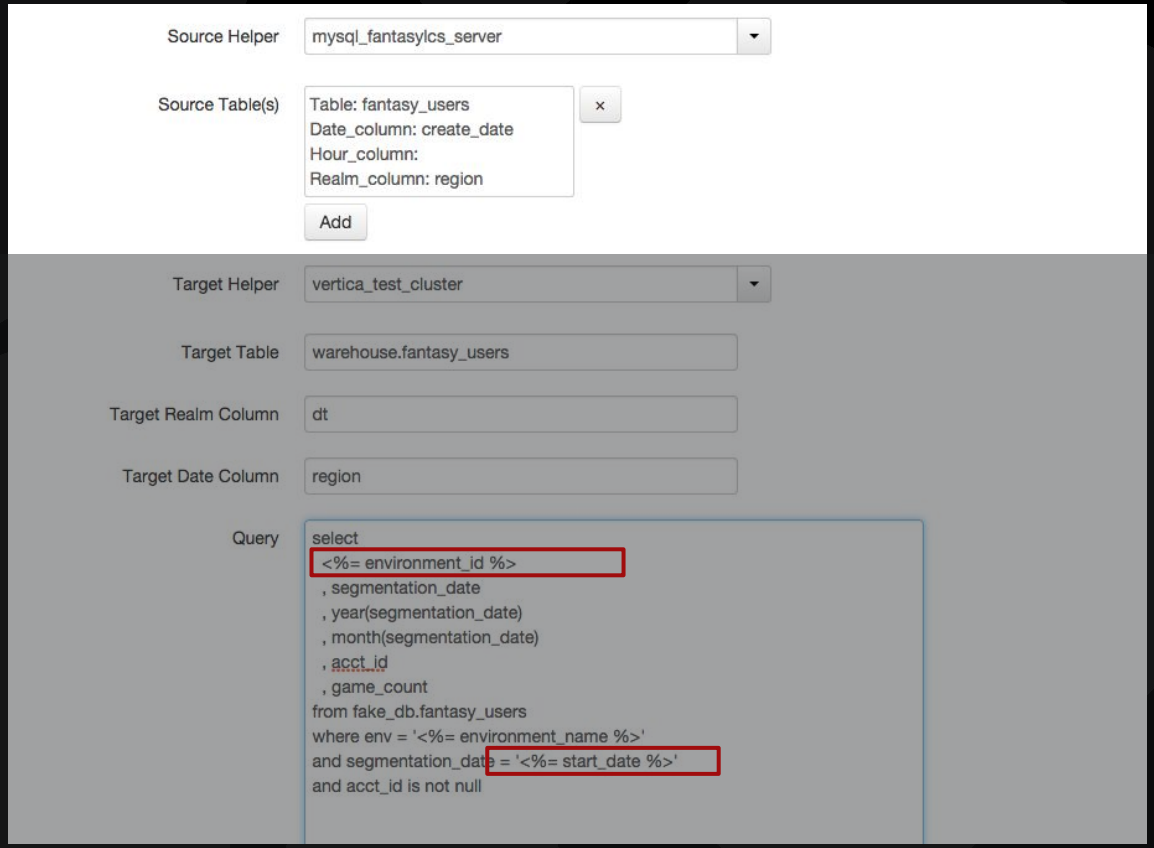
OTHER DATA SOURCES

<REST>

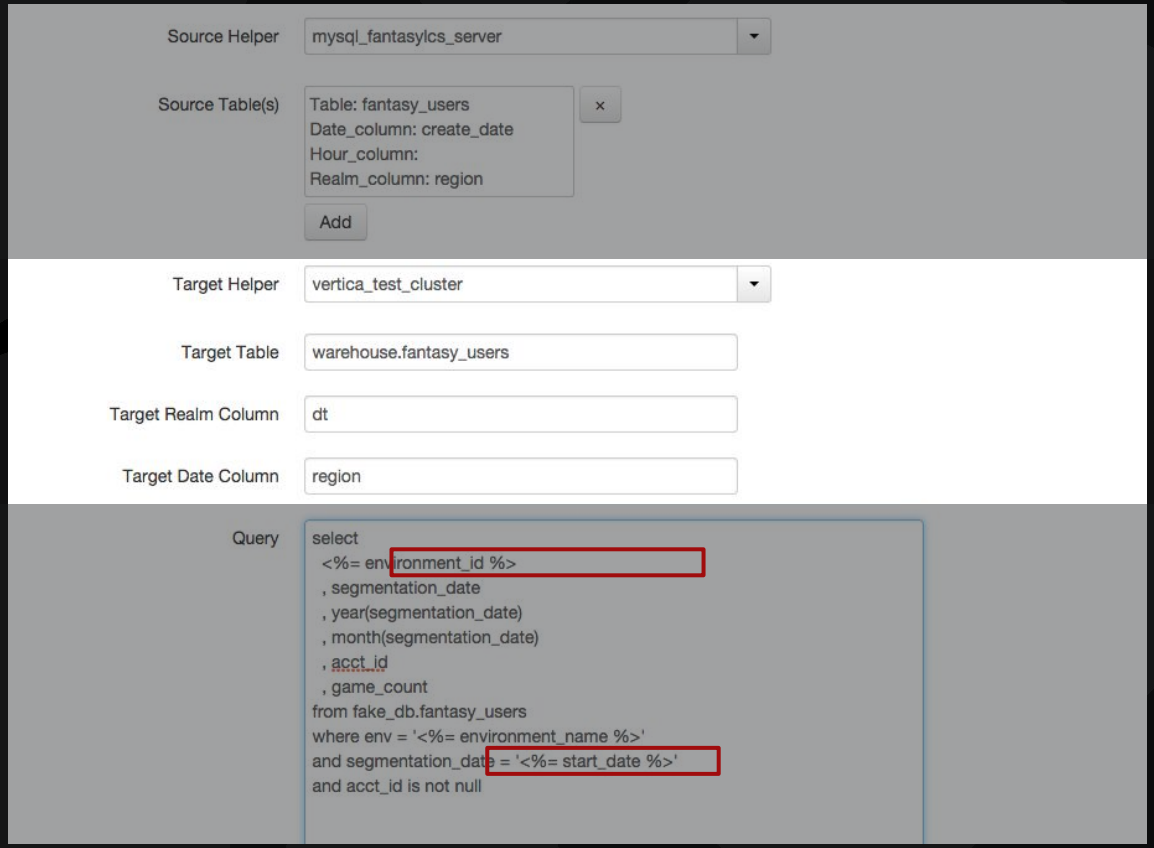
FUETL CAN CONNECT TO

- Amazon S3
- SQS
- (S)FTP
- Hive
- Microsoft SQL Server
- MySQL
- DynamoDB
- Vertica
- Redshift
- REST websites

