

MOVING MOUNTAINS OF PLAYER DATA

SCALABLE INTERNET SERVICES
UCLA/UCSB - NOV 2017

SEAN MALONEY
RIOT GAMES



THAT IS BALONEY



@SEAN_SEANNERY



SMALONEY@riotgames.com

WHO IS THIS GUY?

Team Lead on Riot's big data and real-time services

FUN FACT:

Was a student in this class 6 years ago
Intern at Appfolio



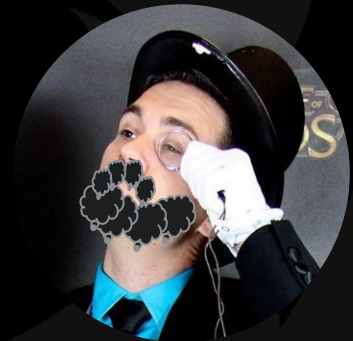
SEAN MALONEY
BIG DATA ENGINEER

WHO IS THIS GUY?

Team Lead on Riot's big data and real-time services

FUN FACT:

Was a student in this class 6 years ago
Intern at Appfolio



SEAN MALONEY
BIG DATA ENGINEER

MOVING **MOUNTAINS** OF DATA

1. INTRODUCTION

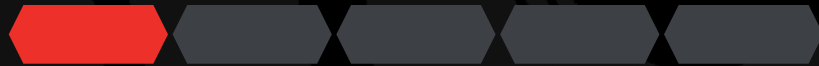
2. THE GAME PLATFORM: OUR MAIN DATA SOURCE

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?

2009
LAUNCH

ONLINE
MULTIPLAYER

WINDOWS
/ OSX

40-50 MIN
GAMES



YOUR CHAMP



THE TEAM



THE BATTLE GROUND



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

OFFENSIVE CHAT DETECTION

Data science team queries all chat messages in game

Sentiment analysis and classification

Jônas has ended ezesa2396's killing spree! (Bounty: 500G)

Jônas (Master Yi): sry

ShadowMaster3000 (Vi): 本本本 your mother 本本本 yi you 本本本 you noob

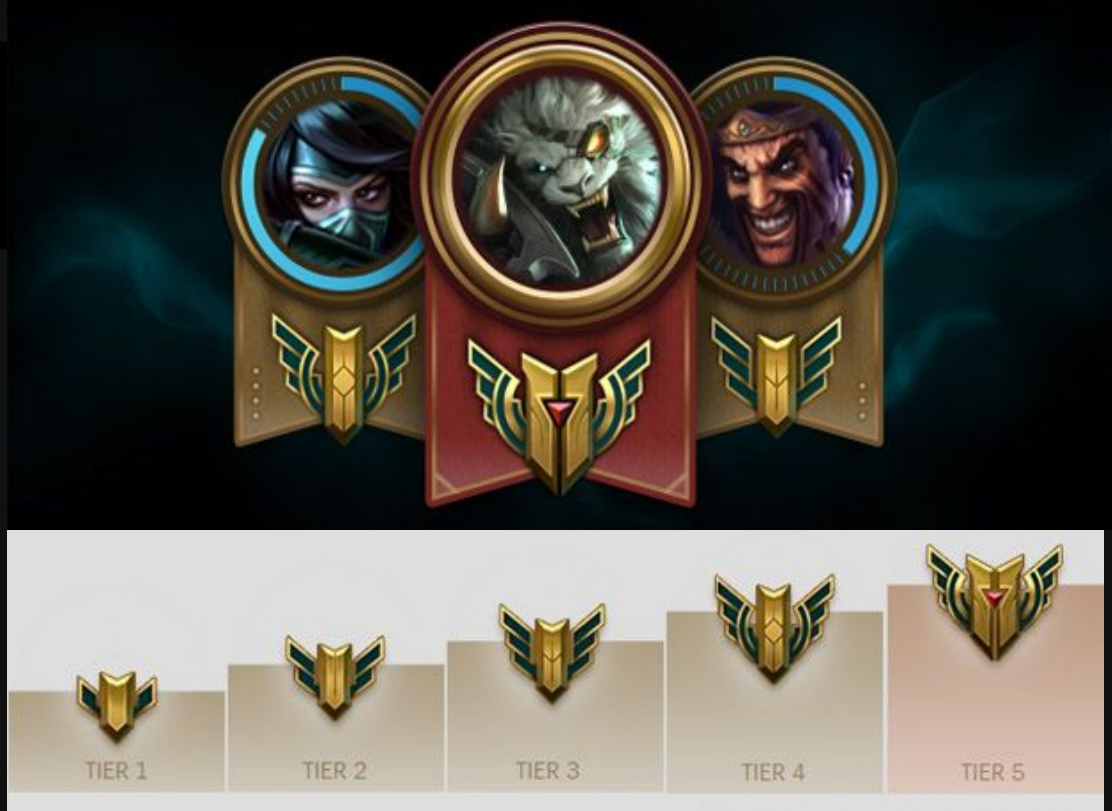
ShadowMaster3000 (Vi): i die and i make ulti and fier and YOPU KILL

Jônas (Master Yi): :DDDDDD

ShadowMaster3000 (Vi): i report you 本本本本



CHAMPION MASTERY



PLAYER SUPPORT

Whoops! Something broke.



POWERED BY
BUGSPLOT

A problem has been encountered and the program needs to close.

Reporting this error will help us make our product better. All information is treated as confidential and is only used only to improve future versions of this program.

Please describe what you were doing before the crash (optional):

Share your contact information with us below if you're okay with us contacting you for more information about this error.

Name: (optional) Email Address: (optional)

[Send Error Report](#) [Don't Send](#) [View Report Details...](#)

**YEAH, IF YOU COULD GO AHEAD
AND PUT IN A HELP DESK TICKET**



THAT'D BE GREAT...

GAME BALANCE

FIRST BLOOD RATE BY CHAMPION

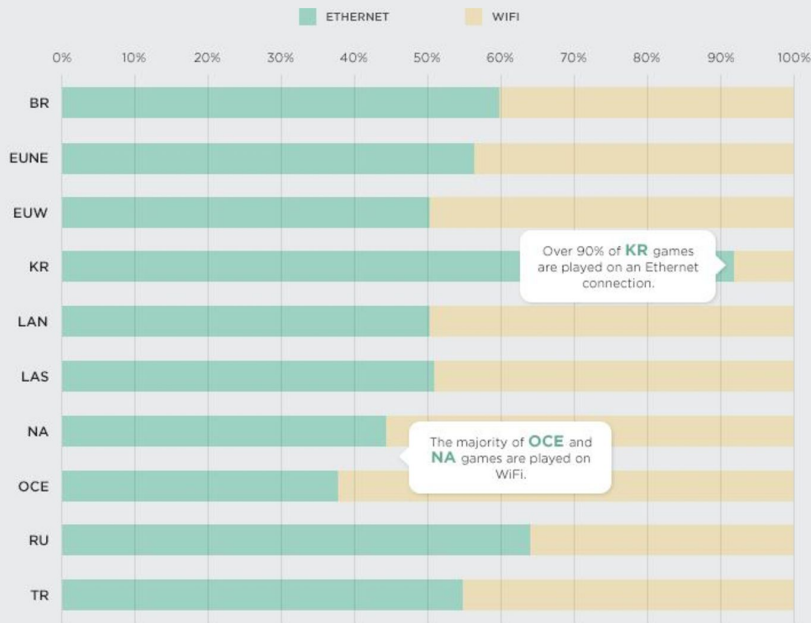
RANK	CHAMPION	FIRST BLOOD RATE
1	TALON	24.9%
2	PANTHEON	20.2%
3	KATARINA	19.8%
4	EVELYNN	19.5%
5	LEBLANC	18.6%
6	LEE SIN	17.9%
7	TRYNDAMERE	17.6%

POPULARITY OF KEYBINDINGS FOR FLASH, WORLDWIDE BY PERCENTAGE OF RECORDED GAMES



LATENCY AND NETWORK

PERCENT OF GAMES PLAYED BY CONNECTION TYPE
BY REGION



Internet Service Provider Leaderboard?

