Scaling Scalability

I do not think “scalable” means what you think it means.

Dmitriy Ryaboy

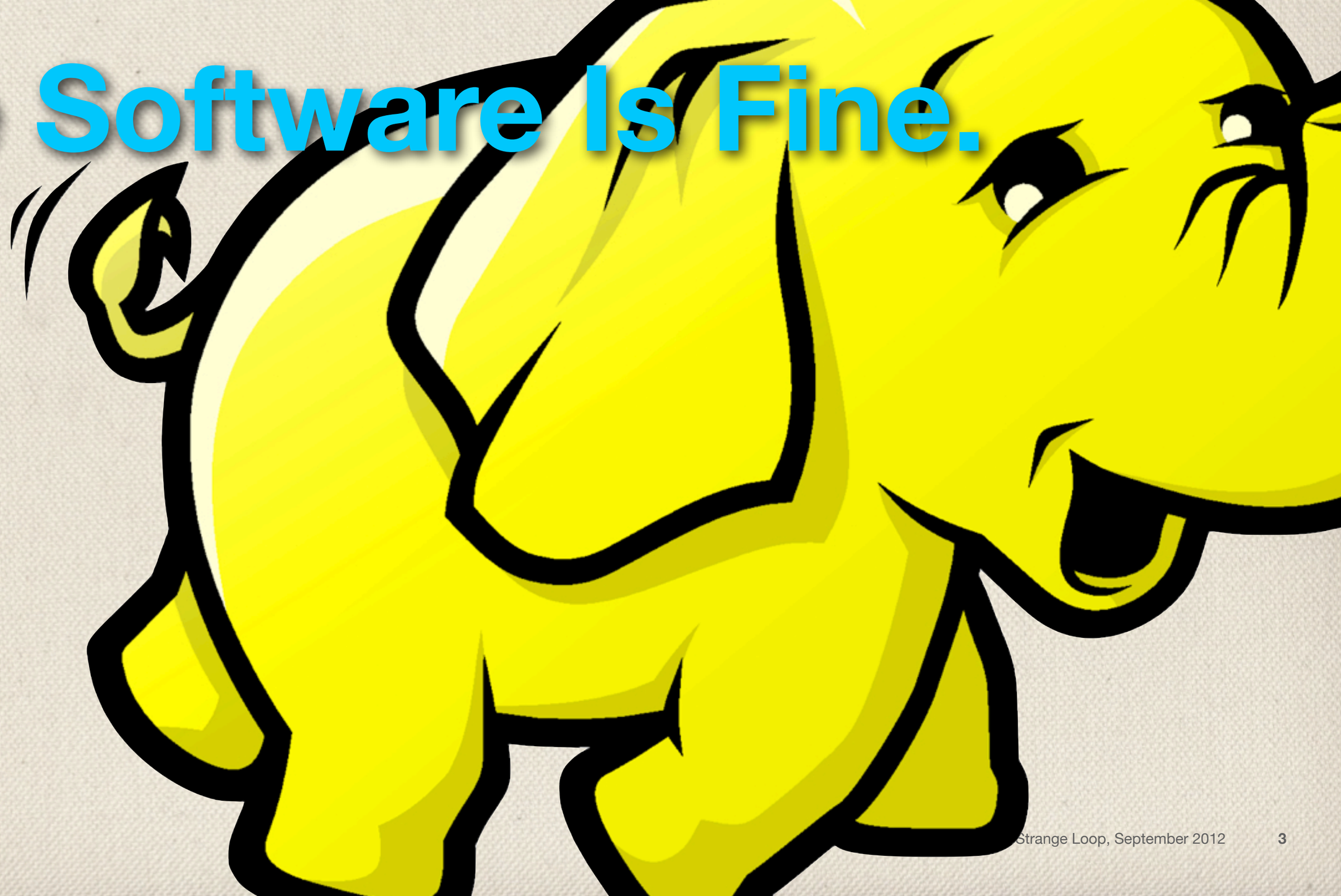
@squarecog

Strange Loop, September 2012

Image Source: http://www.flickr.com/photos/syldavia/48854093/

Add more machines, go bigger, go faster.

The Software Is Fine.



The Software Is Fine.

Hadoop 1.0:

4000 Nodes

96 PB of data

It's fine.

This talk is not about software.

Glass, Concrete, and Steel scale.

That’s not enough to build skyscrapers.

Image Source: http://www.flickr.com/photos/syldavia/48854093/

You are good at making machines do what you want.

It’s not just about you, though.

It’s about People.

‘Soylent Green’ copyright Metro-Goldwyn-Mayer, Inc

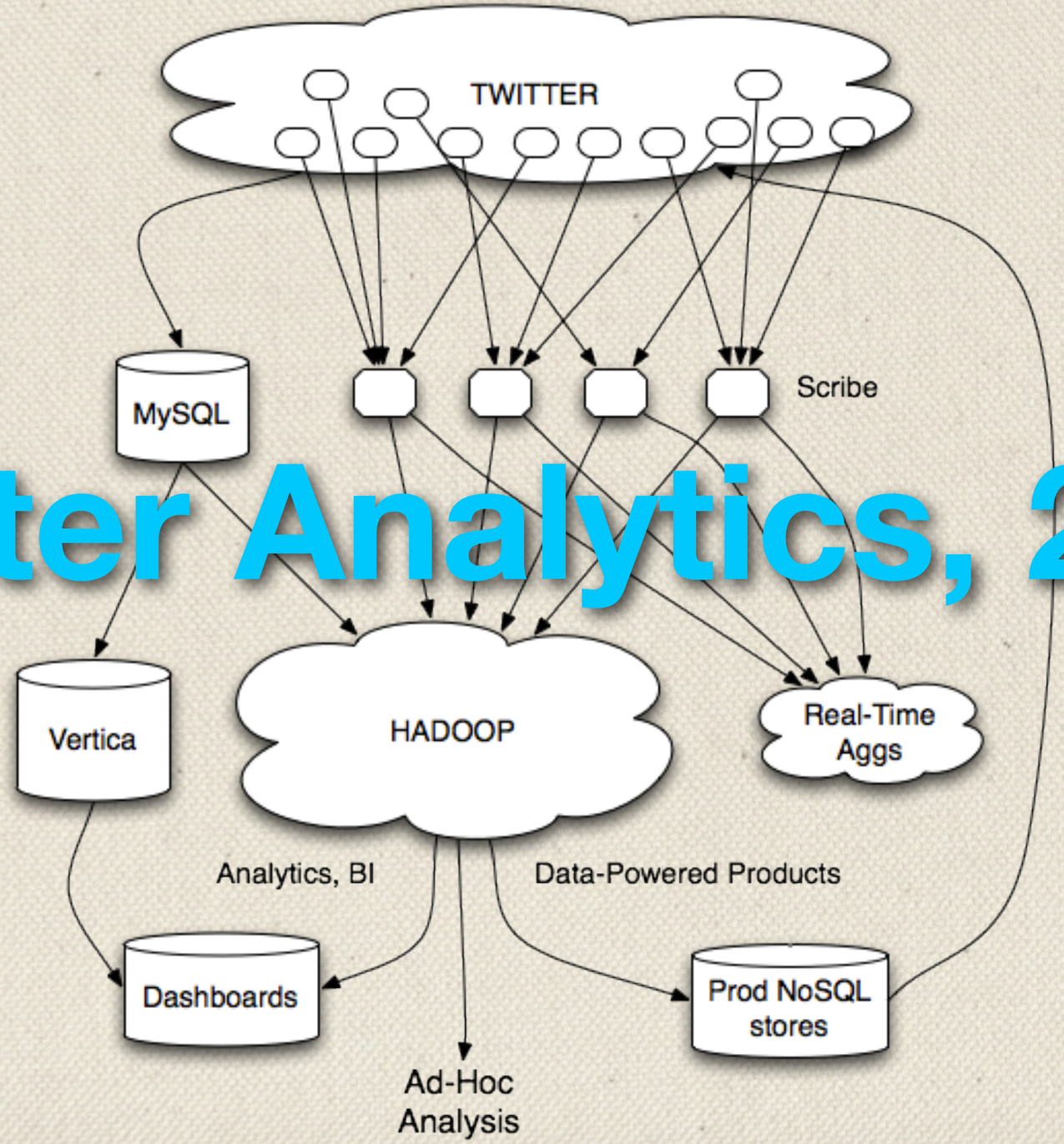
Ok, also some technology.

Ok, also some technology.
In service of People.

Ok, also some technology.

In service of People.