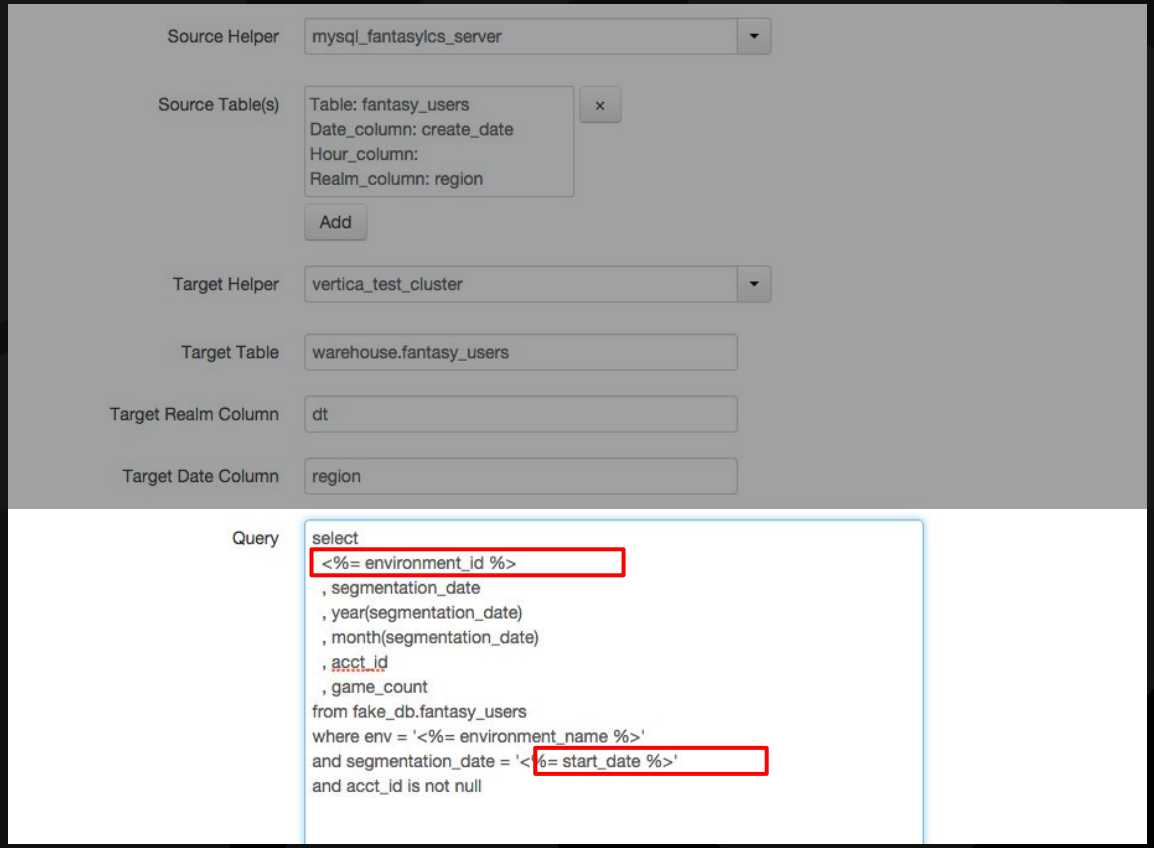
Create an ETL

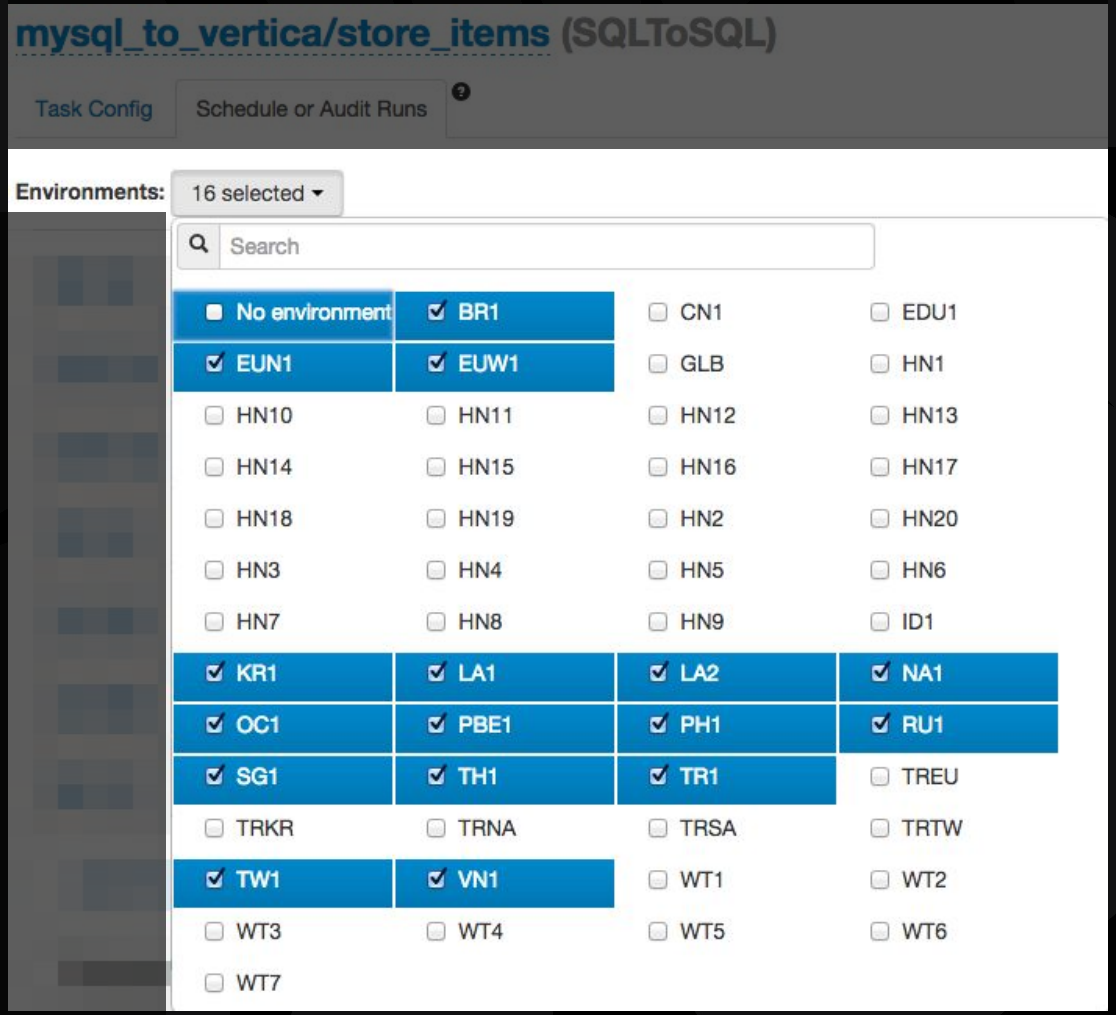


Create an ETL



Create an ETL




mysql_to_vertica/store_items (SQLToSQL)
Task Config
Schedule or Audit Runs
Environments:
16 selected
Search
No environment
BR1
CN1
EDU1
EUN1
EUW1
GLB
HN1
HN10
HN11
HN12
HN13
HN14
HN15
HN16
HN17
HN18
HN19
HN2
HN20
HN3
HN4
HN5
HN6
HN7
HN8
HN9
ID1
KR1
LA1
LA2
NA1
OC1
PBE1
PH1
RU1
SG1
TH1
TR1
TREU
TRKR
TRNA
TRSA
TRTW
TW1
VN1
WT1
WT2
WT3
WT4
WT5
WT6
WT7

mysql_to_vertica/store_items (SQLToSQL)

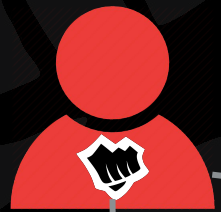
Task Config | Schedule or Audit Runs

Environments: 16 selected

BR1 (9)	success	at 7/9/2015, 4:00:44 PM	Will run again at approximately: **7/9/2015, 10:00:44 PM**	Elapsed: 0:00:10 Median: 0:00:08 Longest: 0:00:33
EUN1 (3)	success	at 7/9/2015, 4:15:40 PM	Will run again at approximately: **7/9/2015, 10:15:40 PM**	Elapsed: 0:00:05 Median: 0:00:05 Longest: 0:00:39
EUW1 (2)	success	at 7/9/2015, 4:20:30 PM	Will run again at approximately: **7/9/2015, 10:20:30 PM**	Elapsed: 0:00:07 Median: 0:00:07 Longest: 0:00:31
KR1 (4)	success	at 7/9/2015, 4:03:15 PM	Will run again at approximately: **7/9/2015, 10:03:15 PM**	Elapsed: 0:00:11 Median: 0:00:12 Longest: 0:00:32
LA1 (37)	success	at 7/9/2015, 4:03:22 PM	Will run again at approximately: **7/9/2015, 10:03:22 PM**	Elapsed: 0:00:06 Median: 0:00:06 Longest: 0:00:26
LA2 (38)	success	at 7/9/2015, 3:58:17 PM	Will run again at approximately: **7/9/2015, 9:58:17 PM**	Elapsed: 0:00:06 Median: 0:00:06 Longest: 0:00:56
NA1 (1)	success	at 7/9/2015, 4:03:08 PM	Will run again at approximately: **7/9/2015, 10:03:08 PM**	Elapsed: 0:00:07 Median: 0:00:08 Longest: 0:00:30

Latest Run | Run History | Live Logs | Overrides

Timestamp	Status	Message	Payload	Interval
7/9/2015, 4:03:07 PM	success	Transferred 2203 rows of data	2203	No Interval
7/9/2015, 4:02:59 PM	running	None	None	No Interval



