

Building real-time analytics applications using



A LinkedIn case study

Engineer jobs in The Bay - Tell us what you love to do and have companies compete over you. Ad ...

Profile card for Kishore Gopalakrishna, Founding Engineer at Stealth Mode Startup Company. Includes 'Who's viewed your profile' and 'Views of your post' statistics.

Post creation area with 'Start a post' button and 'Write an article on LinkedIn' option.

Post by Velimir Radanovic, Architect at Oracle. Features an image of a person at a desk and the text: 'The "Two Weeks" Notice" Approach to Changing Jobs Is Bad for Companies and Employees'.

Post by Jenstep, Inc. with 906 followers. Text: 'From Silicon Valley to South Africa, India to Atlanta, join our fast-paced, high-growth FinTech firm...'.



- What people are talking about now: Trades paying better than college?, Facebook faces global outage, The top talent of top companies, World's largest plane files, Credit scores can alter dating odds.

Job advertisement for Kishore at Pure Storage, listing roles like Software Engineer, Full Stack and Senior Data Scientist.

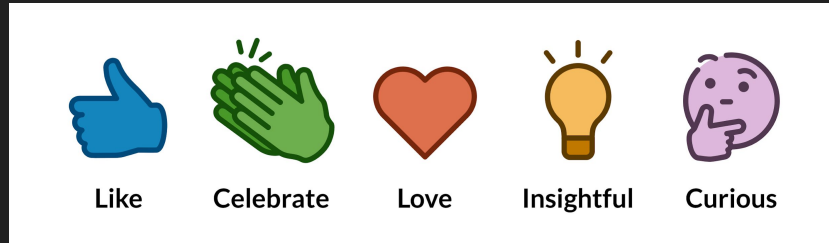
Learning card for 'Origins of Spark' video, part of Apache Spark Essential Training.

