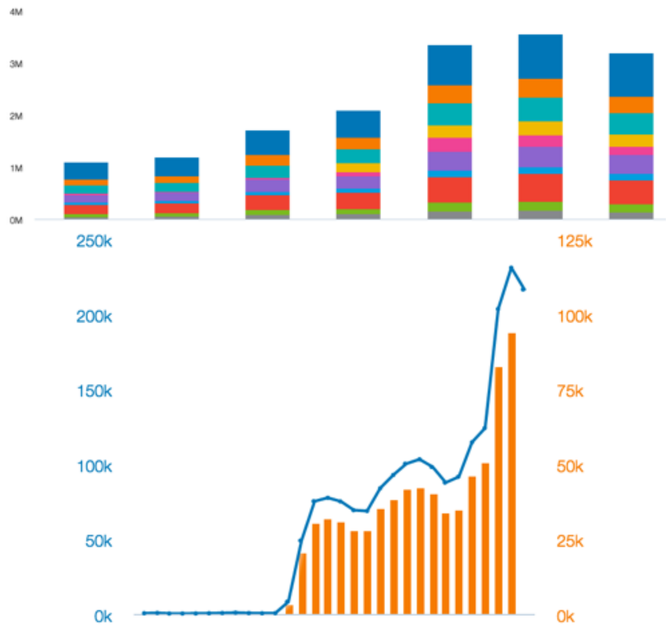
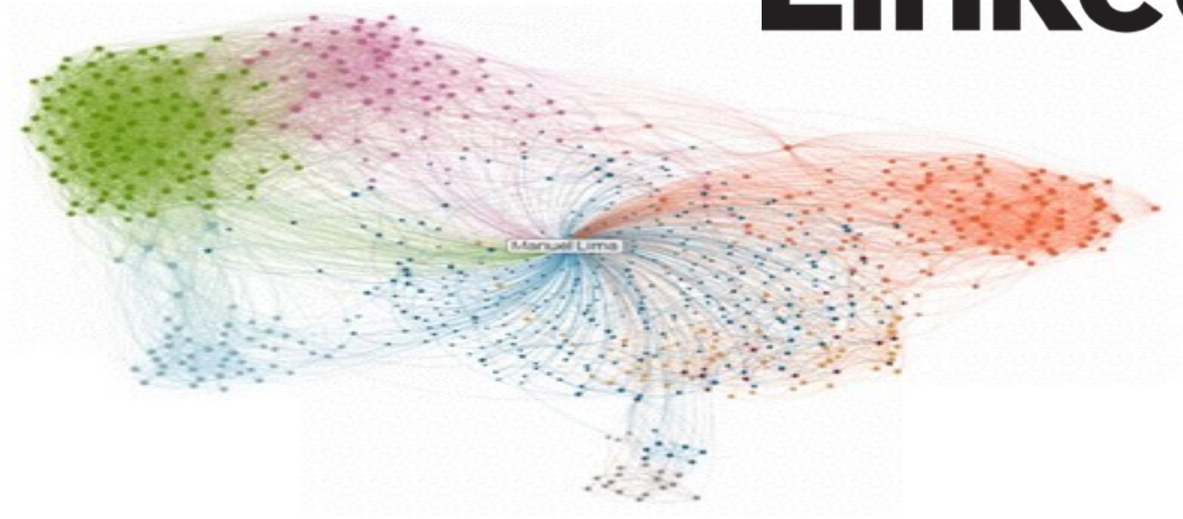
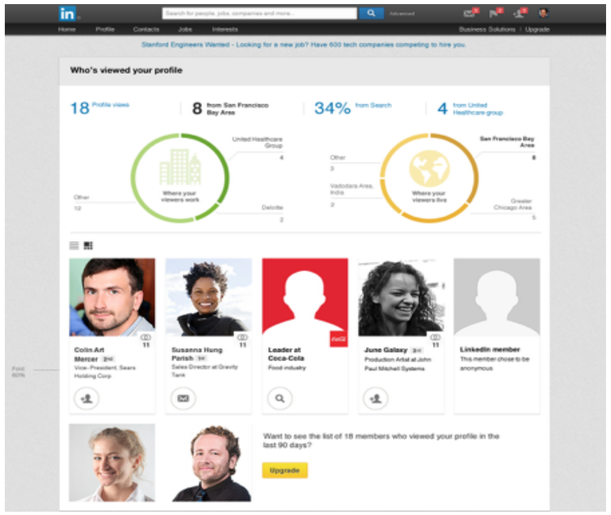




Self-Serve Reporting Platform on Hadoop

Shirshanka Das

Strata Singapore 2015



People you may know

Jiajun Yao
Student at CMU
Connect • Skip

Sumedh Chaudhry
Co-Founder at NewsBytes A...
Connect • Skip

Pulse The news and insights you need to know.

Your News | Top Posts | Discover

Ads You May Be Interested In

Avg Offer for Devs: 136k
Want to move out of your industry? Work with a new stack? Try Hired today!

Bring the "T" back to ETL
Discover a smarter approach to integrate all your data – Download now!

NRIs make huge profits
NRIs are making huge profits through crowd funding in India. Learn more.

Companies you may want to follow

nanobi analytics
Information Technology & Services
✚ Follow

The Chemistry Group
Management Consulting
✚ Follow

One Network Enterprises
Computer Software
✚ Follow

Jobs you may be interested in

Preferences:

Your job activity is private.

vmware
Distributed Systems Engineers- NSX...
Palo Alto, CA, 94301, USA

ancestry.com
Engineering Director, AncestryDNA
San Francisco, California

WD
Senior Director, Content Solution Data...
San Francisco Bay Area

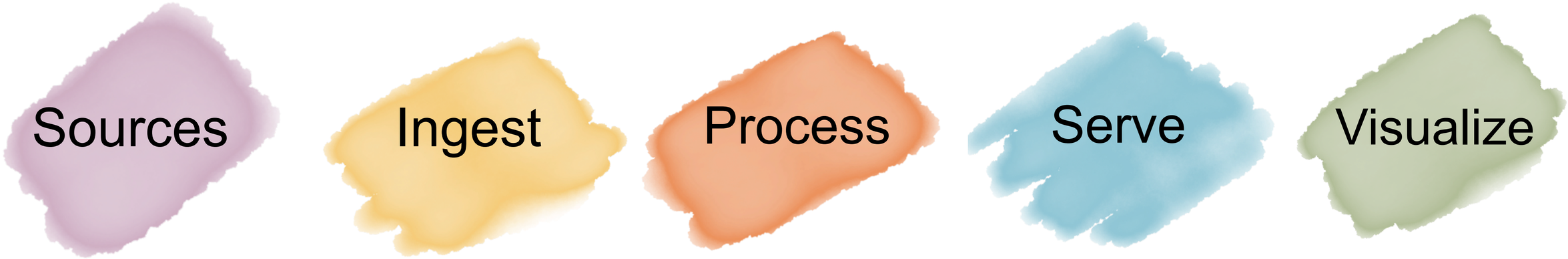
Director - Cloud Infrastructure
San Ramon

EMC²
Director Performance Engineering
Santa Clara, CA

cisco
Director, Engineering
San Jose, CA

Reporting Pipelines

Reporting at LinkedIn: Evolution



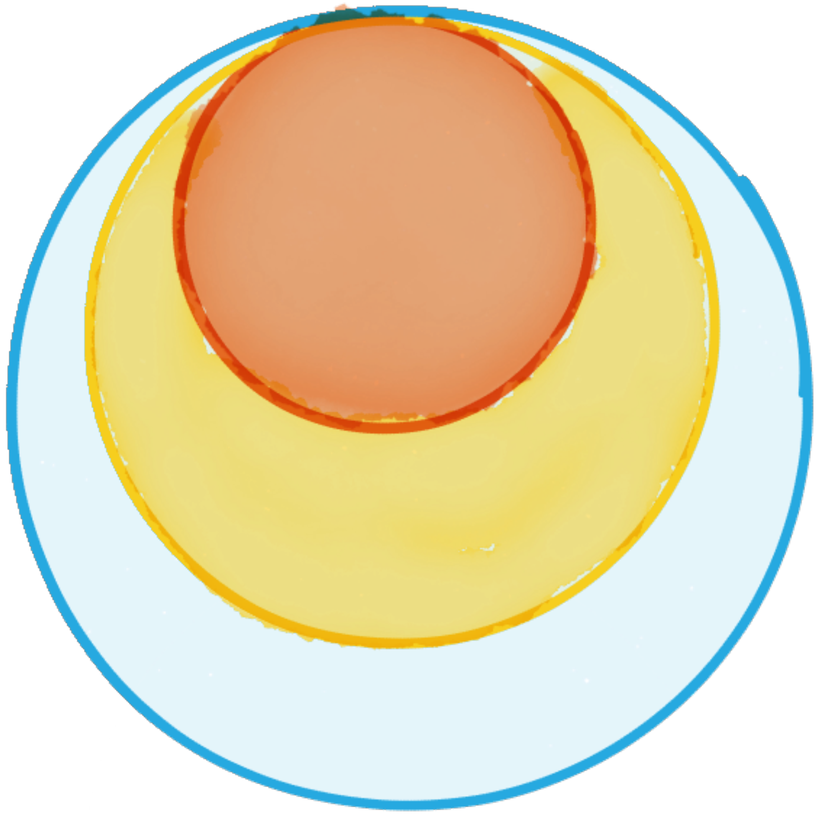
Sources	Ingest	Process	Serve	Visualize
Oracle	INFA + MSTR	+ Scripts	on Teradata	MSTR
Kafka	Custom	Jobs on Hadoop	MySQL	Tableau
Espresso	Custom		Voldemort	Internal Tools
External	Custom		Pinot	

Infra Scale

Number of Hadoop clusters: 12
Total number of machines: ~7k
Largest Cluster: ~3k machines