Great Good OK Poor

SORT BY: SCORE % OF PLAYERS

COMPARE PROVIDERS IN ANN ARBOR, MICHIGAN, UNITED STATES

Synergy Broadband

3% of players in this area

99

Univ. of Michigan

19% of players in this area

98

AT&T U-verse

8% of players in this area

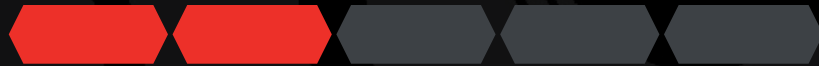
97

Comcast Cable

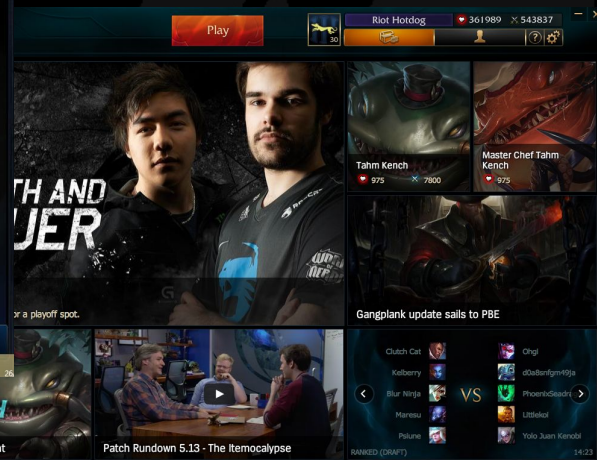
58% of players in this area

96

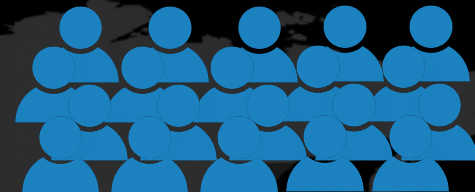
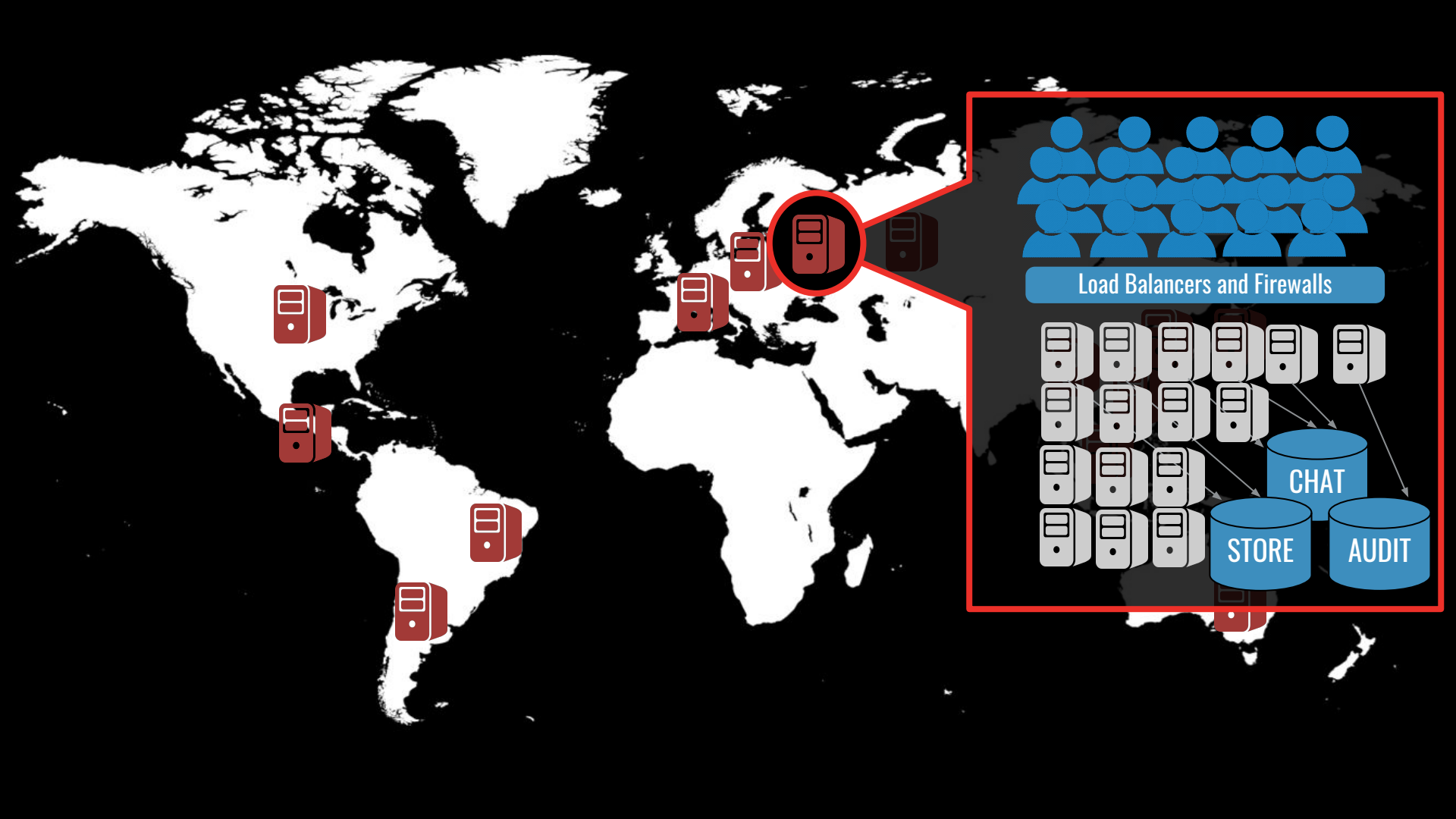
THE GAME PLATFORM



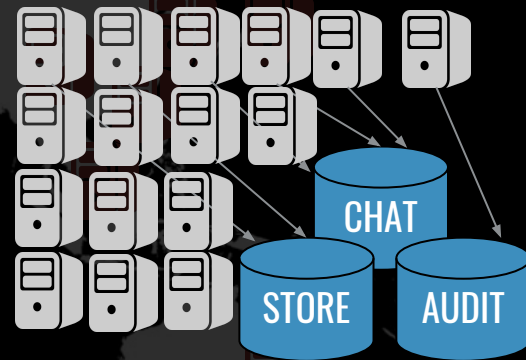
THE CLIENT.