- Member
- Job
- Ad
- Post
- Company
- Course

LinkedIn Activity Data Model



Actor

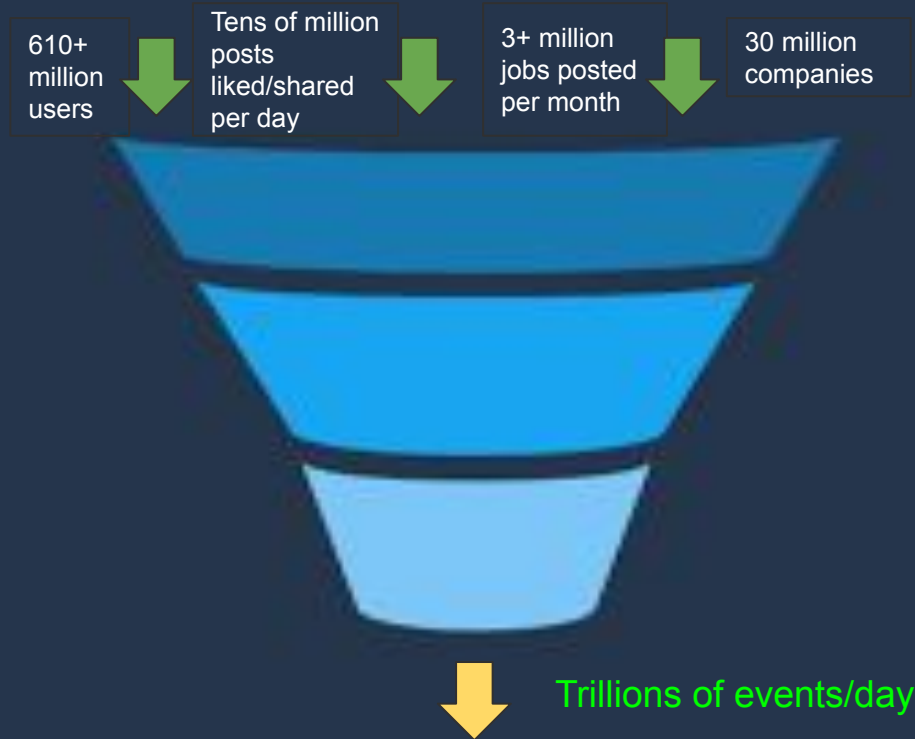


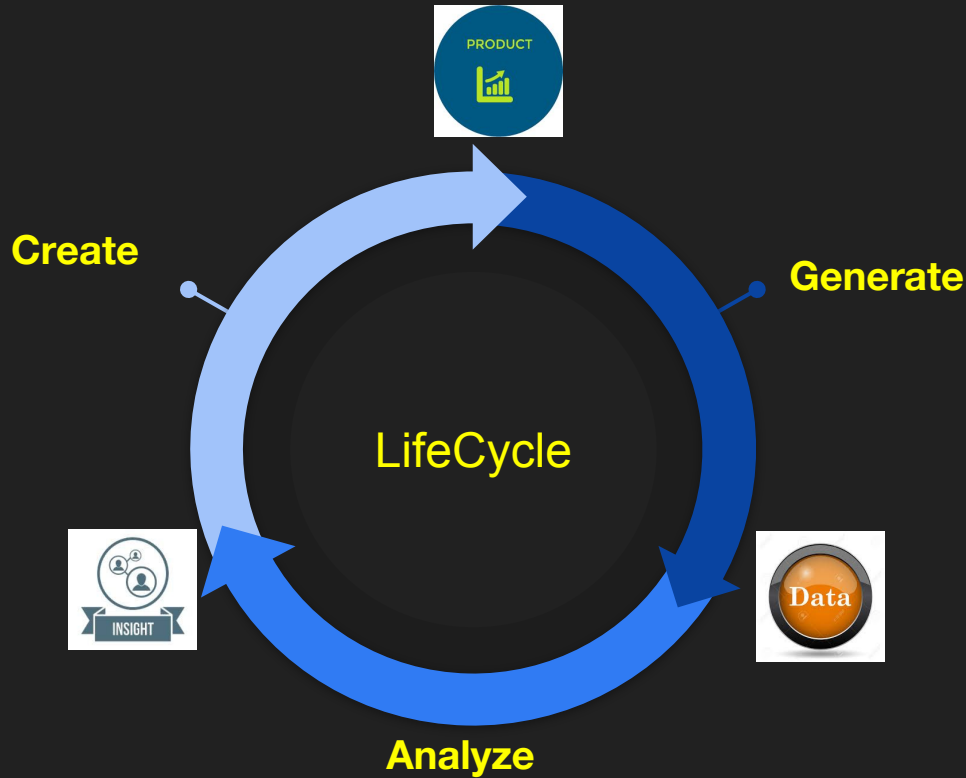
Verb

Member
Job
Ad
Post
Company
Course

Object

Activity Data Scale





What can we do with all the activity data?