Data volume generated per day: XX Terabytes
Total accumulated data: XX Petabytes

People Scale

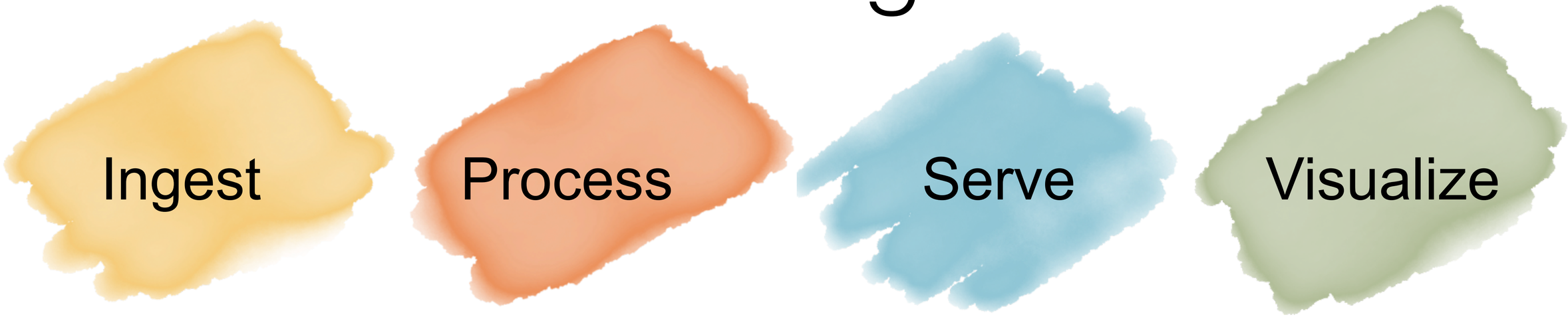


Reporting Platform Team: ~10
Core Warehouse Team: 1x

Data Scientists: 10x
Business Analysts: 10x
Product Managers: 10x

Sales and Marketing: 100x

Challenges



Disjointed efforts, unreliable systems

Unpredictable SLA across all systems

Fragmented data pipelines with inconsistent data

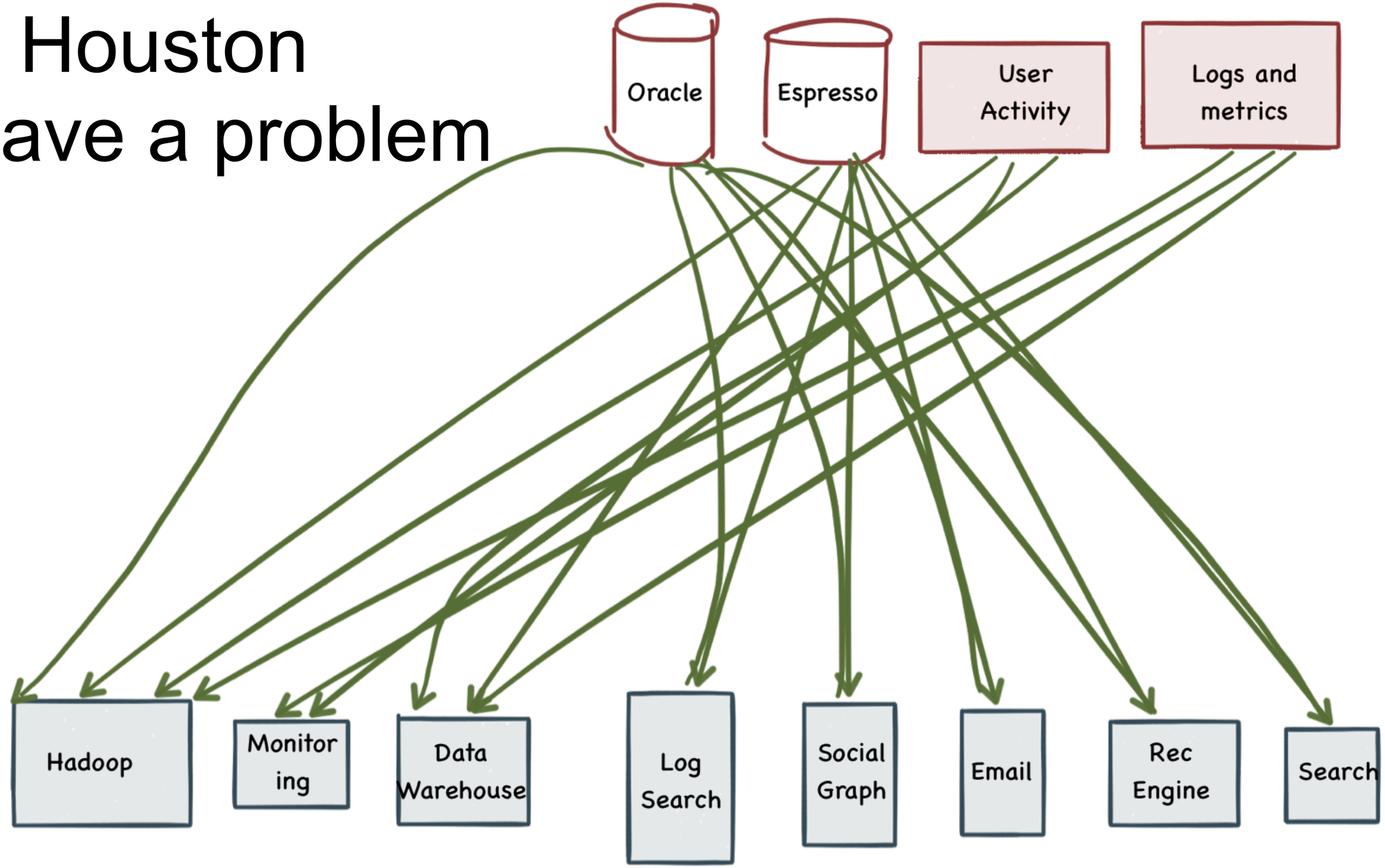
Ingest

Process

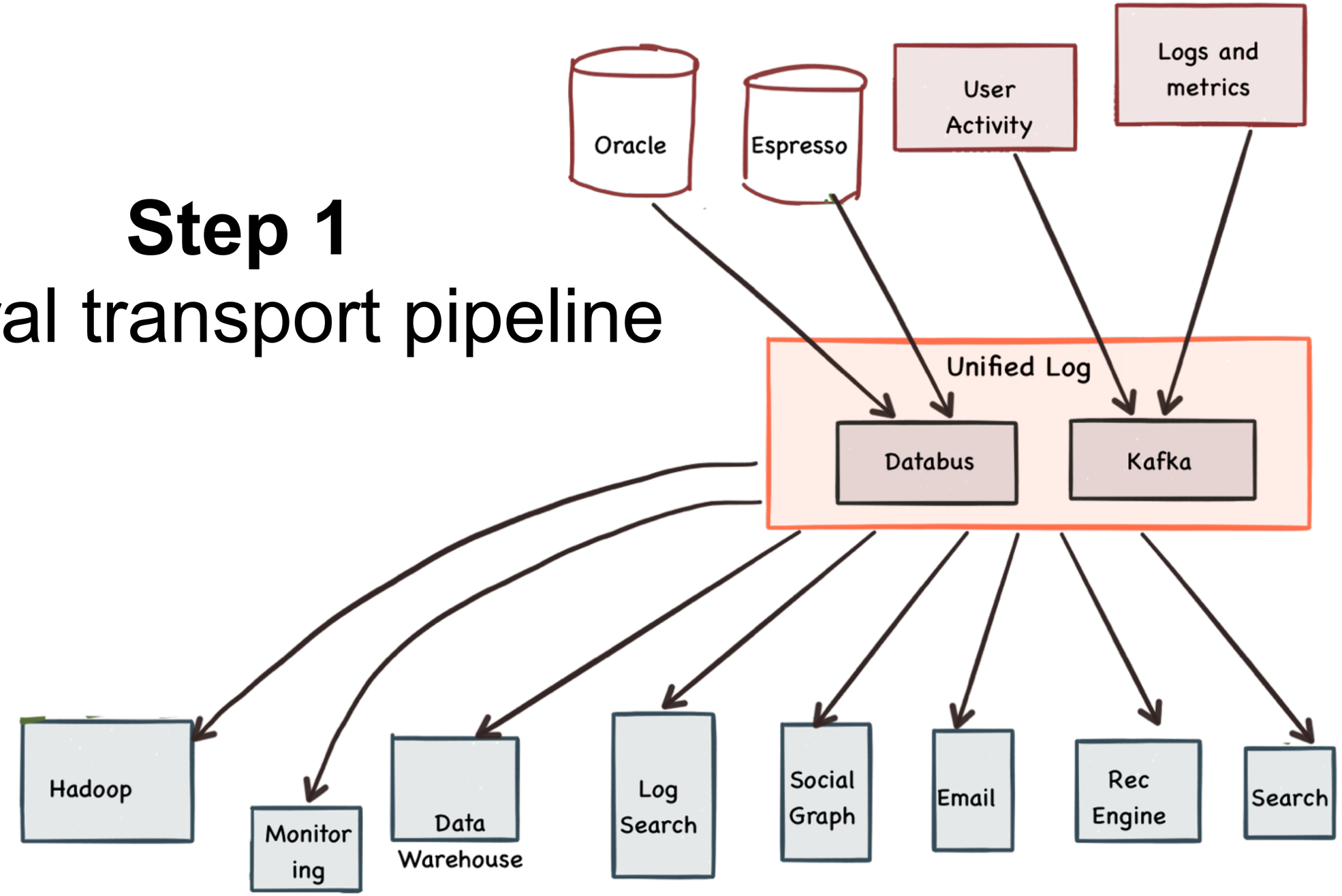
Serve

Visualize

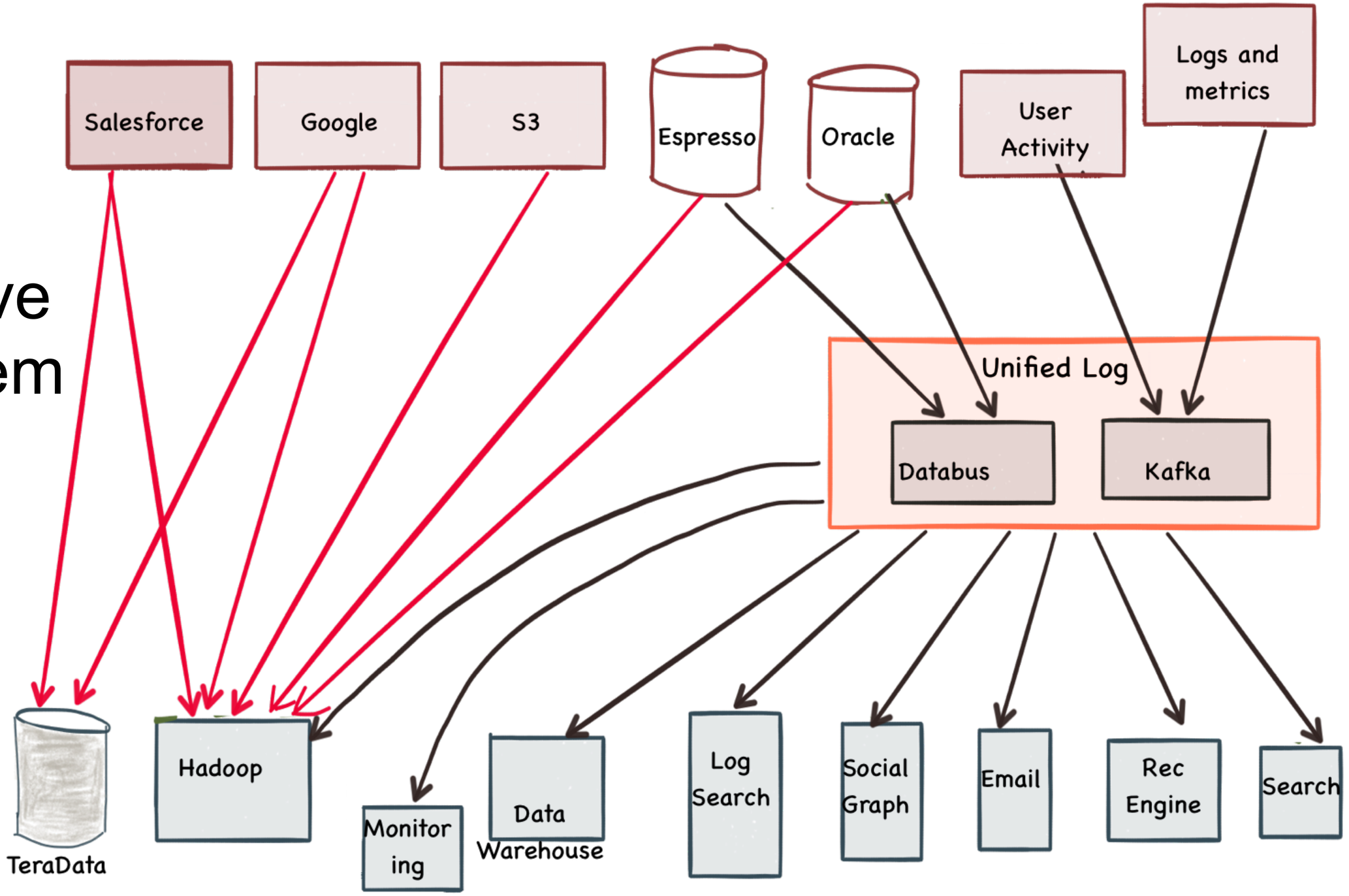
Houston
we have a problem



Step 1
Central transport pipeline

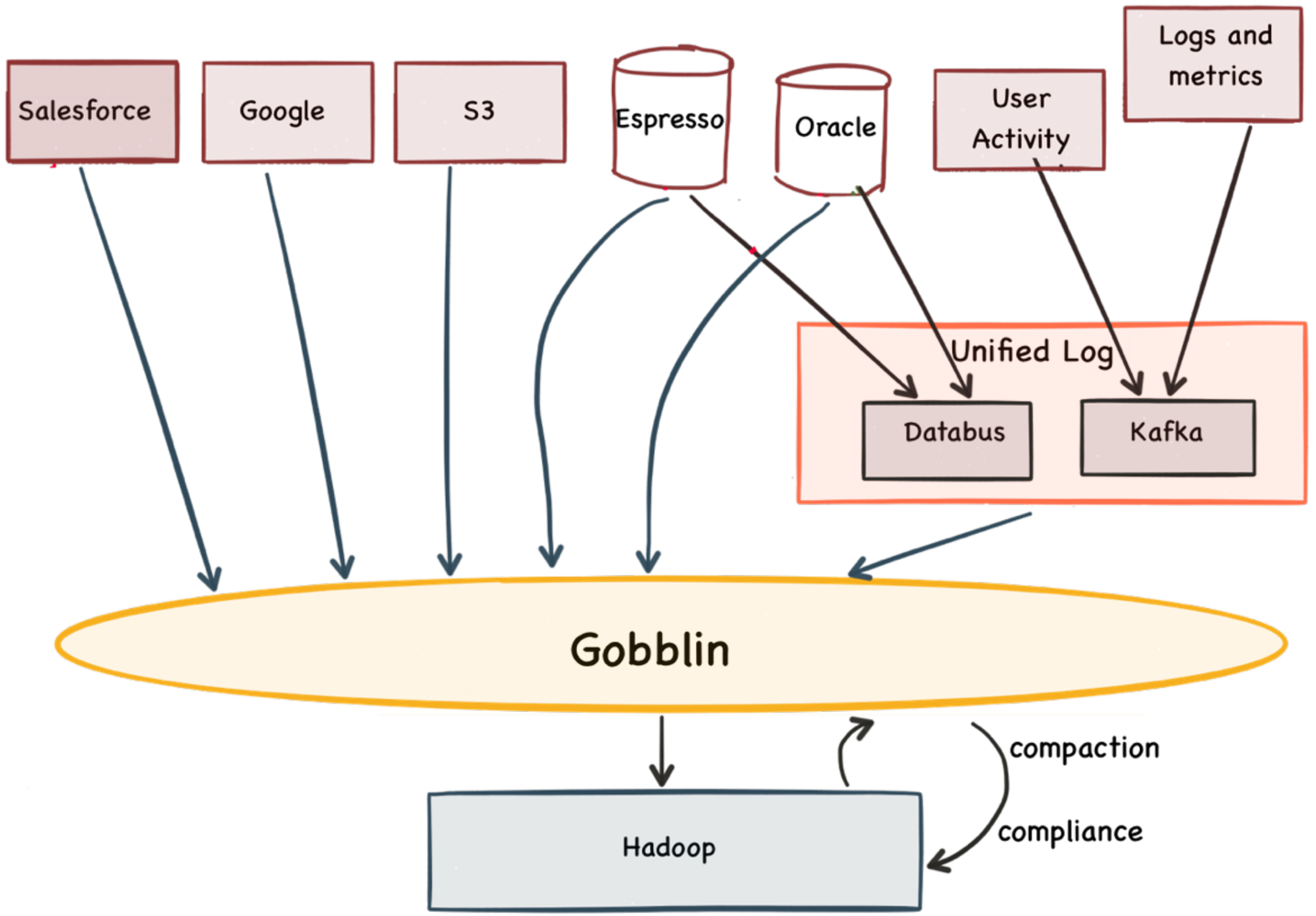


Still have a problem



Step 2

Central Ingestion Framework



Diverse Sources

Stream + Batch

Data Quality

SFTP

REST

JDBC

Open source @ github.com/linkedin/gobblin

In production @ LinkedIn, Intel, Swisscom, NerdWallet

@LinkedIn

~20 distinct source types

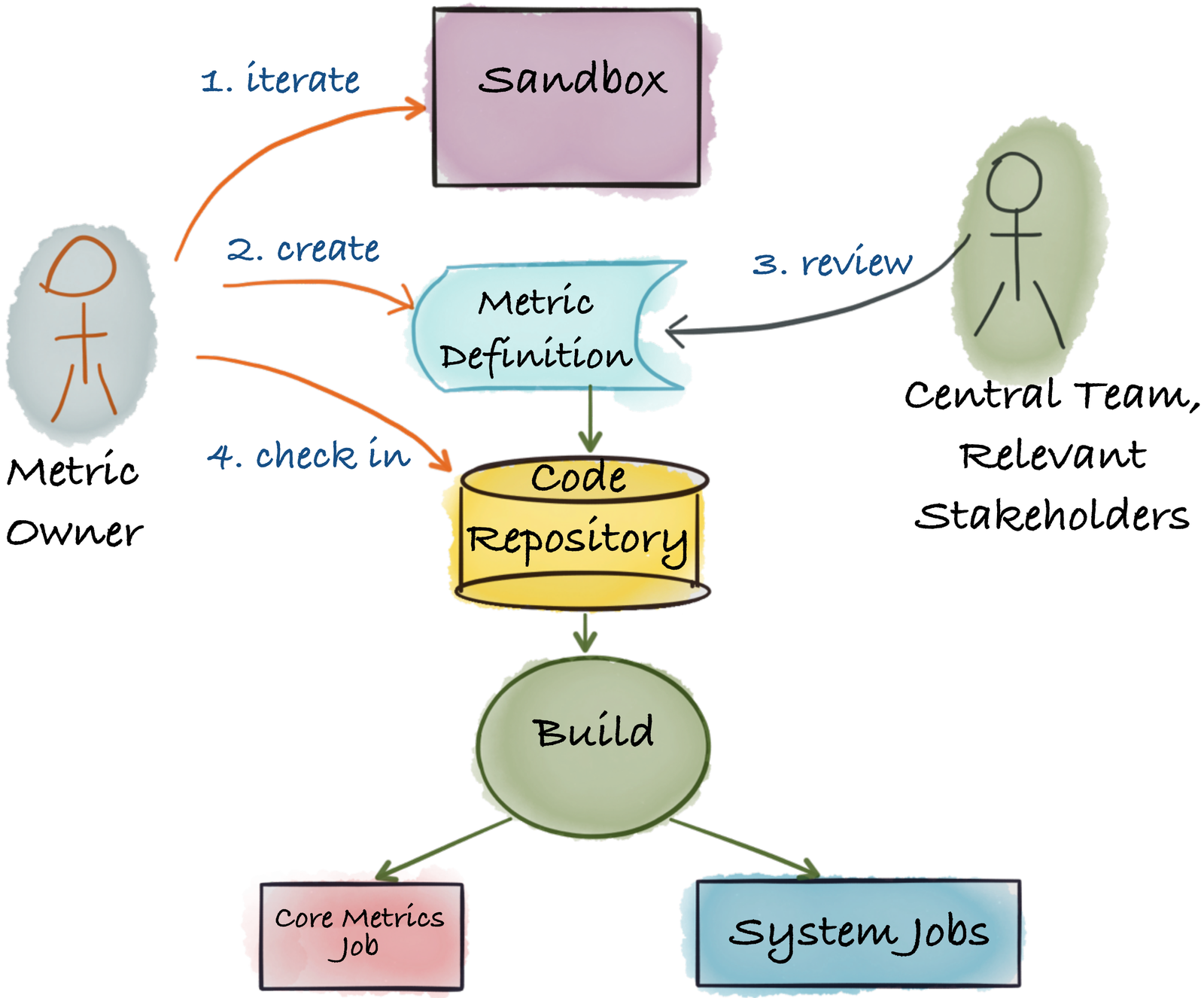
Hundreds of TB per day

Hundreds of datasets

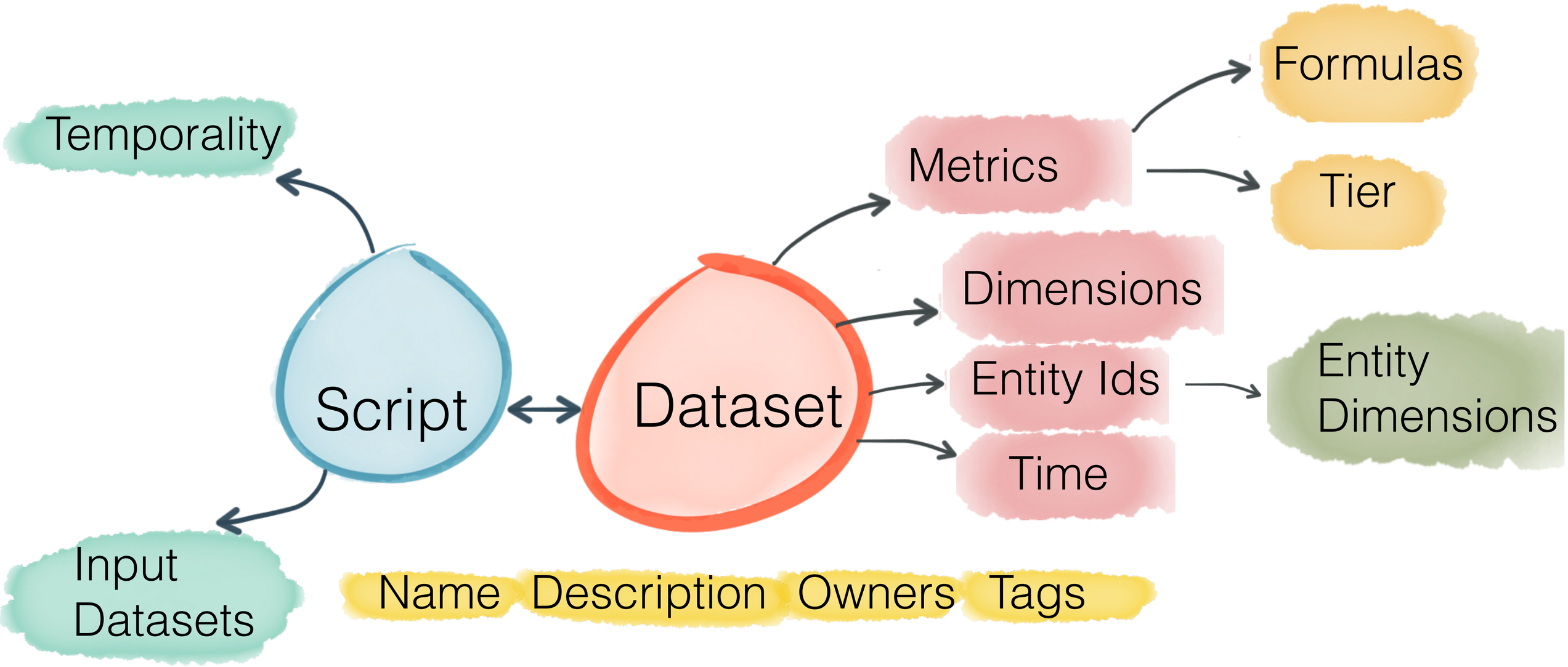
Unified
Metrics
