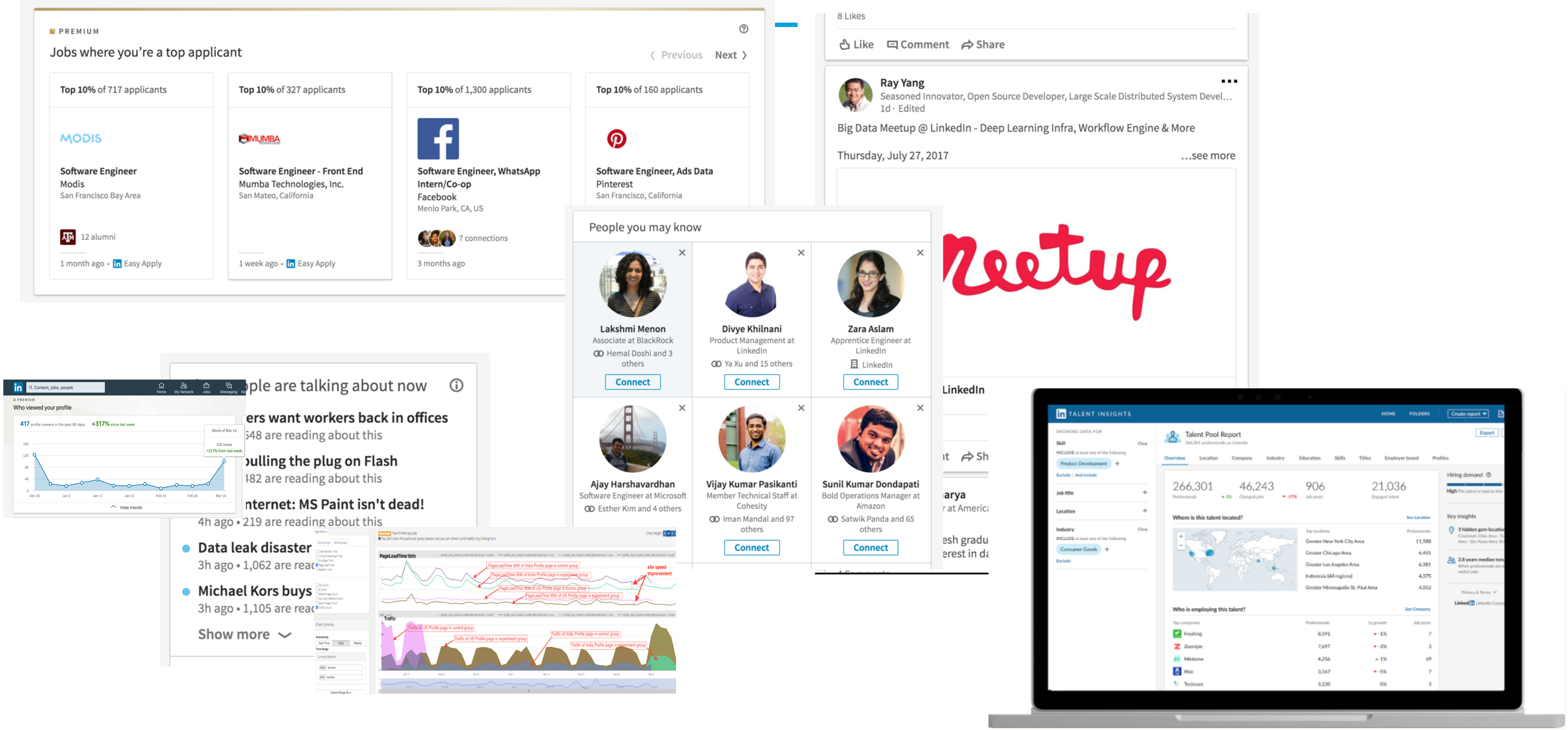
Scaling Data Science Platform as a Service

Ameya Kanitkar

Engineering Manager, Data Science Platform

What’s common amongst these LinkedIn Products?



PREMIUM

Jobs where you’re a top applicant

〈 Previous Next 〉

Top 10% of 717 applicants

Software Engineer
Modis
San Francisco Bay Area
12 alumni
1 month ago · Easy Apply

Top 10% of 327 applicants

Software Engineer - Front End
Mumba Technologies, Inc.
San Mateo, California
1 week ago · Easy Apply

Top 10% of 1,300 applicants

Software Engineer, WhatsApp Intern/Co-op
Facebook
Menlo Park, CA, US
7 connections
3 months ago

Top 10% of 160 applicants

Software Engineer, Ads Data
Pinterest
San Francisco, California

8 Likes

Like Comment Share

Ray Yang
Seasoned Innovator, Open Source Developer, Large Scale Distributed System Devel…
1d · Edited

Big Data Meetup @ LinkedIn - Deep Learning Infra, Workflow Engine & More

Thursday, July 27, 2017 …see more

People you may know

Lakshmi Menon
Associate at BlackRock
Hemal Doshi and 3 others
Connect

Divye Khilnani
Product Management at LinkedIn
Ya Xu and 15 others
Connect

Zara Aslam
Apprentice Engineer at LinkedIn
LinkedIn
Connect

Ajay Harshavardhan
Software Engineer at Microsoft
Esther Kim and 4 others

Vijay Kumar Pasikanti
Member Technical Staff at Cohesity
Iman Mandal and 97 others
Connect

Sunil Kumar Dondapati
Bold Operations Manager at Amazon
Satwik Panda and 65 others
Connect

Content, jobs, people

Home My Network Jobs Messaging

PREMIUM

Who viewed your profile

417 profile viewers in the past 90 days +317% since last week

Week of Mar 14
100 views
+317% from last week

Hide trends

…ple are talking about now

…ers want workers back in offices
…548 are reading about this

…pulling the plug on Flash
…482 are reading about this

…nternet: MS Paint isn't dead!
4h ago • 219 are reading about this

Data leak disaster
3h ago • 1,062 are reac…

Michael Kors buys
3h ago • 1,105 are reac…

Show more

Read-Only Titan Profile lazy ads
You don't own this particular query session, but you can clone it and modify it by clicking here

TALENT INSIGHTS

HOME FOLDERS Create report

SHOWING DATA FOR
Skill Clear
INCLUDE at least one of the following
Product Development
Exclude | And include
Job title
Location
Industry Clear
INCLUDE at least one of the following
Consumer Goods
Exclude

Talent Pool Report
266,301 professionals on LinkedIn

Export

Overview Location Company Industry Education Skills Titles Employer brand Profiles

266,301 Professionals ▲5%
46,243 Changed jobs ▼-17%
906 Job posts
21,036 Engaged talent

Hiring demand
High This talent is hard to hire

Where is this talent located? See Location

Top locations	Professionals
Greater New York City Area	11,588
Greater Chicago Area	6,451
Greater Los Angeles Area	6,381
Indonesia (All regions)	4,375
Greater Minneapolis-St. Paul Area	4,052