Webapp

View
- backbone.js
- Bootstrap CSS

Task / Helper / Controllers

Scheduler Process

Worker Process

Command Line Tool

Core Libraries

Task Service

Environment Service

Helper Service

Tasks

Helpers

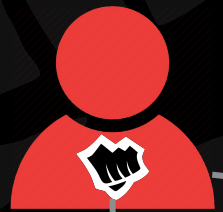
Task DAO

Env. Task DAO

Environment DAO

Env. Helper DAO

Helper DAO



Webapp

View
- backbone.js
- Bootstrap CSS

Scheduler Process

Worker Process

Command Line Tool

Task / Helper / Controllers

Core Libraries

Task Service

Environment Service

Helper Service

Tasks

Helpers

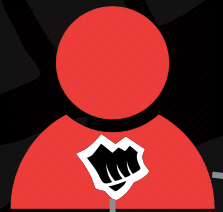
Task DAO

Env. Task DAO

Environment DAO

Env. Helper DAO

Helper DAO



Webapp

View
- backbone.js
- Bootstrap CSS

Task / Helper / Controllers

Scheduler Process

Worker Process

Command Line Tool

Core Libraries

Task Service

Environment Service

Helper Service

Tasks

Helpers

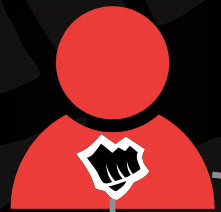
Task DAO

Env. Task DAO

Environment DAO

Env. Helper DAO

Helper DAO



Webapp

View
- backbone.js
- Bootstrap CSS

Scheduler Process

Worker Process

Command Line Tool

Task / Helper / Controllers

Core Libraries

Task Service

Environment Service

Helper Service

Tasks

Helpers

Task DAO

Env. Task DAO

Environment DAO

Env. Helper DAO

Helper DAO

FuETL STATISTICS

5213

ACTIVE REGIONAL ETLS

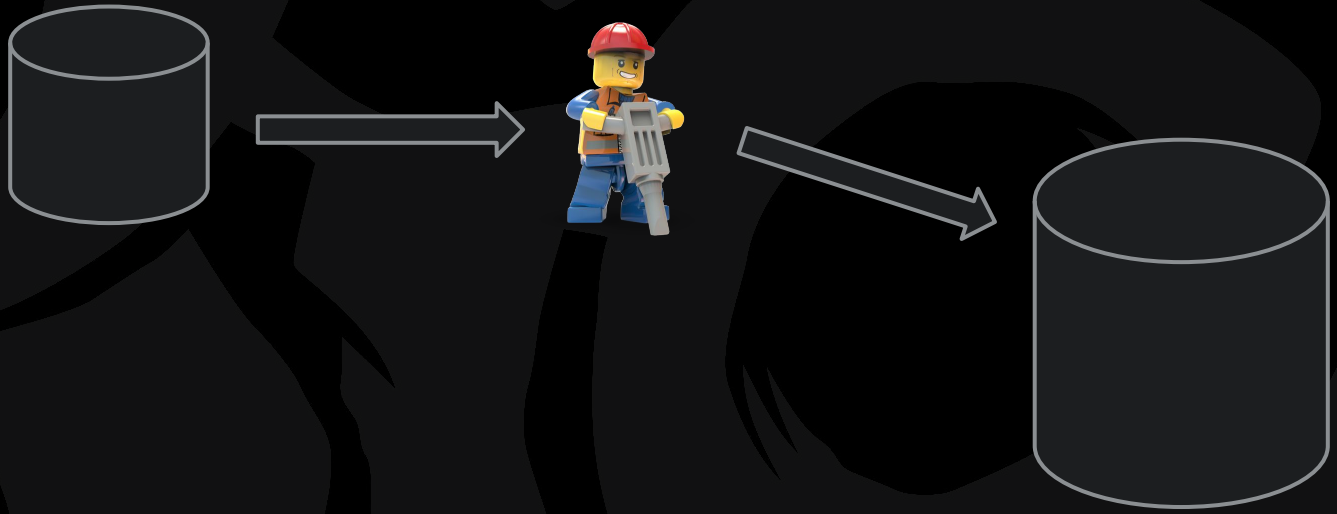
23125

DAILY ETL RUNS

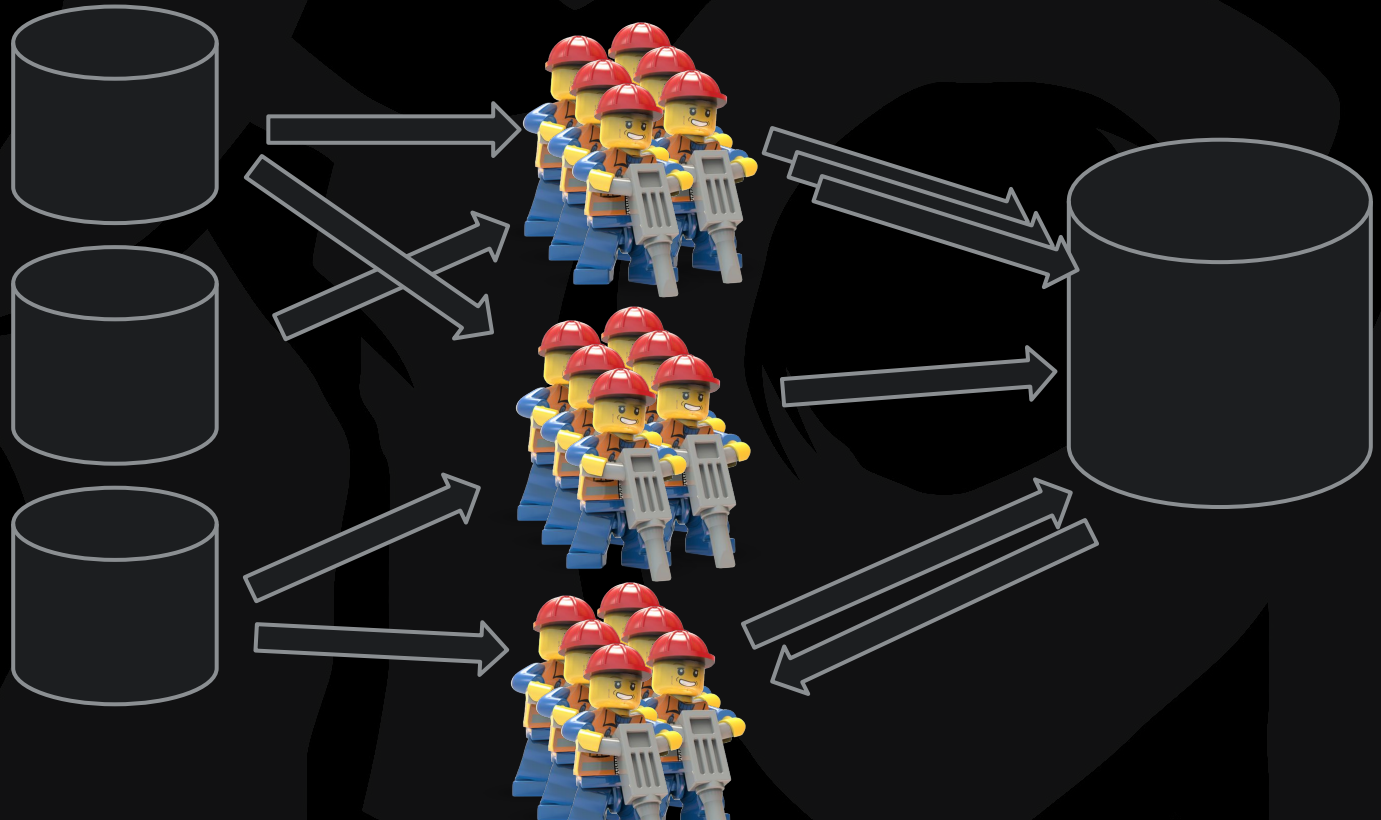
14 TB

DATA MOVED DAILY

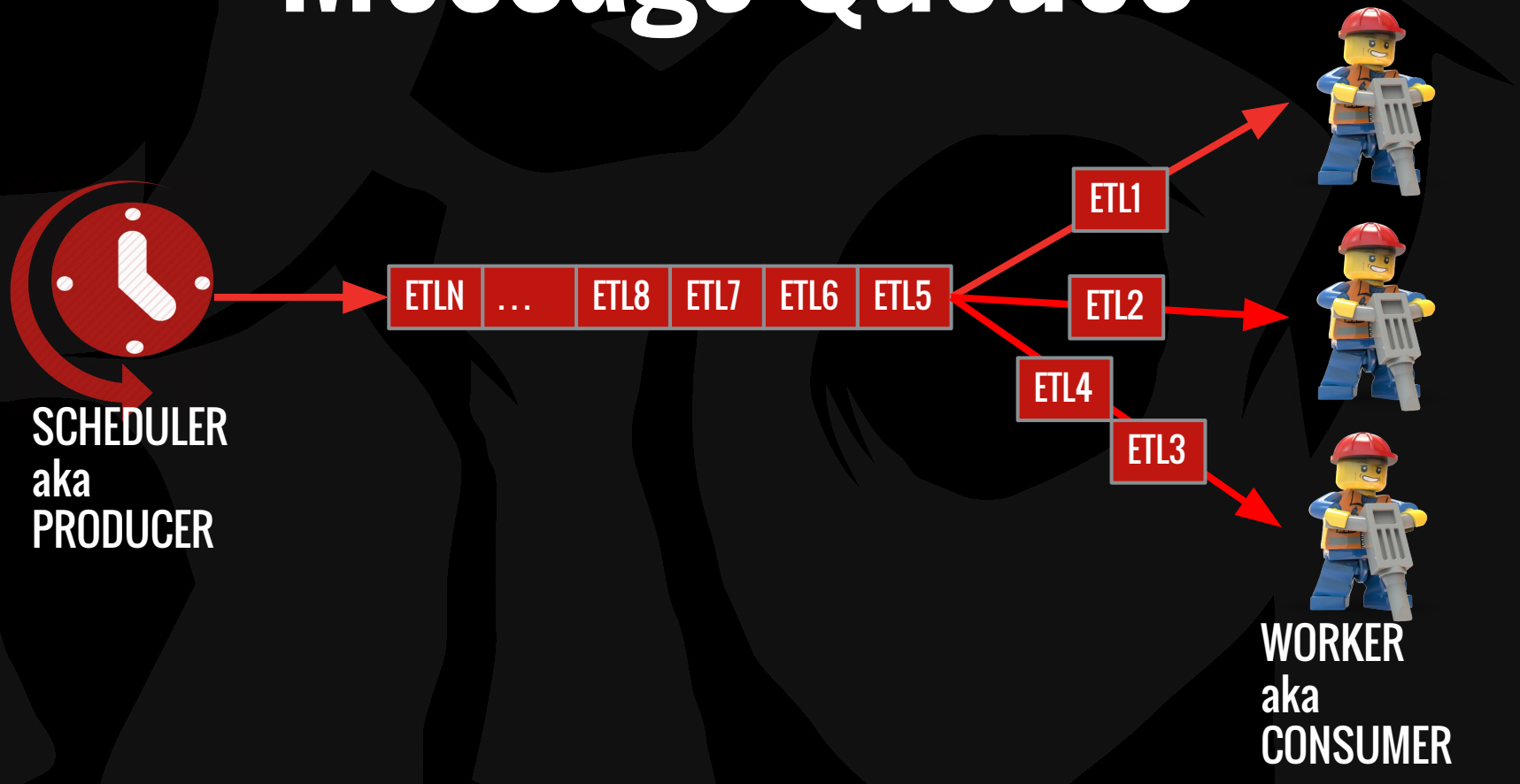
FuETL SCALING



FuETL SCALING



Message Queues



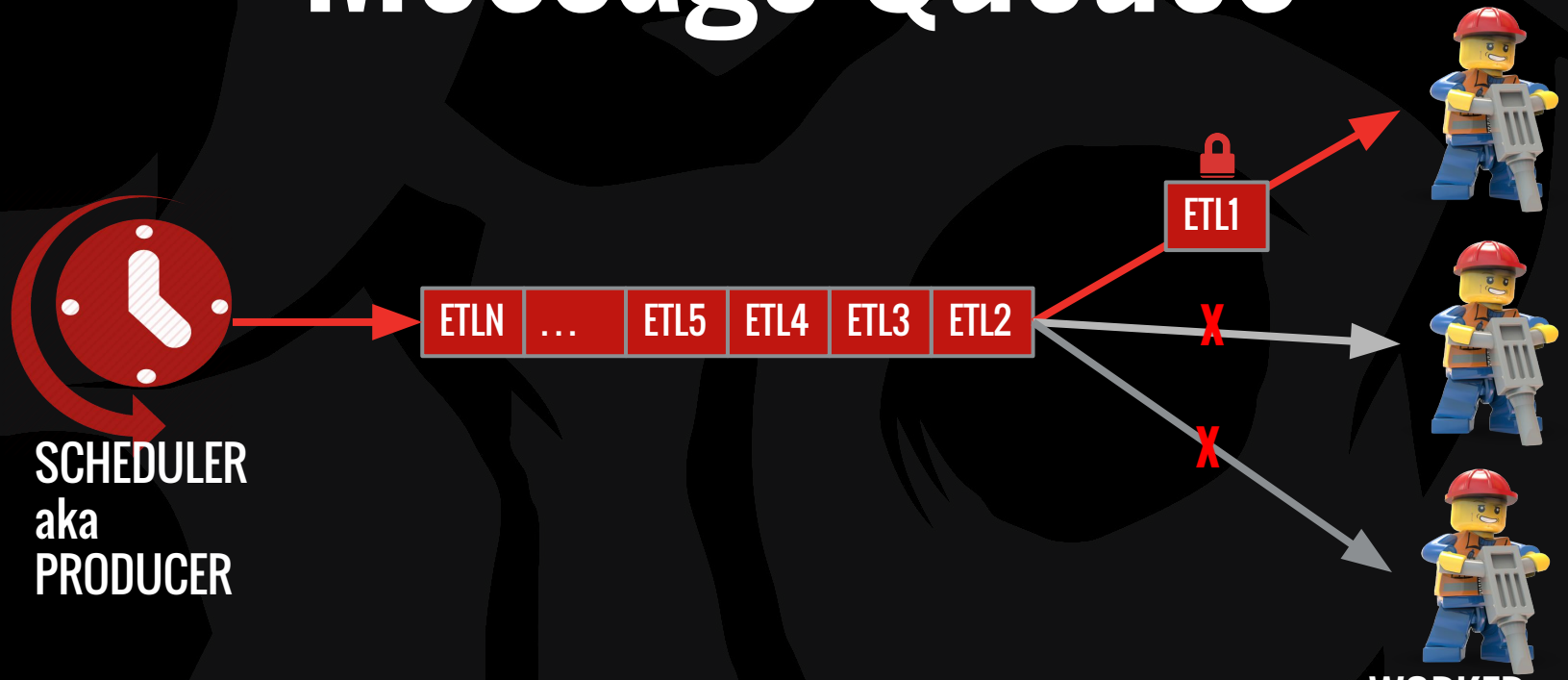
Message Queues

- Redundancy
- Delivery Guarantees
- Easy to Scale
- Asynchronous Communication
- Abstraction / Decoupling