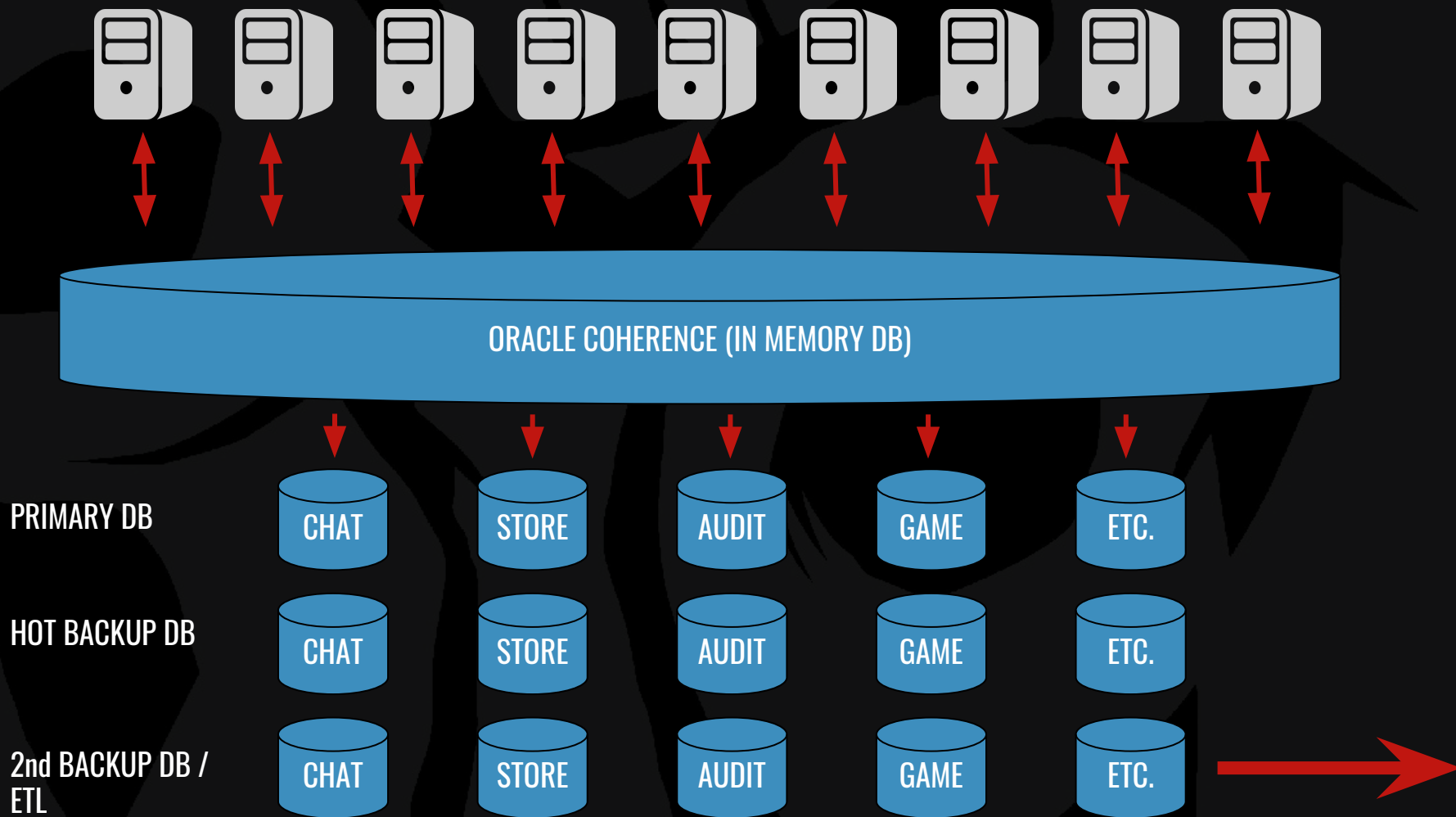


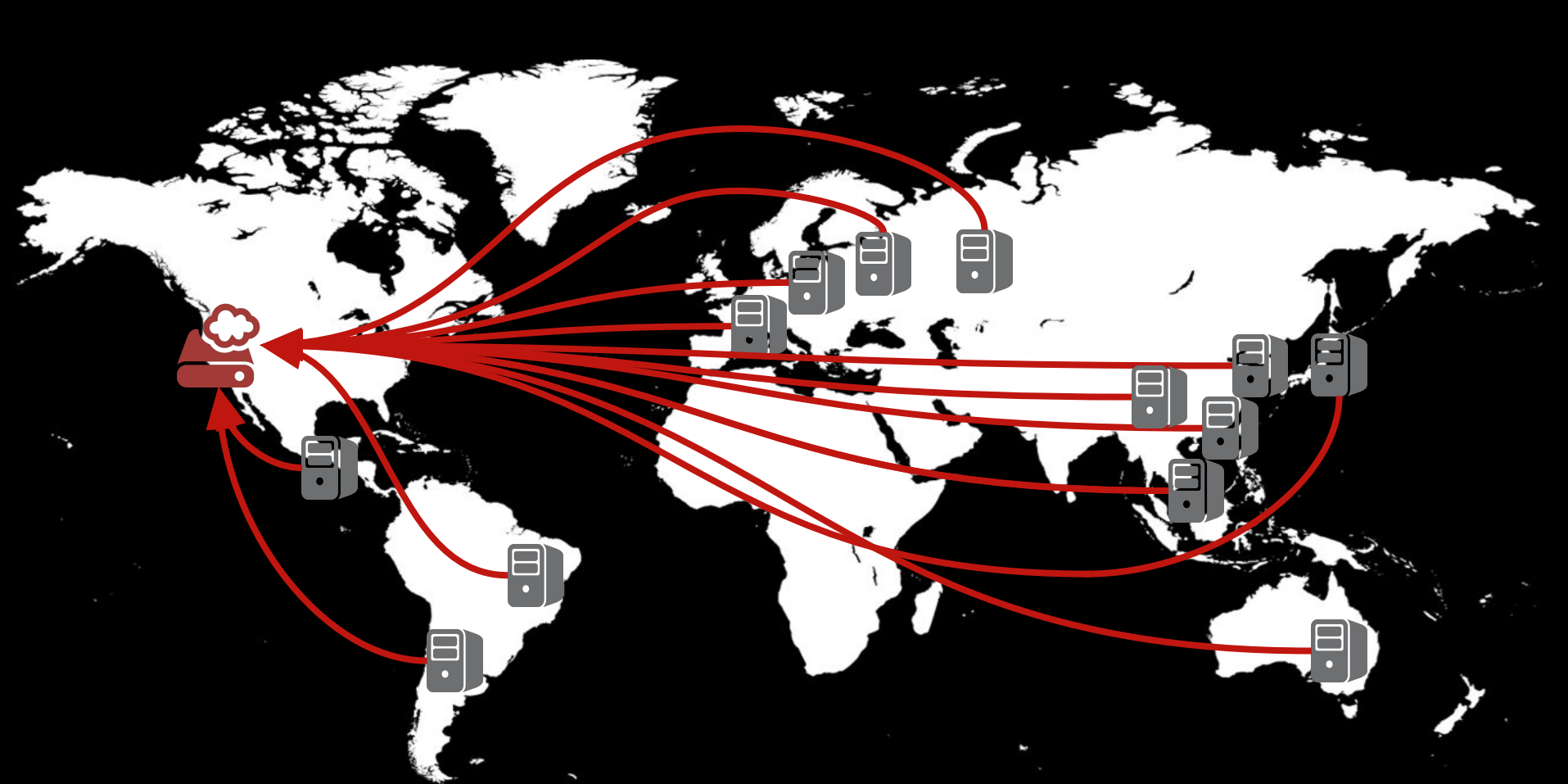




Load Balancers and Firewalls







DATA INGESTION



INGESTION

PULL-BASED / ETL



FuETL

- OLTP game data
- External Data Sources

PUSH-BASED



- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE



DATA AUDITING



QUERY / VIEWS

AGGREGATE QUERIES



BATCH QUERIES



SINGLE-ROW QUERIES



VIZ. TOOLS



INGESTION

PULL-BASED / ETL



FuETL

- OLTP game data
- External Data Sources

PUSH-BASED



- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE



DATA AUDITING



QUERY / VIEWS

AGGREGATE QUERIES



BATCH QUERIES



SINGLE-ROW QUERIES



VIZ. TOOLS





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

Source Helper

Source Table(s)
Date_column: create_date
Hour_column:
Realm_column: region

Target Helper

Target Table

Target Realm Column

Target Date Column

Query

Create an ETL

Source Helper: mysql_fantasylcs_server

Source Table(s):
Table: fantasy_users
Date_column: create_date
Hour_column:
Realm_column: region
Add

Target Helper: vertica_test_cluster

Target Table: warehouse.fantasy_users

Target Realm Column: dt

Target Date Column: region

Query

```
select
  <%= environment_id %>
  , segmentation_date
  , year(segmentation_date)
  , month(segmentation_date)
  , acct_id
  , game_count
from fake_db.fantasy_users
where env = '<%= environment_name %>'
and segmentation_date = '<%= start_date %>'
and acct_id is not null
```

Create an ETL

Source Helper: mysql_fantasylcs_server

Source Table(s):
Table: fantasy_users
Date_column: create_date
Hour_column:
Realm_column: region
Add

Target Helper: vertica_test_cluster

Target Table: warehouse.fantasy_users

Target Realm Column: dt

Target Date Column: region

Query

```
select
  <%= environment_id %>
  , segmentation_date
  , year(segmentation_date)
  , month(segmentation_date)
  , acct_id
  , game_count
from fake_db.fantasy_users
where env = '<%= environment_name %>'
and segmentation_date = '<%= start_date %>'
and acct_id is not null
```

mysql_to_vertica/store_items (SQLToSQL)

Task Config

Schedule or Audit Runs

2

Environments: 16 selected ▾

Q Search

<input type="checkbox"/> No environment	<input checked="" type="checkbox"/> BR1	<input type="checkbox"/> CN1	<input type="checkbox"/> EDU1
<input checked="" type="checkbox"/> EUN1	<input checked="" type="checkbox"/> EUW1	<input type="checkbox"/> GLB	<input type="checkbox"/> HN1
<input type="checkbox"/> HN10	<input type="checkbox"/> HN11	<input type="checkbox"/> HN12	<input type="checkbox"/> HN13
<input type="checkbox"/> HN14	<input type="checkbox"/> HN15	<input type="checkbox"/> HN16	<input type="checkbox"/> HN17
<input type="checkbox"/> HN18	<input type="checkbox"/> HN19	<input type="checkbox"/> HN2	<input type="checkbox"/> HN20
<input type="checkbox"/> HN3	<input type="checkbox"/> HN4	<input type="checkbox"/> HN5	<input type="checkbox"/> HN6
<input type="checkbox"/> HN7	<input type="checkbox"/> HN8	<input type="checkbox"/> HN9	<input type="checkbox"/> ID1
<input checked="" type="checkbox"/> KR1	<input checked="" type="checkbox"/> LA1	<input checked="" type="checkbox"/> LA2	<input checked="" type="checkbox"/> NA1
<input checked="" type="checkbox"/> OC1	<input checked="" type="checkbox"/> PBE1	<input checked="" type="checkbox"/> PH1	<input checked="" type="checkbox"/> RU1
<input checked="" type="checkbox"/> SG1	<input checked="" type="checkbox"/> TH1	<input checked="" type="checkbox"/> TR1	<input type="checkbox"/> TREU
<input type="checkbox"/> TRKR	<input type="checkbox"/> TRNA	<input type="checkbox"/> TRSA	<input type="checkbox"/> TRTW
<input checked="" type="checkbox"/> TW1	<input checked="" type="checkbox"/> VN1	<input type="checkbox"/> WT1	<input type="checkbox"/> WT2
<input type="checkbox"/> WT3	<input type="checkbox"/> WT4	<input type="checkbox"/> WT5	<input type="checkbox"/> WT6
<input type="checkbox"/> WT7			

mysql_to_vertica/store_items (SQLToSQL)

[Task Config](#)[Schedule or Audit Runs](#)

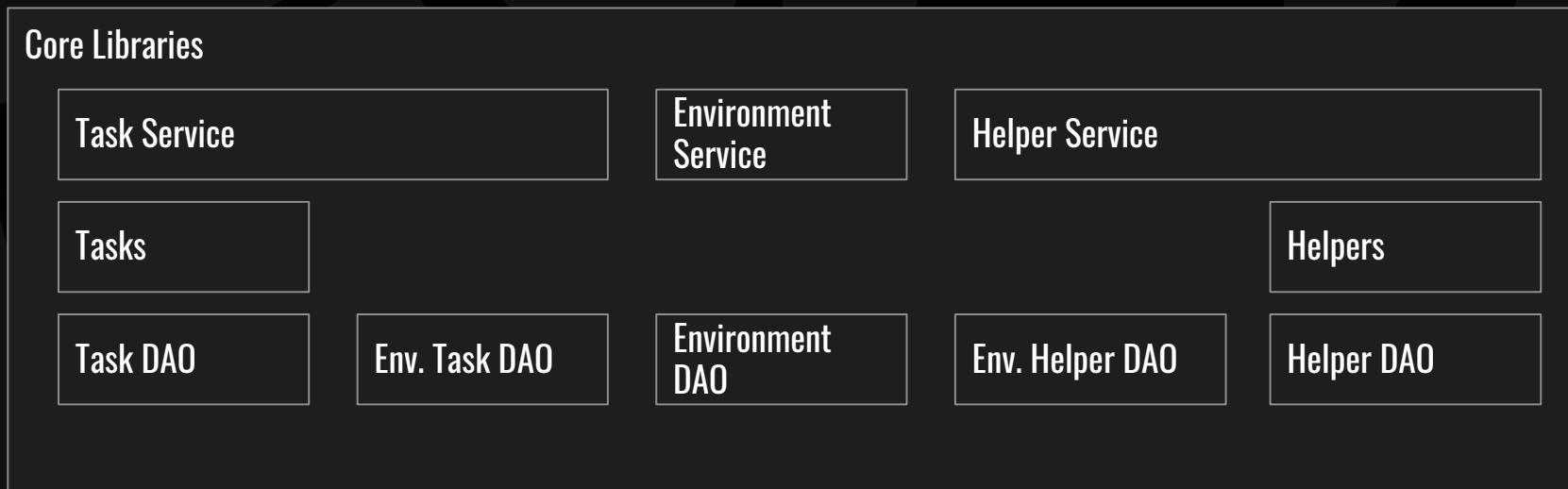
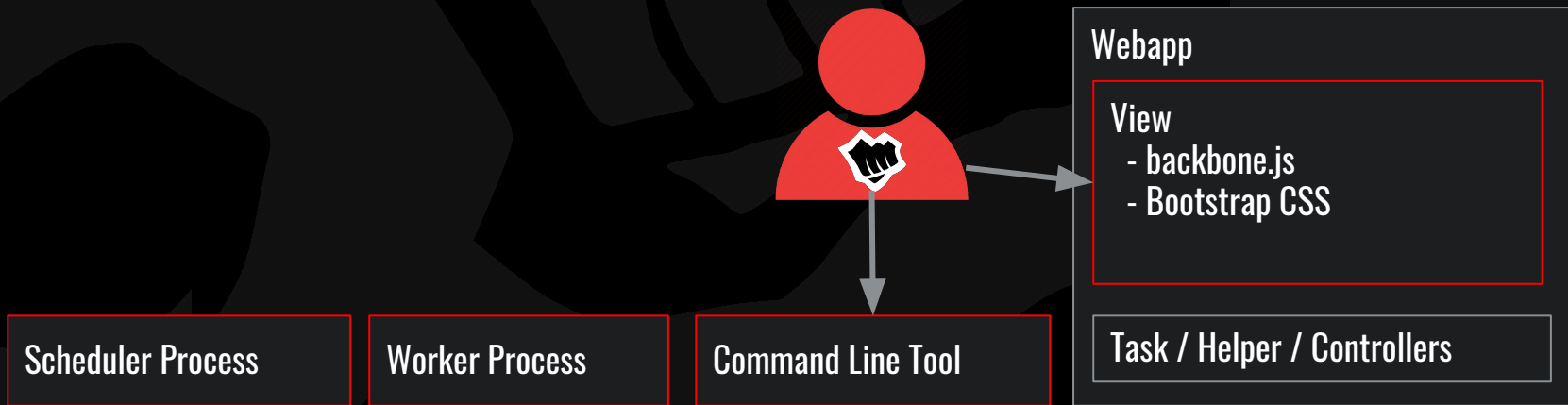
9