Who is employing this talent? See Company

Top companies	Professionals	1y growth	Job posts
Freshing	8,591	▼-1%	7
Zoomjax	7,697	▼-3%	3
Mintome	4,256	▲1%	69
Itkix	3,567	▼-5%	7
Techcore	3,220	0%	3

Key insights
3 hidden gem locatio…
Cincinnati, Ohio Area - To… Area - São Paulo Area, Br…
2.8 years median ten…
When professionals are … switch jobs

Privacy & Terms

160
120
80
40
0
Dec 20
Jan 3
Jan 17
Jan 31
Feb 14
Feb 28
Mar 14
Chart Height: S M L
PageLoadTime 90th
PageLoadTime 90th of India Profile page in control group
PageLoadTime 90th of India Profile page in experiment group
PageLoadTime 90th of US Profile page in control group
PageLoadTime 90th of US Profile page in experiment group
site speed improvement
Traffic
Traffic of US Profile page in control group
Traffic of US Profile page in experiment group
Traffic of India Profile page in control group
Traffic of India Profile page in experiment group
Oct 17
Oct 24
Oct 31
Nov 7
Nov 14
Nov 21
Nov 28
Dec 5
Weekly
Meetup
LinkedIn

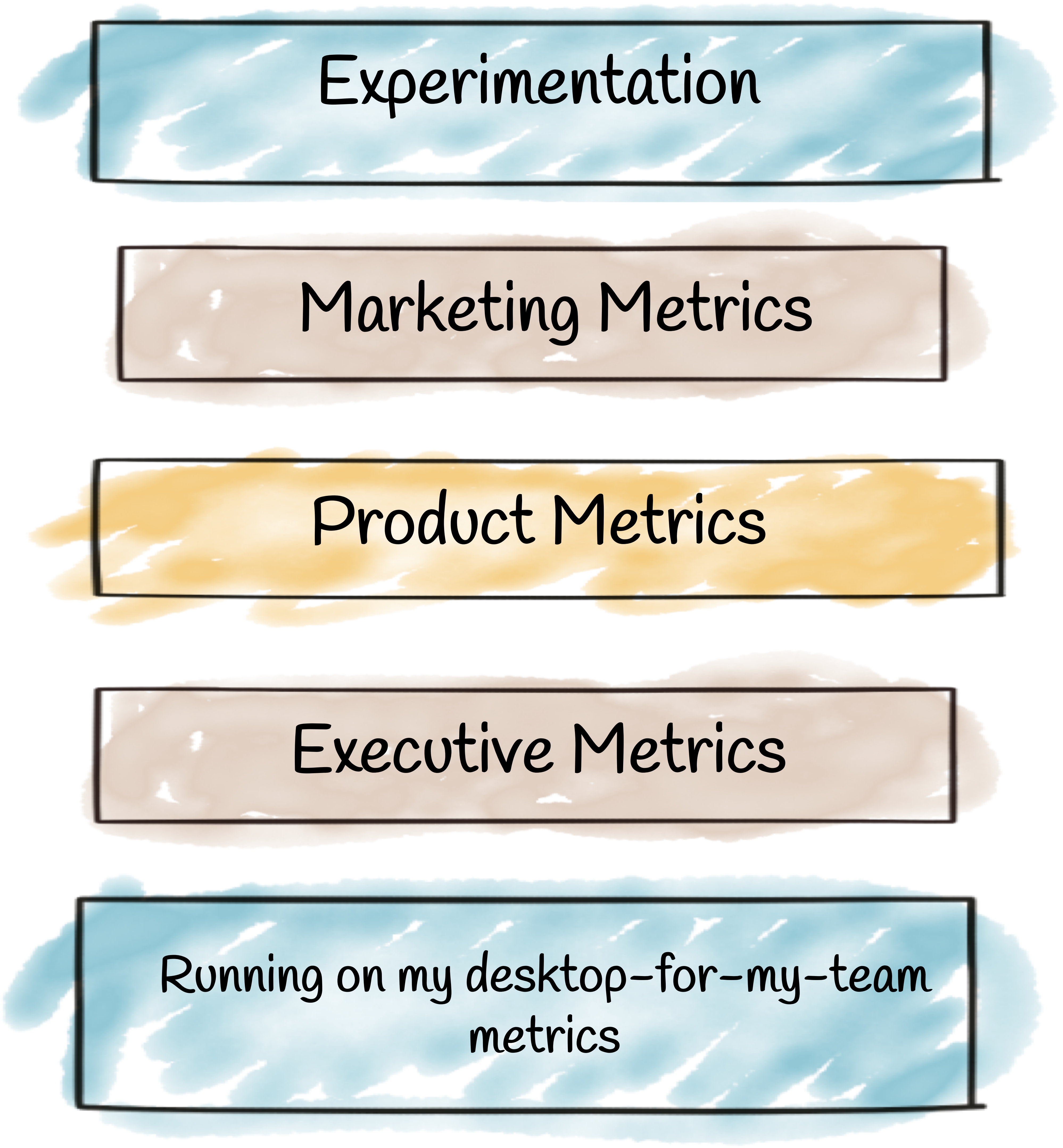
This Talk

is about Metrics

- Linkedin has a data driven culture
- We measure and monitor everything
- Metrics and how we compute and serve them is at the heart of our data science platform as a service