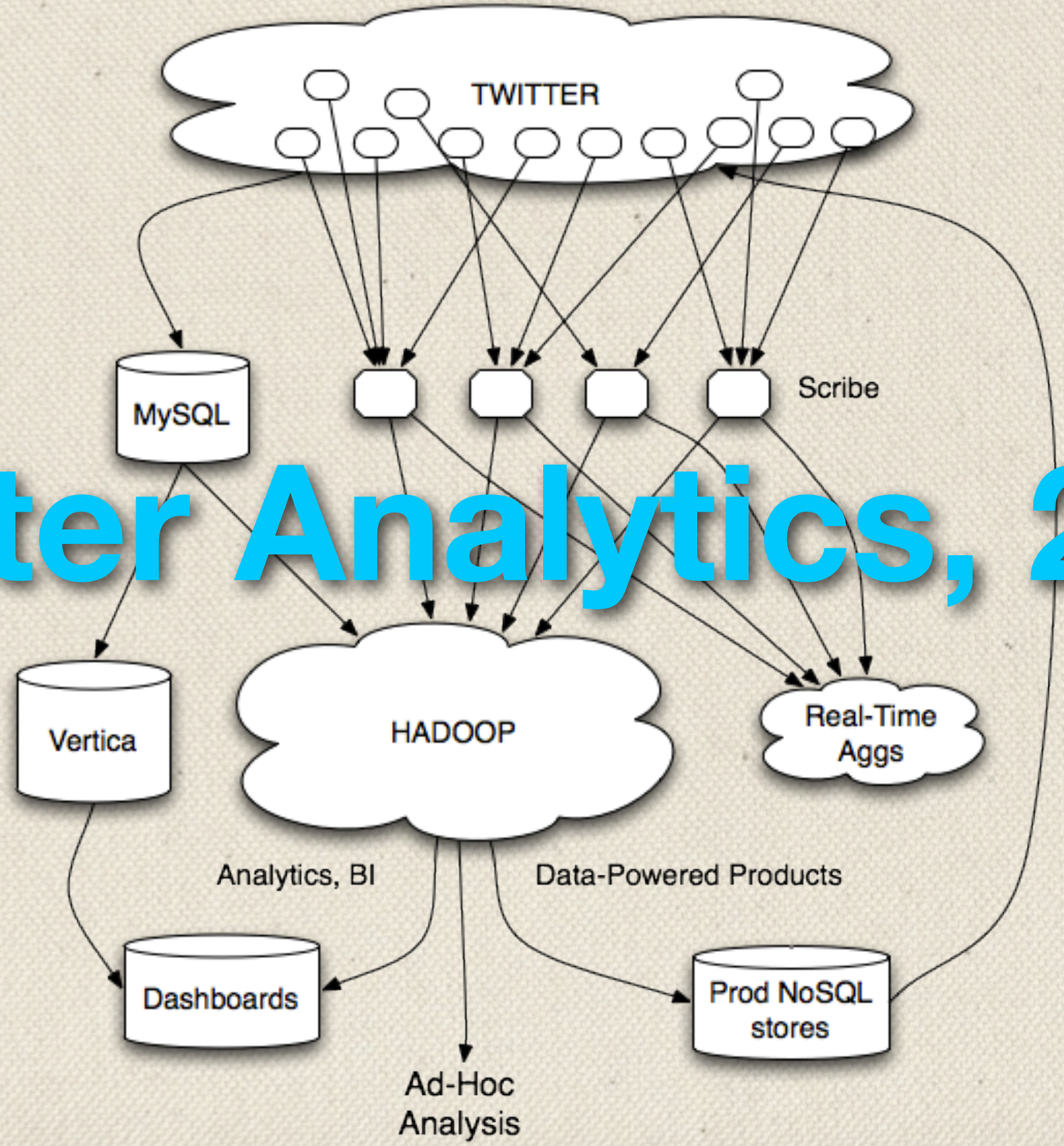
Twitter Analytics, 2010



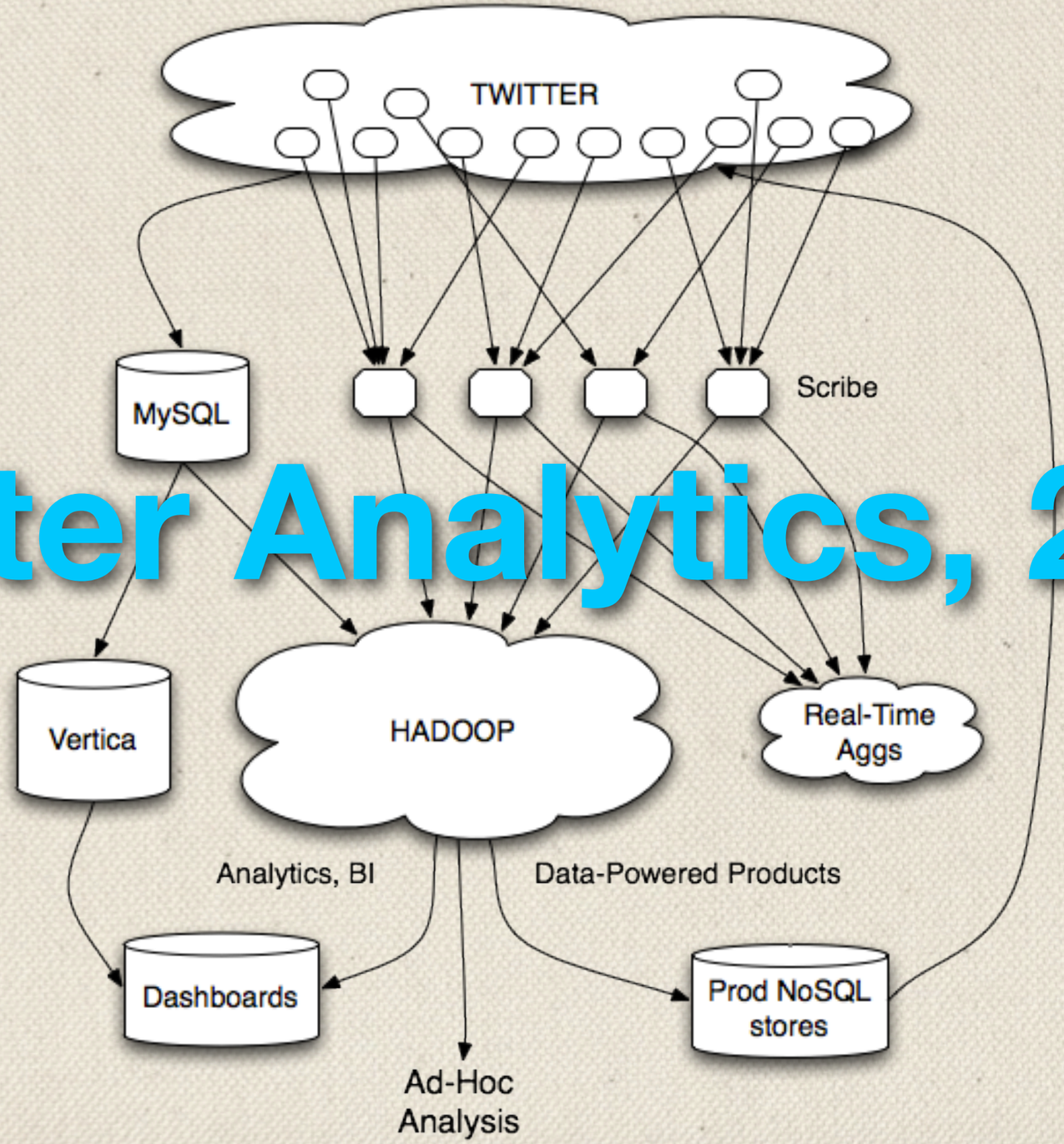
100 People
30 Hadoop nodes
1 team, 4 People use Analytics stack daily

Twitter Analytics, 2012

Twitter Analytics, 2012



Twitter Analytics, 2012



1000+ People
3k+ Hadoop nodes
Dozen of teams, $O(10^4)$ daily MR jobs

100+ TBs ingested daily.
Hundreds of engineers.
Hundreds of log categories.
Thousands of daily jobs.
Chaos potential: high.

IMHO.

IMHO.

Not representative of opinions of Twitter, Inc.

IMHO.

Not representative of opinions of Twitter, Inc.

Void where prohibited by law.

IMHO.

Not representative of opinions of Twitter, Inc.

Void where prohibited by law.

Management is not responsible for misuse.

IMHO.

Not representative of opinions of Twitter, Inc.

Void where prohibited by law.

Management is not responsible for misuse.

If someone asks, I probably wasn't here.

Data Needs a Schema

If data is a goldmine, a schema is a headlight

Image Source: http://images.lightstrade.com/2012/7/21/m6347848491852900 63.jpg

Option 1: System.err.println

```
  public static final Pattern twitterLogPattern = Pattern

.compile("^(\\w+\\s+\\d+\\s+\\d+:\\d+:\\d+)\\s+([^@]+?)@(\\S+)\\s+(\\S+):\\s+(\\S+)\\s+(\\S+
)\\s+((?:\\S+?,\\s+)*(?:\\S+?))\\s+(\\S+)\\s+(\\S+)\\s+\\[([^\\]]+)\\]\\s+\"(\\w+)\\s+([^\"\\
\\]*(?:\\\\.[^\"\\\\]*)*)\\s+(\\S+)\"\\s+(\\S+)\\s+(\\S+)\\s+\"([^\"\\\\]*(?:\\\\.[^\"\\\\]*)
*)\"\\s+\"([^\"\\\\]*(?:\\\\.[^\"\\\\]*)*)\"\\s+([\\d-]+)?\\s*(\\d+)(.*)$");
```

Let’s not.

Option 2: MySQL as a log

```
create table `my_audit_log` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `created_at` datetime,
  `user_id` int(11),
  `action` varchar(256),
  -- ......
  )  ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

When you have a hammer, everything looks like a thumb.

Option 3: JSON!

This works. For a little while.

```
> stuff = { "thing_int" : 100 }
  ▶ Object
> stuff.thing_int
  100
> stuff.thing_int = "definitely not an int"
  "definitely not an int"
> stuff.thing_int
  "definitely not an int"
> // Define "type safety"
  undefined
>
```

```
{ "token": 945842,
  "feature_enabled": "super_special",
  "user_id": 2299229,
  "page": "null",
  "info": { "email": "me@place.com"} }
```

String “null”? Null? Nil?

```
{ "token": 945842,
  "feature_enabled": "super_special",
  "user_id": 2299229,
  "page": "null",
  "info": { "email": "me@place.com"} }
```

String “null”? Null? Nil?

That’s a screen name. It happens to be numeric for this user.

```
{ "token": 945842,
  "feature_enabled": "super_special",
  "user_id": 2299229,
  "page": "null",
  "info": { "email": "me@place.com"} }
```

String “null”? Null? Nil?

That’s a screen name. It happens to be numeric for this user.

```
{ "token": 945842,
  "feature_enabled": "super_special",
  "user_id": 2299229,
  "page": "null",
  "info": { "email": "me@place.com"} }
```

What other fields are possible?

String “null”? Null? Nil?

That’s a screen name. It happens to be numeric for this user.

```
{ "token": 945842,
  "feature_enabled": "super_special",
  "user_id": 2299229,
  "page": "null",
  "info": { "email": "me@place.com"} }
```

What other fields are possible?

Should be a list. This will get fixed tomorrow.

String “null”? Null? Nil?

That’s a screen name. It happens to be numeric for this user.