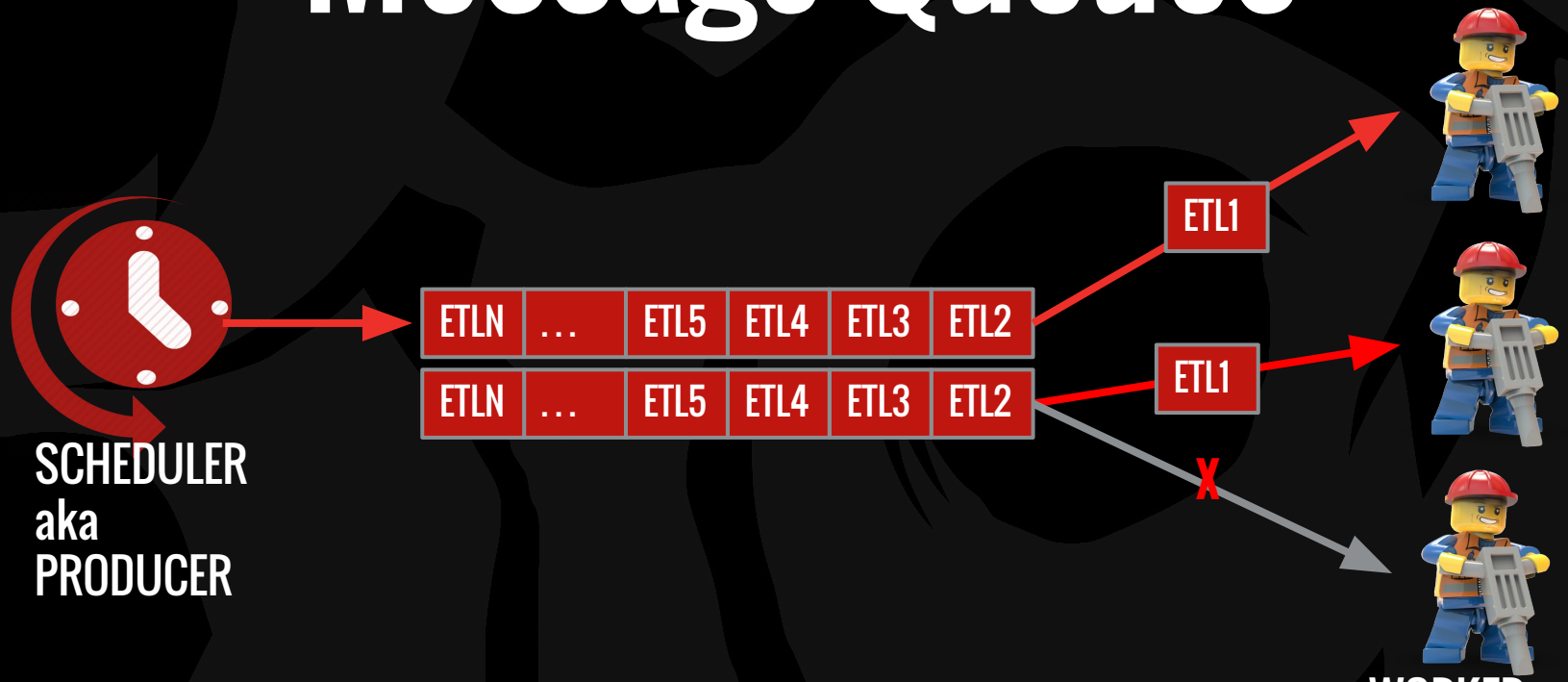
Message Queues

- Amazon Simple Queue Service
- Apache ActiveMQ
- RabbitMQ
- HornetQ
- Microsoft MQ (MSMQ)

Message Queues



Message Queues

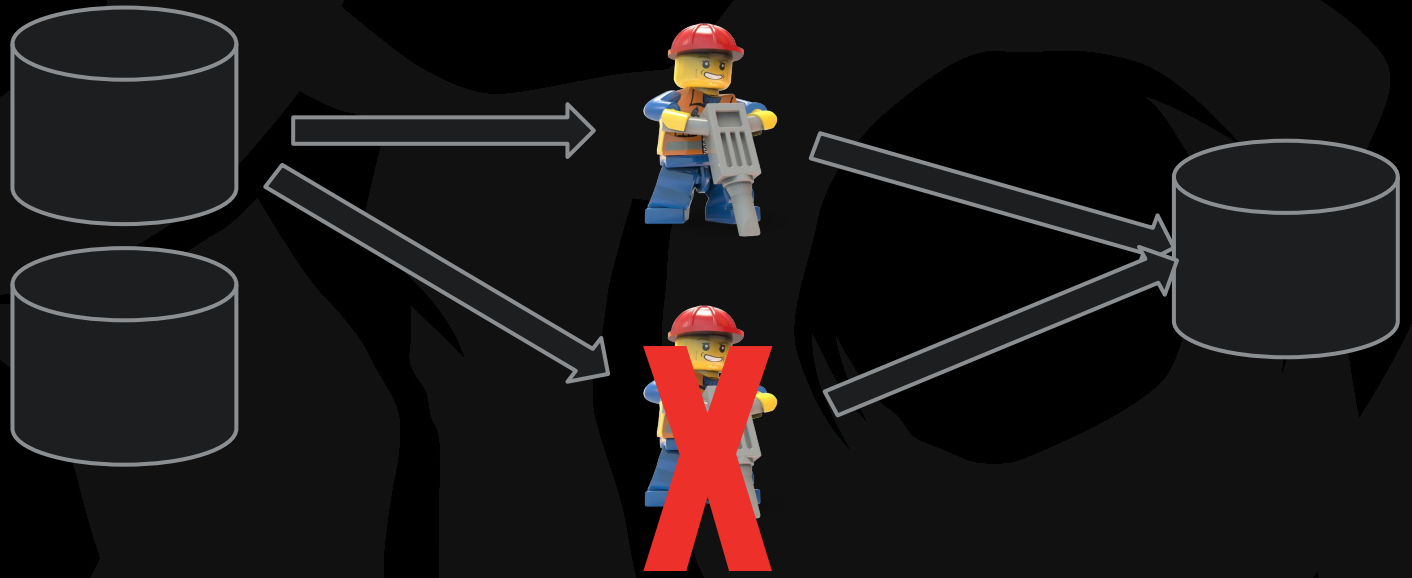


What will happen

In the big data / OLAP world….
(hint: no primary key validation)

```
INSERT INTO games_played
(SELECT * FROM games_played_na
 WHERE date >= '2015-10-25')
```

KEEPING INTEGRITY



INGESTION

PULL-BASED / ETL

FuETL
- OLTP game data
- External Data Sources

PUSH-BASED

kafka

- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE

amazon web services™
S3 Simple Storage Service

DATA AUDITING

Wasp

QUERY / VIEWS

AGGREGATE QUERIES

VERTICA
An HP Company

BATCH QUERIES

HIVE

SINGLE-ROW QUERIES

amazon DynamoDB

VIZ. TOOLS

+ableau

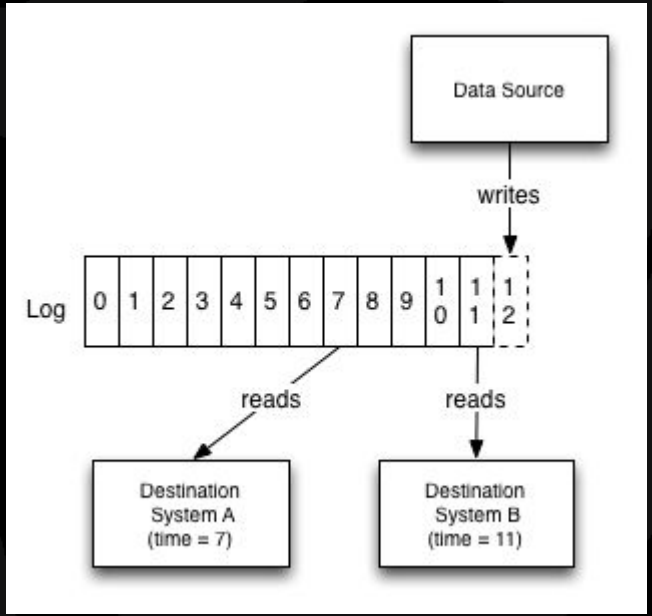
platfora
You should know.

Kafka

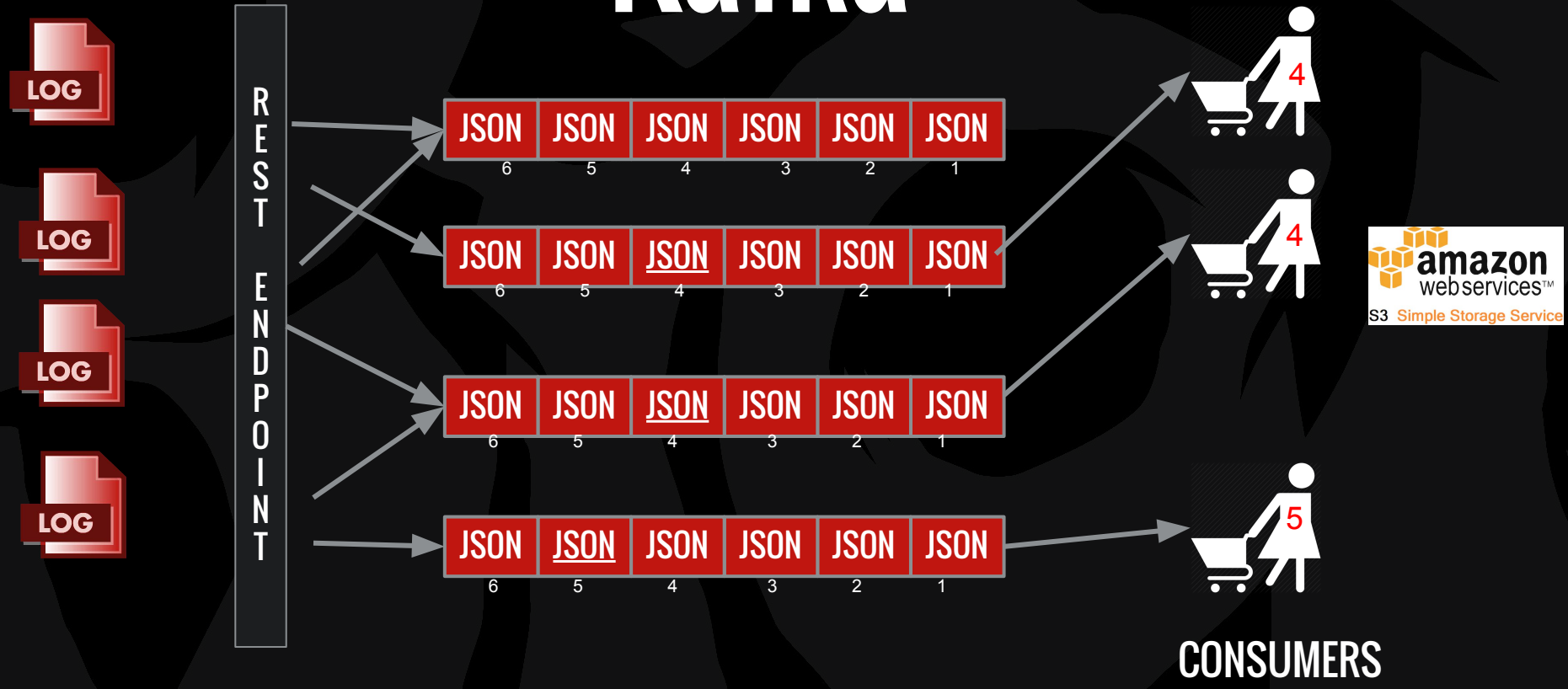
- The new hotness in big data
- Open-source project maintained by Confluent
- Very fast distributed message queue
- Data is replicated across “partitions” to ensure no loss