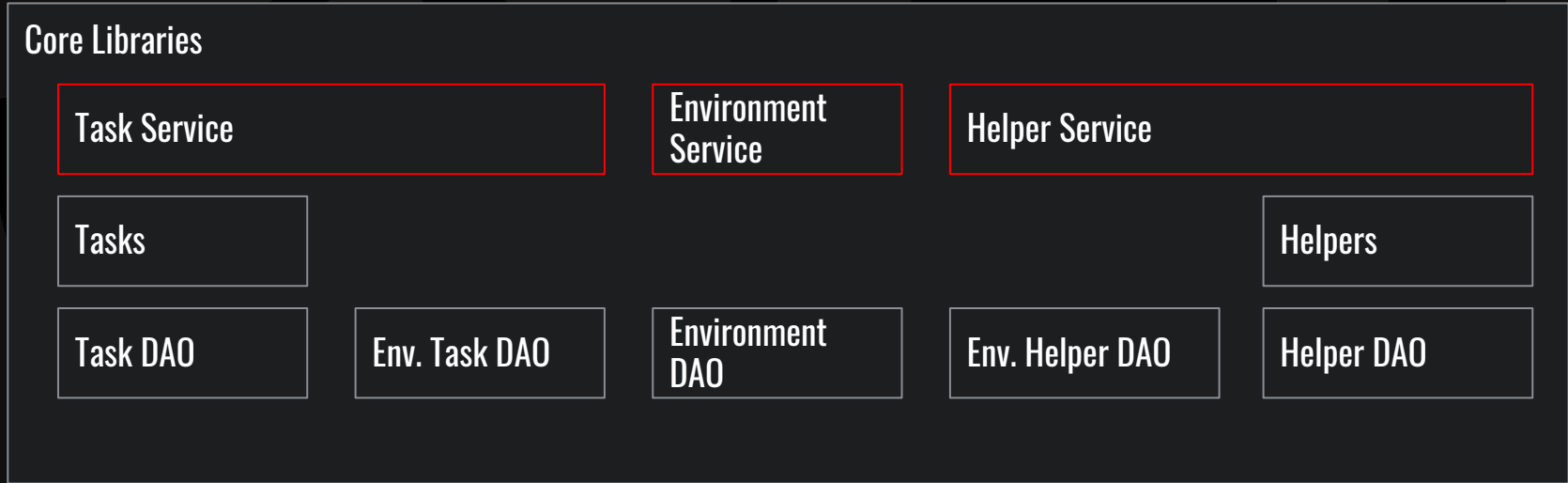
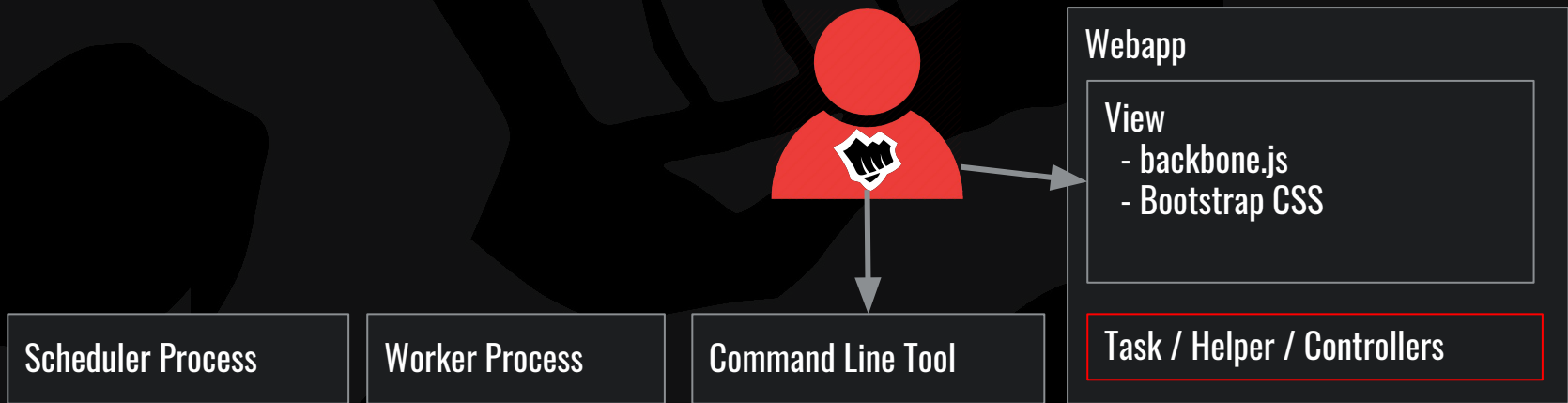
Environments: 16 selected ▾

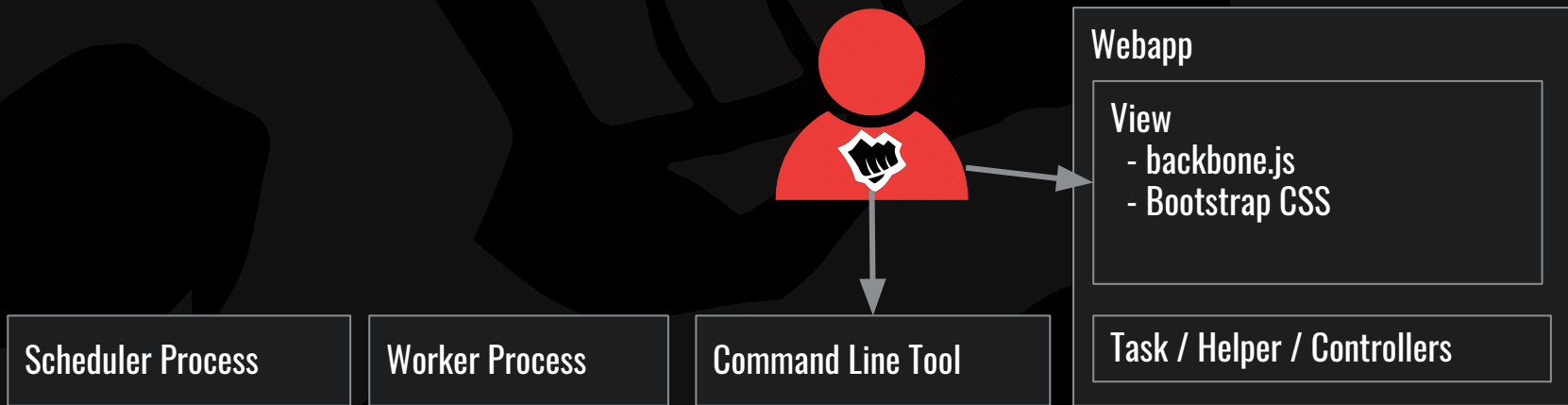
BR1 (9)	success	at 7/9/2015, 4:00:44 PM	Will run again at approximately: 7/9/2015, 10:00:44 PM	↑	Elapsed: 0:00:10 Median: 0:00:08 Longest: 0:00:33
EUN1 (3)	success	at 7/9/2015, 4:15:40 PM	Will run again at approximately: 7/9/2015, 10:15:40 PM	↑	Elapsed: 0:00:05 Median: 0:00:05 Longest: 0:00:39
EUW1 (2)	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







Core Libraries

Task Service

Environment
Service

Helper Service

Tasks

Helpers

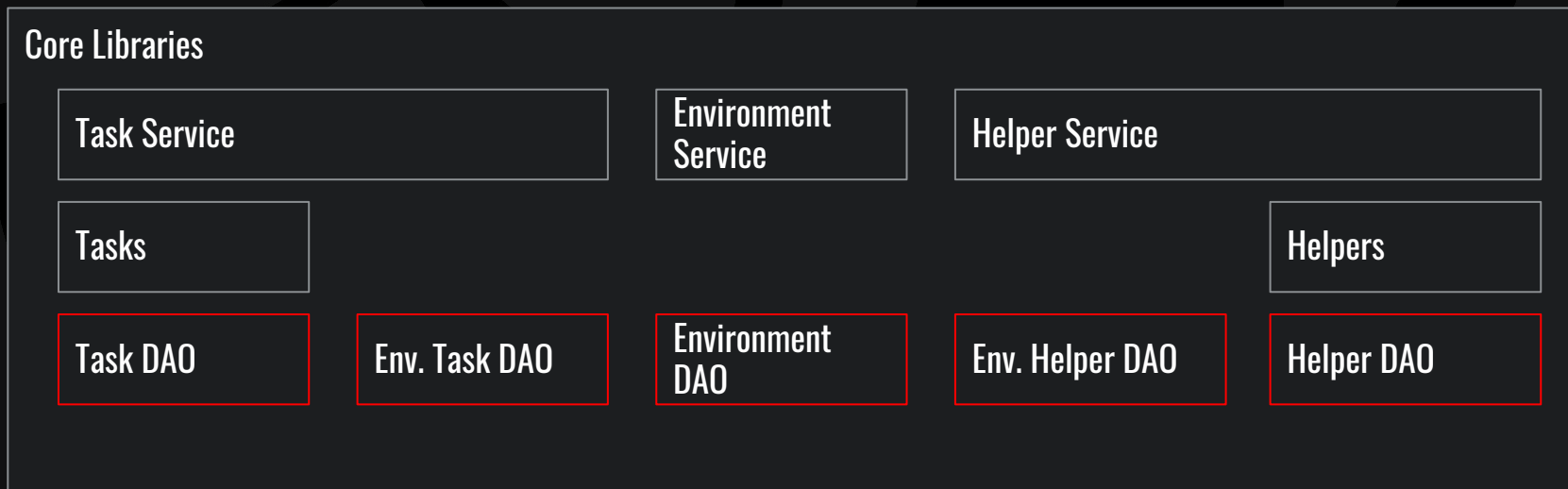
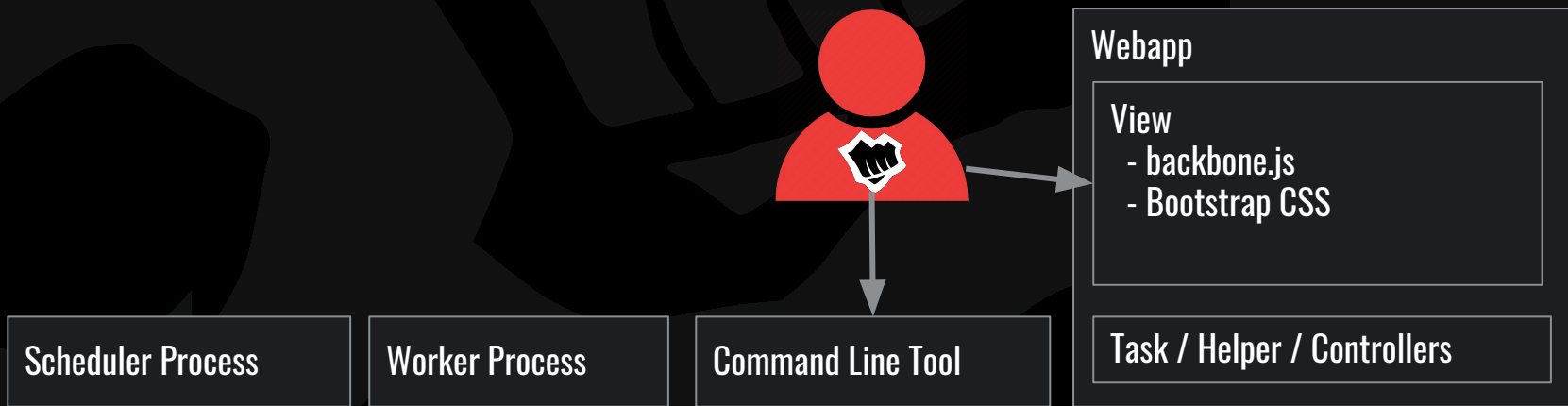
Task DAO