```
{ "token": 945842,
  "feature_enabled": "super_special",
  "user_id": 2299229,
  "page": "null",
  "info": { "email": "me@place.com"} }
```

What other fields are possible?

Should be a list. This will get fixed tomorrow.

Do they all look like this, or are some completely different?

JSON Doesn’t Scale

(For logging and long-term data analysis)

Use Thrift/Avro/Protobufs

(For logging and long-term data analysis)

Thrift example (Zipkin)

```
// some event took place, either one by the framework or by the user
struct Annotation {
  1: i64 timestamp                  // microseconds from epoch
  2: string value                   // what happened at the timestamp?
  3: optional Endpoint host         // host this happened on
  4: optional i32 duration          // how long did the operation take? microseconds
}

struct Span {
  1: i64 trace_id                   // unique trace id, use for all spans in trace
  3: string name,                   // span name, rpc method for example
  4: i64 id,                        // unique span id, only used for this span
  5: optional i64 parent_id,                // parent span id
  6: list<Annotation> annotations, // list of all annotations/events that occured
  9: optional bool debug = 0
}
```

Thrift, Avro, Protobufs.

Efficient serialization, codegen in many languages, type safety, sane evolution path.

Logical / Physical separation

Thrift / Proto / Avro defines a contract

Frees you to improve storage, compression, etc.

Data Discovery Service

If a schema is a headlamp, a metadata service is mine lighting.

http://www.westpacled.com/products/led-mining-lights/

Data Discovery Service

Or maybe a map.

http://en.wikipedia.org/wiki/File:Carta_Marina.jpeg

Scandinavian Sea Monsters
Data Discovery Service
Or maybe a map.
http://en.wikipedia.org/wiki/File:Carta_Marina.jpeg
CARTA MARINA ET DESCRIPTIO SEPTEMTRIONALIVM TERRARVM AC MIRABILIVM RERVM IN EIS CONTENTARVM DILIGENTISSIME ELABORATA
ISLANDIA
MARE GLACIALE
FINMARCHIA
LAPPIA
CARELIA
BOTNIA
MARE BOTNICVM
SCANDIA
NORVEGIA
SVECIA
MARE SVETICVM
FINLANDIA VEL TAVASTIA
MOSCOVIE PARS
LIVONIA
RVSSIA ALBA
MARE DEVCALIDONICVM
SCOCIE PARS
TILE
ORCADES
OCCIDENS
OCEANVS SCITHICVS

Data Discovery Service

Or maybe a map.

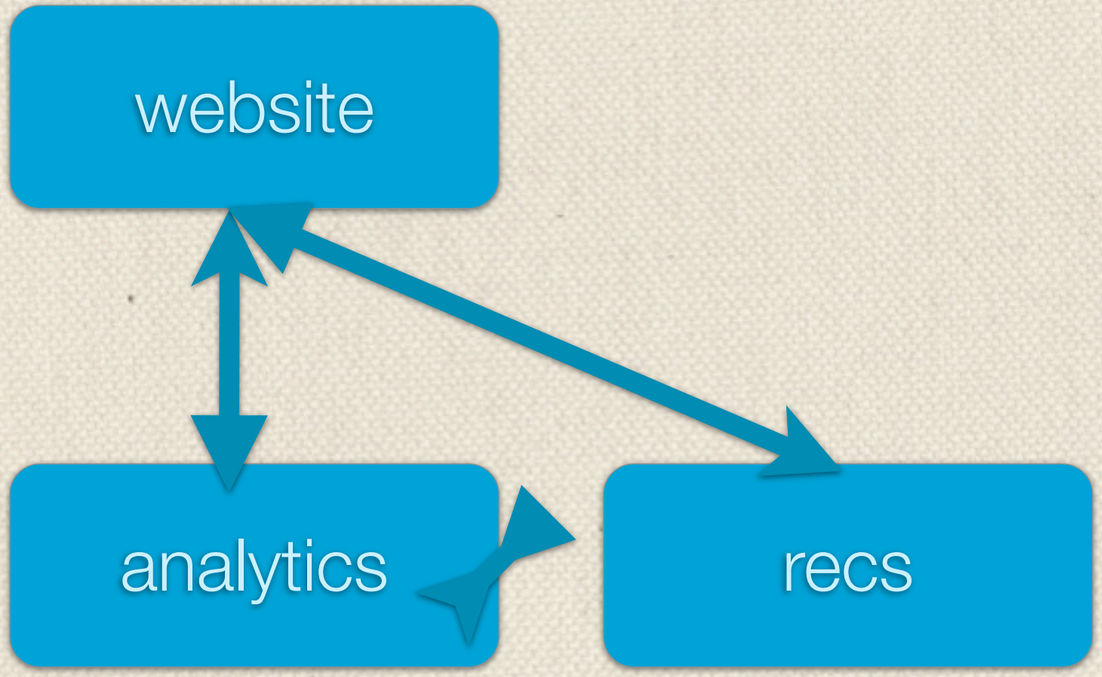
http://en.wikipedia.org/wiki/File:Carta_Marina.jpeg

Make it easy to find stuff.

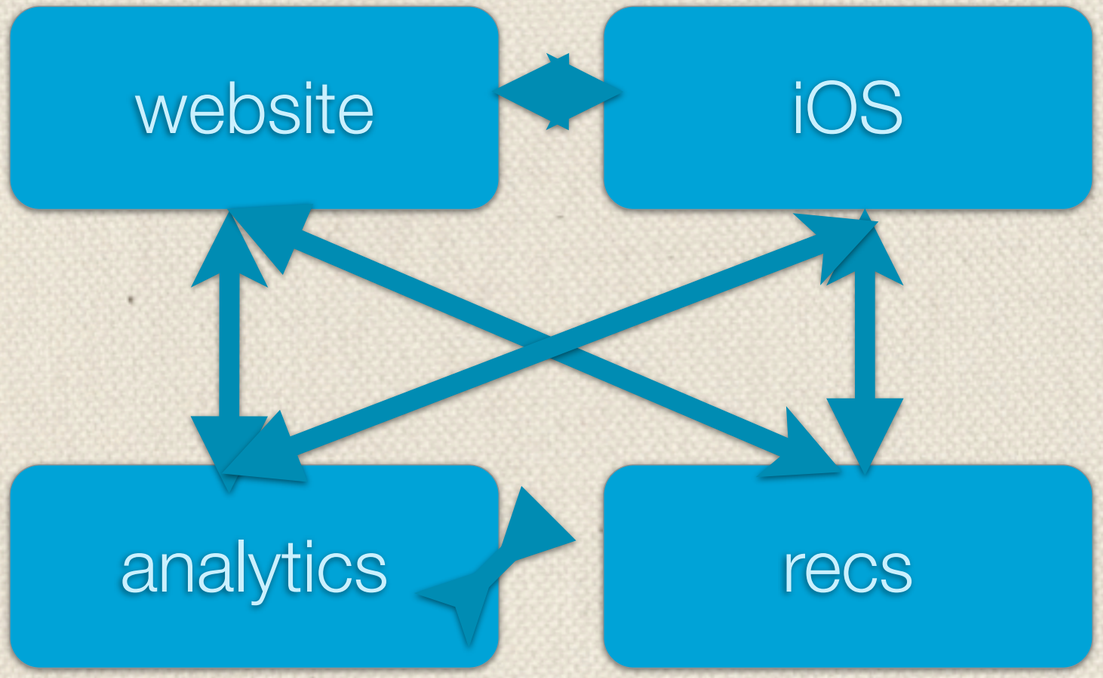
Reduce cross-talk

Reduce cross-talk

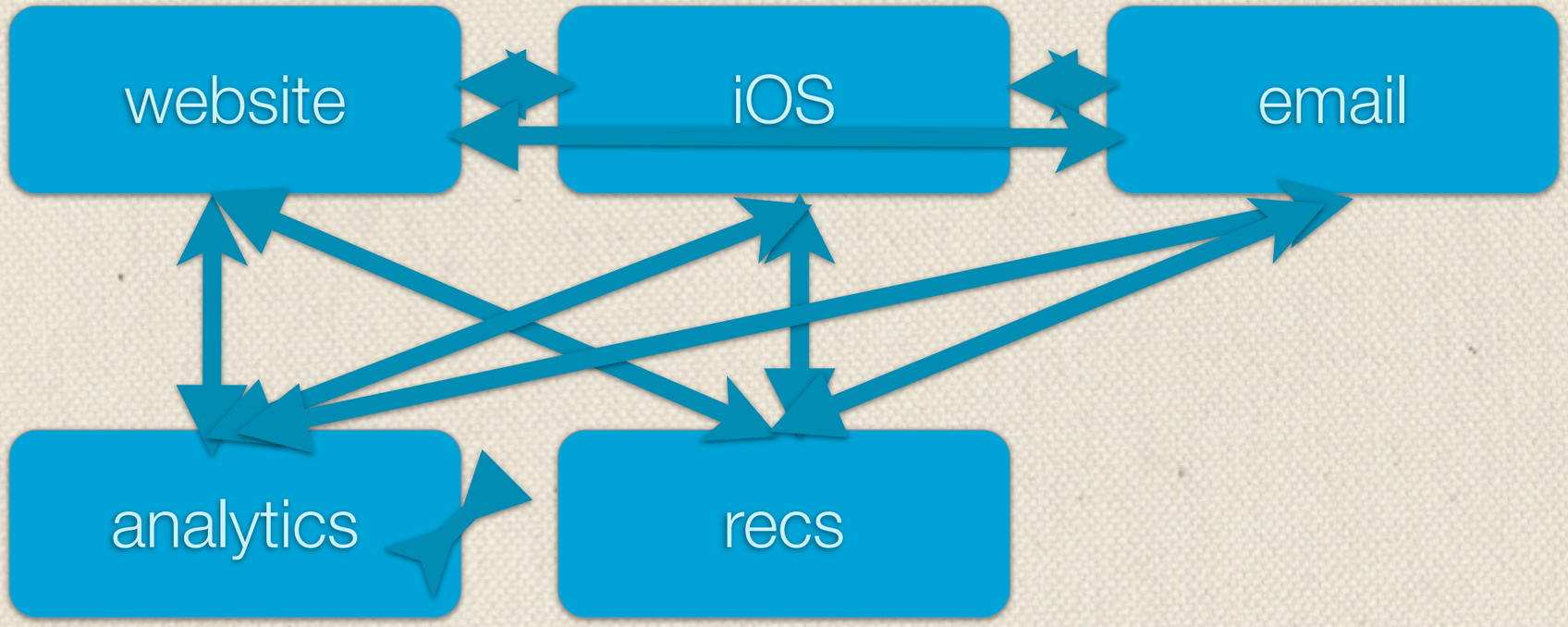
Reduce cross-talk



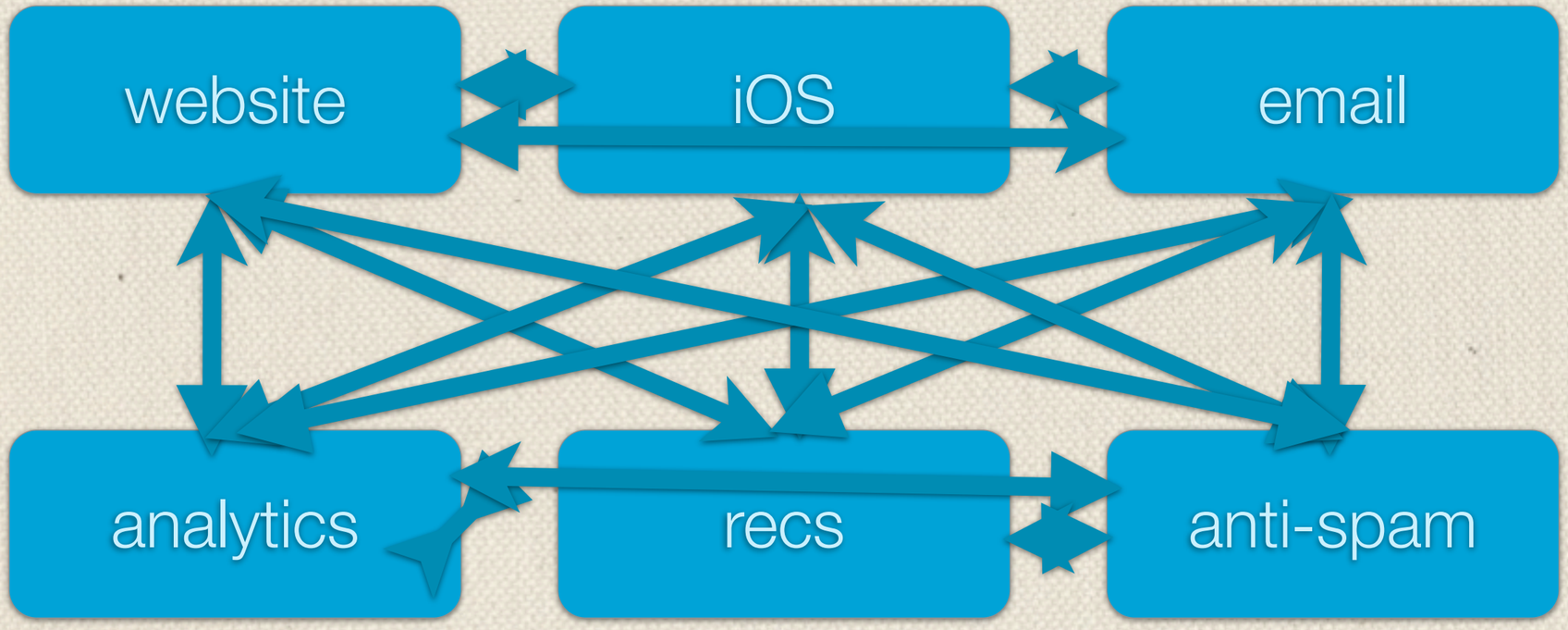
Reduce cross-talk



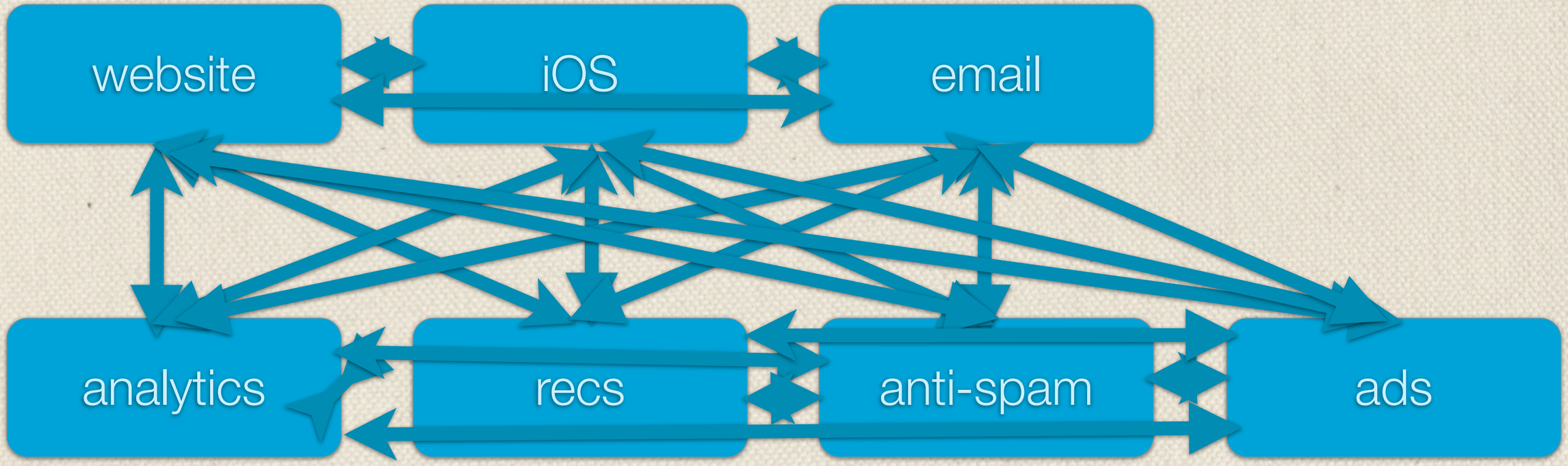
Reduce cross-talk



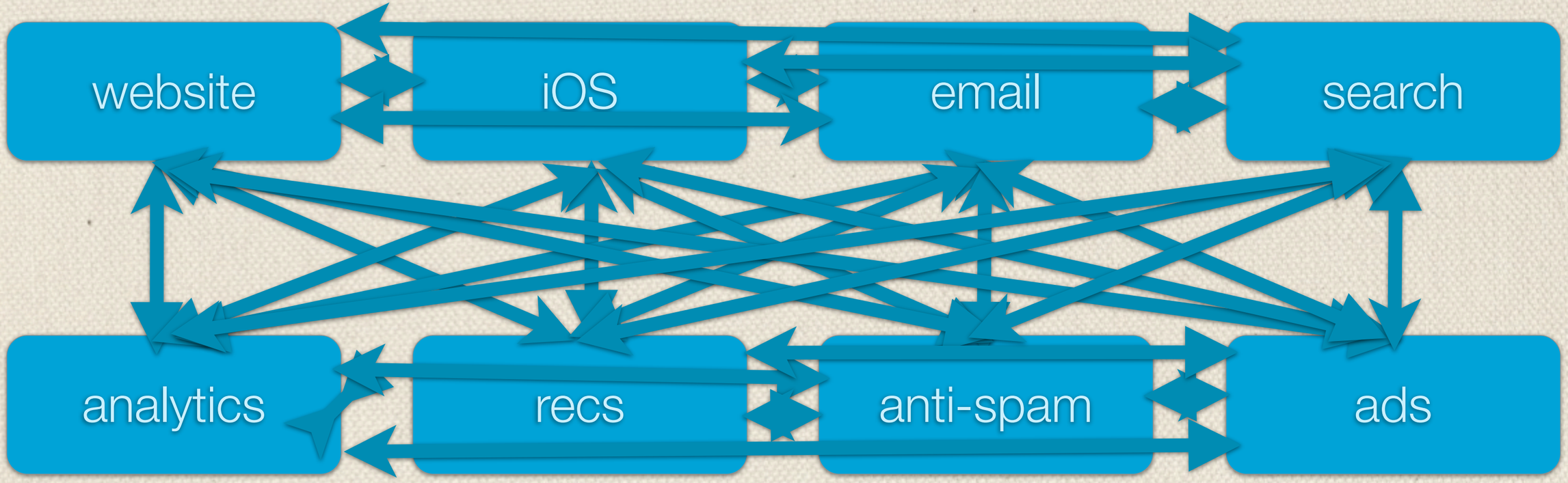
Reduce cross-talk



Reduce cross-talk