Kafka

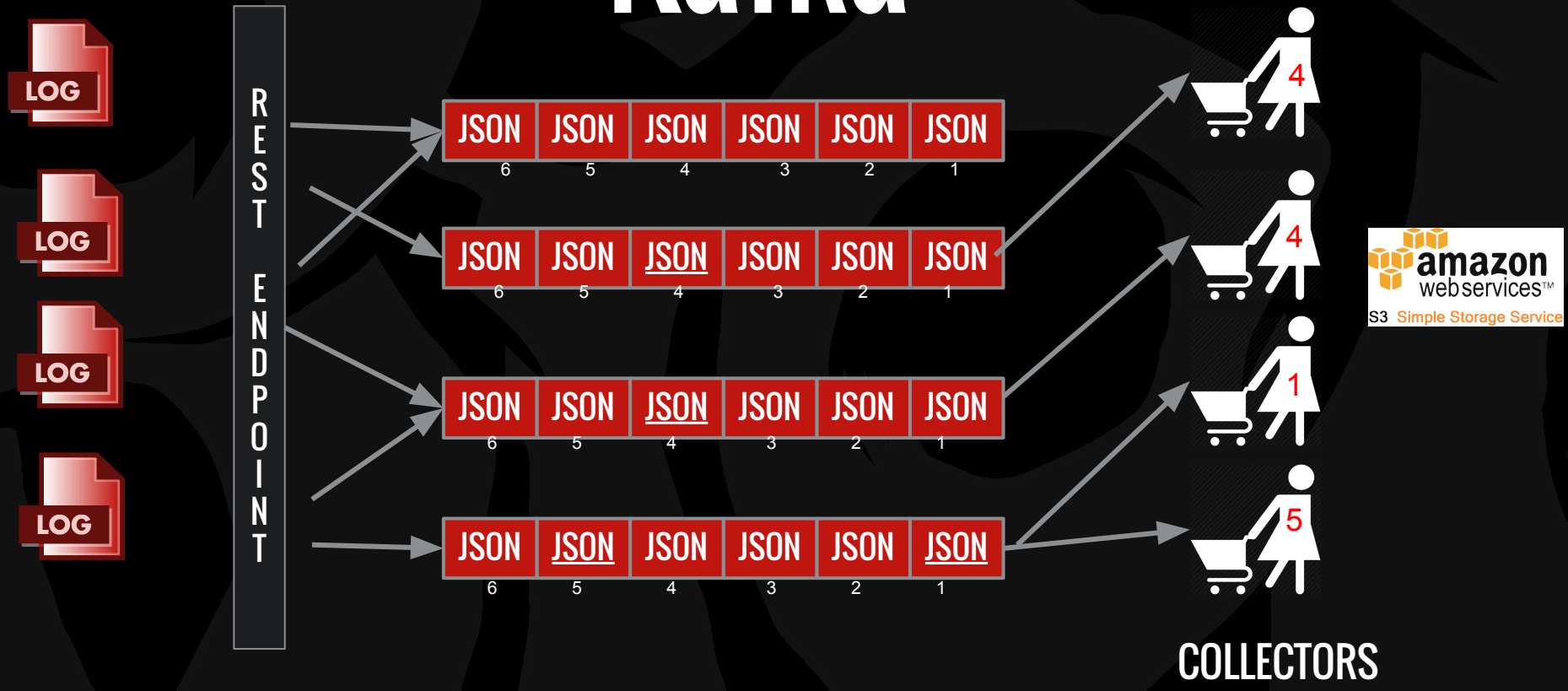
Has a DB Commit Log
(ooh revolutionary)



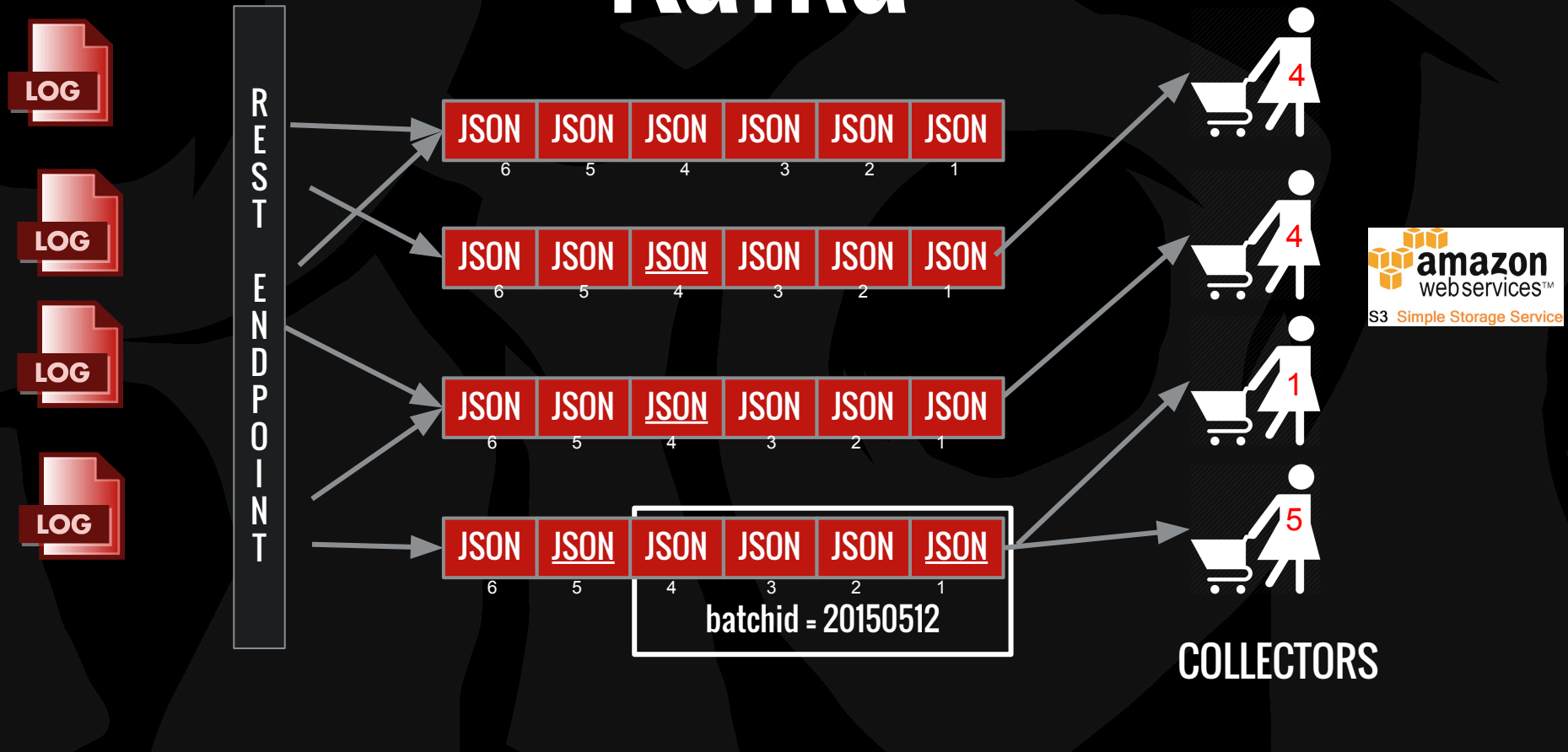
Kafka


LOG
LOG
LOG
LOG
REST ENDPOINT
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
4
4
5
CONSUMERS
amazon web services™
S3 Simple Storage Service

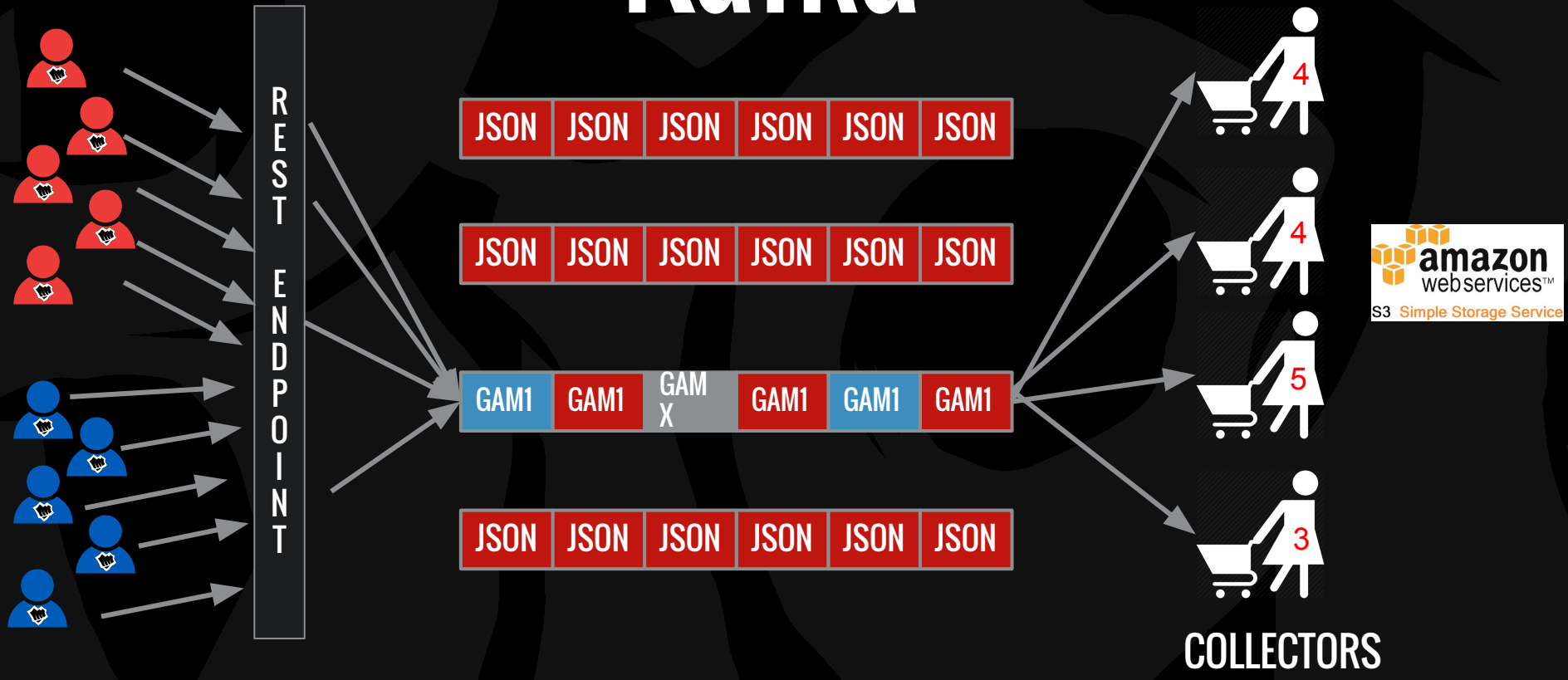
Kafka


LOG
LOG
LOG
LOG
REST ENDPOINT
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
4
4
1
5
amazon web services™
S3 Simple Storage Service
COLLECTORS

Kafka


LOG
LOG
LOG
LOG
REST ENDPOINT
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
JSON 6 JSON 5 JSON 4 JSON 3 JSON 2 JSON 1
batchid = 20150512
4
4
1
5
COLLECTORS
amazon web services™
S3 Simple Storage Service

Kafka


REST ENDPOINT
JSON JSON JSON JSON JSON JSON
JSON JSON JSON JSON JSON JSON
GAM1 GAM1 GAM X GAM1 GAM1 GAM1
JSON JSON JSON JSON JSON JSON
4
4
5
3
COLLECTORS
amazon web services™
S3 Simple Storage Service

Idempotency

Idempotent - an operation that will produce the same results if executed once or multiple times

EXAMPLE:

Non-Idempotent:
- `x = x * 5;`
- Submitting a purchase

Idempotent:
- `abs( abs(x) ) = abs(X)`
- Cancelling a purchase

Idempotent?