Platform

Requirements

Workflow


1. iterate
Sandbox
2. create
Metric Definition
3. review
Central Team, Relevant Stakeholders
Metric Owner
4. check in
Code Repository
Build
Core Metrics Job
System Jobs

Metric Definition



An example: video play analysis

```
name: "video"

description: “Metrics for video tracking”

label: “video”

tags: [flagship, feed]

owners: [jdoe, jsmith]

enabled: true

retention: 90d

timestamp: timestamp

frequency: daily

script: video_play.pig

output_window: 1d
```

```
input_datasets
[
  {
    name: actionsRaw
    path: Tracking.ActionEvent
    range: 1d
    }
]
```

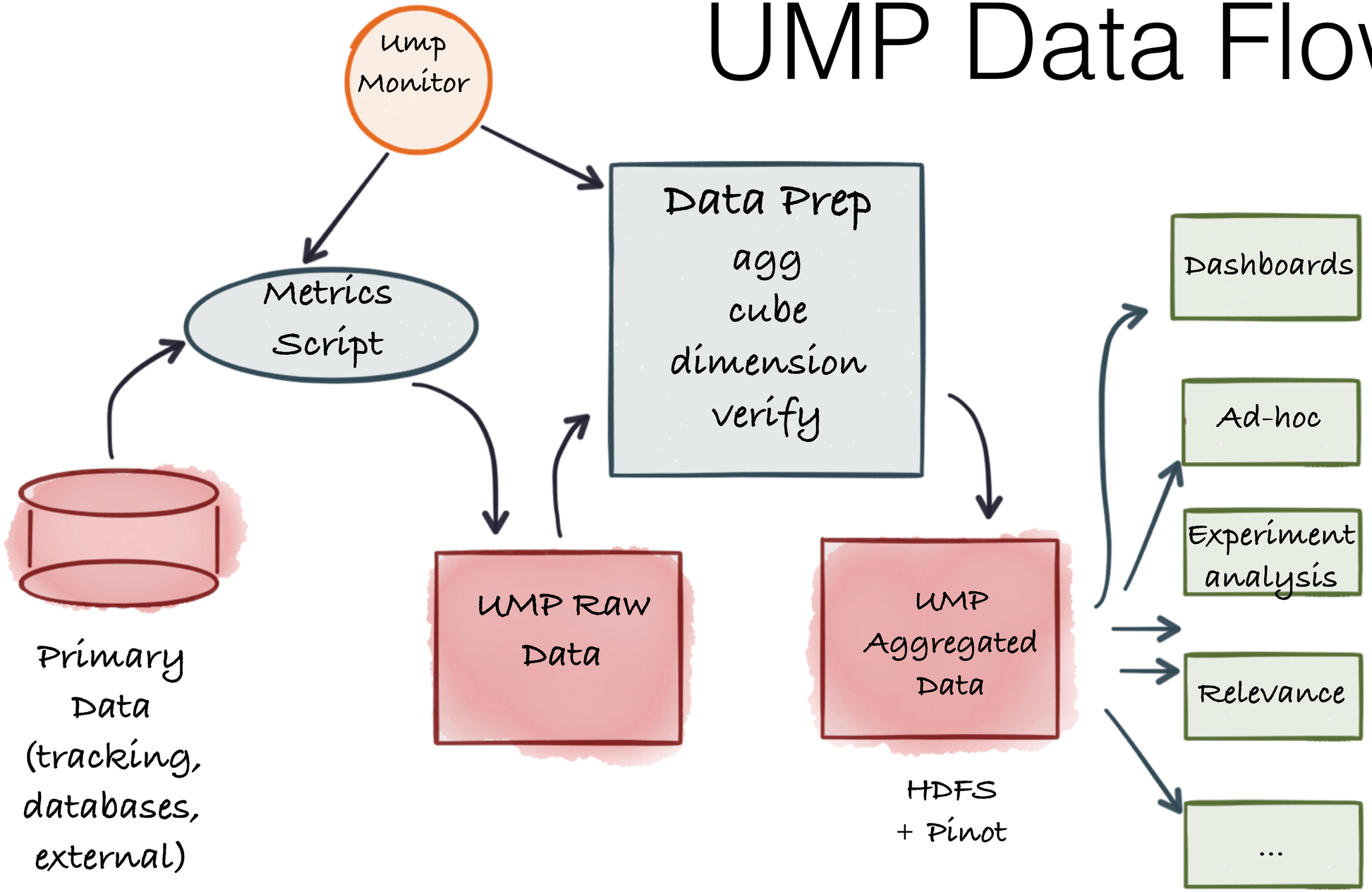
```
dimensions:[
   {
     name: platform
     doc: “phone, tablet or desktop"
   }
   {
     name: action_type
     doc: “click play or auto-play“
   }
 ]
```

An example contd…

```
entity_ids: [
    {
       name: member_id
       category: member
       }
       {
       name:video_id
       category: video
       }
   ]
```

```
metrics: [
         name: unique_viewers
        doc: “Count of unique viewers”
         formula: “unique(member_id)”
         tier: 2
         good_direction: "up"
       }
       {
         name: play_actions
         doc: “Sum of play actions"
         tier: 2
         formula: “sum(play_actions)"
         good_direction: "up"
       }
   ]
```

UMP Data Flow



UMP by the numbers

First version in production since early 2014
Significant redesign in 2015

Total amount of data being scanned per day: Hundreds of TBs
Total number of metrics being computed: 2k+
Total number of scripts: ~ 400
Number of authors for these metrics: ~ 200
Maximum number of dimensions per dataset: ~ 30
Number of people responsible for upkeep of pipeline: 2

Learnings so far

Ease of onboarding

Hard when you have > 1000 users with different skill sets
Need great UX to complement developer friendly alternatives

Single source of truth

Not just a technology challenge
Organization needs to rally around it

Operability

Multi-tenant Hadoop pipeline with SLA-s and QoS: hard
Cost 2 Serve: Managing metrics lifecycle is important

The Next Big Things

Bridging streaming and batch
Code-free metrics
Sessions, Funnels, Cohorts
Open source