Once Upon a Time

there were great many metrics platforms



Experimentation

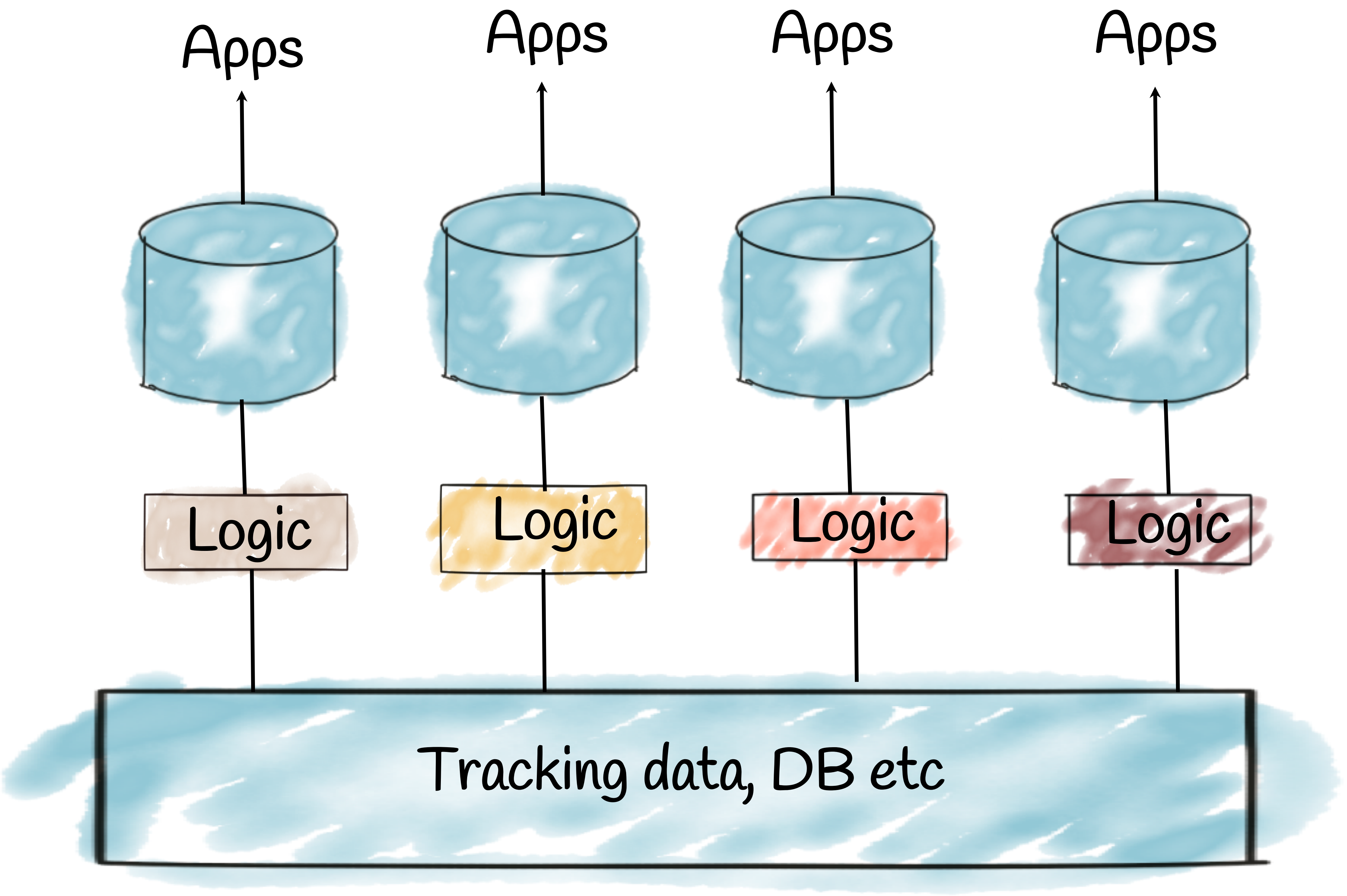
Marketing Metrics

Product Metrics

Executive Metrics

Running on my desktop-for-my-team metrics

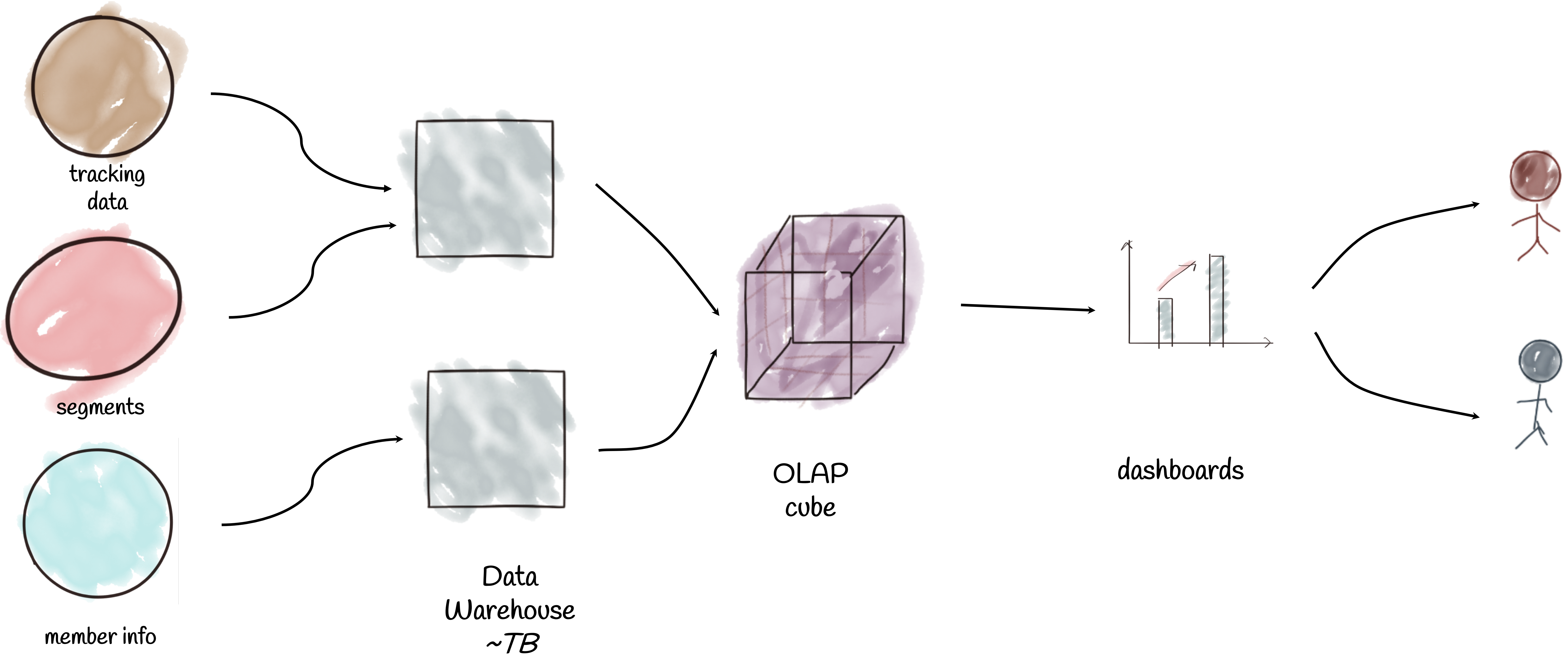
Which was bad



- Inconsistent Data
- Lack of trust
- Closed systems with data and knowledge in silos