Pinot @ LinkedIn

50+ site facing use cases



Hundreds of thousands of QPS at peak



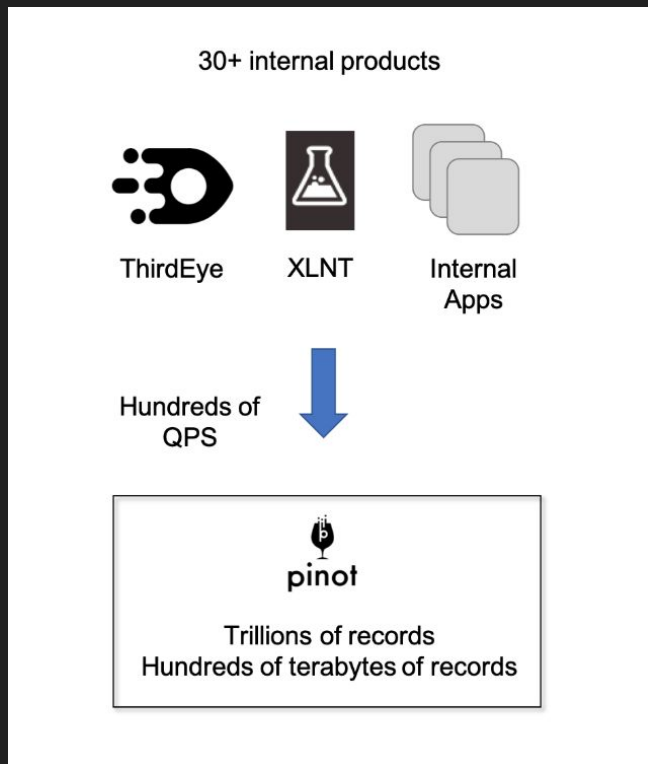
10s to 100s of milliseconds query time



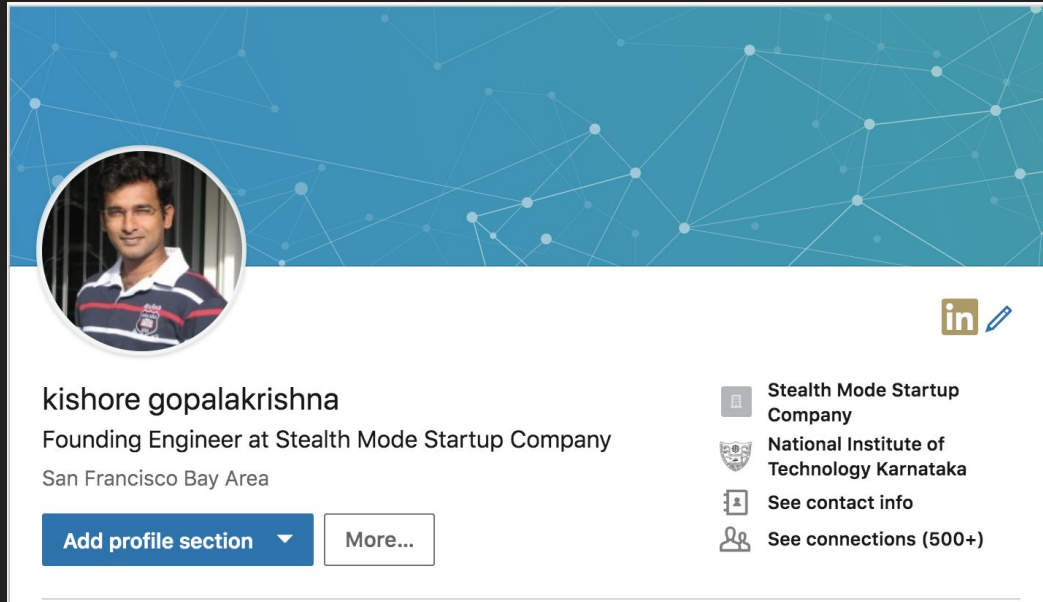
pinot

Hundreds of billions of records
Tens of terabytes of data

Pinot @ LinkedIn



Who Am I



A LinkedIn profile card for kishore gopalakrishna. The header is a blue banner with a white network graph pattern. On the left is a circular profile picture of a man in a striped polo shirt. To the right of the picture is a white box containing the text: "kishore gopalakrishna", "Founding Engineer at Stealth Mode Startup Company", and "San Francisco Bay Area". Below this text are two buttons: "Add profile section" (blue) and "More..." (white). To the right of the profile picture is a small LinkedIn icon with a pencil. Below the profile picture is a list of four items: "Stealth Mode Startup Company" (with a company icon), "National Institute of Technology Karnataka" (with a university crest icon), "See contact info" (with a contact card icon), and "See connections (500+)" (with a person icon).

Profile picture of kishore gopalakrishna

kishore gopalakrishna
Founding Engineer at Stealth Mode Startup Company
San Francisco Bay Area

[Add profile section](#) [More...](#)

- Stealth Mode Startup Company
- National Institute of Technology Karnataka
- See contact info
- See connections (500+)



ThirdEye



Why am I finding it so painful

Published on December 18, 2018

[Edit article](#)

[View stats](#)



Kishore Gopalakrishna
Founding Engineer at Stealth Mode Startup
Company
[1 article](#)



6,809



446



57



0

Why am I finding it so painful 57 comments



[6,809 article views](#)

5 reshares



1,376 views from people at LinkedIn

Google	263
Microsoft	186
Facebook	169
Uber	126



1,889 have the job title Software Developer

Technology Manager	785
Product Manager	227
Engineer	223
Product Development Engineer	155



3,704 views from San Francisco Bay Area

Greater Seattle Area	382
Bengaluru Area, India	368
Greater New York City Area	242
Greater Los Angeles Area	86




Your article was found through LinkedIn.com

	152
	34
	1


[Show more](#)





Use case 1: Article Analytics

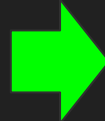


Why am I finding it so painful

Published on December 18, 2018 [Edit article](#) | [View stats](#)