Ingest

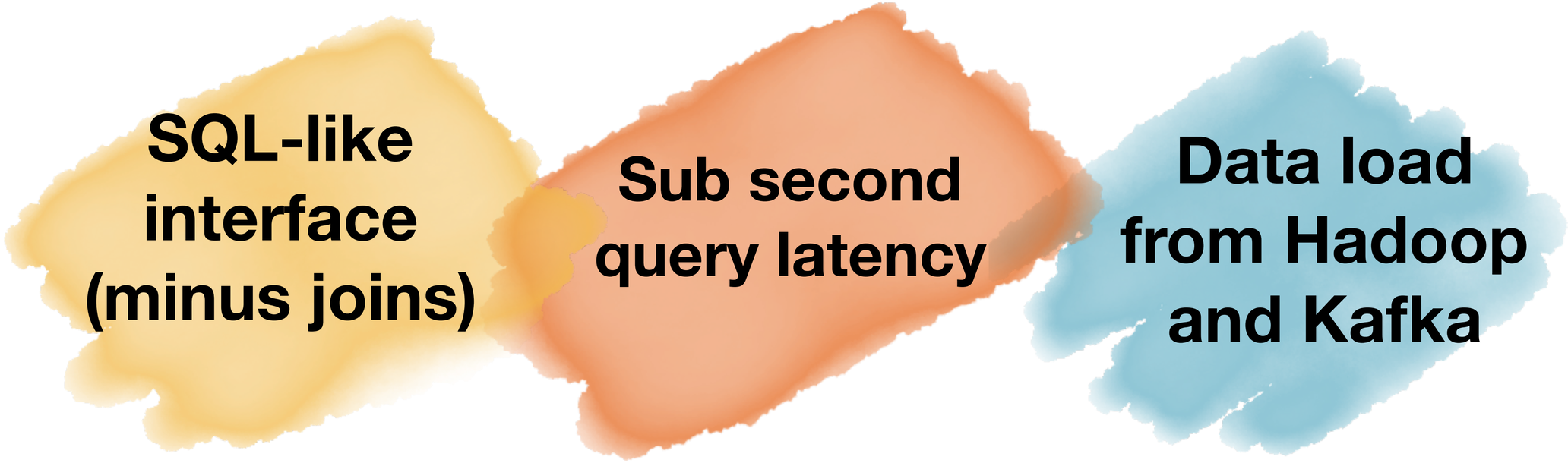
Process

Serve

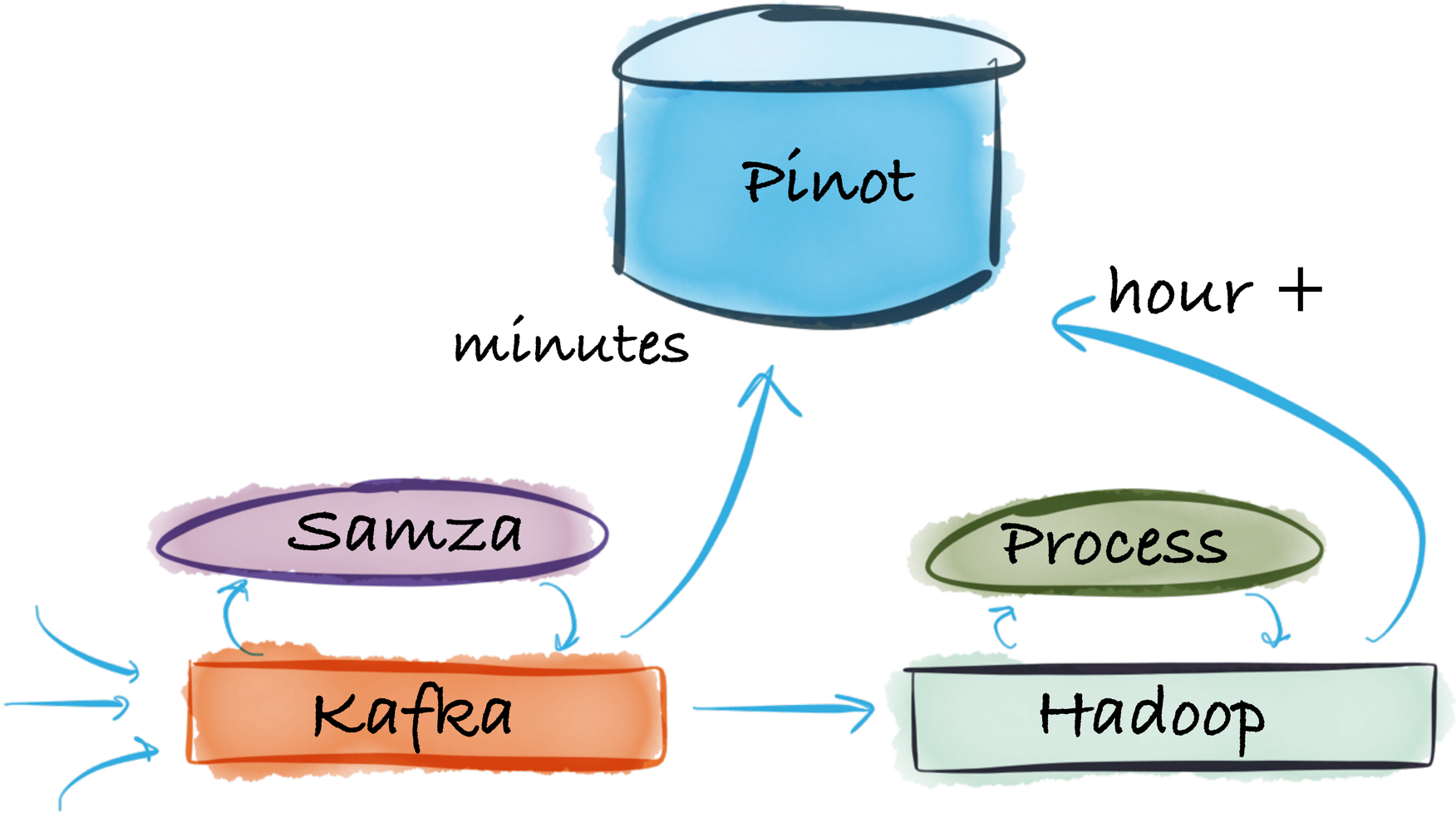
Visualize

Pinot

Capabilities



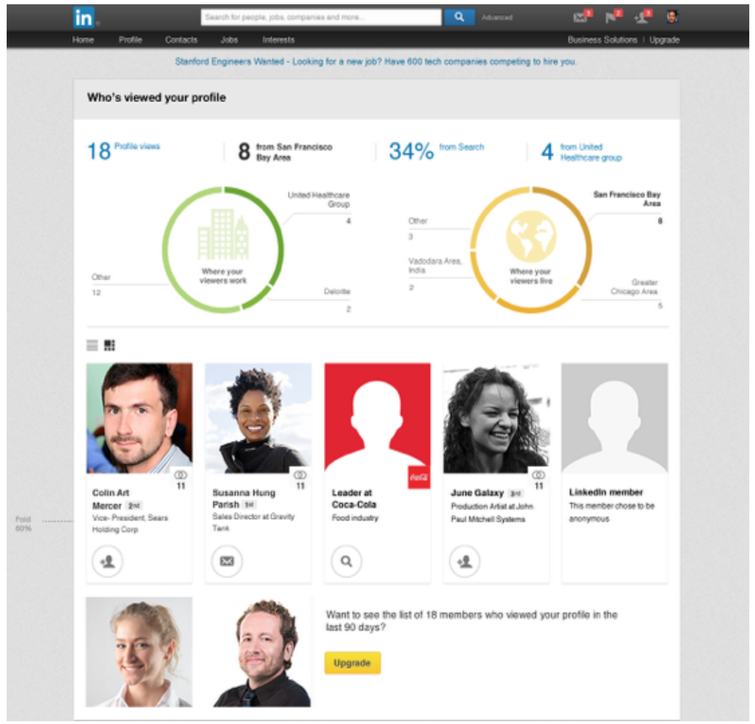
Pinot Data Flow



Pinot@LinkedIn

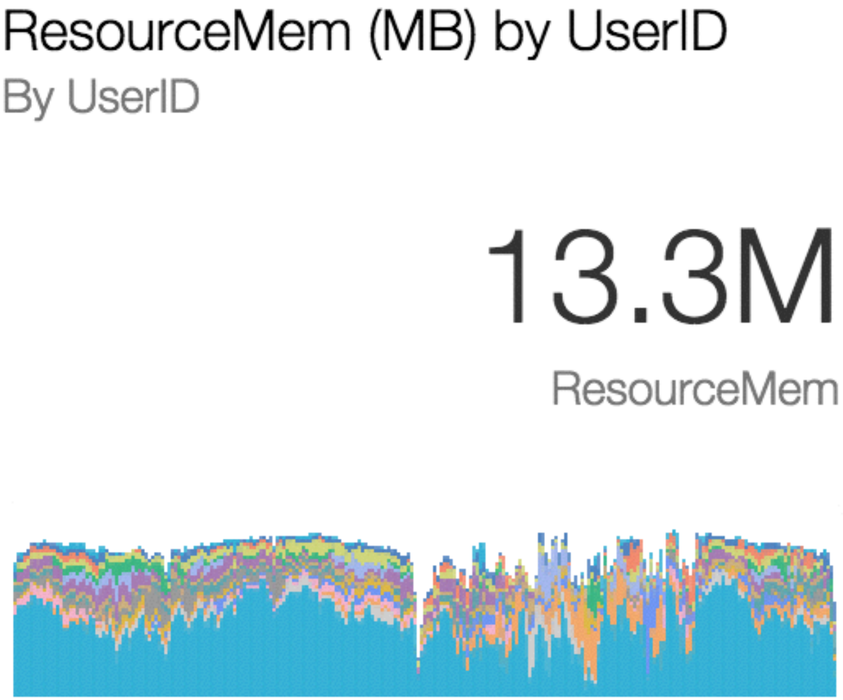
In production since 2012

Open source @ github.com/linkedin/pinot



Site-facing Apps

Reporting dashboards



Monitoring

Ingest

Process

Serve

Visualize

Raptor

Standardize Visualization

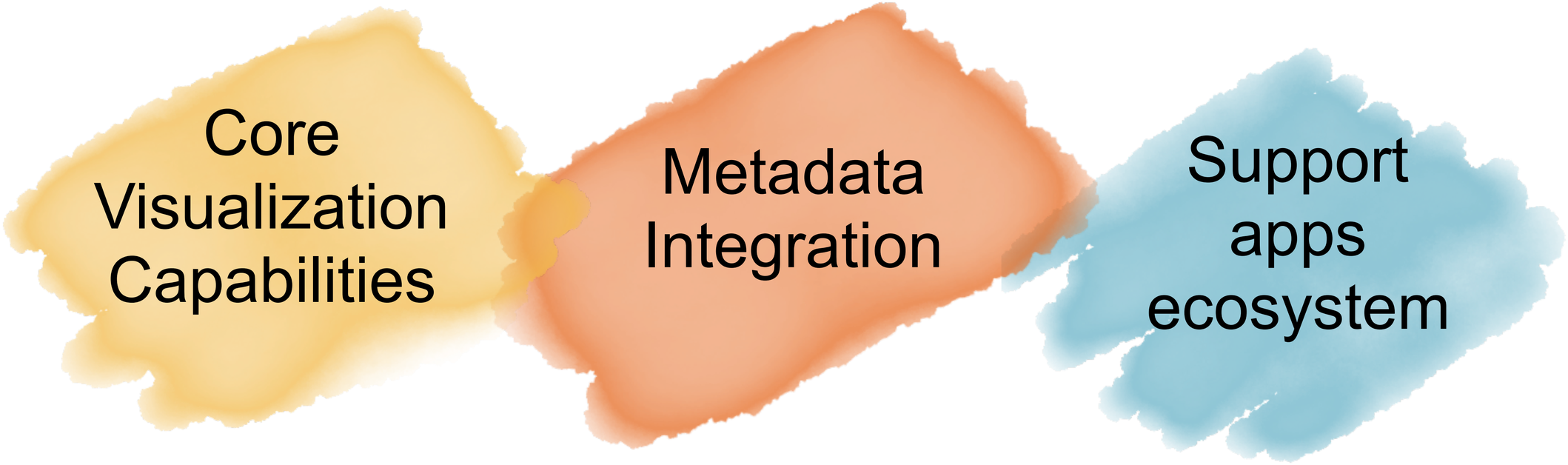
Leverage

- Standalone app, with support for embedding
- Can use existing analytics backend: Pinot

Strategic

- Reduces dependency on 3rd party BI tools
- Closer integration with LinkedIn's ecosystem of experimentation, anomaly detection solutions