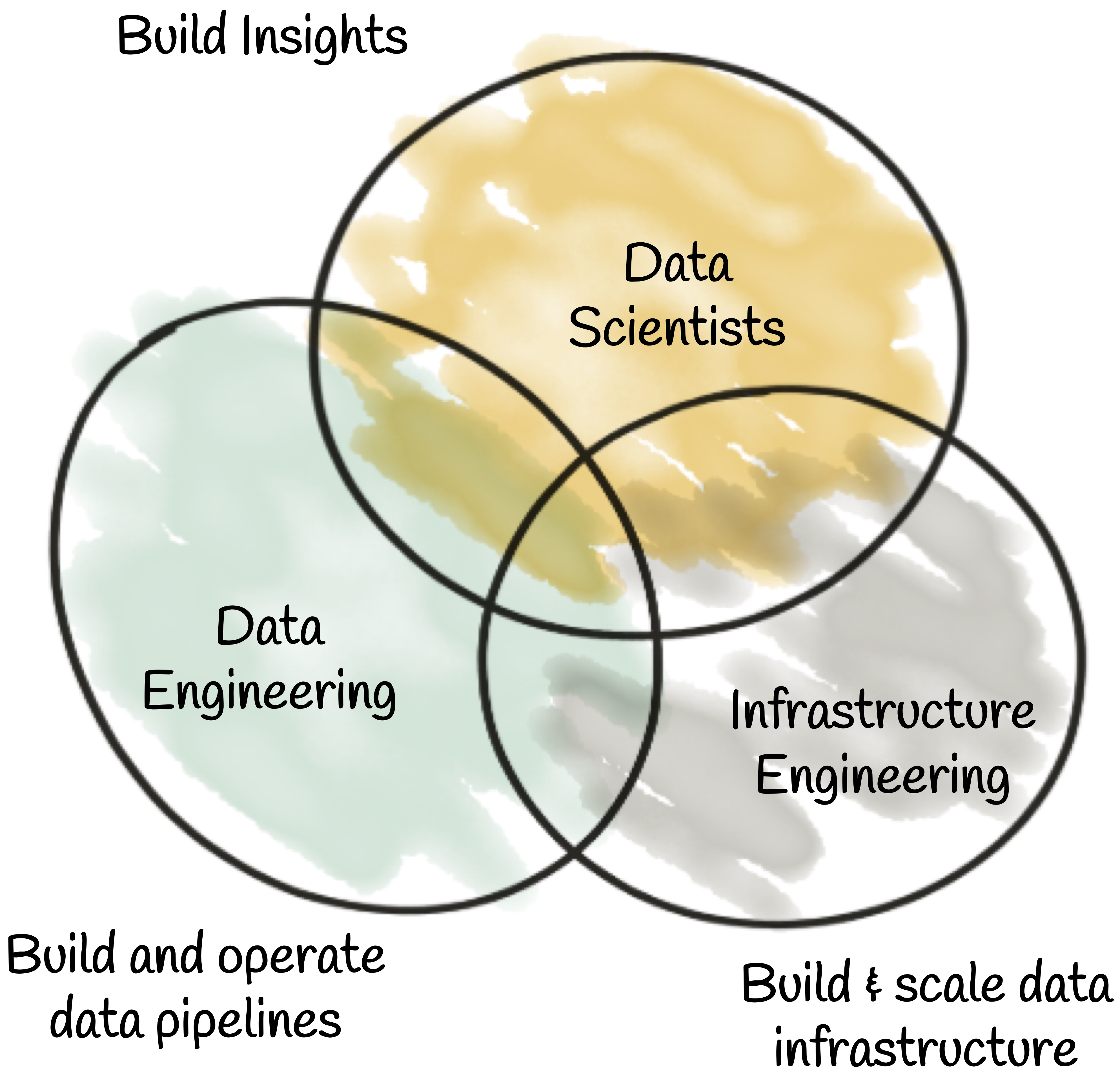
Challenge 1: Data Consistency

Challenge 2: Engineering Productivity & Operations



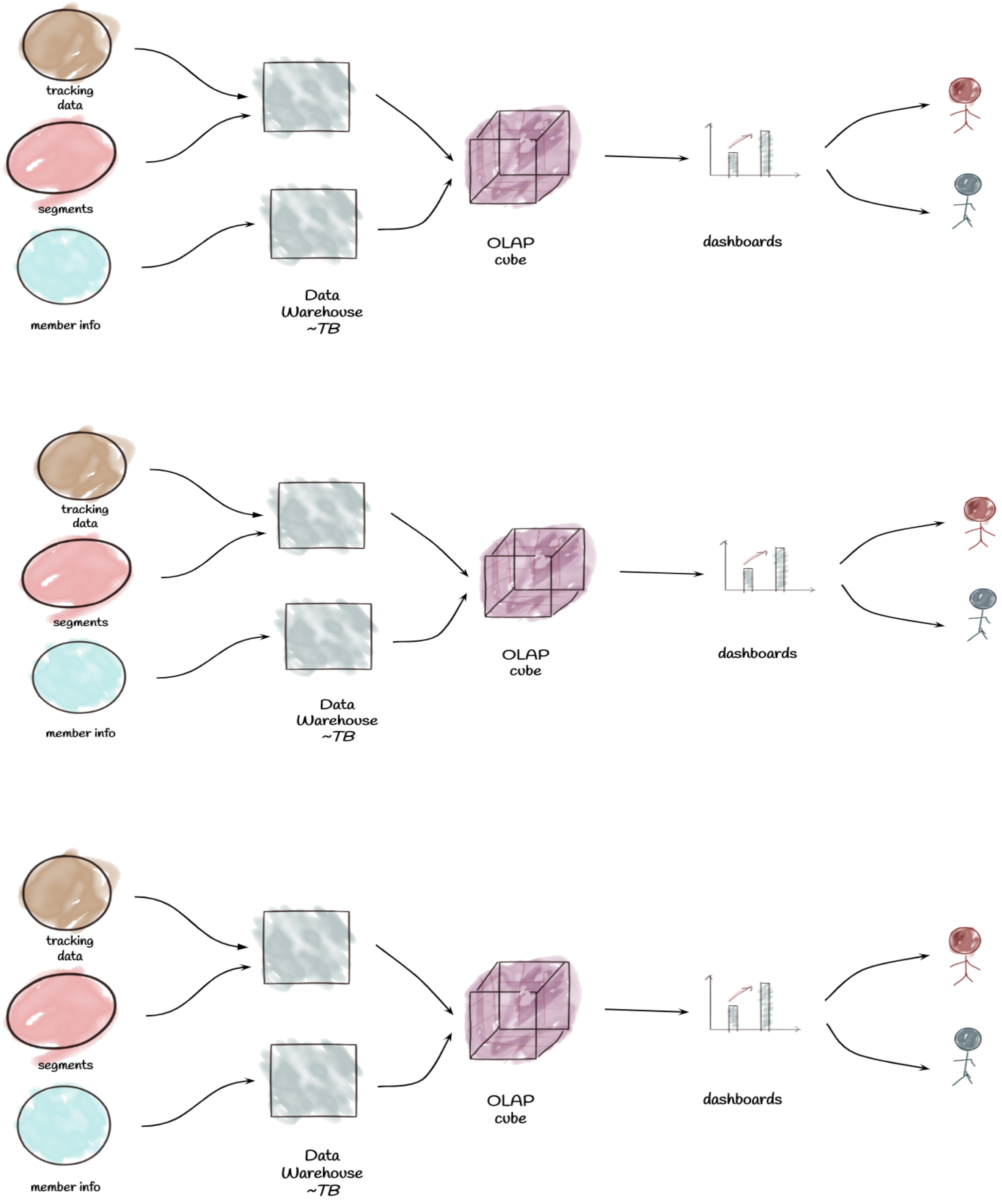
- Identify data sources
- Build data pipelines
- Data Modeling
- Insights & Analysis
- Data Governance and Access management

Challenges: Engineering Productivity (Contd..)