Reduce cross-talk



Data FAQs

Data FAQs

Where is the data?
How do I read it?
Who produces it?
Known anomalies?
Known derivations?

HCatalog

HCatalog

Hive’s metastore, for all.

HCatalog

Hive's metastore, for all.
"show, describe tables" for Hadoop

HCatalog

Hive's metastore, for all.
"show, describe tables" for Hadoop

Maintains nested record abstraction

HCatalog

Hive's metastore, for all.
"show, describe tables" for Hadoop

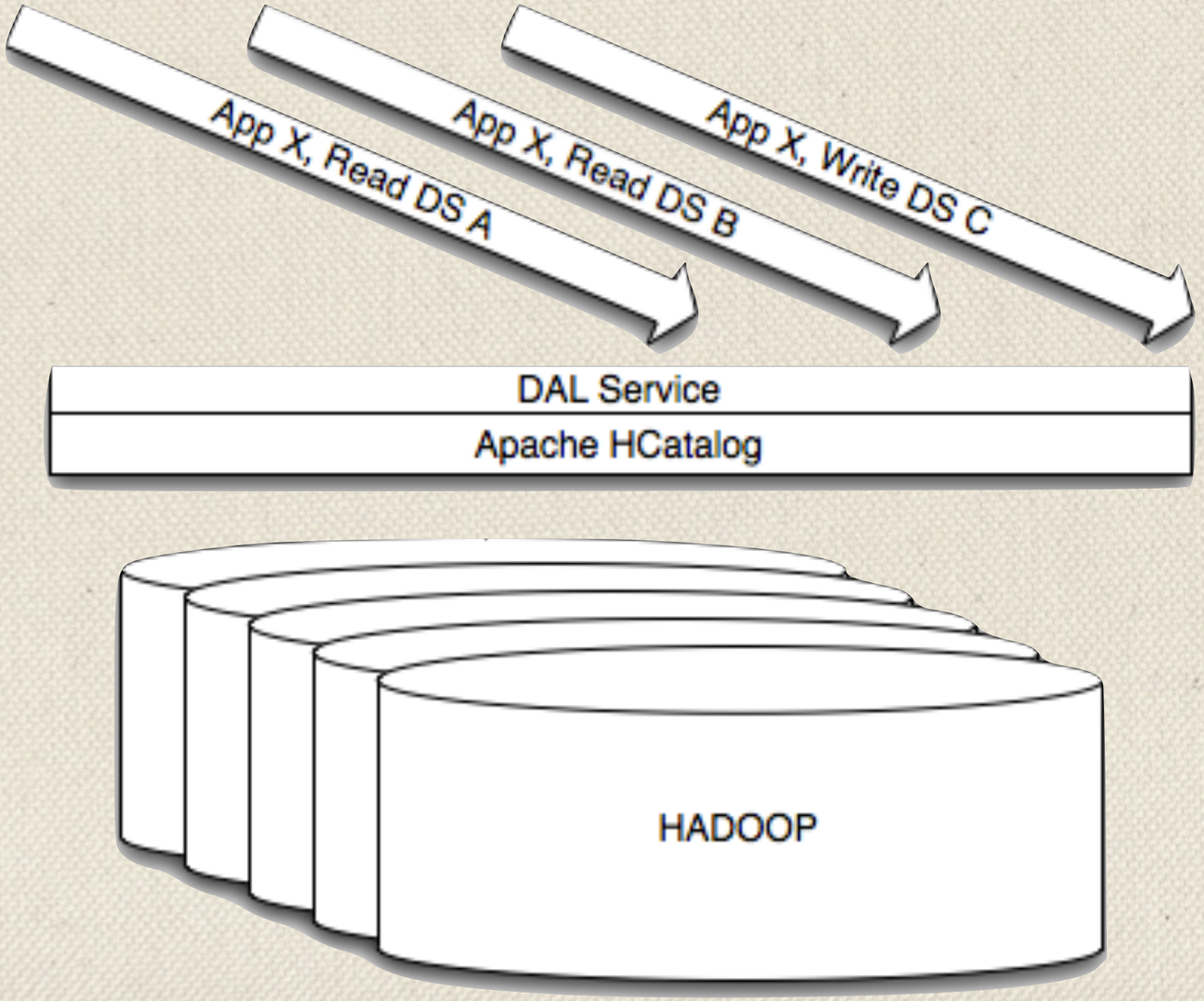
Maintains nested record abstraction
Simplifies format migrations

HCat Answers

HCat Answers

Where is the data?
How do I read it?

DAL: metadata++



DAL

DAL

Records audit information

DAL

Records audit information

Which app did what, and to whom.

DAL

Records audit information

Which app did what, and to whom.

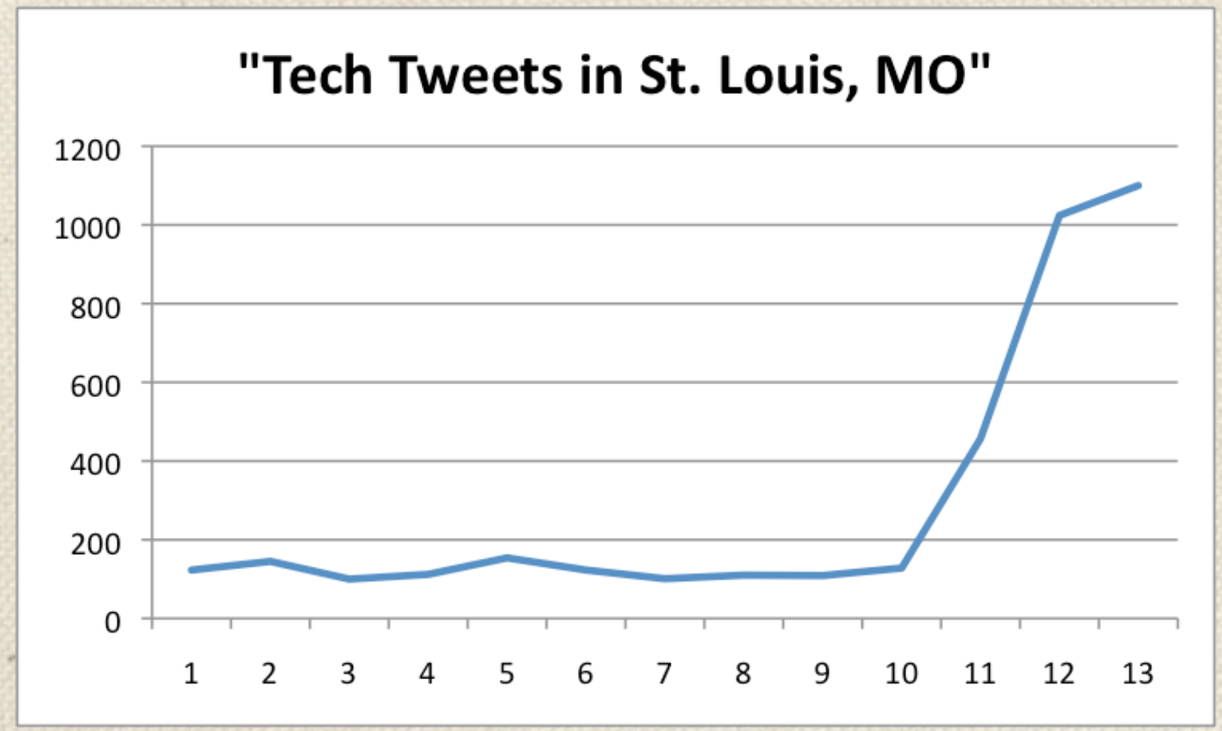
Automated discovery of dependencies and ownership

DAL Answers

DAL Answers

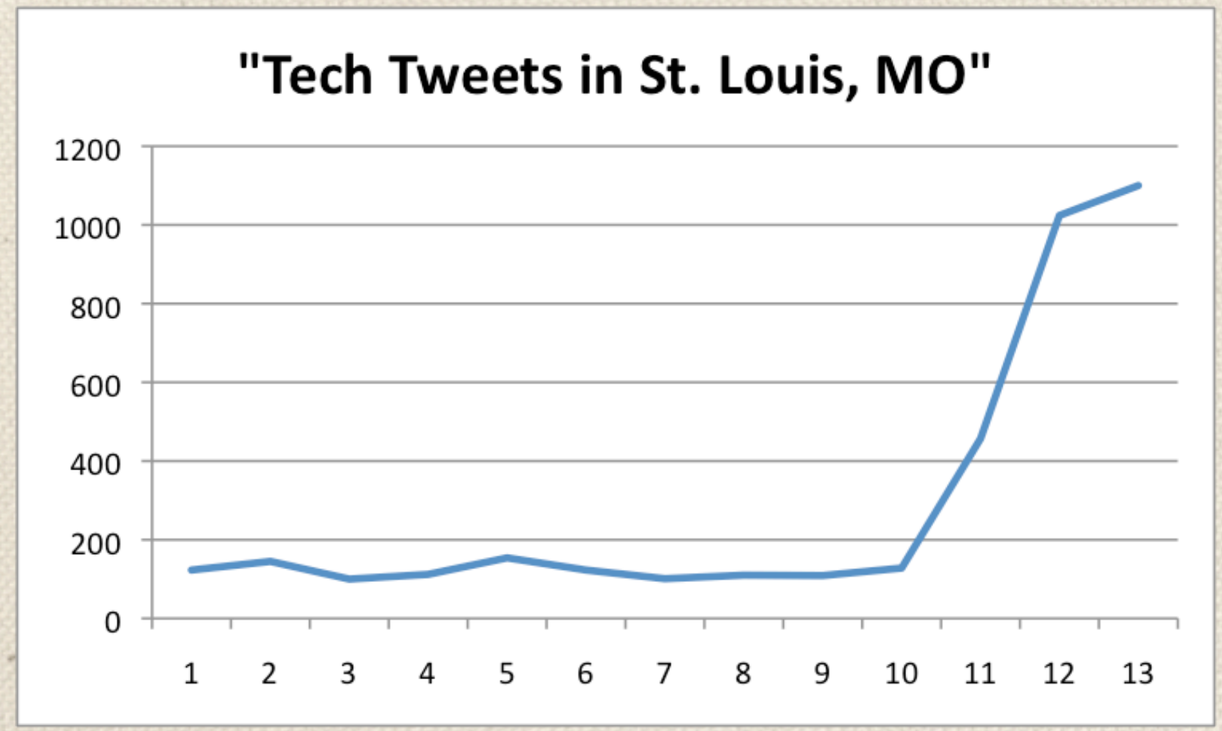
Who produces it?
Known derivations?
Known Dependencies?

Deep Integration



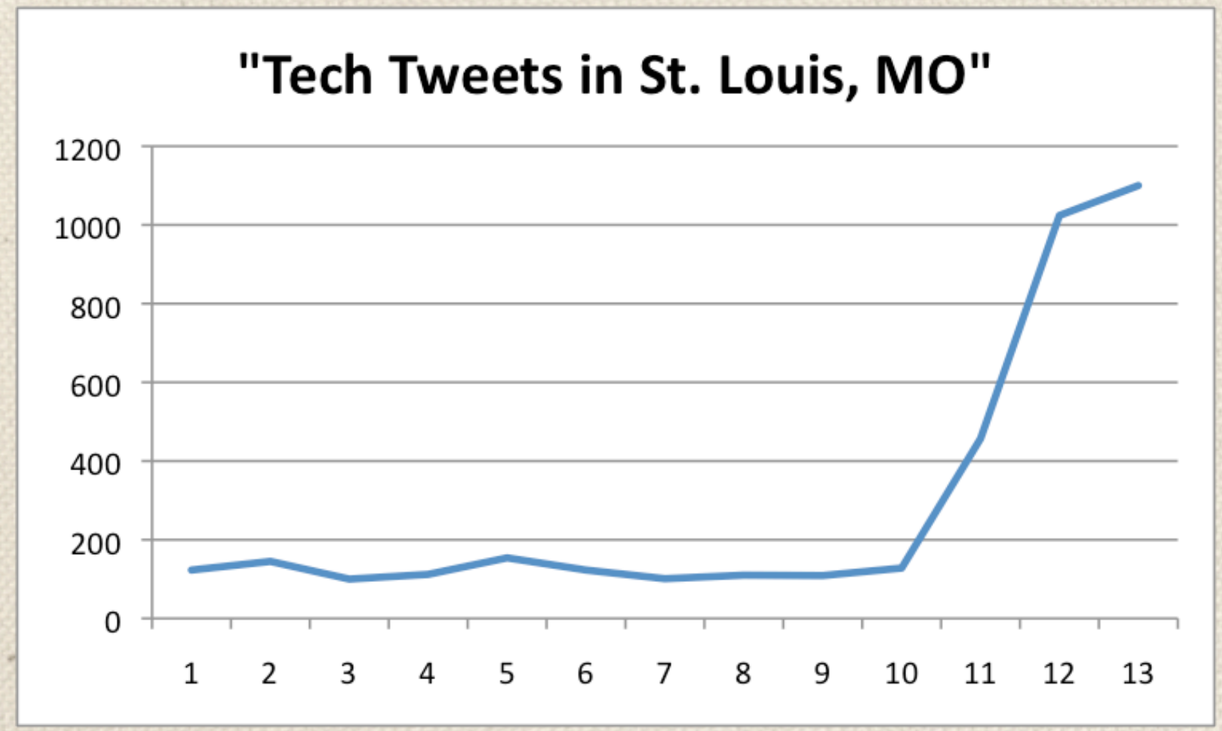
Deep Integration

Map app ids to Git, CI, deploy systems.



Deep Integration

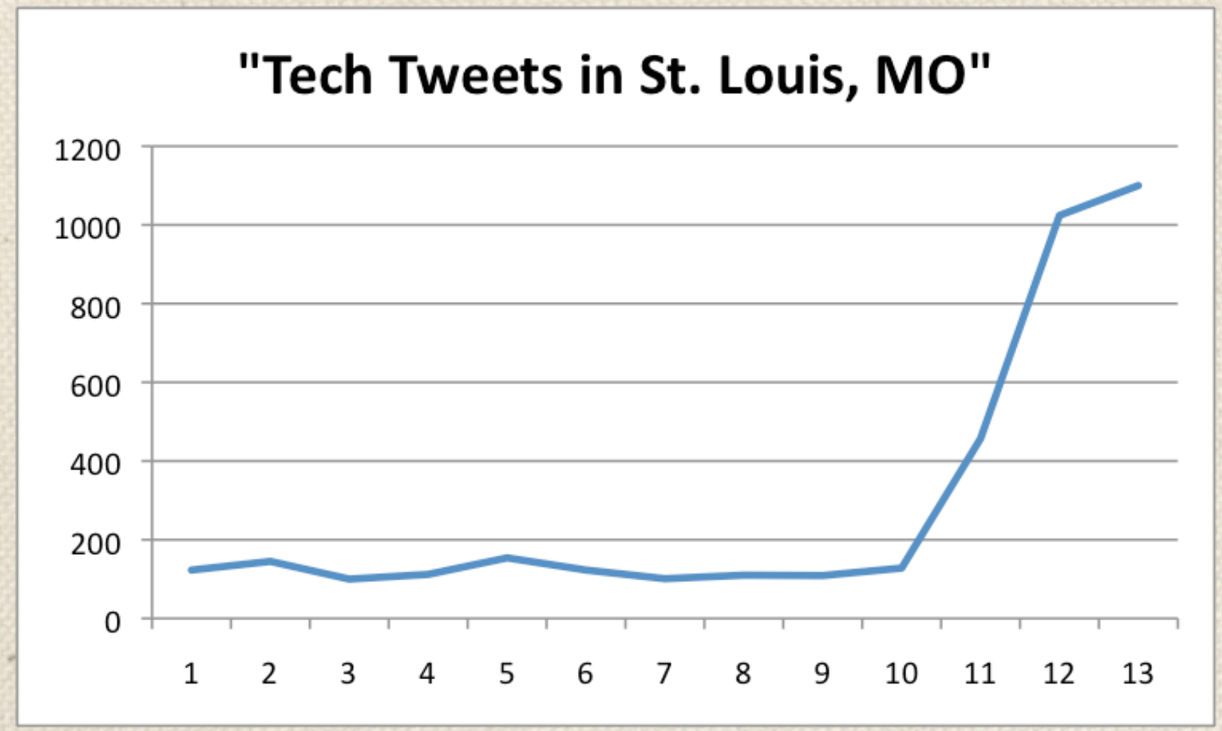
Map app ids to Git, CI, deploy systems.



Classification bug?

Deep Integration

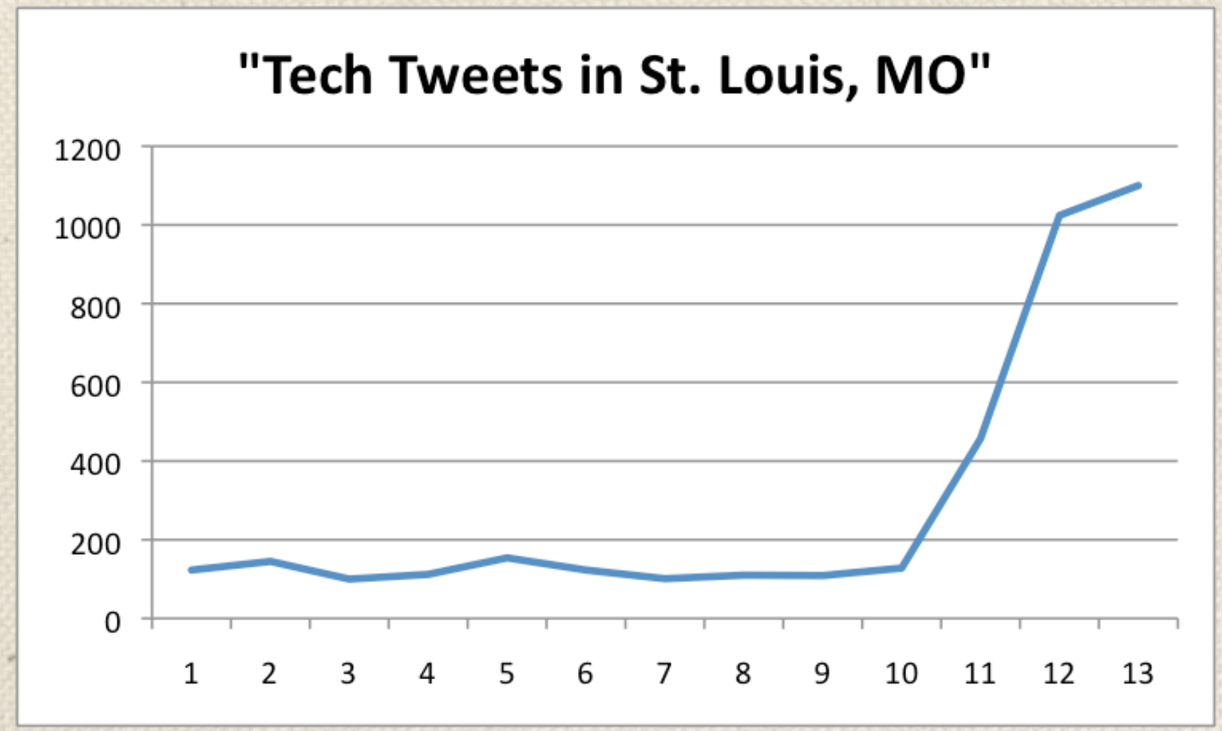
Map app ids to Git, CI, deploy systems.