Requirements



Raptor 1.0

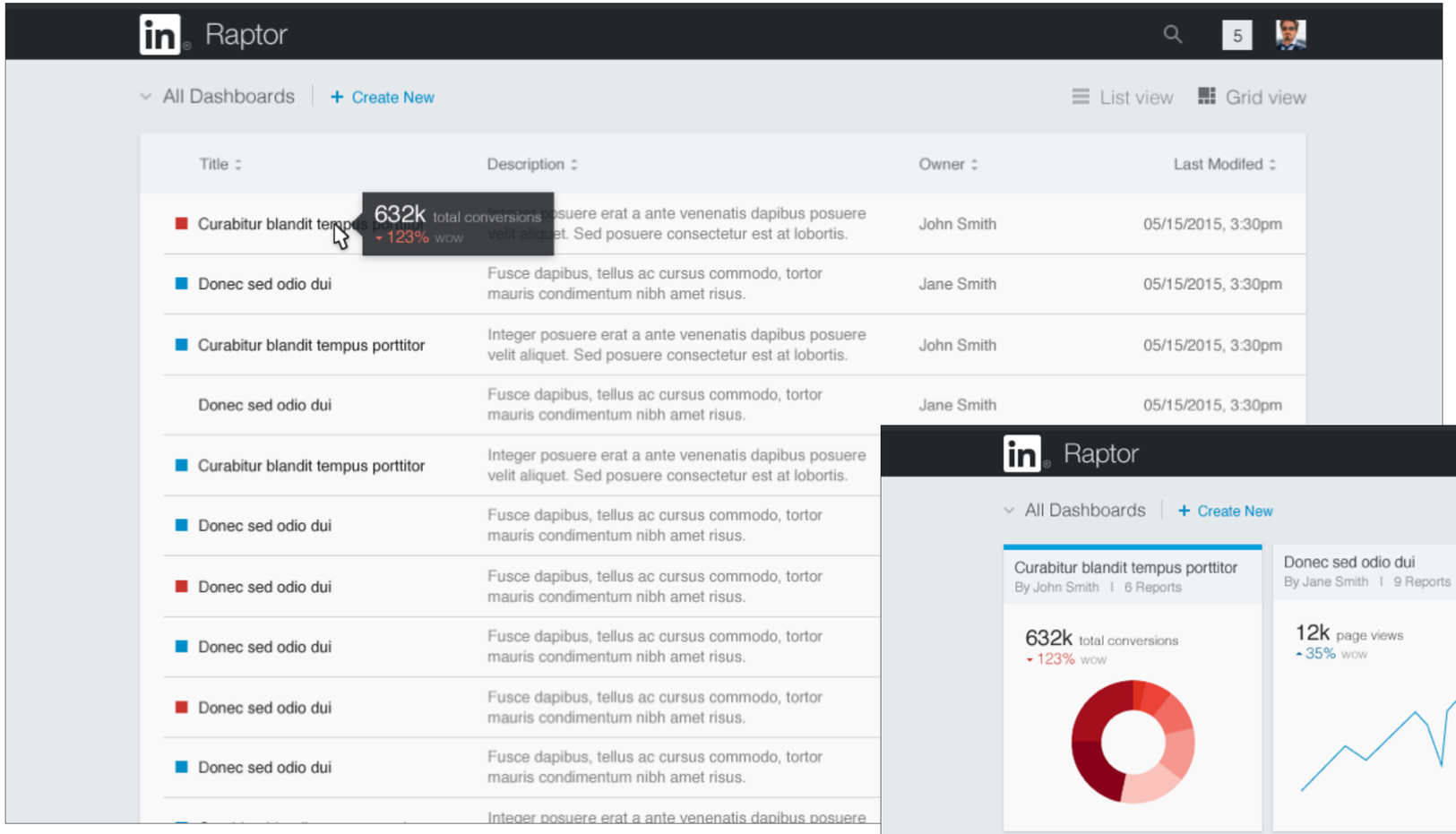
First version built by 3 engineers in a quarter

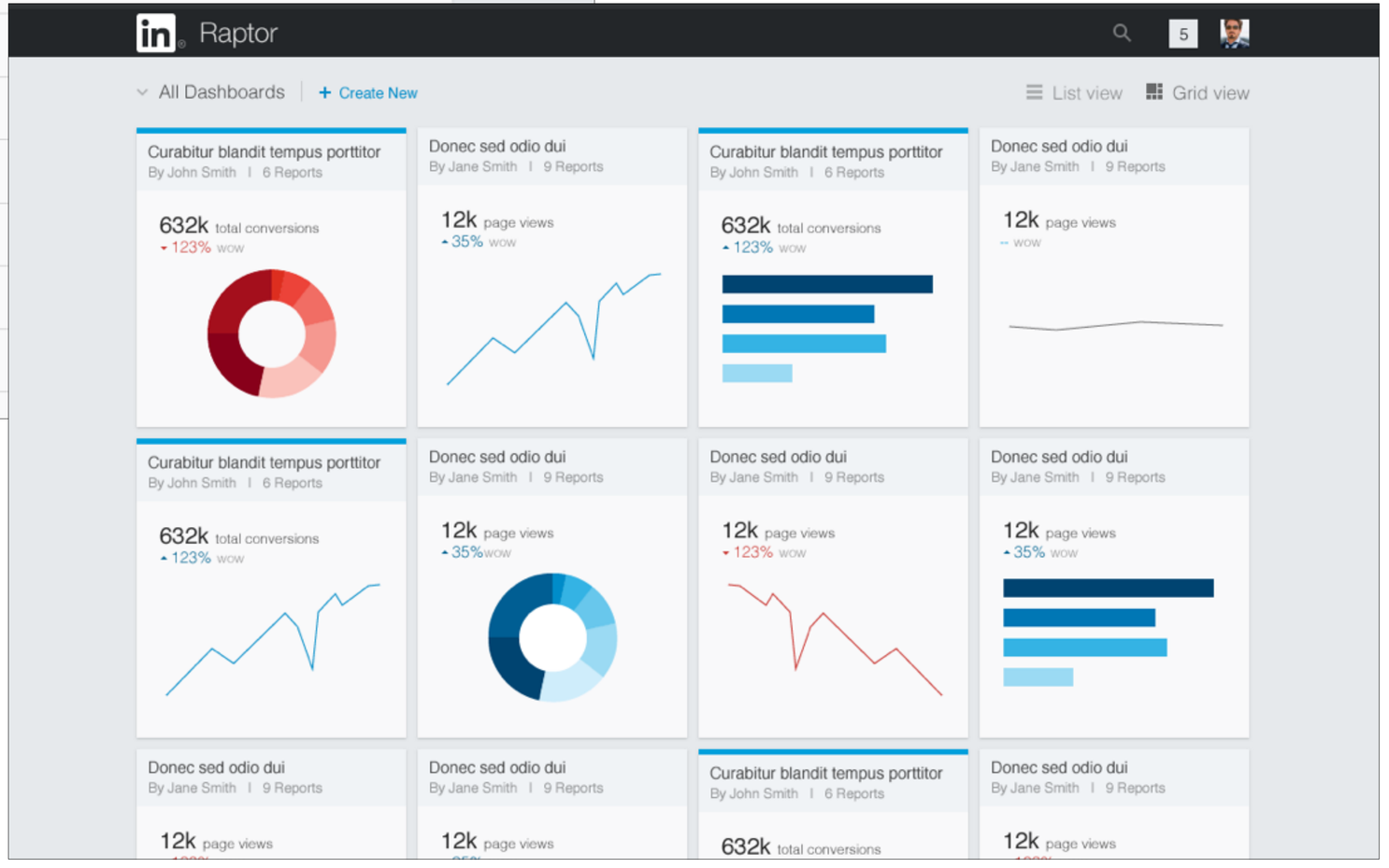
Features

- Integration with UMP, Pinot
- Time series, bar charts, …
- Create, Publish, Clone, Discover Dashboards

Numbers

- Number of dashboards: ~100
- Weekly unique users: ~400


Raptor
5
All Dashboards
+ Create New
List view
Grid view
Title
Description
Owner
Last Modifed
Curabitur blandit tempus porttitor
632k total conversions
123% wow
Integer posuere erat a ante venenatis dapibus posuere velit aliquet. Sed posuere consectetur est at lobortis.
John Smith
05/15/2015, 3:30pm
Donec sed odio dui
Fusce dapibus, tellus ac cursus commodo, tortor mauris condimentum nibh amet risus.
Jane Smith
05/15/2015, 3:30pm
Curabitur blandit tempus porttitor
Integer posuere erat a ante venenatis dapibus posuere velit aliquet. Sed posuere consectetur est at lobortis.
John Smith
05/15/2015, 3:30pm
Donec sed odio dui
Fusce dapibus, tellus ac cursus commodo, tortor mauris condimentum nibh amet risus.
Jane Smith
05/15/2015, 3:30pm
Curabitur blandit tempus porttitor
Integer posuere erat a ante venenatis dapibus posuere velit aliquet. Sed posuere consectetur est at lobortis.
Donec sed odio dui
Fusce dapibus, tellus ac cursus commodo, tortor mauris condimentum nibh amet risus.
Donec sed odio dui
Fusce dapibus, tellus ac cursus commodo, tortor mauris condimentum nibh amet risus.
Donec sed odio dui
Fusce dapibus, tellus ac cursus commodo, tortor mauris condimentum nibh amet risus.
Donec sed odio dui
Fusce dapibus, tellus ac cursus commodo, tortor mauris condimentum nibh amet risus.
Donec sed odio dui
Fusce dapibus, tellus ac cursus commodo, tortor mauris condimentum nibh amet risus.
Integer posuere erat a ante venenatis dapibus posuere


Raptor
5
All Dashboards
+ Create New
List view
Grid view
Curabitur blandit tempus porttitor
By John Smith | 6 Reports
632k total conversions
123% wow
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views
35% wow
Curabitur blandit tempus porttitor
By John Smith | 6 Reports
632k total conversions
123% wow
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views
-- wow
Curabitur blandit tempus porttitor
By John Smith | 6 Reports
632k total conversions
123% wow
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views
35%wow
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views
123% wow
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views
35% wow
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views
Curabitur blandit tempus porttitor
By John Smith | 6 Reports
632k total conversions
Donec sed odio dui
By Jane Smith | 9 Reports
12k page views

Raptor

5

I'm searching for a metric named email in all dashboards.

Curabitur blandit tempus porttitor
By John Smith | 6 Reports

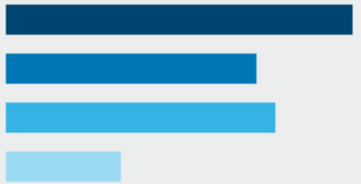
632k total conversions
123% wow

Donec sed odio dui
By Jane Smith | 9 Reports

12k page views
35% wow

Curabitur blandit tempus porttitor
By John Smith | 6 Reports

632k total conversions
123% wow



Donec sed odio dui
By Jane Smith | 9 Reports

12k page views
-- wow

Curabitur blandit tempus porttitor
By John Smith | 6 Reports

632k total conversions
123% wow

Donec sed odio dui
By Jane Smith | 9 Reports

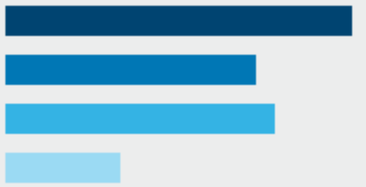
12k page views
35%wow

Donec sed odio dui
By Jane Smith | 9 Reports

12k page views
123% wow

Donec sed odio dui
By Jane Smith | 9 Reports

12k page views
35% wow



Donec sed odio dui
By Jane Smith | 9 Reports

Donec sed odio dui
By Jane Smith | 9 Reports

Curabitur blandit tempus porttitor
By John Smith | 6 Reports

Donec sed odio dui
By Jane Smith | 9 Reports

Curabitur blandit tempus porttitor | Dashboard

Auto-saved; 05/20/15 03:50pm Update Cancel

Cras justo odio dapibus ac

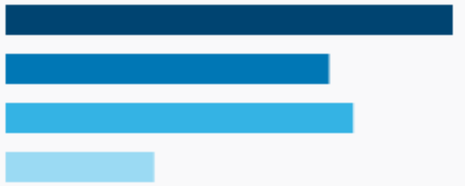
632k total conversions

Vestibulum id ligula

- Tristique ipsum
- Vulputate elit
- Porta tortor lorem
- Justo ridiculus
- Pellentesque ligula
- Euismod vulputate
- Venenatis fringilla