Build Insights
Data Scientists
Data Engineering
Infrastructure Engineering
Build and operate data pipelines
Build & scale data infrastructure

- Speed can slow down due to multiple teams are involved
- Data scientists productivity declines as they need to be dependent on other teams to build necessary insights

Challenges: Operations (Contd..)



- Building and maintaining thousands of data pipelines is messy

- Consistently maintaining metadata, data lineage, data dependencies, SLA requires specialized systems and needs to be developed across all teams

Solution

Our Approach

“Provide trusted repository of metrics, and build a self-serve platform for sustainable life cycle of metrics”

Wish List

What did we want?

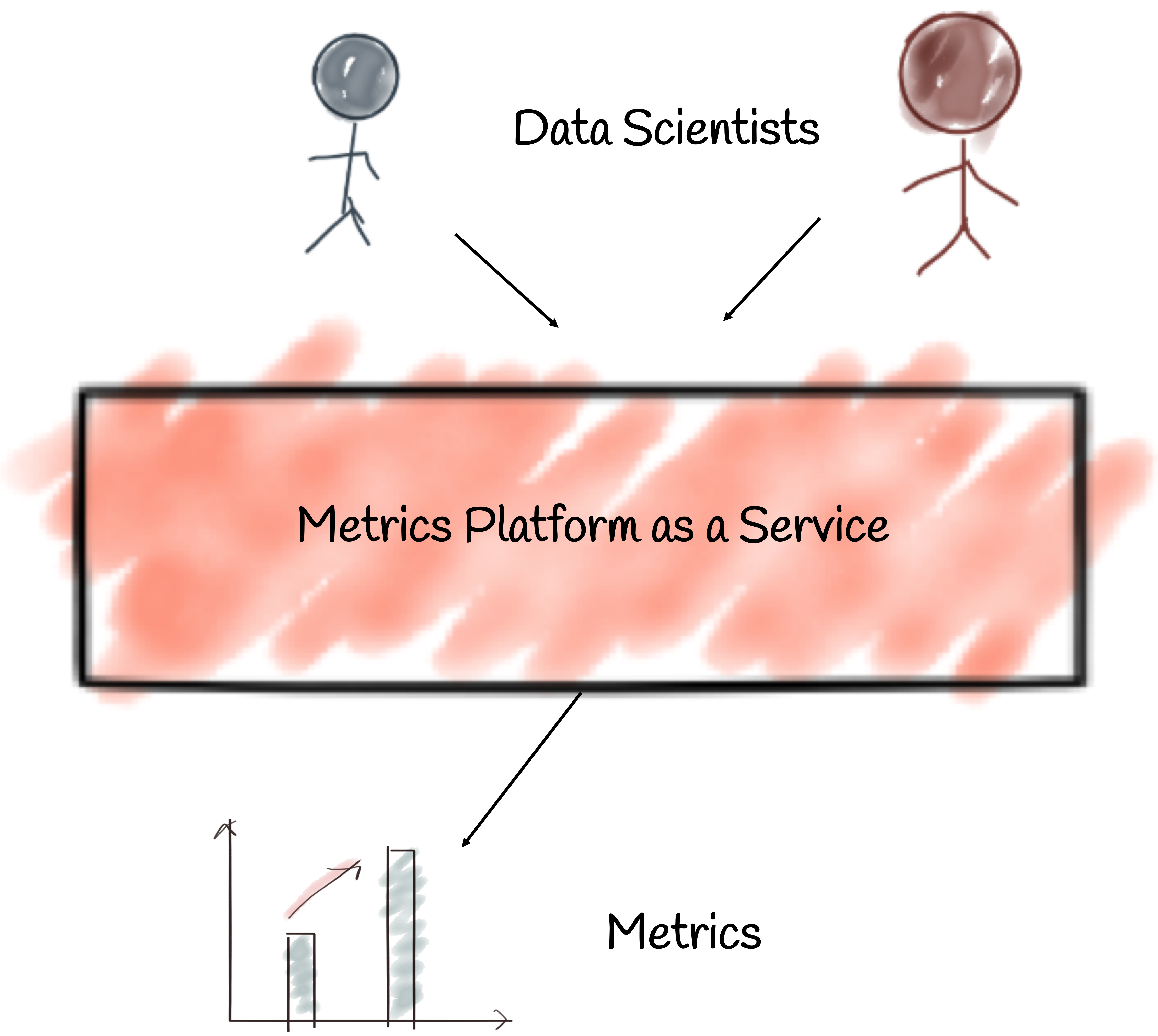
- Same and consistent metrics/ insights in all data applications
- Same metadata everywhere
- Single definition of entity dimensions
- Single definition of event dimensions
- No duplication of metrics
- Allow data scientists to focus on their core skills/ job
- Increase engineering productivity, simplify and optimize operations