In the transactional OLTP world….

```
INSERT INTO games_played
(SELECT * FROM games_played_na
 WHERE date >= '2015-10-25')
```

Idempotent?

In the big data / OLAP world….

```
INSERT INTO games_played
(SELECT * FROM games_played_na
 WHERE date >= '2015-10-25')
```

Idempotency

Use application logic to make idempotent

```
msg = queue.pop;
if (processed_games.contains( msg.game_id )
{
    return; //do nothing
else {
    process_game(msg);
}
```

INGESTION

PULL-BASED / ETL

FuETL
- OLTP game data
- External Data Sources

PUSH-BASED

kafka

- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE

amazon web services™
S3 Simple Storage Service

DATA AUDITING

Wasp

QUERY / VIEWS

AGGREGATE QUERIES

VERTICA
An HP Company

BATCH QUERIES

HIVE

SINGLE-ROW QUERIES

amazon DynamoDB

VIZ. TOOLS

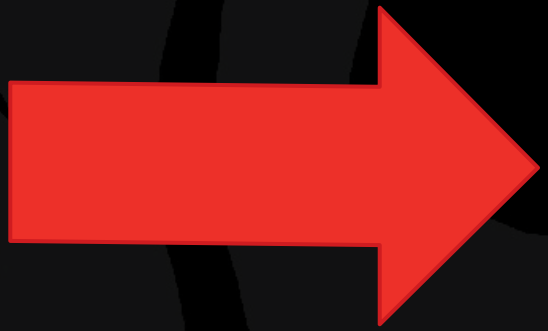
+ableau

platfora
You should know.

AMAZON S3 STRUCTURE

HIVE

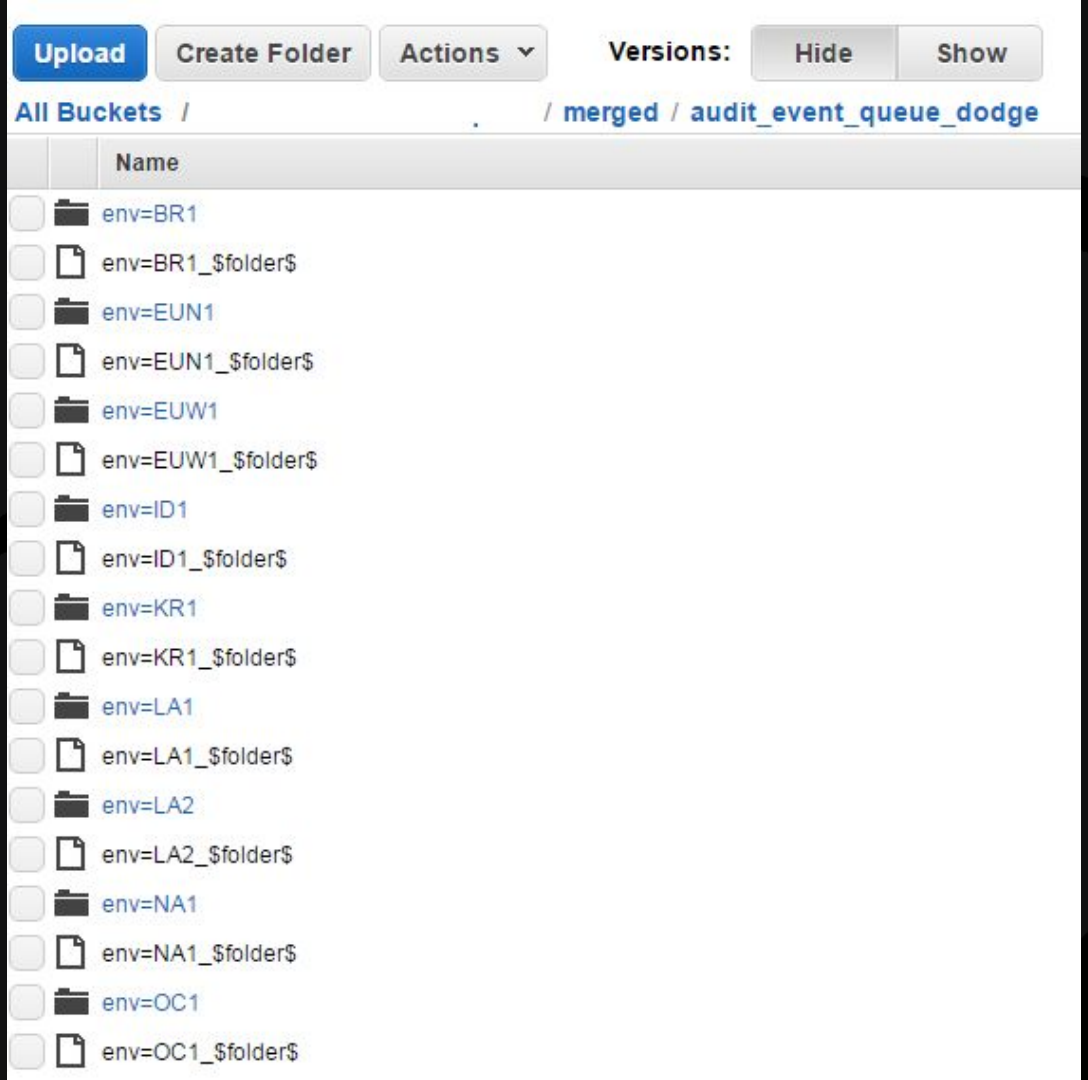
- schema1
  - table1
    - env
    - dt
    - time
  - table2
  - table3
- schema2
  - table1
  - ...
- schema3
- schema4

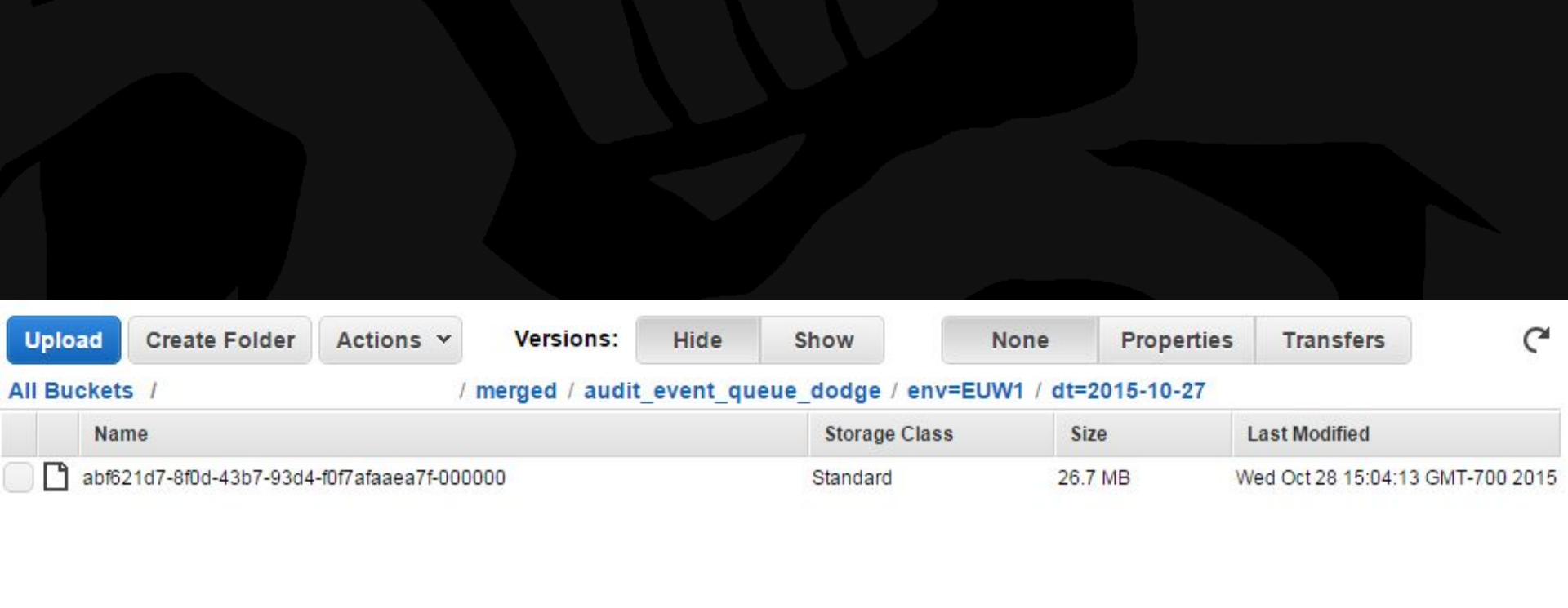


AMAZON S3

```
s3n://datawarehouse/
   schema1/
      table1/
         env/
            dt/
               time/
      table2/
      table3/
   schema2/

s3n://telemetrydata/
   application1/
      table1/
         env/
         dt/
      table2/
   application2/
```


Upload
Create Folder
Actions
Versions:
Hide
Show
All Buckets /
/ merged / audit_event_queue_dodge
Name
env=BR1
env=BR1_$folder$
env=EUN1
env=EUN1_$folder$
env=EUW1
env=EUW1_$folder$
env=ID1
env=ID1_$folder$
env=KR1
env=KR1_$folder$
env=LA1
env=LA1_$folder$
env=LA2
env=LA2_$folder$
env=NA1
env=NA1_$folder$
env=OC1
env=OC1_$folder$


Upload
Create Folder
Actions
Versions:
Hide
Show
None
Properties
Transfers
All Buckets /
/ merged / audit_event_queue_dodge / env=EUW1 / dt=2015-10-27
Name
Storage Class
Size
Last Modified
abf621d7-8f0d-43b7-93d4-f0f7afaaea7f-000000
Standard
26.7 MB
Wed Oct 28 15:04:13 GMT-700 2015

INGESTION

PULL-BASED / ETL

FuETL
- OLTP game data
- External Data Sources

PUSH-BASED

kafka

- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE

amazon web services™
S3 Simple Storage Service

DATA AUDITING

Wasp

QUERY / VIEWS

AGGREGATE QUERIES

VERTICA
An HP Company

BATCH QUERIES

HIVE

SINGLE-ROW QUERIES

amazon DynamoDB

VIZ. TOOLS

+ableau

platfora
You should know.