 **Kishore Gopalakrishna**
Founding Engineer at Steath Mode Startup Company
1 article

 6,809  446  57  0



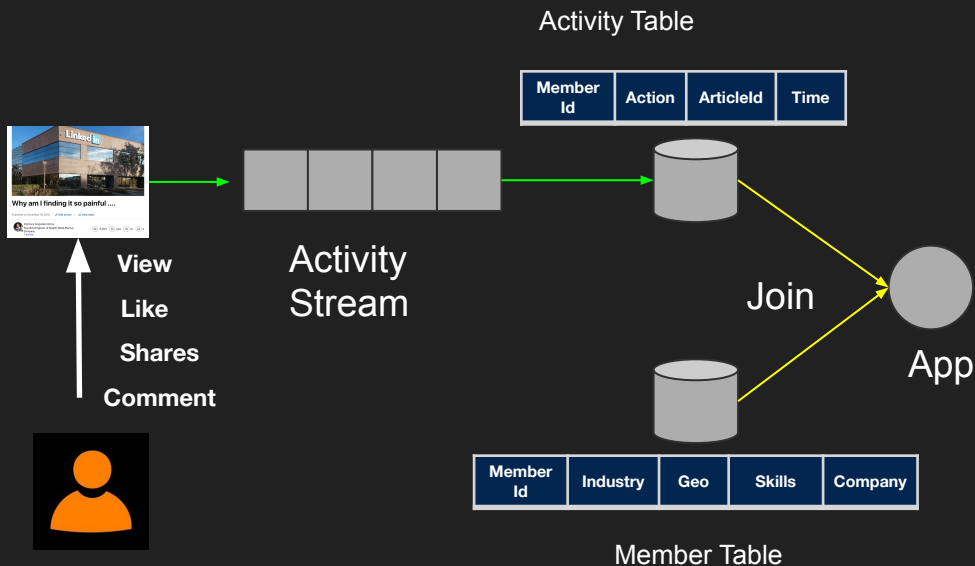
Why am I finding it so painful 57 comments

6,809 article views 5 reshares

 1,376 views from people at LinkedIn	 1,889 have the job title Software Developer	 3,704 views from San Francisco Bay Area	 Your article was found through LinkedIn.com
Google 263	Technology Manager 785	Greater Seattle Area 382	152
Microsoft 186	Product Manager 227	Bengaluru Area, India 368	34
Facebook 169	Engineer 223	Greater New York City Area 242	1
Uber 126	Product Development Engineer 155	Greater Los Angeles Area 86	

[Show more](#)

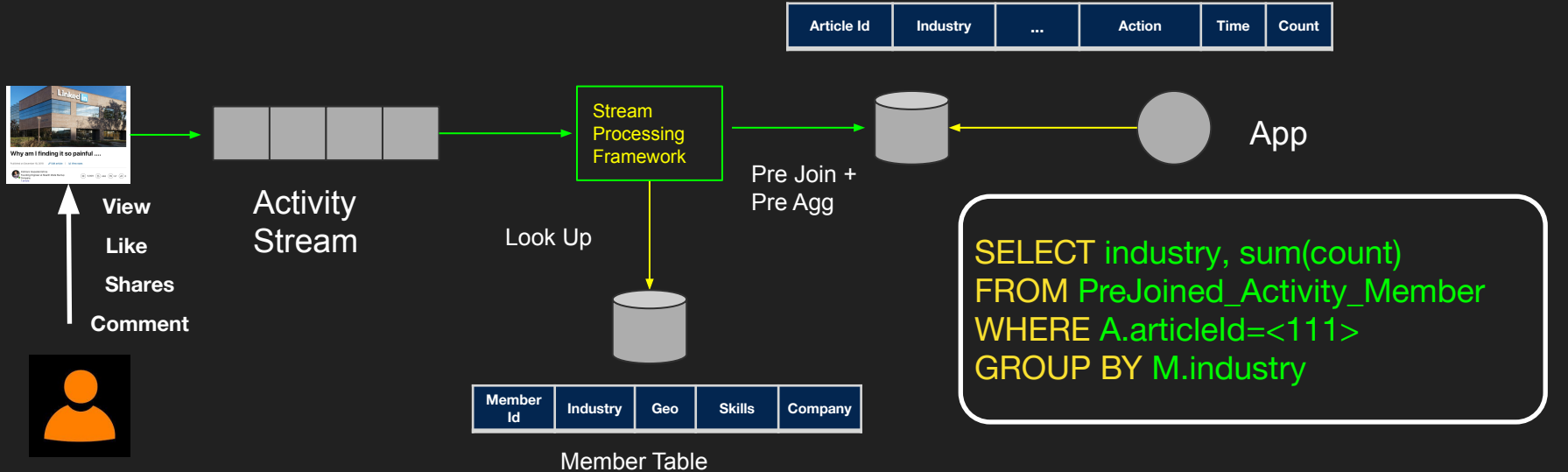
Option 1: Join on the Fly



```
SELECT M.industry, count(*)  
FROM Activity as A  
INNER join Member as M  
ON A.memberId = M.memberId  
WHERE A.articleId=<111>  
GROUP BY M.industry
```