Classification bug?
Geo-coding bug?

Deep Integration

Map app ids to Git, CI, deploy systems.



Classification bug?
Geo-coding bug?
#StrangeLoop?

Would be nice.

Would be nice.

Known anomalies?

Elastic != Unconstrained

Dmitriy Ryaboy
@squarecog

Map-Reduce: a powerful abstraction that helps programmers create non-performant code and run it at extremely large scale.

Reply Delete Favorite

26 RETWEETS | **10** FAVORITES

6:52 PM - 14 Dec 11 · Embed this Tweet

Make it easy to see cost.

Make it easy to see cost.
A Data Scientist's job is insights.

Make it easy to see cost.
A Data Scientist's job is insights.
A Product Engineer's job is product.

Make it easy to see cost.
A Data Scientist's job is insights.
A Product Engineer's job is product.
If they see egregious cost, they'll fix.

Make it easy to see cost.
A Data Scientist's job is insights.
A Product Engineer's job is product.
If they see egregious cost, they'll fix.
If they don't, you'll have to keep buying servers.

Cloudera Manager

cloudera manager | Services | Hosts | Activities | Logs | Events | Reports | Search | Support | Help | natty

Sep 25 8:10 AM | Sep 25 8:20 AM | Sep 25 8:30 AM | Sep 25 8:40 AM | Sep 25 8:50 AM | Sep 25 9:00 AM | Sep 25 9:10 AM | Sep 25 9:20 AM | Sep 25 9:30 AM | 30m | 1h | 2h | 6h | 12h | 1d

Activities : mapreduce1 »

(September 25 2012, 9:00 AM - 9:30 AM PDT)