Env. Task DAO

Environment
DAO

Env. Helper DAO

Helper DAO



FuETL STATISTICS



5213

**ACTIVE REGIONAL
ETLS**



23125

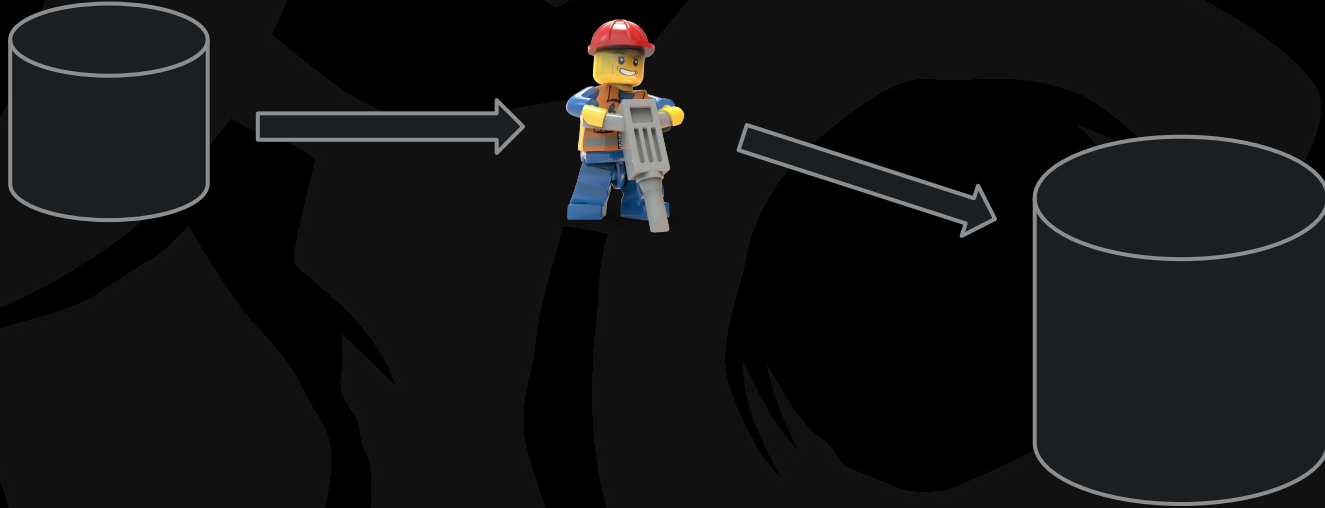
DAILY ETL RUNS



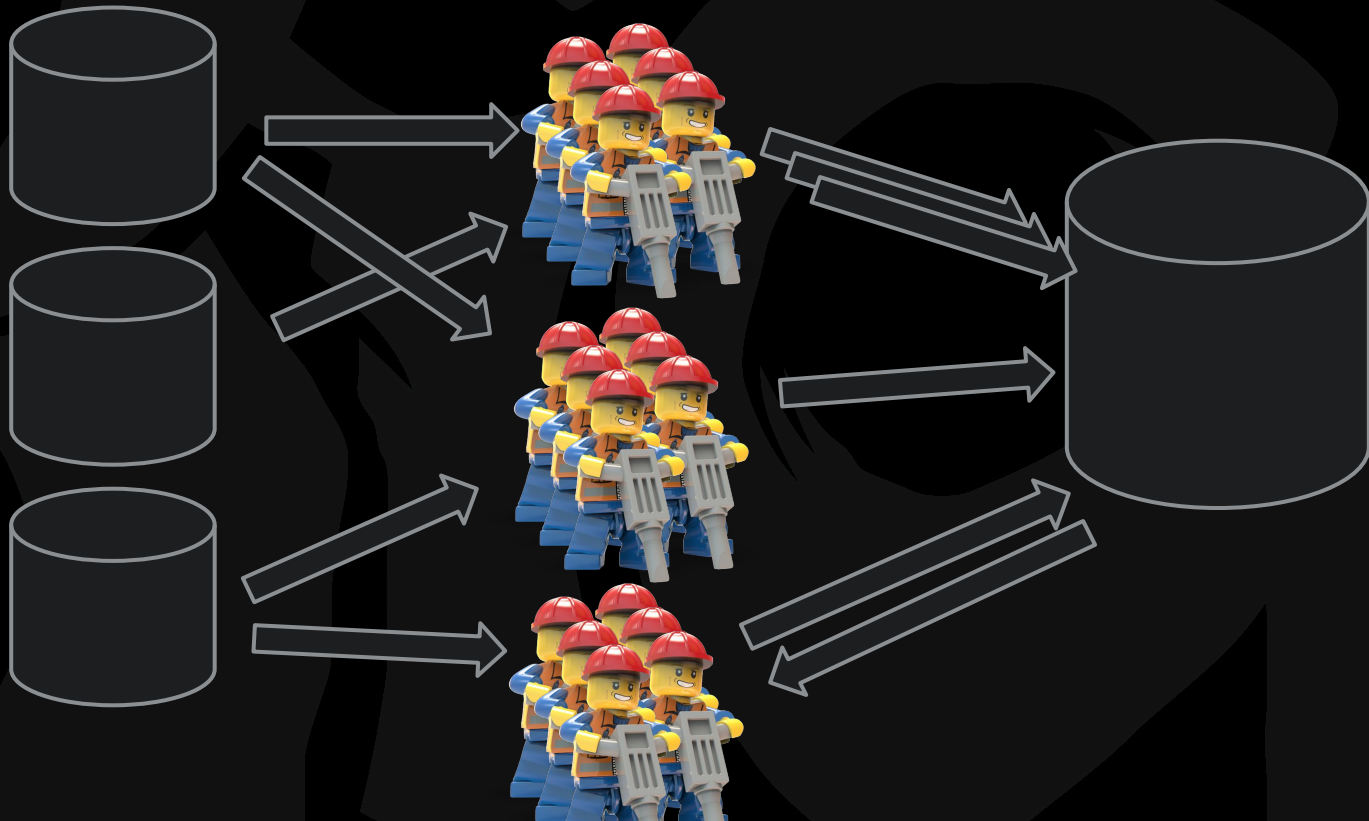
14 TB

DATA MOVED DAILY

FuETL SCALING



FuETL SCALING



Message Queues



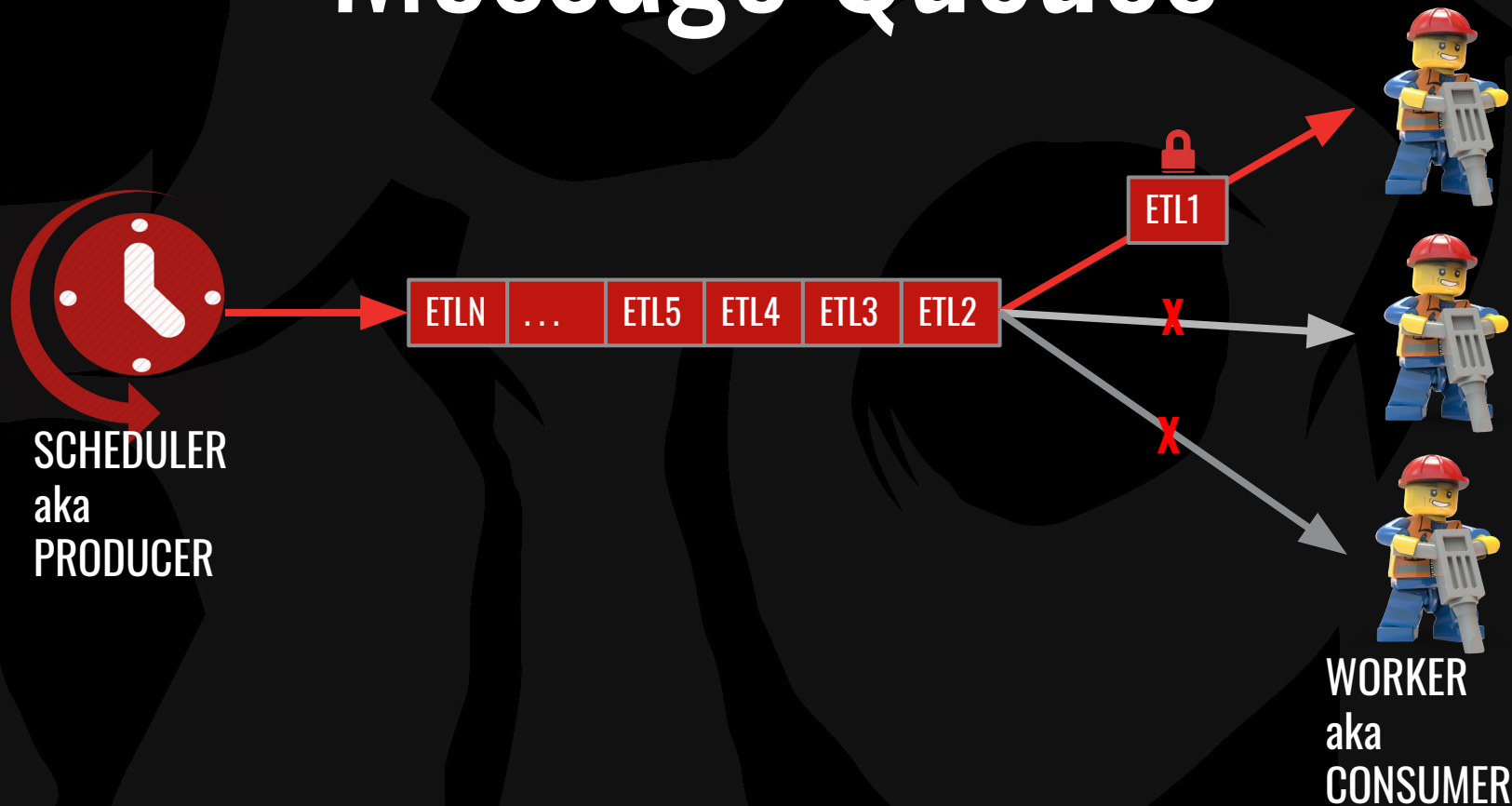
Message Queues

- 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



SCHEDULER
aka
PRODUCER



ETL1

ETL1

X