Aenean lacinia bibendum nulla

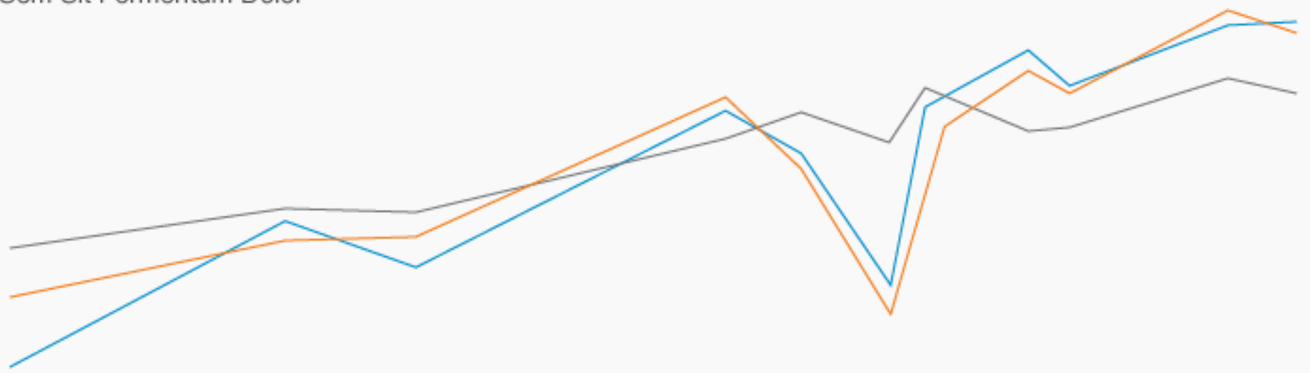
12k page views



Vestibulum id ligula

- Tristique ipsum
- Vulputate elit
- Porta tortor lorem
- Justo ridiculus

Sem Sit Fermentum Dolor



12k page views

▲35% wow

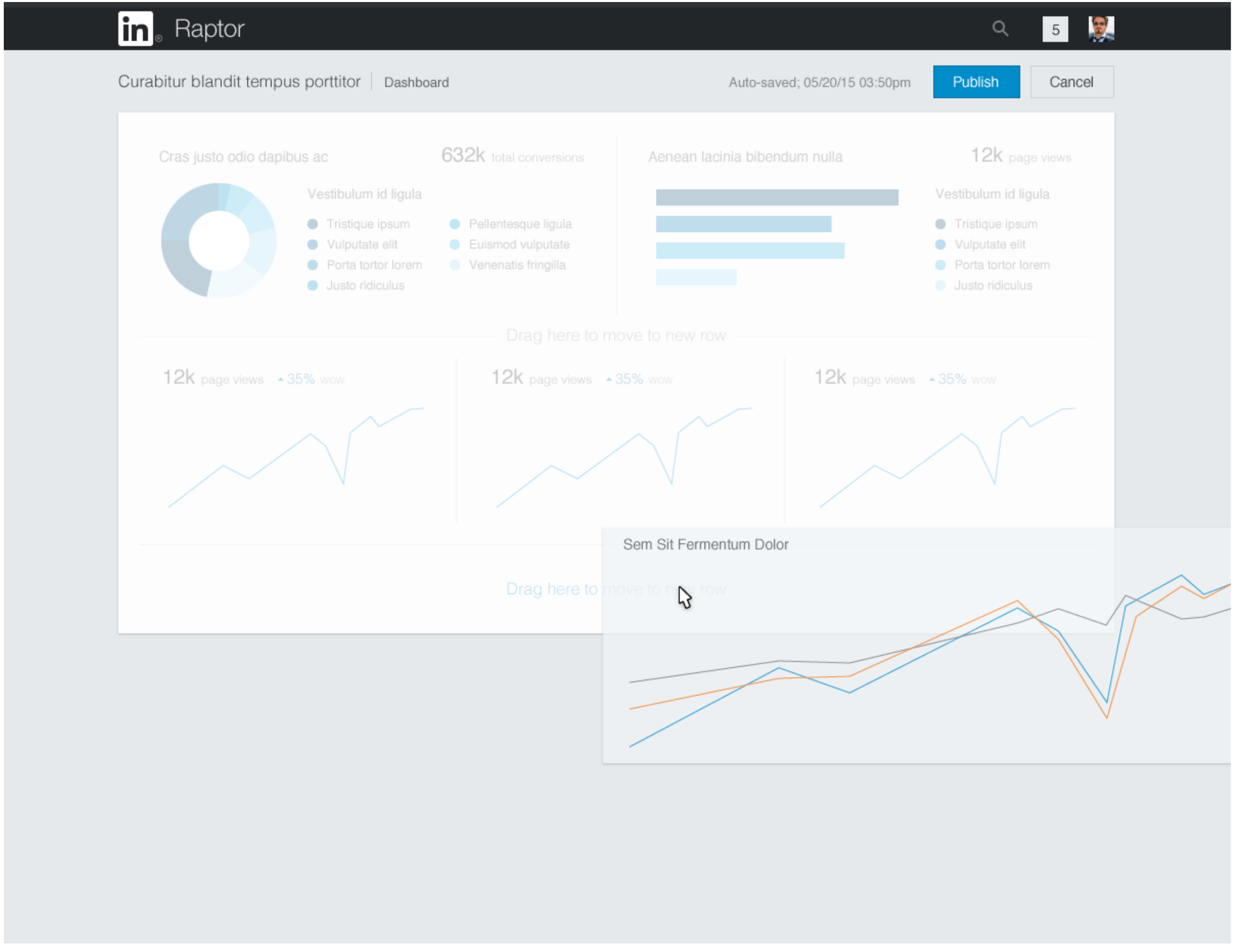
Vehicula Vulputate Ornare

- Tristique ipsum
- Vulputate elit
- Porta tortor lorem

12k page views ▲35% wow

12k page views ▲35% wow

12k page views ▲35% wow


Raptor
5
Curabitur blandit tempus porttitor
Dashboard
Auto-saved; 05/20/15 03:50pm
Publish
Cancel
Cras justo odio dapibus ac
632k total conversions
Vestibulum id ligula
Tristique ipsum
Vulputate elit
Porta tortor lorem
Justo ridiculus
Pellentesque ligula
Euismod vulputate
Venenatis fringilla
Aenean lacinia bibendum nulla
12k page views
Vestibulum id ligula
Tristique ipsum
Vulputate elit
Porta tortor lorem
Justo ridiculus
Drag here to move to new row
12k page views 35% wow
12k page views 35% wow
12k page views 35% wow
Sem Sit Fermentum Dolor

Curabitur blandit tempus porttitor | Dashboard

Auto-saved; 05/20/15 03:50pm Update Cancel

Cras justo odio dapibus ac

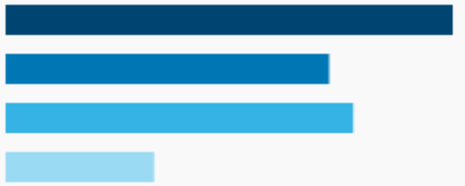
632k total conversions

Vestibulum id ligula

- Tristique ipsum
- Vulputate elit
- Porta tortor lorem
- Justo ridiculus
- Pellentesque ligula
- Euismod vulputate
- Venenatis fringilla