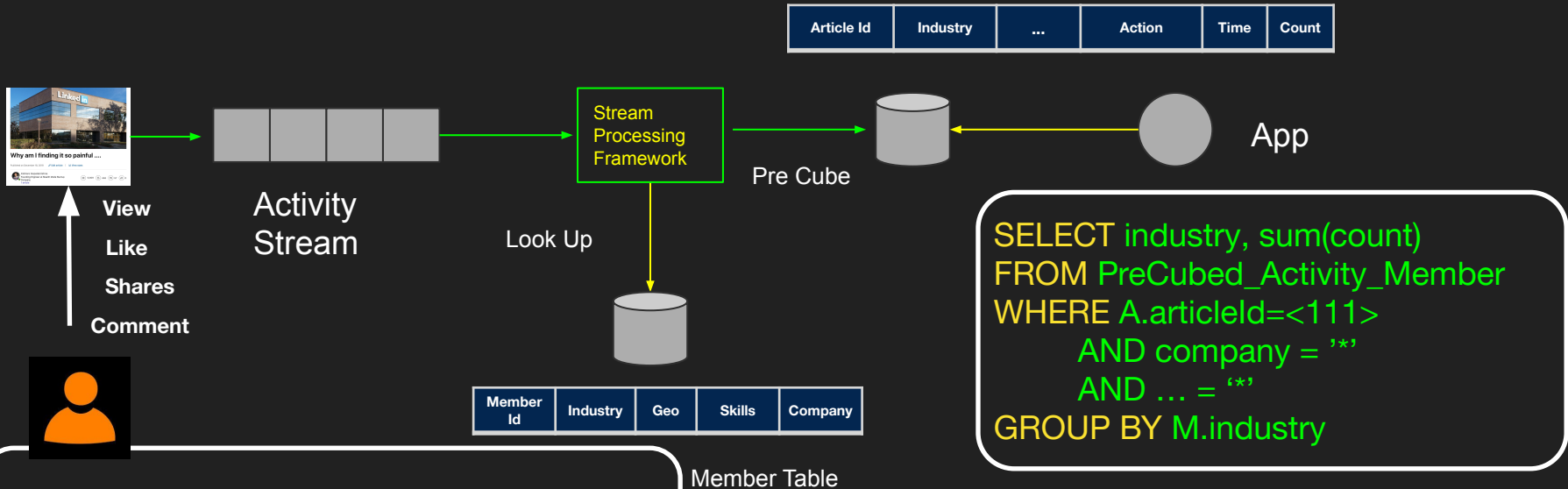
- REALTIME (Depending on storage)
- High Latency