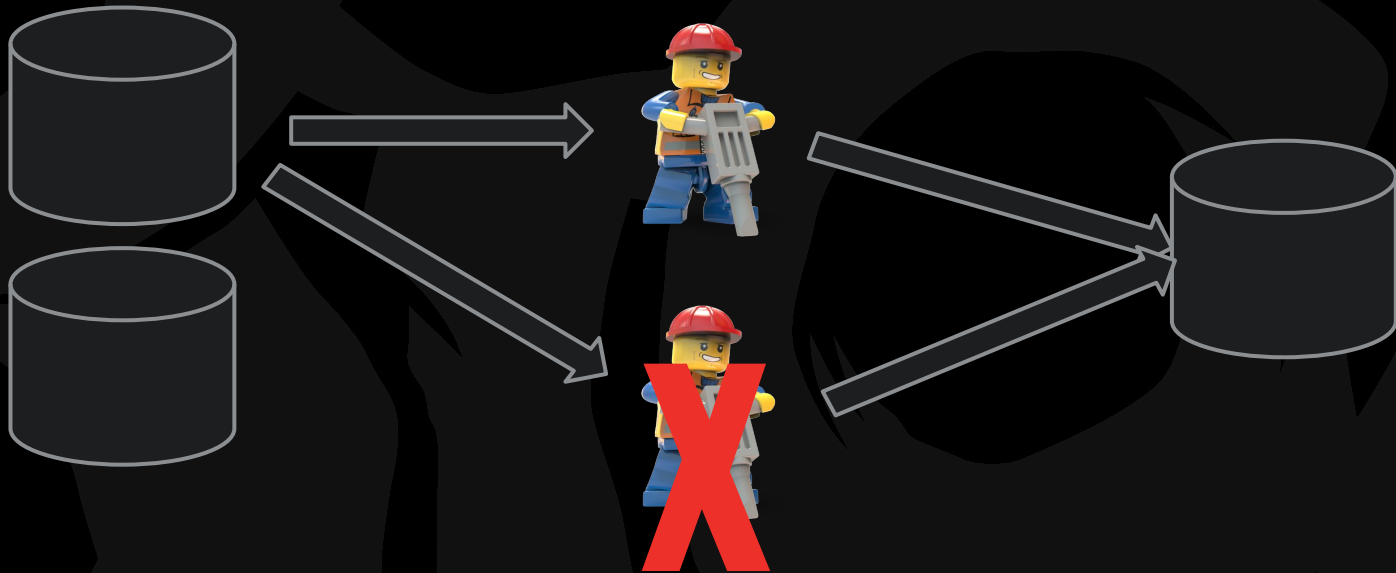
WORKER
aka
CONSUMER

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



DATA AUDITING



QUERY / VIEWS

AGGREGATE QUERIES



BATCH QUERIES



SINGLE-ROW QUERIES



VIZ. TOOLS



Kafka

The new hotness in big data

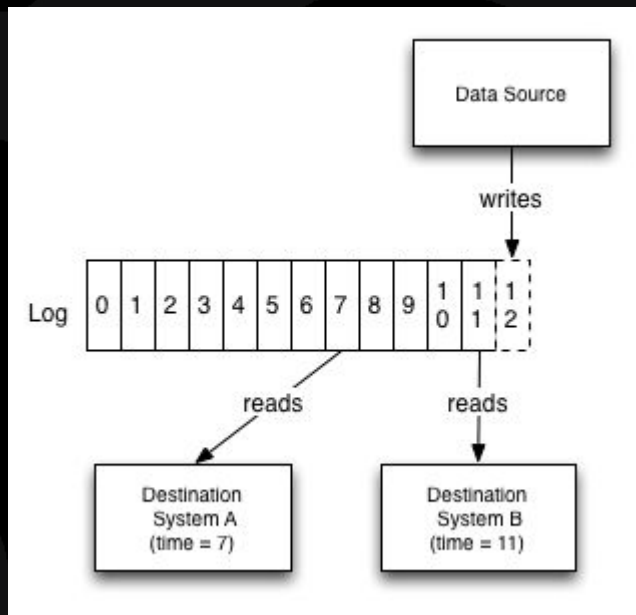
Open-source project maintained by
Confluent

Very fast distributed message queue

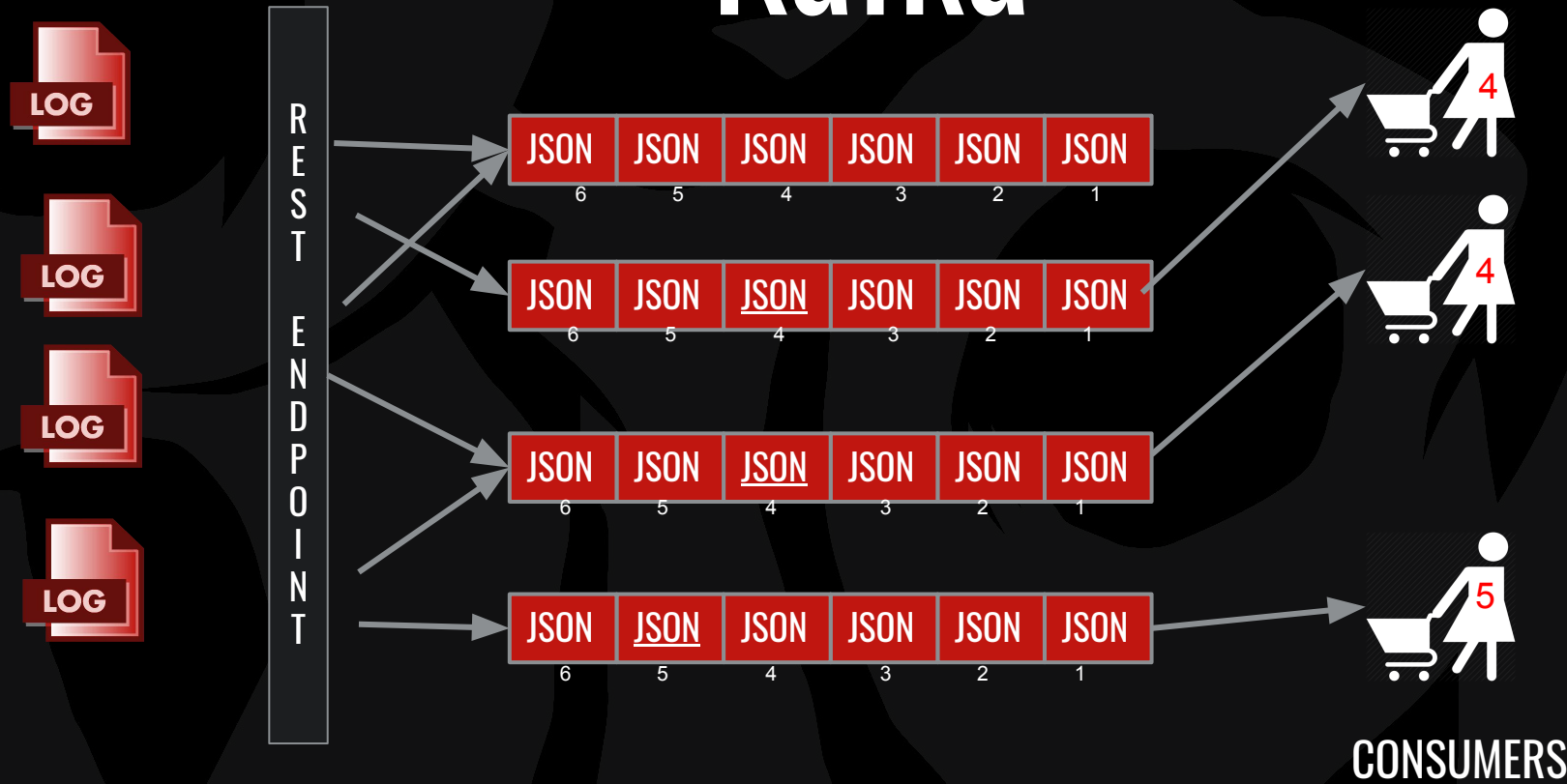
Data is replicated across “partitions” to
ensure no loss