Solution

Metrics Platform as Service

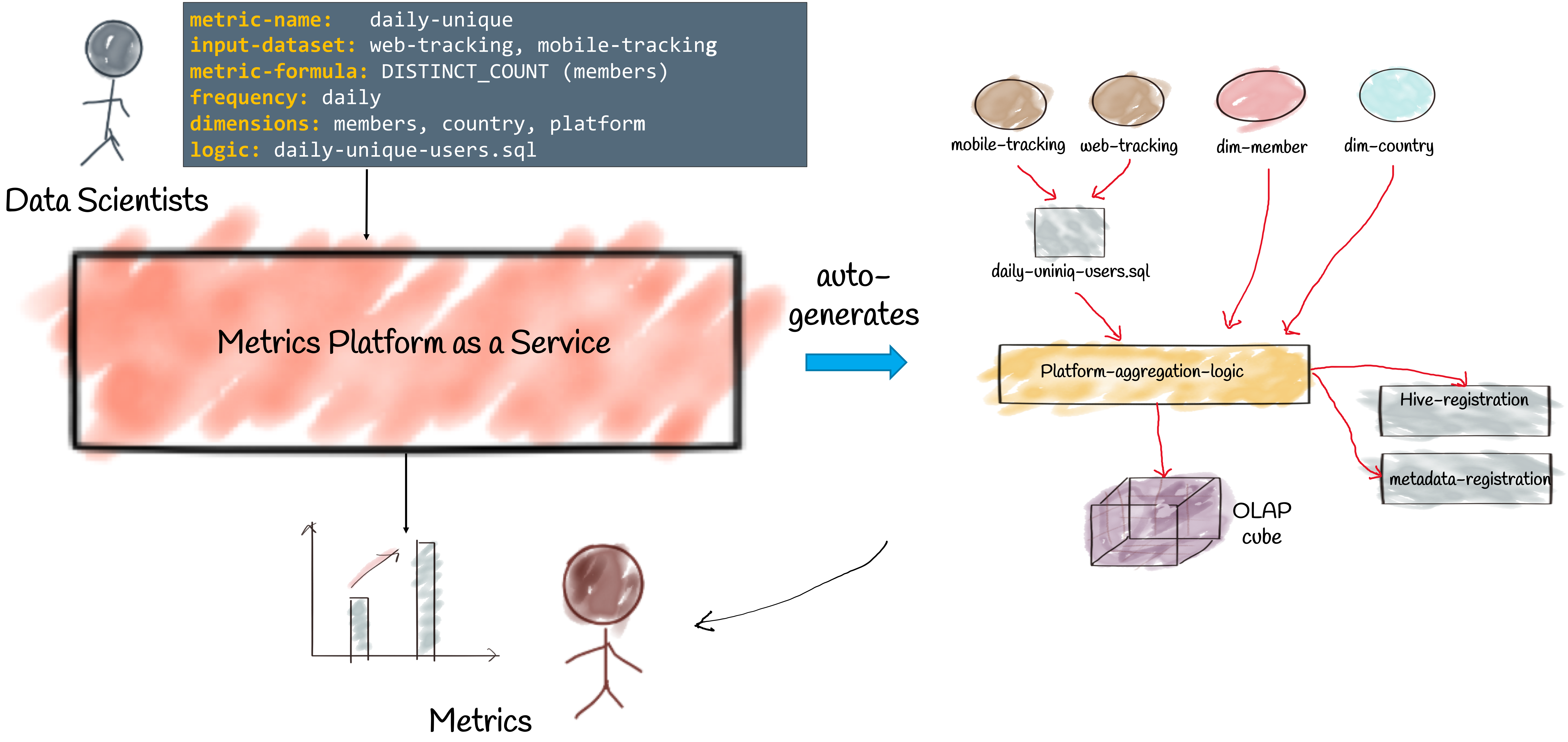
“Platform that builds, manages and scales all metrics across all applications at LinkedIn”

Metrics Platform as a Service



- Data scientists build metrics logic (config + code) into the central platform
- Platform automatically generates necessary data pipelines and centrally operate
- Platform computes and publishes metrics results into dashboard/ downstream apps

Metrics Platform as a Service



```
metric-name:   daily-unique
input-dataset: web-tracking, mobile-tracking
metric-formula: DISTINCT_COUNT (members)
frequency: daily
dimensions: members, country, platform
logic: daily-unique-users.sql
```

Scaling Metrics Platform

Our journey in evolving our platforms as a de-facto metrics platform for the entire company