hudson_20120925084545_d1bcc5a8-6eec-480a-8555-58d290fd6ad9 | Children | Details | Compare | Zoom To Duration

Activities | All Activities

Status	Type	ID	Name	User	Duration	Map Progress	Reduce Progress	Current User CPUs	Current System CPUs	Residen memor
✓		job_201209241157_0039	SELECT SUM(cast....100(Stage-6)	hudson	37m, 49s	100.0%	100.0%	0	0	0
✓		job_201209241157_0040	SELECT SUM(cast....100(Stage-7)	hudson	21s	100.0%	100.0%	0	0	0
✓		job_201209241157_0041	SELECT SUM(cast....100(Stage-1)	hudson	15.9s	100.0%	100.0%	0	0	0
✓		job_201209241157_0042	SELECT SUM(cast....100(Stage-2)	hudson	16.8s	100.0%	100.0%	0	0	0
✓		job_201209241157_0043	SELECT SUM(cast....100(Stage-3)	hudson	15.3s	100.0%	100.0%	0	0	0
✓		job_201209241157_0044	SELECT SUM(cast....100(Stage-4)	hudson	16.4s	100.0%	100.0%	0	0	0

Charts

Tasks Running hudson_20120925084545_d1bcc5a8-6eec-480a-8555-58d290fd6ad9 Running maps hudson_20120925084545_d1bcc5a8-6eec-480a-8555-58d290

1,200 maps | 800 maps | 400 maps | 0 maps | Sep 25 9:10 AM | Sep 25 9:20 AM

CPU Usage hudson_20120925084545_d1bcc5a8-6eec-480a-8555-58d290fd6ad9 Current User CPUs hudson_20120925084545_d1bcc5a8-6eec-480a-8555-58d29

0.8 CPUs / s | 0.4 CPUs / s | 0.0 CPUs / s | Sep 25 9:10 AM | Sep 25 9:20 AM

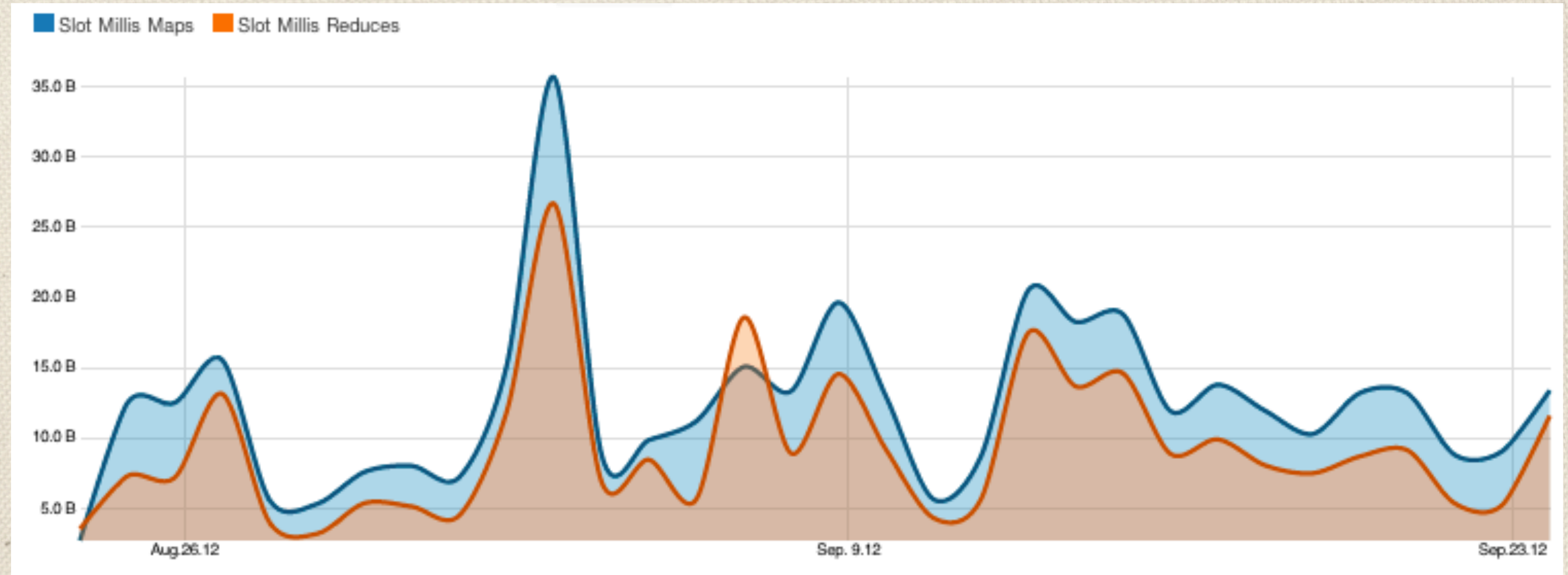
Memory Usage hudson_20120925084545_d1bcc5a8-6eec-480a-8555-58d290fd6ad9 Resident memory