Kafka

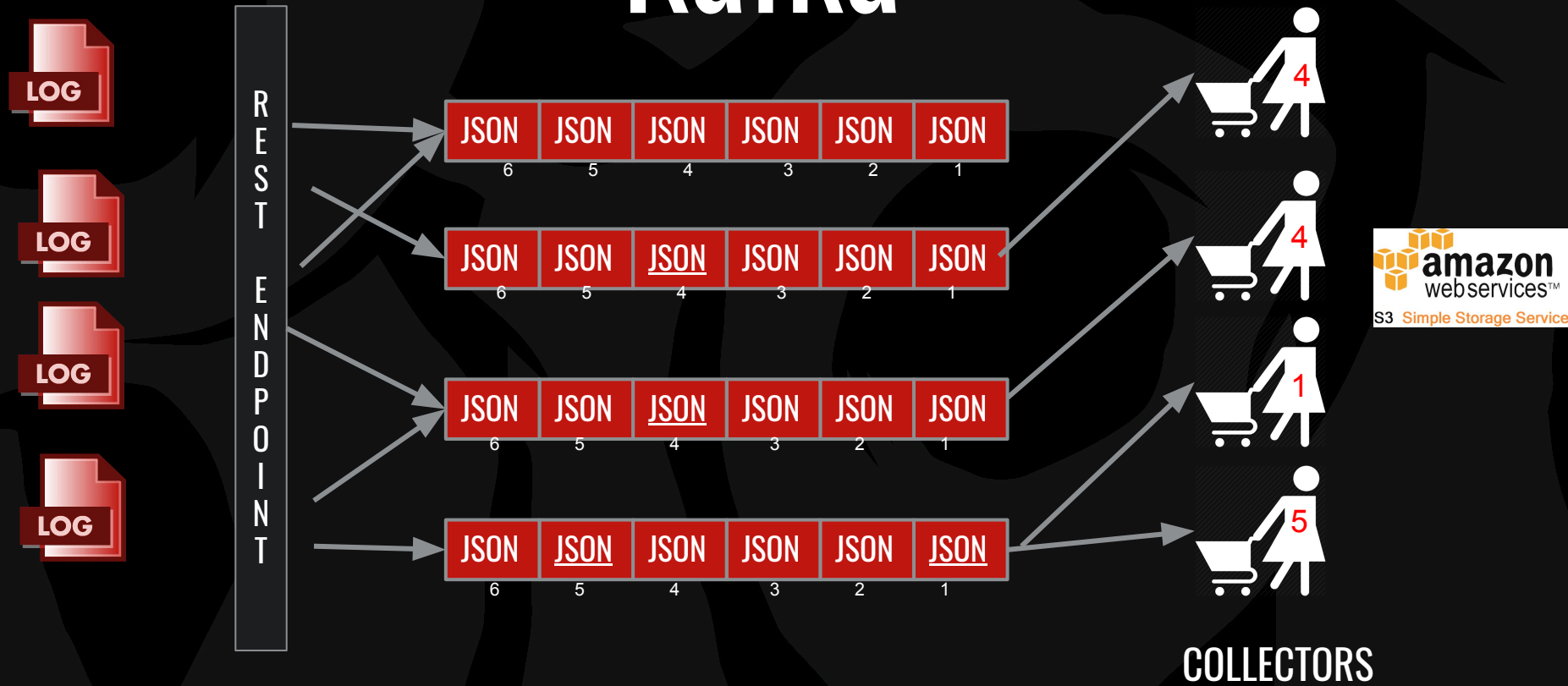
Has a DB Commit Log
(ooh revolutionary)



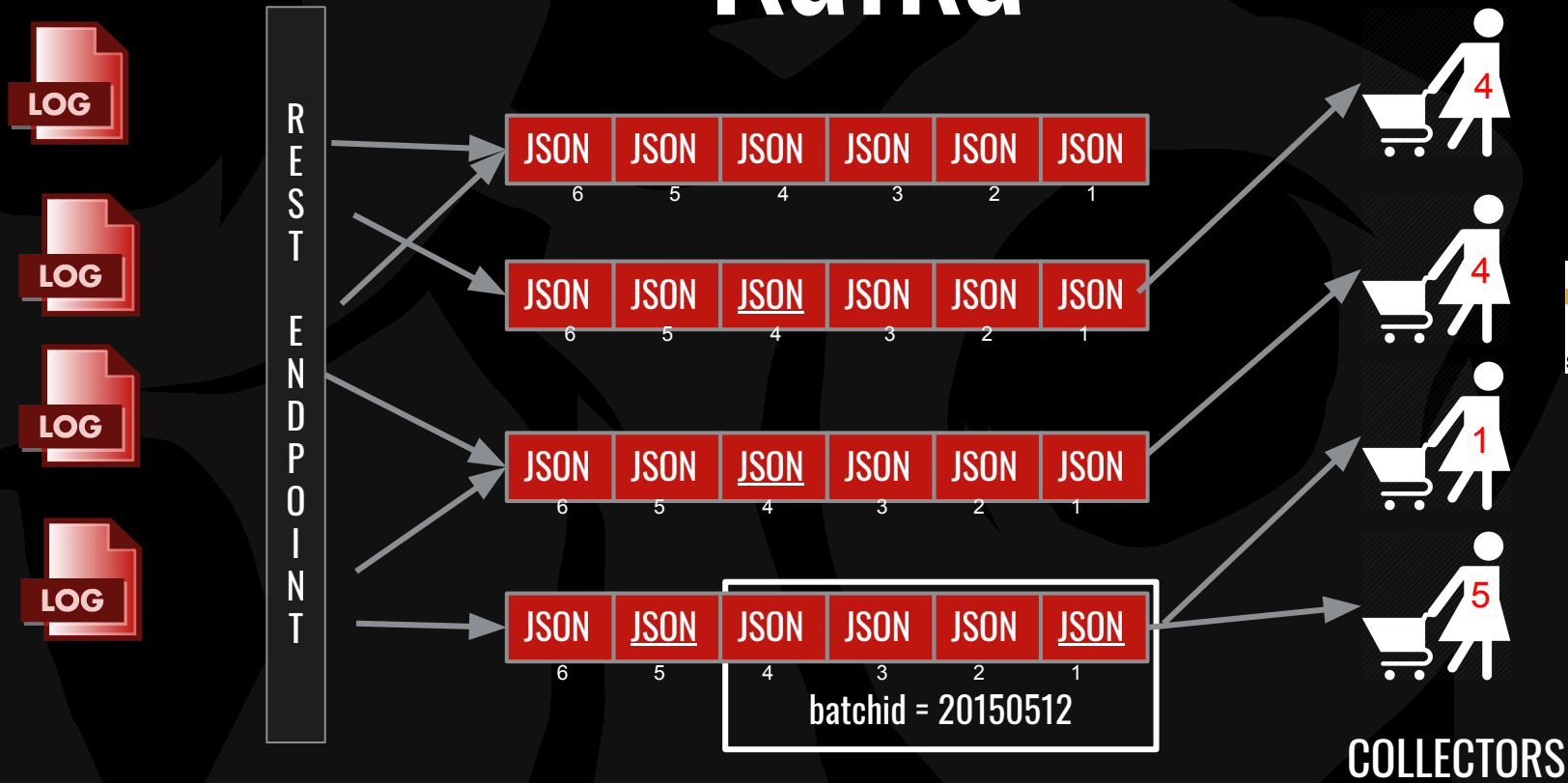
Kafka



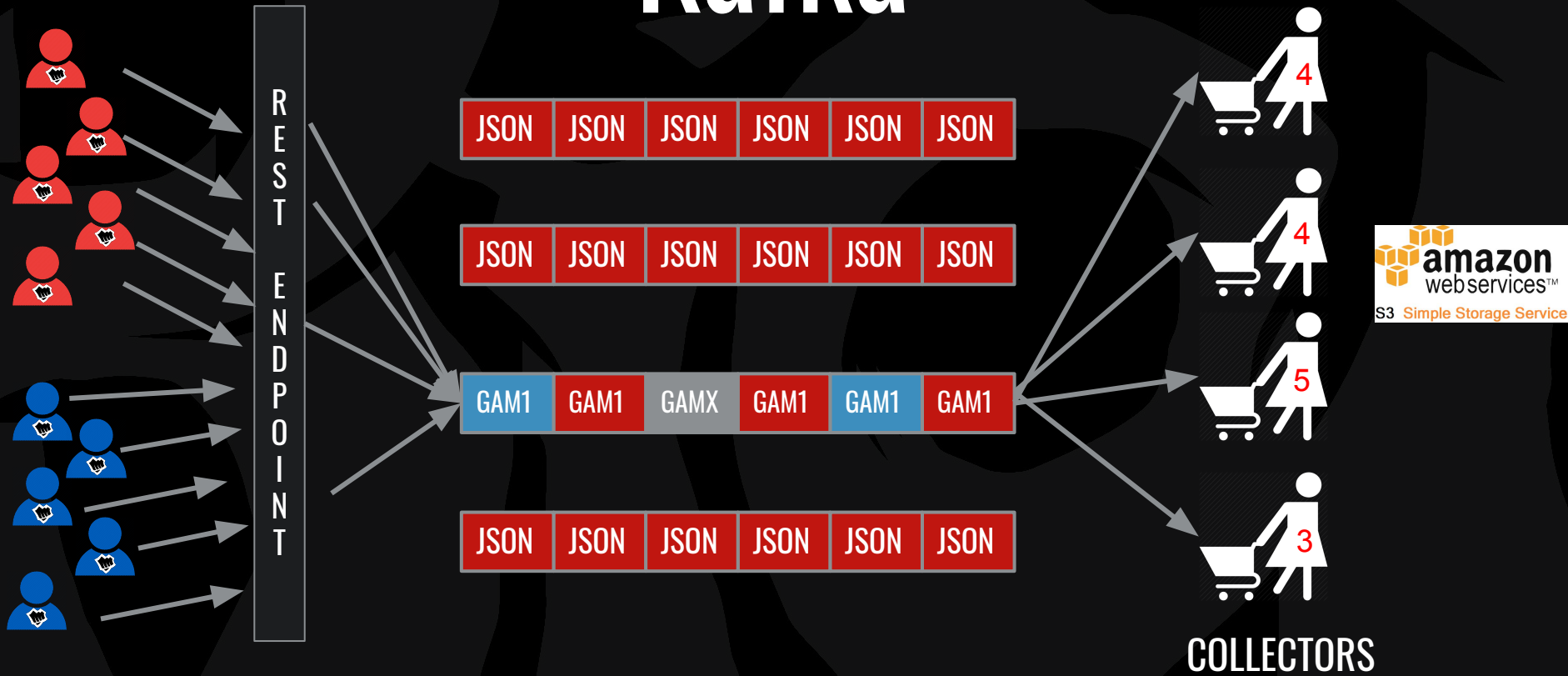
Kafka



Kafka



Kafka



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:

- $\text{abs}(\text{abs}(x)) = \text{abs}(x)$
- 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



- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE



DATA AUDITING



QUERY / VIEWS

AGGREGATE QUERIES



BATCH QUERIES



SINGLE-ROW QUERIES



VIZ. TOOLS



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
<input type="checkbox"/>	 env=BR1
<input type="checkbox"/>	 env=BR1_\$folder\$
<input type="checkbox"/>	 env=EUN1
<input type="checkbox"/>	 env=EUN1_\$folder\$
<input type="checkbox"/>	 env=EUW1
<input type="checkbox"/>	 env=EUW1_\$folder\$
<input type="checkbox"/>	 env=ID1
<input type="checkbox"/>	 env=ID1_\$folder\$
<input type="checkbox"/>	 env=KR1
<input type="checkbox"/>	 env=KR1_\$folder\$
<input type="checkbox"/>	 env=LA1
<input type="checkbox"/>	 env=LA1_\$folder\$
<input type="checkbox"/>	 env=LA2
<input type="checkbox"/>	 env=LA2_\$folder\$
<input type="checkbox"/>	 env=NA1
<input type="checkbox"/>	 env=NA1_\$folder\$
<input type="checkbox"/>	 env=OC1
<input type="checkbox"/>	 env=OC1_\$folder\$