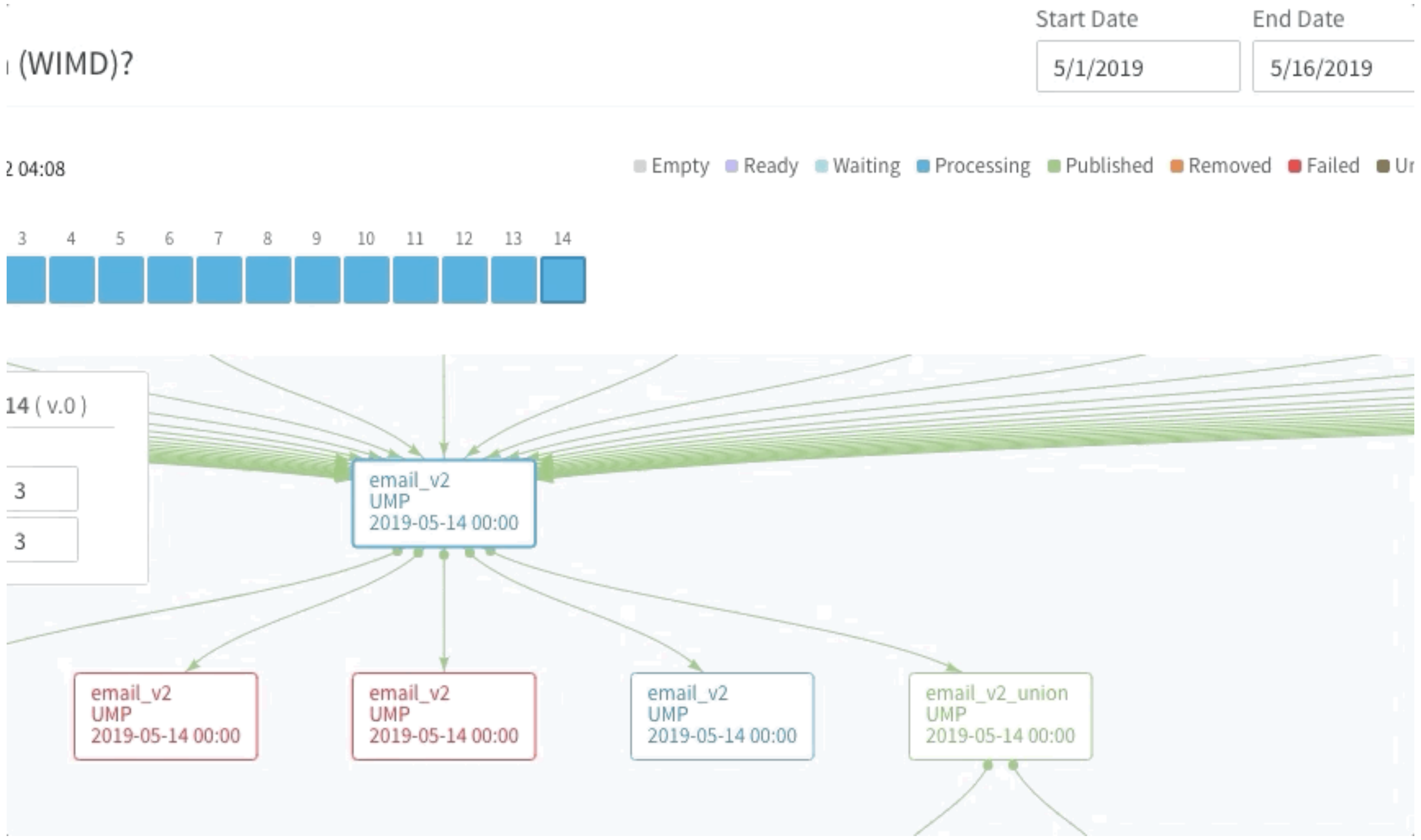
Scaling Governance is not Easy

metric-defs

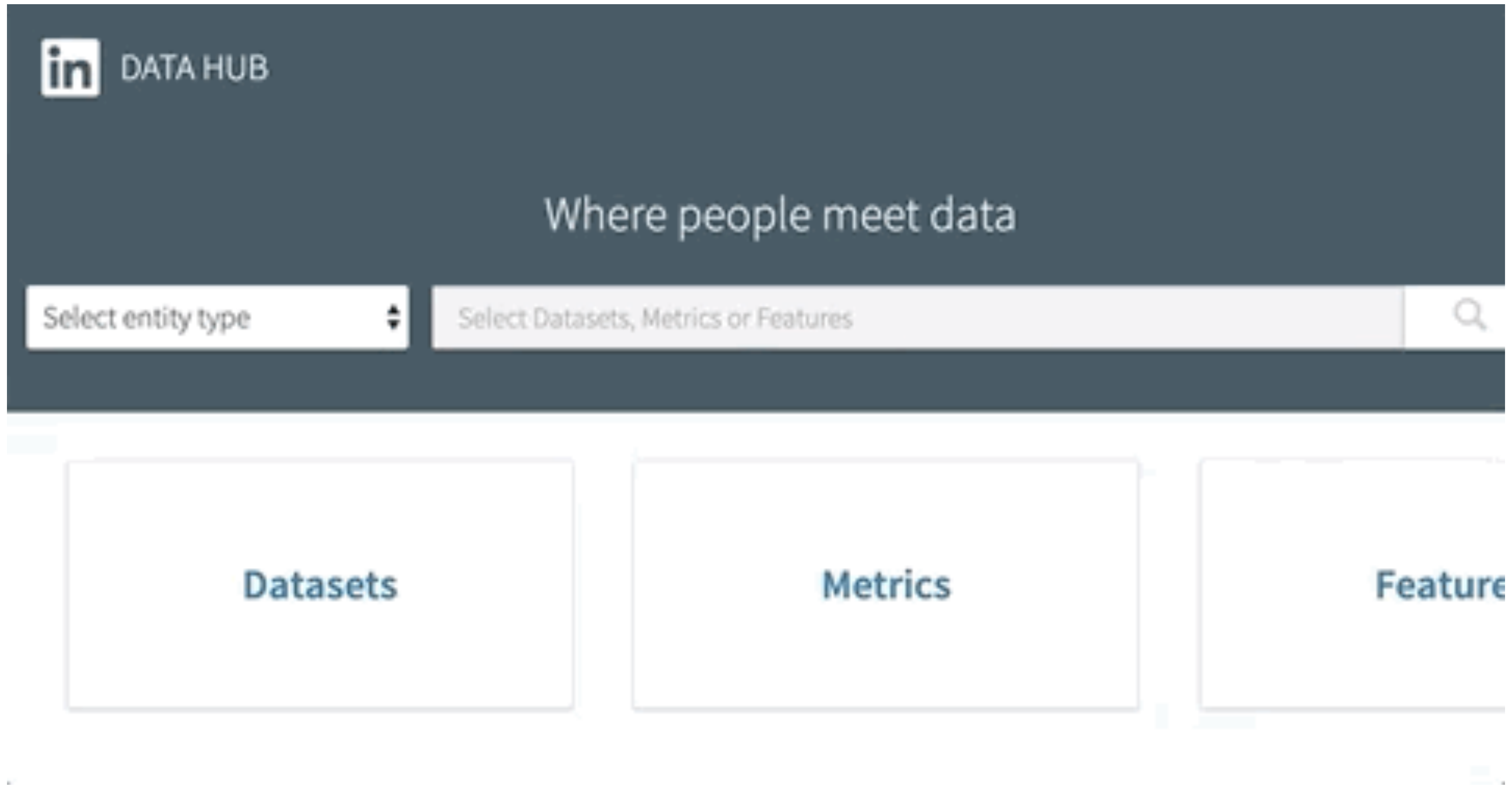
Single source/repository for all metrics

ACL's

ACL Management by Subject Matter Expertise



Tools for data lineage



Search & Discovery experience for your datasets and metrics

Lessons along the way

Implementing default use cases should be simple, but more complex use cases possible

1. Added support for various compute engines: Pig, Hive, Spark, Spark SQL, Presto
2. Templatize common patterns such as cubing, multi datacenter availability, percentile support
3. Build downstream integrations
4. Explore plug-in architecture to avoid making platform team a bottleneck

Lessons along the way

Leverage economies of scale of being the central platform

Simplified Compliance

- Platform managed compliance
- Data Retention and cleanup
- Data Quality Checks

Self-Serve & Operational Tooling

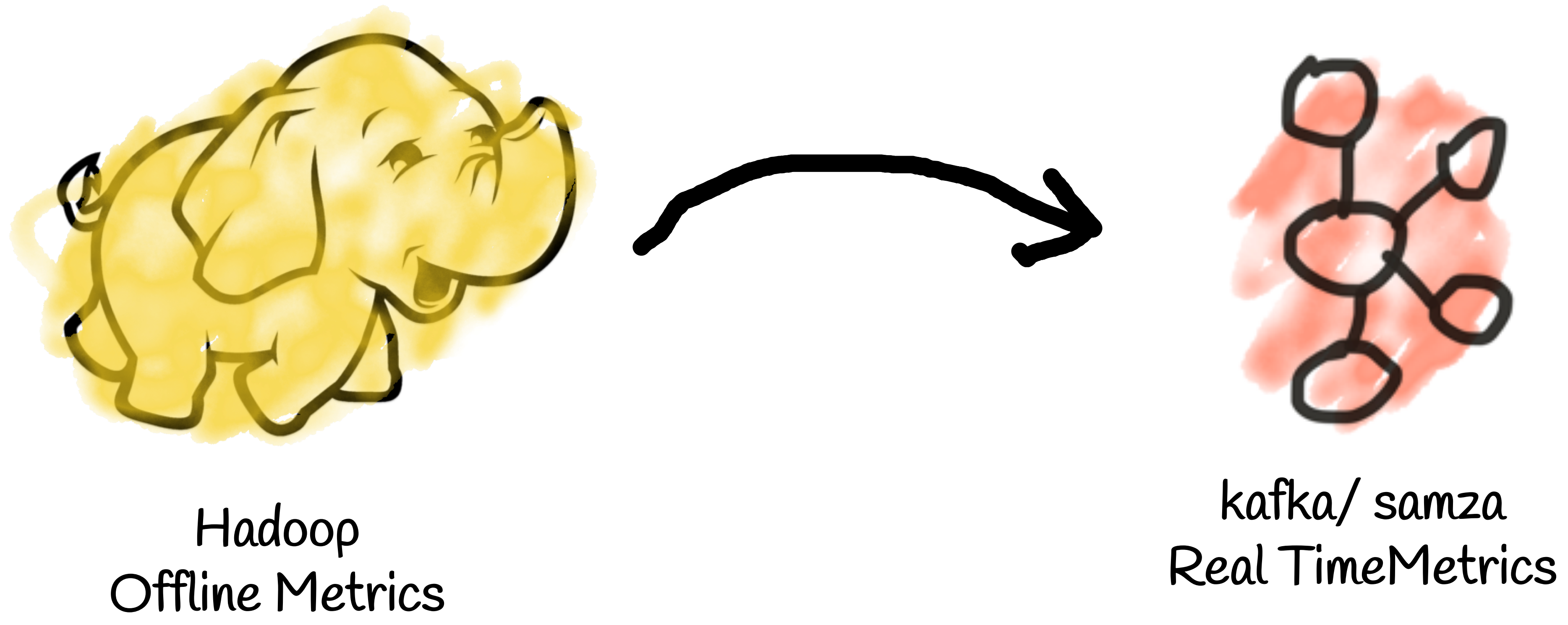
- Flow Management
- Backfills Portal
- Self-heal