Aenean lacinia bibendum nulla

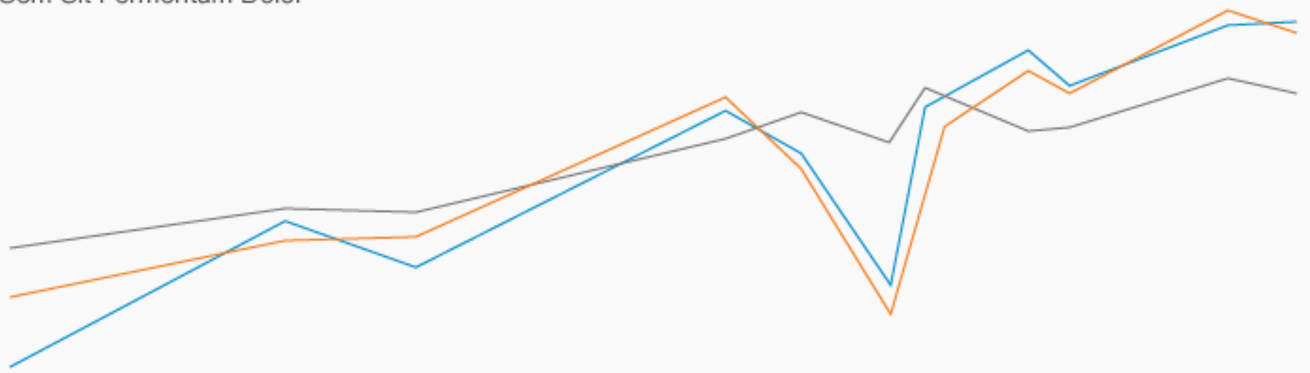
12k page views



Vestibulum id ligula

- Tristique ipsum
- Vulputate elit
- Porta tortor lorem
- Justo ridiculus

Sem Sit Fermentum Dolor



12k page views

▲35% wow

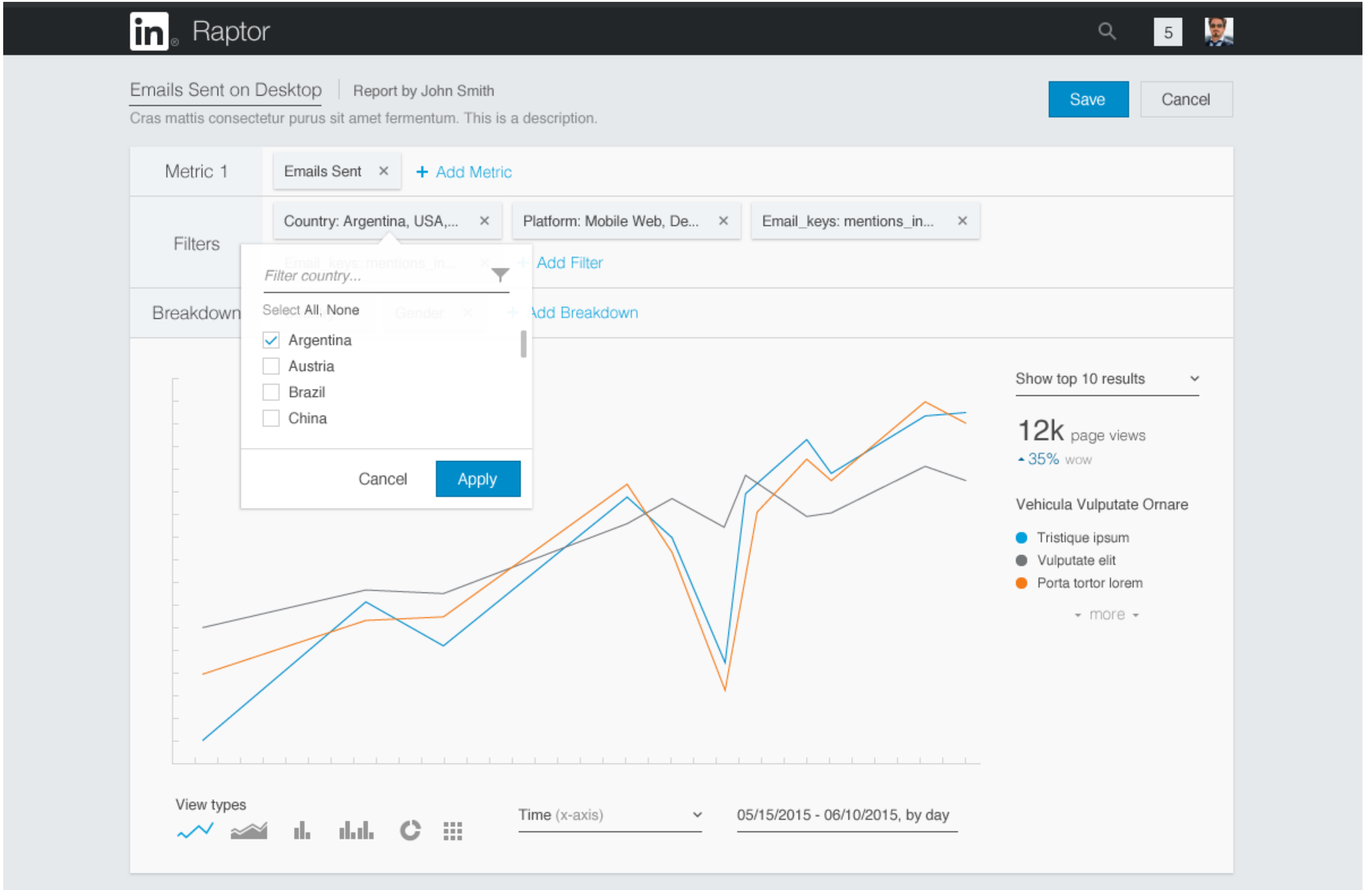
Vehicula Vulputate Ornare

- Tristique ipsum
- Vulputate elit
- Porta tortor lorem

12k page views ▲35% wow

12k page views ▲35% wow

12k page views ▲35% wow


Raptor
5
Emails Sent on Desktop
Report by John Smith
Cras mattis consectetur purus sit amet fermentum. This is a description.
Save
Cancel
Metric 1
Emails Sent
+ Add Metric
Filters
Country: Argentina, USA,...
Platform: Mobile Web, De...
Email_keys: mentions_in...
+ Add Filter
Filter country...
Select All, None
Argentina
Austria
Brazil
China
Cancel
Apply
Breakdown
Gender
+ Add Breakdown
Show top 10 results
12k page views
35% wow
Vehicula Vulputate Ornare
Tristique ipsum
Vulputate elit
Porta tortor lorem
more
View types
Time (x-axis)
05/15/2015 - 06/10/2015, by day

Raptor

5

Emails Sent on Desktop | Report by John Smith

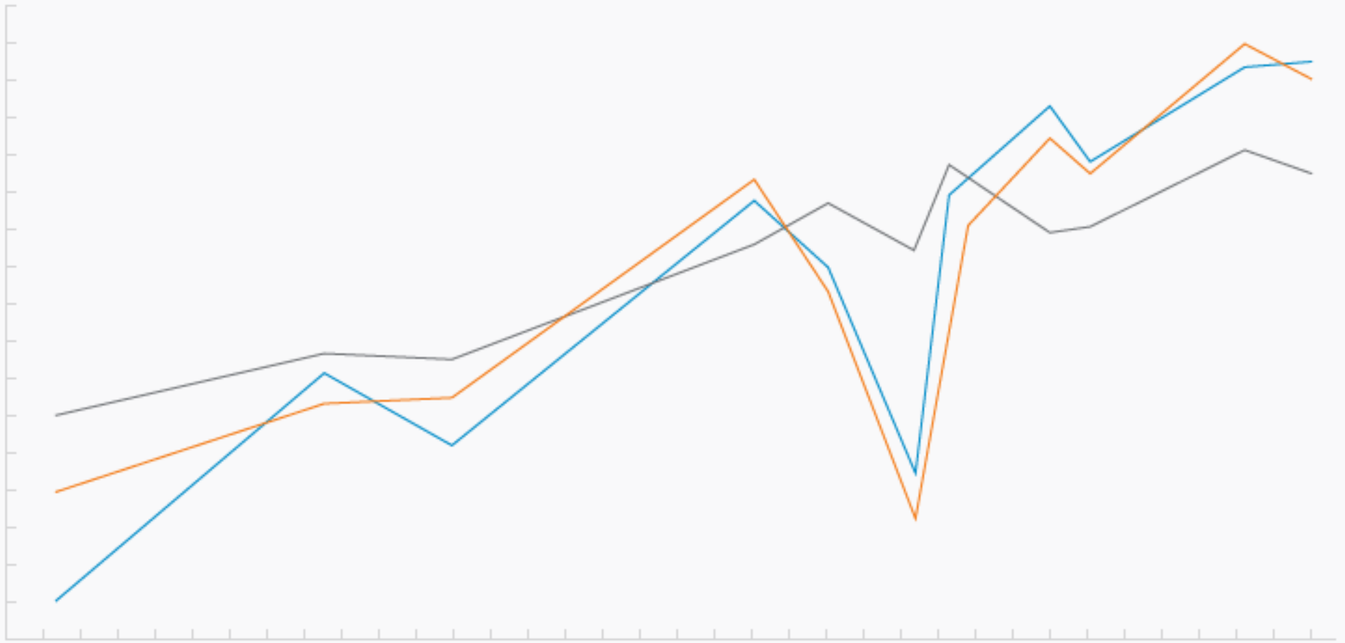
Cras mattis consectetur purus sit amet fermentum. This is a description.

Save Cancel

Metric 1: Emails Sent × + Add Metric

Filters: Country: Argentina, USA,... × | Platform: Mobile Web, De... × | Email_keys: mentions_in... × | Email_keys: mentions_in... × + Add Filter

Breakdown: Country × | Gender × + Add Breakdown



Show top 10 results

12k page views

▲35% wow

Vehicula Vulputate Ornare

- Tristique ipsum
- Vulputate elit
- Porta tortor lorem

more

View types

Time (x-axis)

05/15/2015 - 06/10/2015, by day

The Future for Raptor

Social Collaboration features

Intelligence

- Anomaly detection
- Dashboards You May Like

Embedding into data products

Open Source

A Few Good Hammers

Ingest

Process

Unified
Metrics
Platform

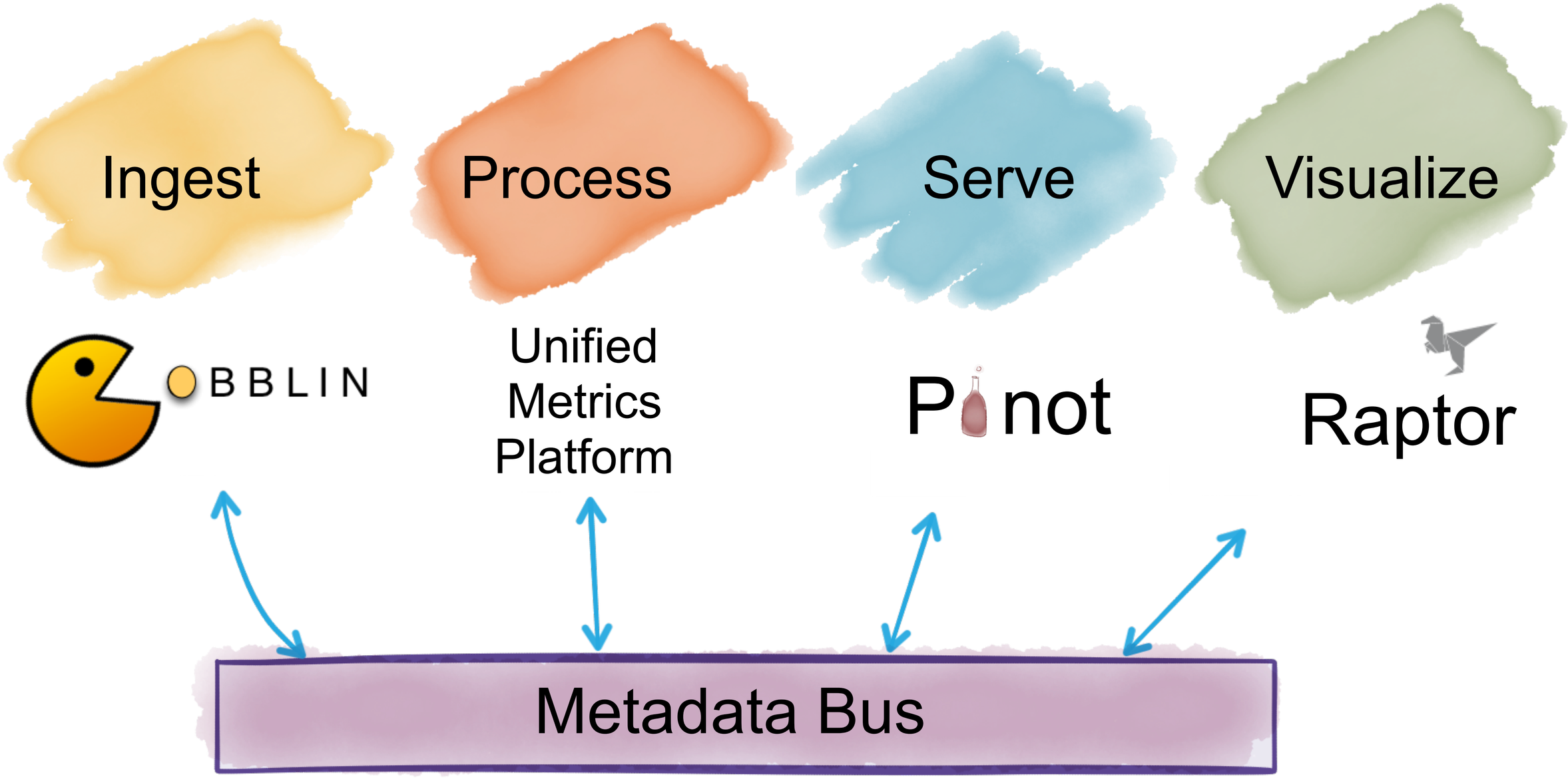
Serve

Pinot

Visualize

Raptor

What we're excited about



Metadata driven e2e Optimizations

Dynamic prioritization of data ingest

Surface source data quality issues in dashboard

Surface backfill status on dashboard

Cascading deprecation of dashboards, computation and data sources through lineage

Catch me offline to chat about…

What we're doing for

- Views on Hadoop
- Data Quality
- Metadata

Shirshanka Das

@shirshanka