Upload

Create Folder

Actions ▾

Versions:

Hide

Show

None


Properties

Transfers

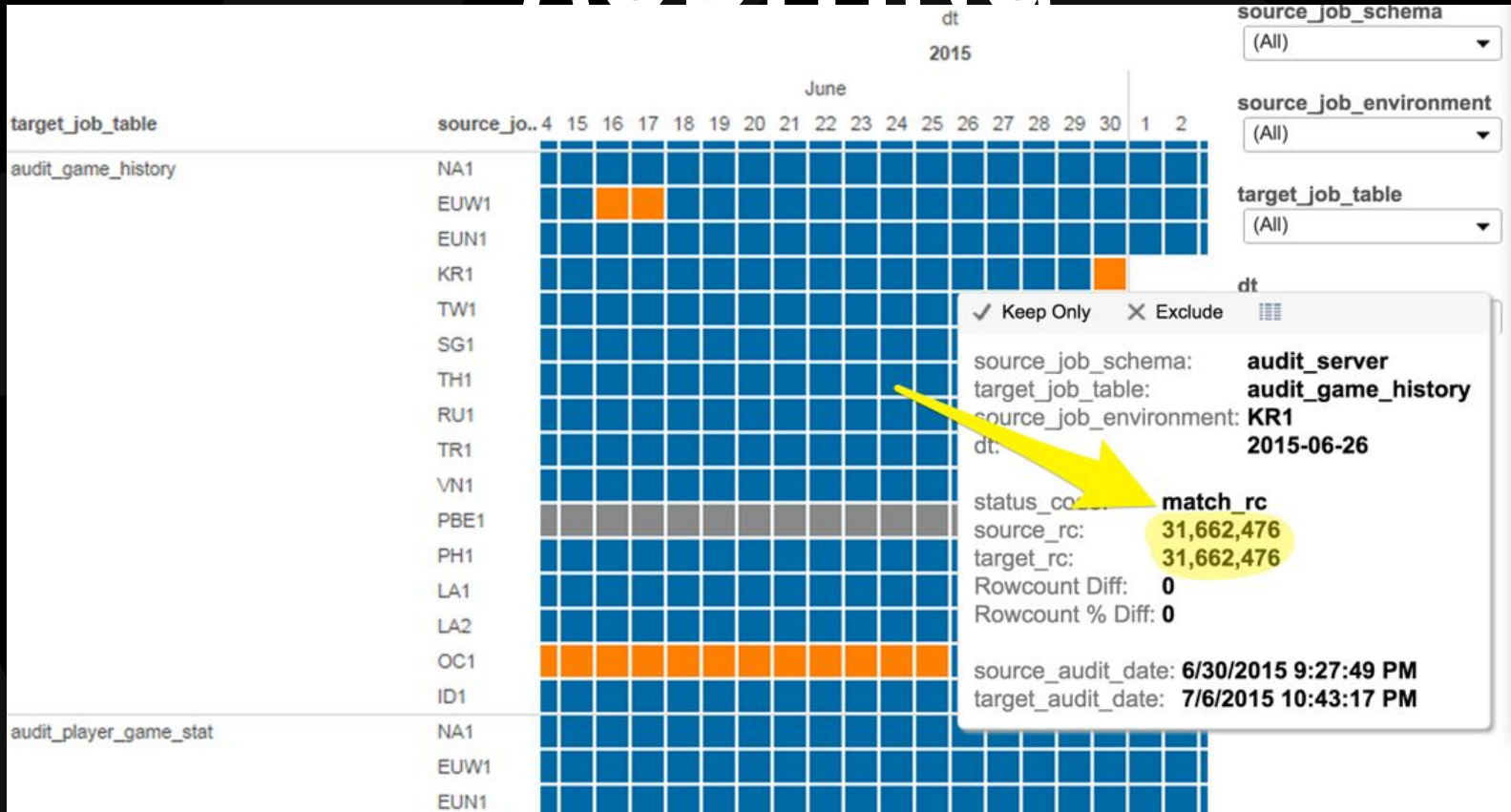
↻

All Buckets /

/ merged / audit_event_queue_dodge / env=EUW1 / dt=2015-10-27

	Name	Storage Class	Size	Last Modified
<input type="checkbox"/>	 abf621d7-8f0d-43b7-93d4-f0f7afaaea7f-000000	Standard	26.7 MB	Wed Oct 28 15:04:13 GMT-700 2015

AUDITING



audit_game_history

June 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 1 2 3 4 5

NA1
EUW1
EUN1
KR1
TW1
SG1
BR1
TH1
RU1
TR1
VN1
PBE1
PH1
LA1
LA2
OC1
ID1

source_job_schema: audit_server
target_job_table: audit_game_hist
source_job_environment: OC1
dt: 2015-06-16

status_code: diff_rc
source_rc: 1,796,278
target_rc: 1,796,653
Rowcount Diff: -375
Rowcount % Diff: 0.02

source_audit_date: 7/13/2015 9:00:03 PM
target_audit_date: 7/13/2015 8:59:19 PM

```
source_job_schema:      audit_server
target_job_table:       audit_game_history
source_job_environment: OC1
dt:                     2015-06-16
```

```
status_code:      diff_rc
source_rc:        1,796,278
target_rc:        1,796,653
Rowcount Diff:    -375
Rowcount % Diff:  0.02
```

```
source_audit_date: 7/13/2015 9:00:03 PM
target_audit_date: 7/13/2015 8:59:19 PM
```

INGESTION

PULL-BASED / ETL



FuETL

- OLTP game data
- External Data Sources

PUSH-BASED



- Anything pushed to it
- Server logs

STORAGE

MASTER WAREHOUSE



DATA AUDITING



QUERY / VIEWS

AGGREGATE QUERIES



BATCH QUERIES



SINGLE-ROW QUERIES



VIZ. TOOLS





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.

INGESTION

PULL-BASED / ETL



- FuETL**
- OLTP game data
 - External Data Sources

PUSH-BASED



- HONU**
- Anything pushed to it
 - Server logs

STORAGE

MASTER WAREHOUSE



DATA AUDITING



QUERY / VIEWS

AGGREGATE QUERIES



BATCH QUERIES



SINGLE-ROW QUERIES



VIZ. TOOLS



[illegible]

BATCH

[illegible]

OLAP

[illegible]

POINT

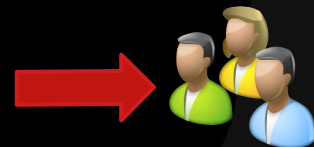
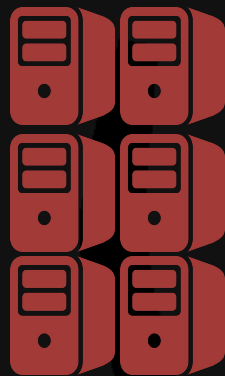
SCALING IN AWS



SCALING

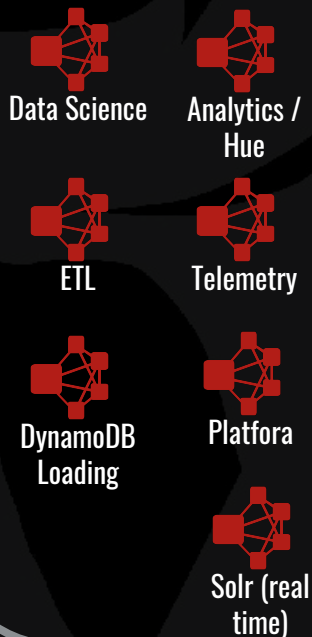
RESOURCE CONTENTION

Hive .08 pre YARN, immature resource scheduling



AWS Infrastructure Today

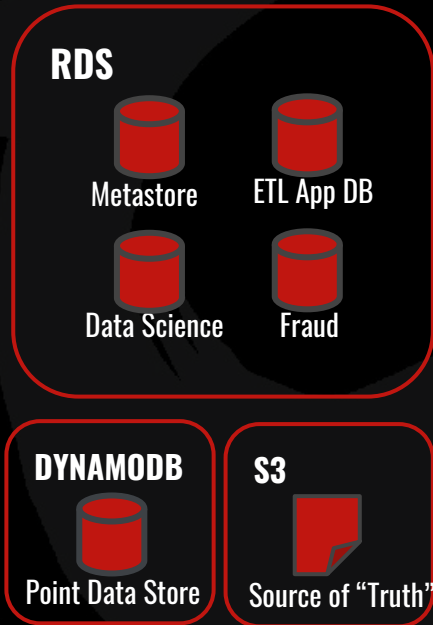
EMR



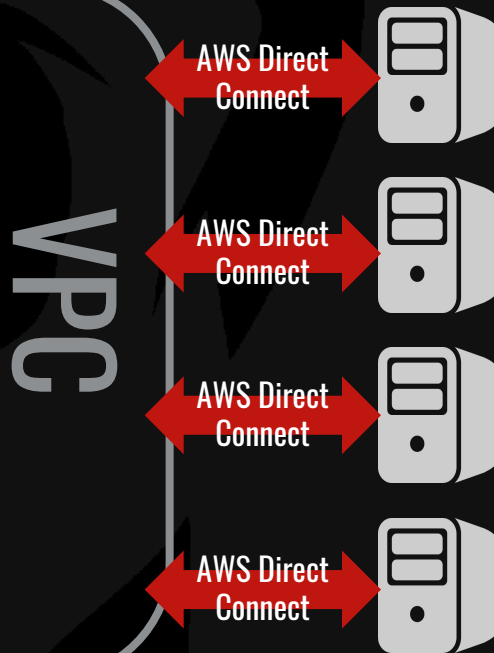
EC2



Storage



Networking



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

QUESTIONS?

ENGINEERING BLOG

<http://engineering.riotgames.com>

CAREERS

<http://www.riotgames.com/careers>

SEAN MALONEY



SMALONEY @
[riotgames.com](mailto:smaloney@riotgames.com)



@SEAN_SEANNERY