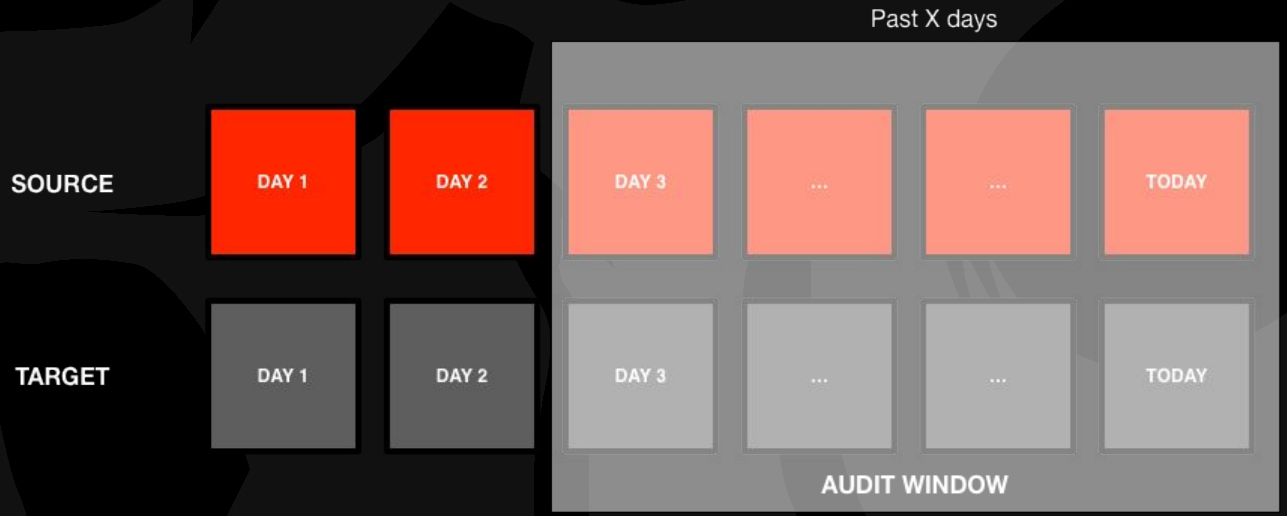
Warehouse
Auditing
Service
Platform

REST micro-service built with Java and docker.

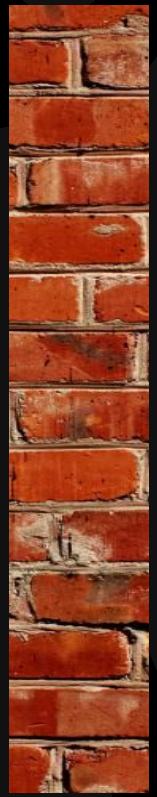
Source and target comparison.

Reports and visualizations we can use to find problems.