745.1GiB | 558.8GiB | 372.5GiB | 186.3GiB | 0.0B | Sep 25 9:10 AM | Sep 25 9:20 AM

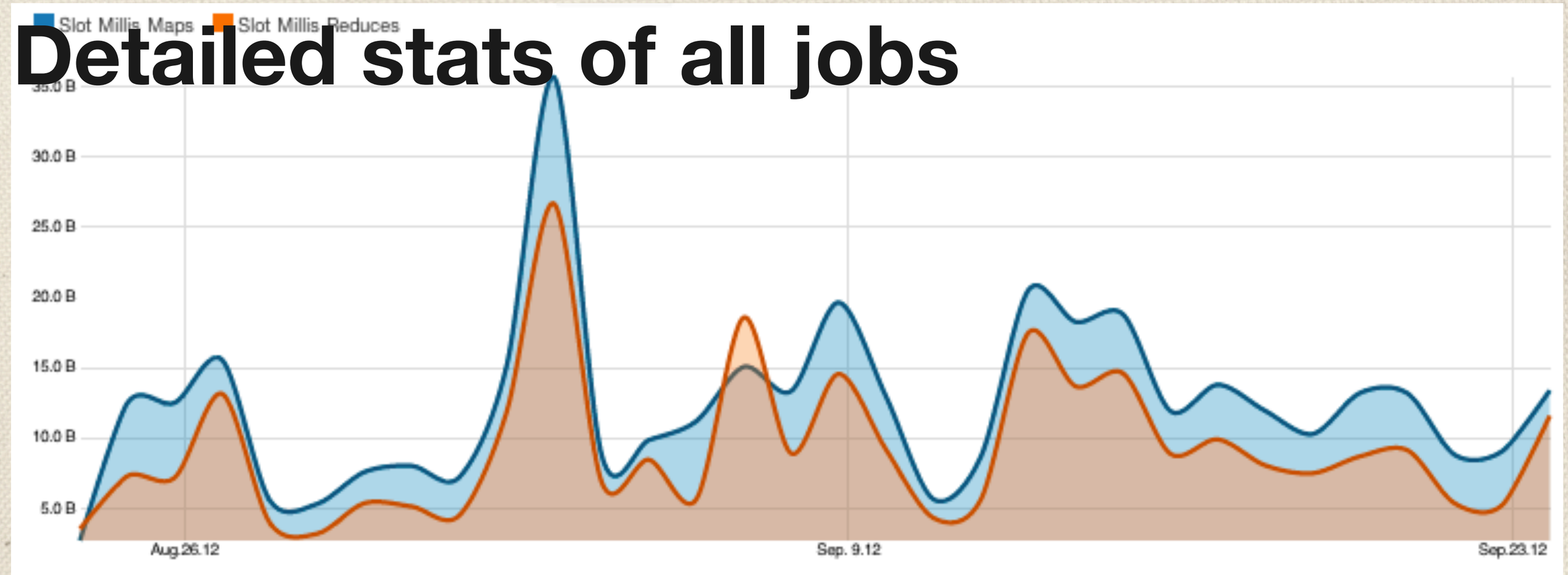
Events

Event Type All | Severity All | Hosts Enter a host name | Keywords Enter keywords (no wildcards | Alerts Only | Search

Custom Tooling

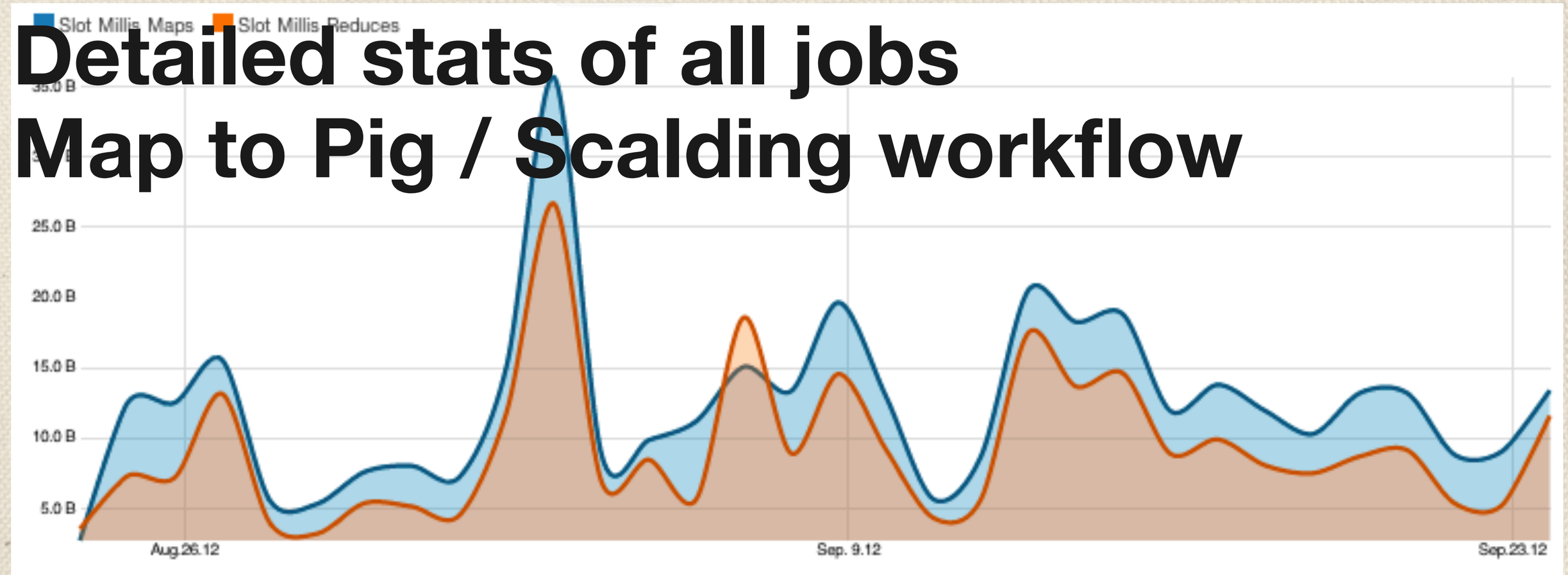


Custom Tooling



Detailed stats of all jobs

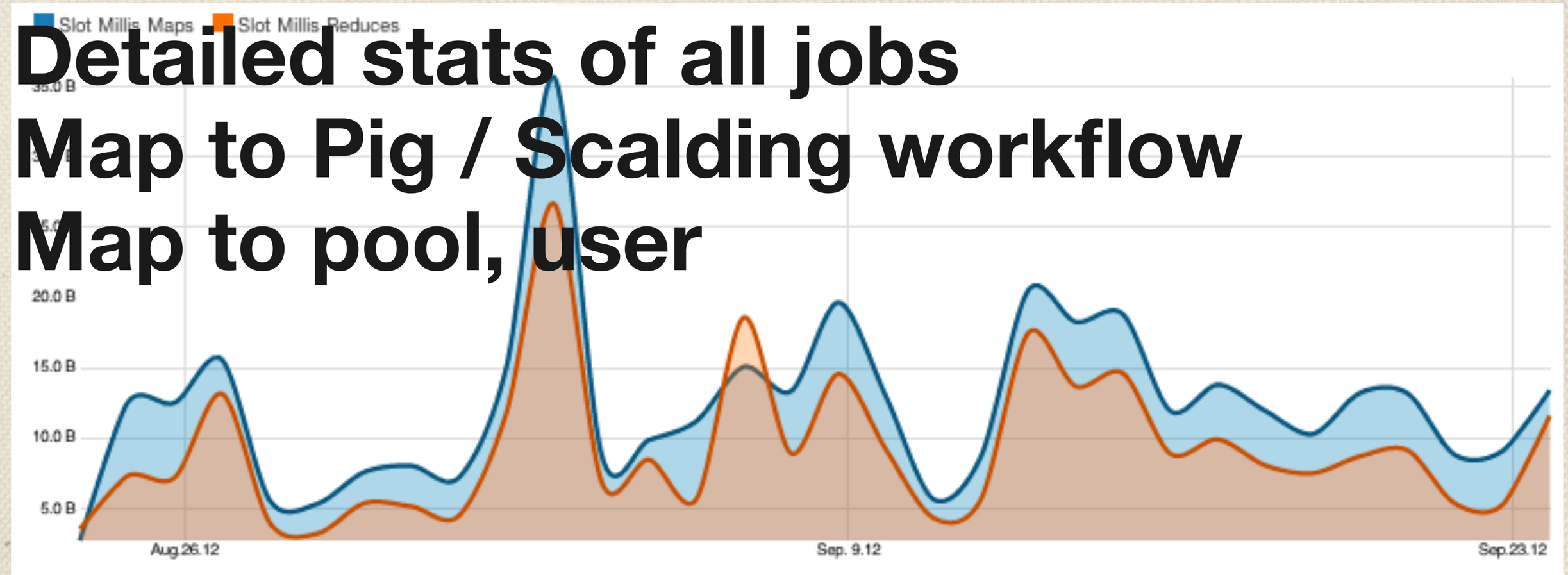
Custom Tooling



Detailed stats of all jobs

Map to Pig / Scalding workflow

Custom Tooling

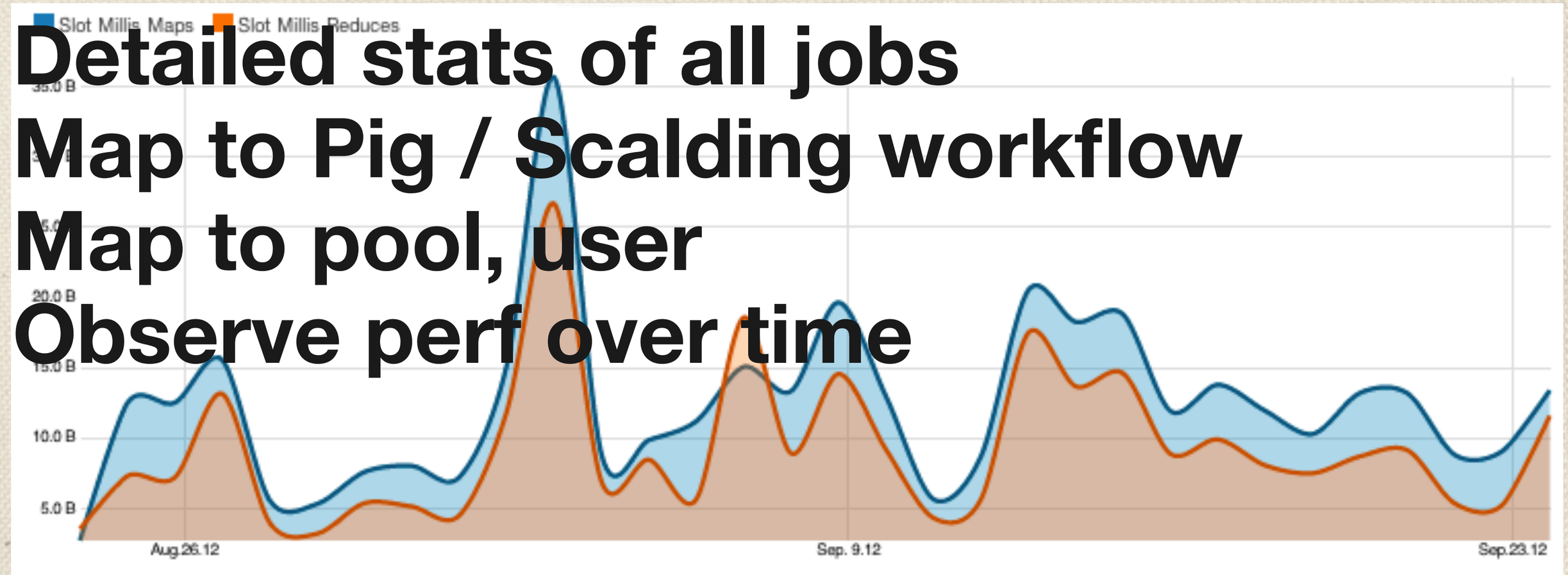


Detailed stats of all jobs

Map to Pig / Scalding workflow

Map to pool, user

Custom Tooling



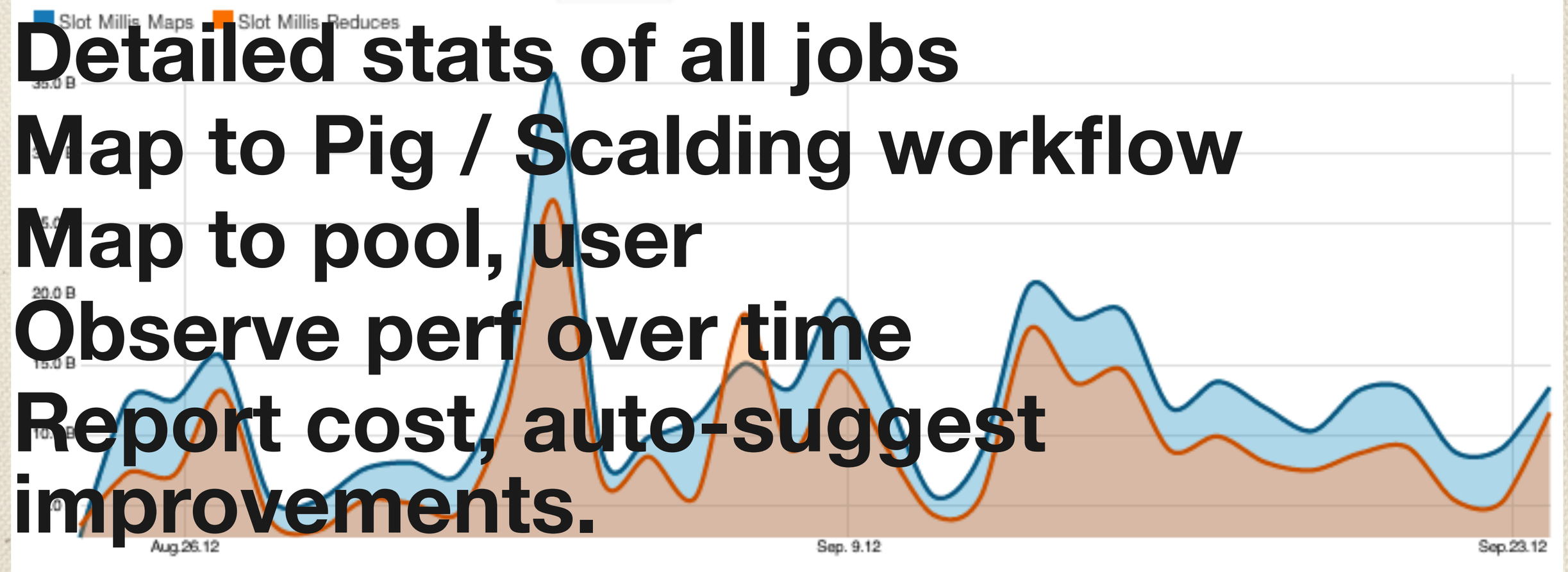
Detailed stats of all jobs

Map to Pig / Scalding workflow

Map to pool, user

Observe perf over time

Custom Tooling



Detailed stats of all jobs

Map to Pig / Scalding workflow

Map to pool, user

Observe perf over time

Report cost, auto-suggest improvements.

You are a SPOF.

Your team is a potential bottleneck.
Get out of the way.

Replication is not a Backup

Build tools, not frameworks

Summary

Summary

Data loves a schema

Summary

Data loves a schema

Make data sources discoverable

Summary

Data loves a schema

Make data sources discoverable

Make costs visible

Questions?

Image Source: http://disintegrator.co.uk/images/foto24.html