Option 2: Pre Join + Pre Aggregate



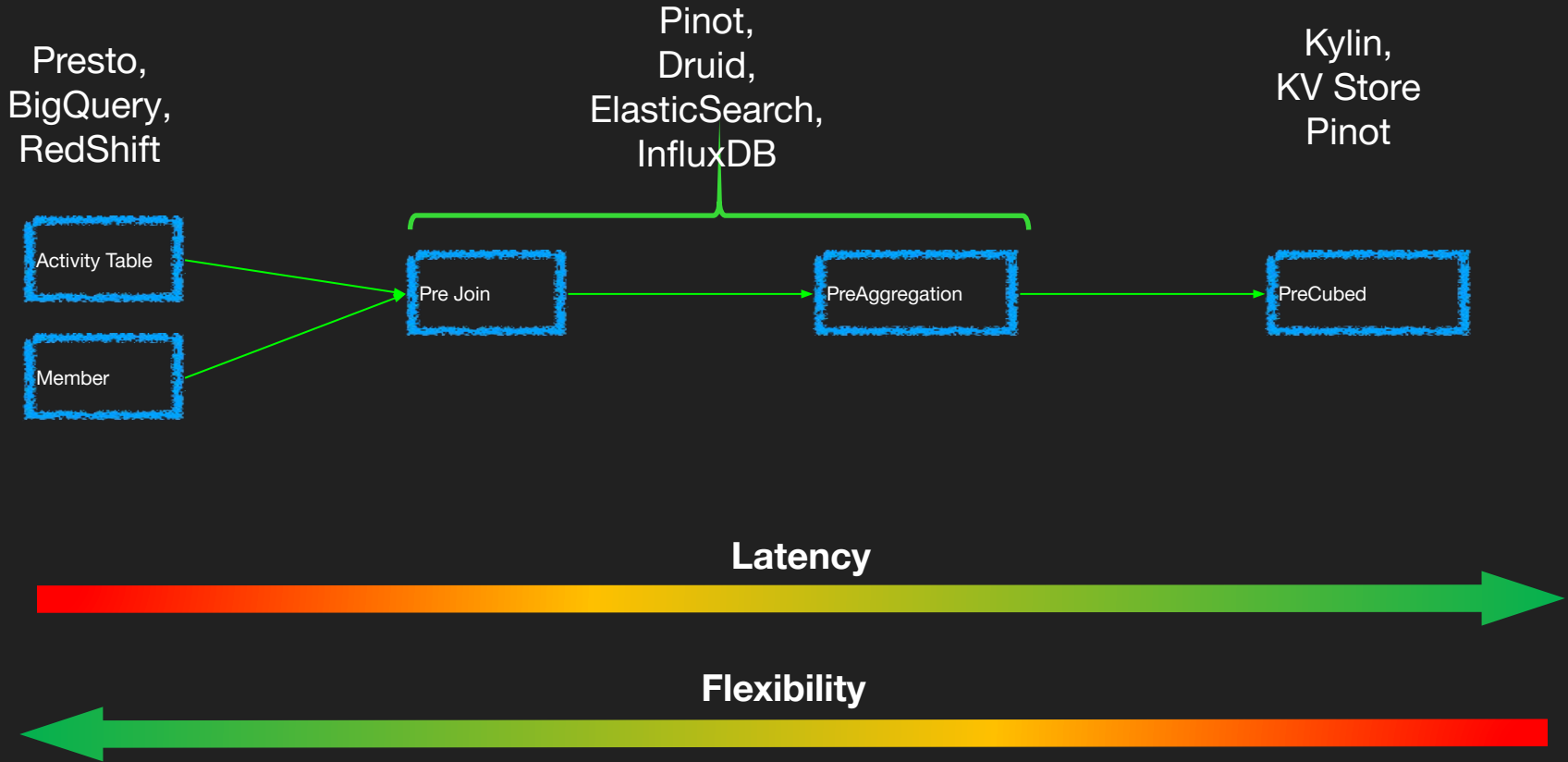
- Near real-time ingestion
- Low latency (unpredictable*)