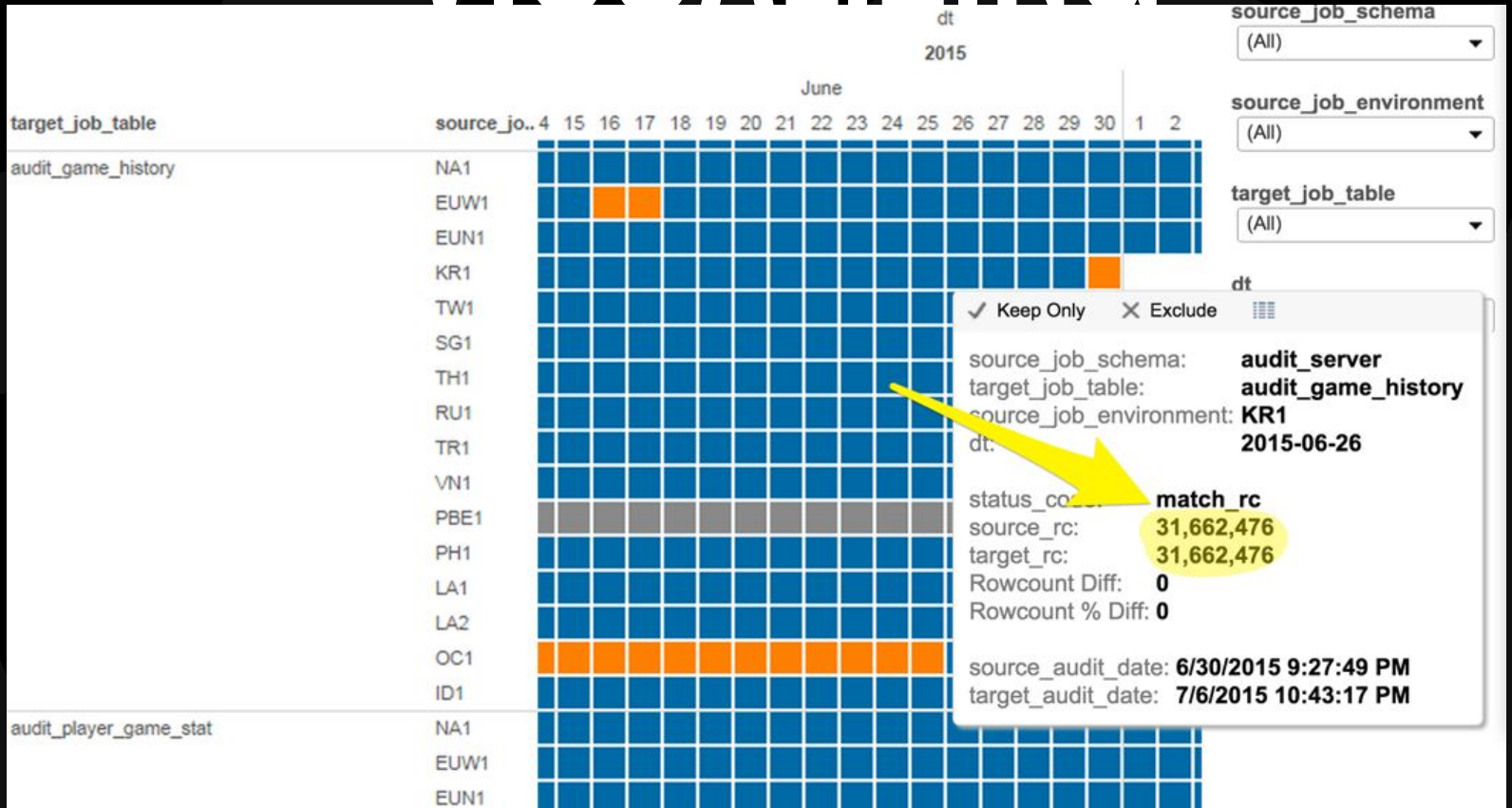
HOW TO AUDIT


Past X days
SOURCE
DAY 1
DAY 2
DAY 3
...
...
TODAY
TARGET
DAY 1
DAY 2
DAY 3
...
...
TODAY
AUDIT WINDOW

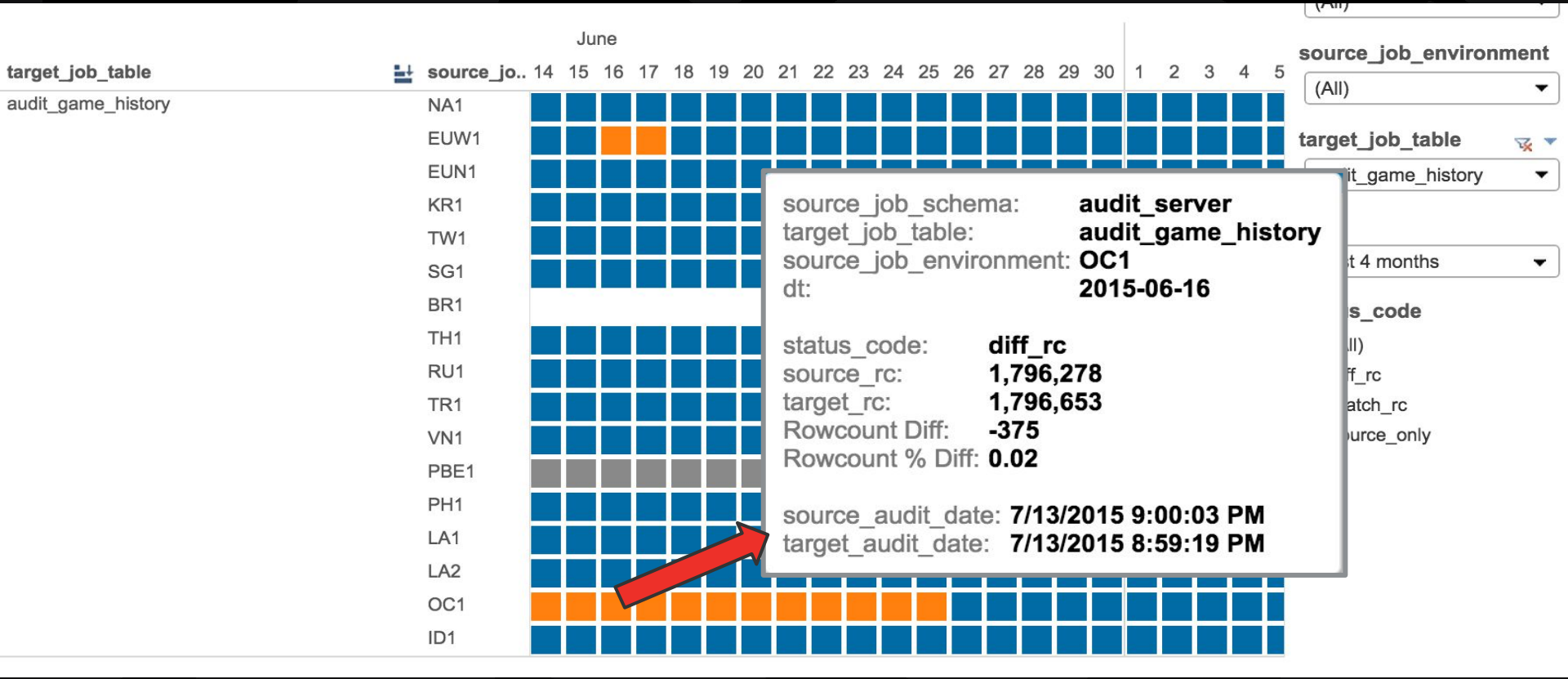
HOW TO AUDIT



VISUALIZING



VISUALIZING

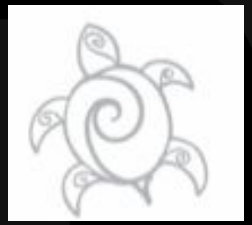


INGESTION

PULL-BASED / ETL

FuETL
- OLTP game data
- External Data Sources

PUSH-BASED



HONU
- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE

amazon web services™
S3 Simple Storage Service

DATA AUDITING

Wasp

QUERY / VIEWS

AGGREGATE QUERIES

VERTICA
An HP Company

BATCH QUERIES

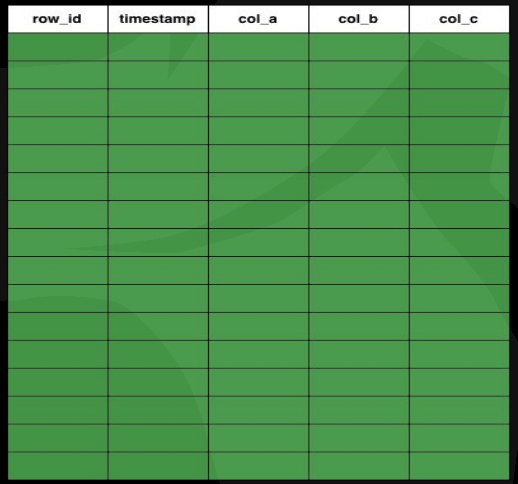
SINGLE-ROW QUERIES

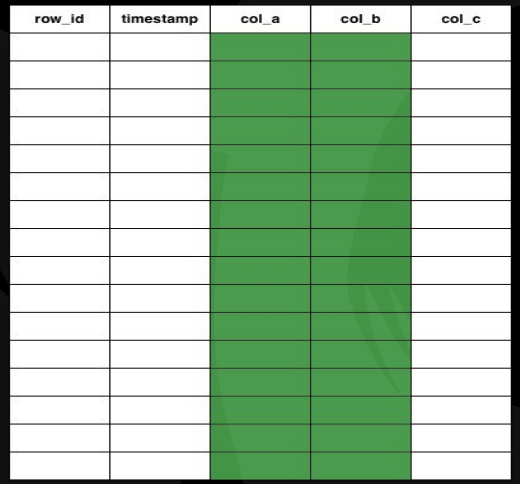
amazon DynamoDB

VIZ. TOOLS

+ableau

platfora
You should know.





BATCH

OLAP

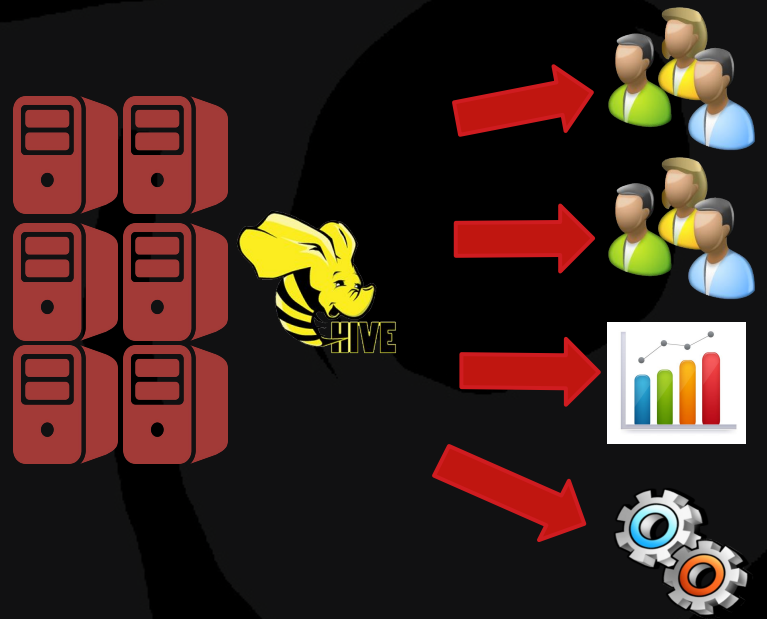
POINT

SCALING IN AWS

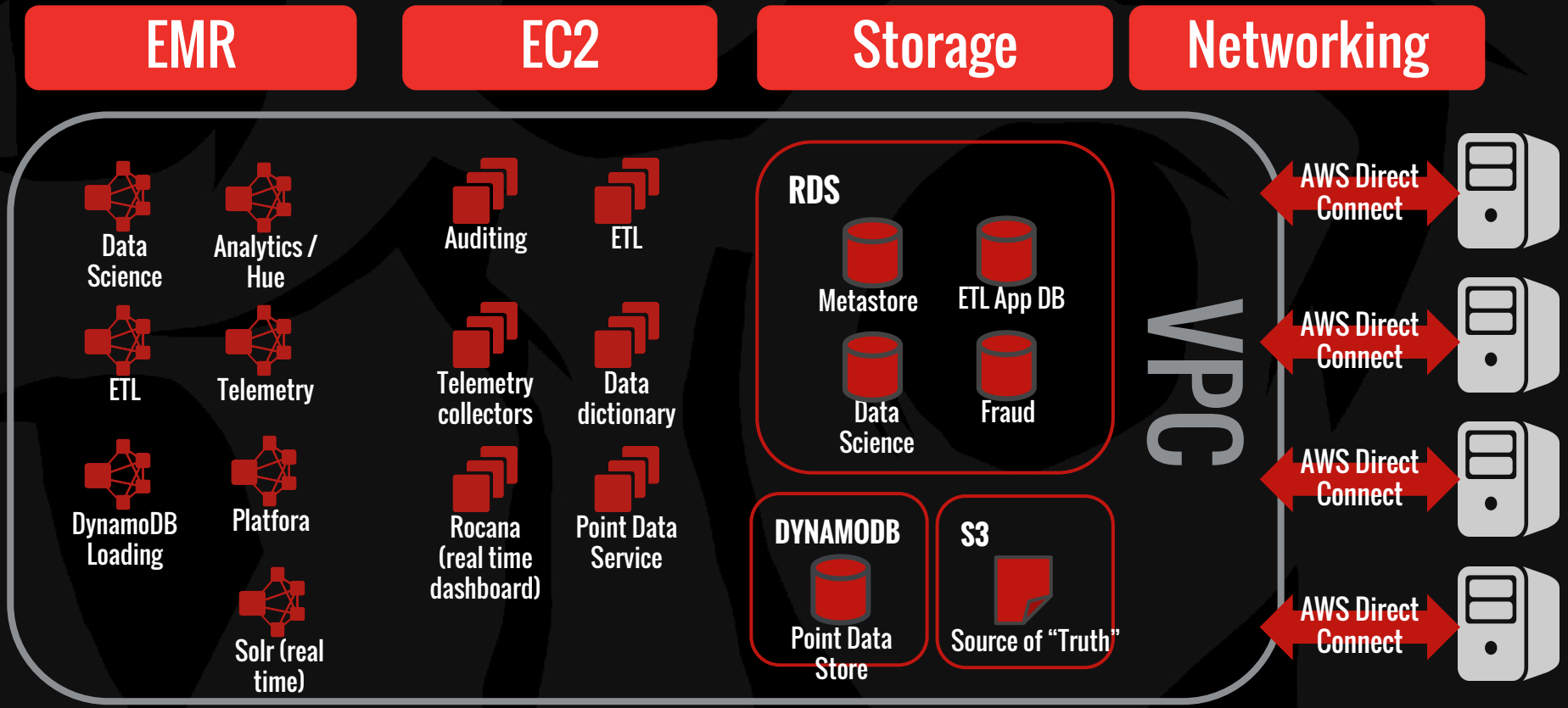
SCALING

RESOURCE CONTENTION

Hive .08 pre YARN, immature resource scheduling



AWS Infrastructure Today



CONCLUSION

SEAN'S PRO TIPS OF THE DAY

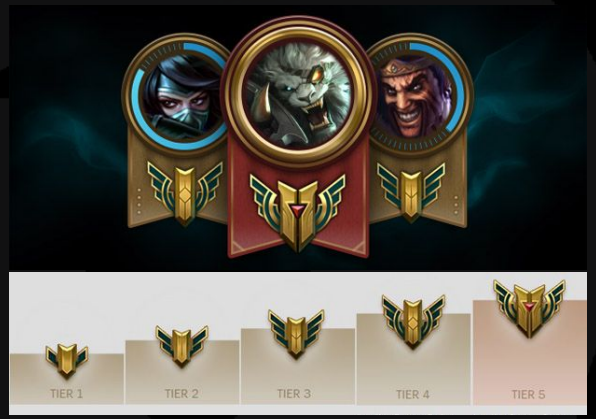
DO

- Keep idempotency in mind and use MQ architecture
- Get an auditing solution for DW accuracy
- Prepare for multiple data access patterns
- Allocate time for tuning AWS infrastructure

DON'T

- Don't underestimate simple problems in big data.
- Don't forget to track cost. AWS bills can surprise you
- Don't wait. Create S3 permissions and naming standards early
- Don't stop. Believing

CHAMPION MASTERY

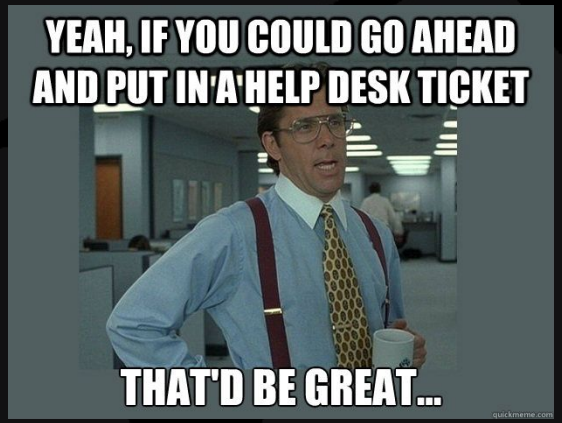


Custom rewards for mastering different champions

Intensive query that spans every game that every player has played

Improves player engagement

PLAYER SUPPORT



Full copy of our data warehouse in DynamoDB

Hive->DynamoDB Dynamic Partition

Support can answer questions faster than ever.

OFFENSIVE CHAT DETECTION

Data science team queries all chat messages in game

Sentiment analysis and classification

Identifies negative, offensive players and mutes them automatically.