Lessons along the way

Innovations are a lot more impactful with platforms

Batch -> Realtime Convergence

- Automatic Hadoop MR -> Kafka/ Samza convergence via apache calcite

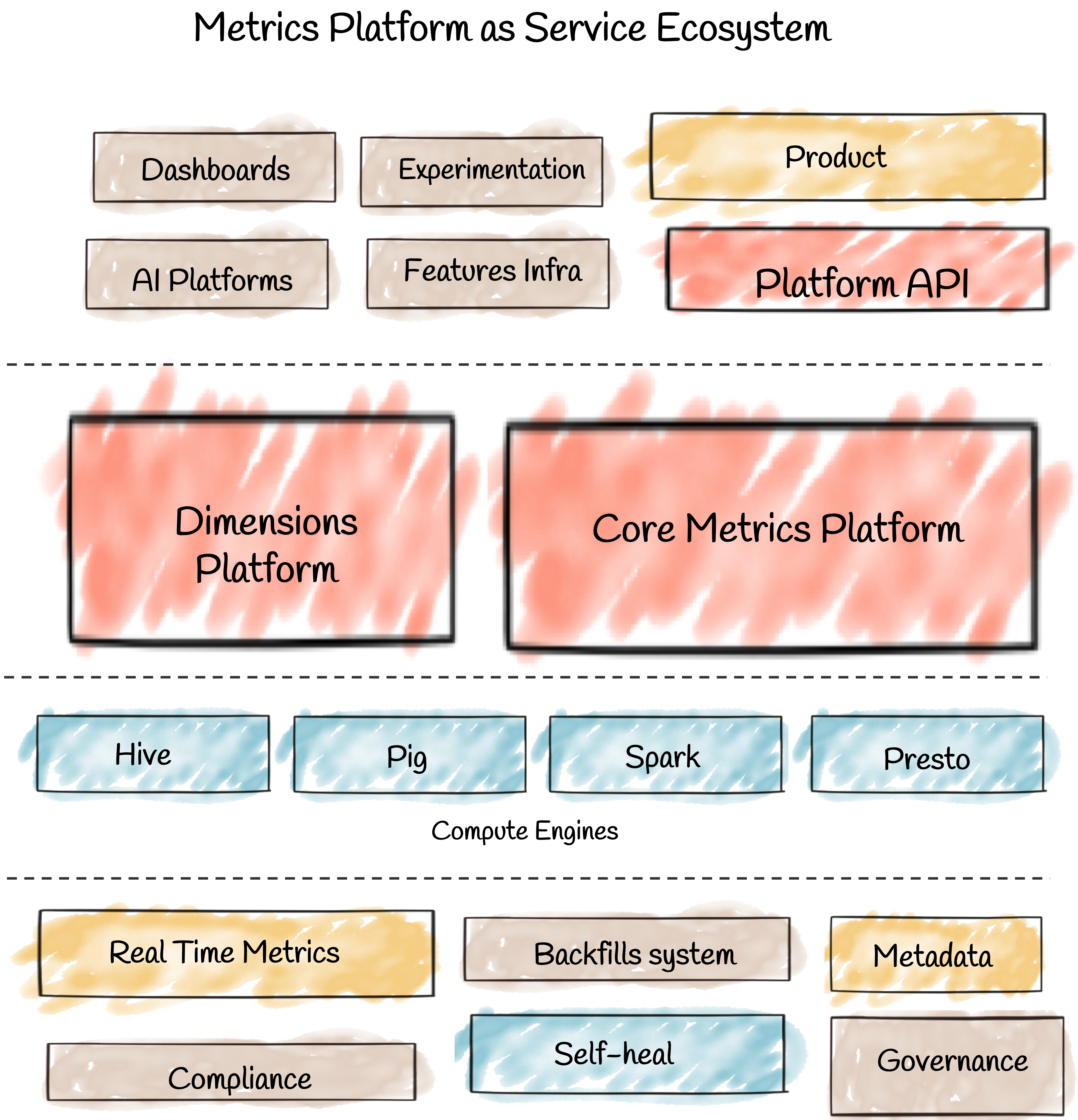


Invest in scaling governance

Build for customizations

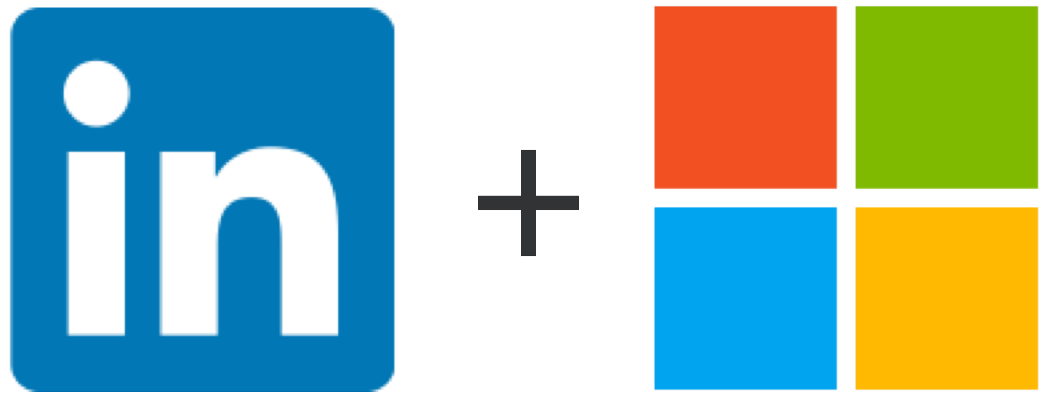
Leverage economies of scale

Evolve and Innovate



Thank you

akanitka@linkedin.com



in
+