Option 3: Pre Join + Pre Cube + Pre Agg

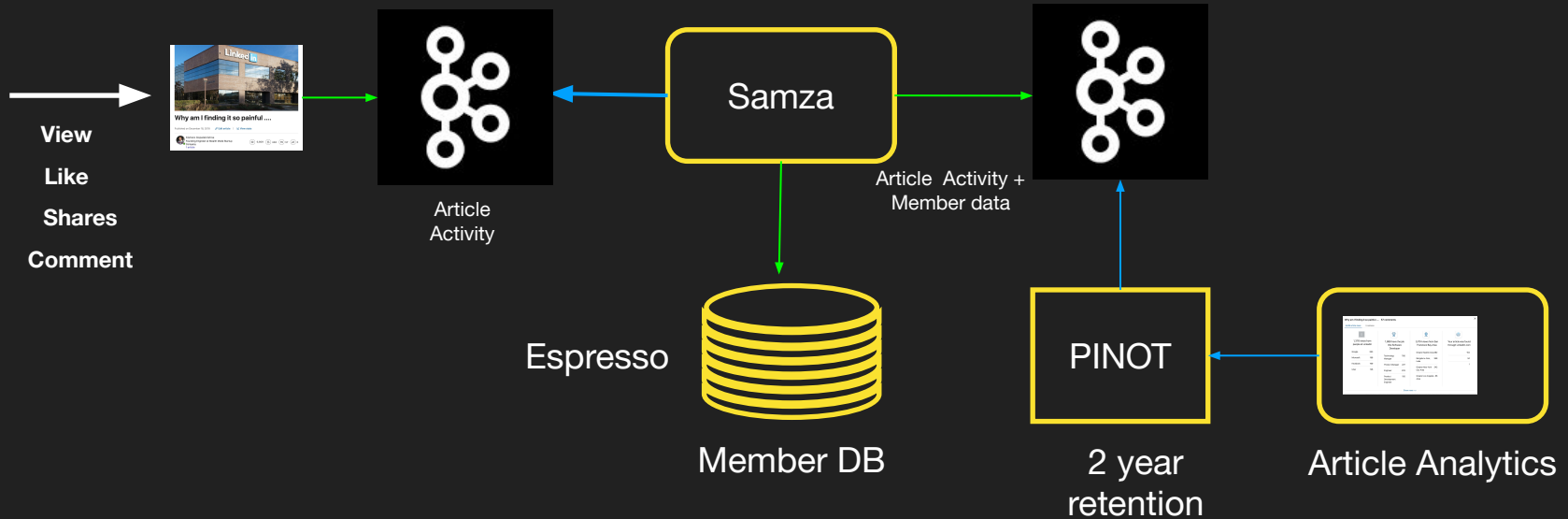


- Very fast (mostly lookup)
- Batch (Hourly/Daily)
- Extra storage (Curse of dimensionality)
- Re-bootstrap on schema changes
- Limited query capability

Comparison



Publisher Analytics Architecture

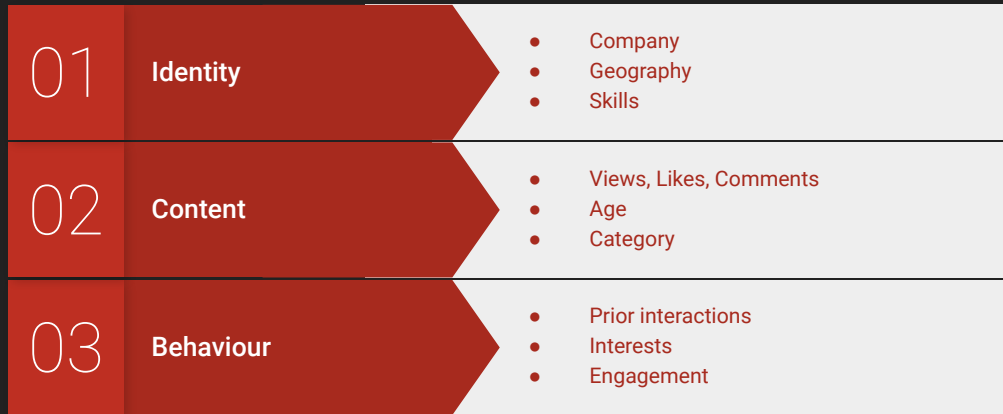




Can we use the activities data to improve the feed?

Feed Relevance

Rank the feed based on relevance



Companies apply to you... Tell us what you love to do and have companies compete over you. Ad ...

PREMIUM

kishore gopalakrishna
Founding Engineer at Stealth Mode Startup Company

Who's viewed your profile 905
Views of your post 8,730

See all Premium features

Recent

- Distributed Systems D...
- SQL on Hadoop
- LinkedIn Alumni Netw...
- The LI Massage Team
- Real Time Analytics

Groups

- Distributed Systems D...
- SQL on Hadoop
- LinkedIn Alumni Netw...

Show more

Followed Hashtags

Discover more

Start a post

Write an article on LinkedIn

Sort by: Recent

Rupesh Dabhir likes this

Punit Singh Soni • 2nd
CEO - Suki, Ex-Flipkart, Motorola, Google

"Instead of stable truth, I chose unstable possibilities" - Murakami

I like this a lot. It's all about embracing those unstable elements that surround us, enriching our life story. And nothing is more unstable than the life of a startup entrepreneur.

507 Likes • 23 Comments

Like Comment Share

Rupesh Dabhir likes this

Avik Das • 2nd
Software Engineer at LinkedIn

I've been helping a friend understand dynamic programming (DP for short), so I've been on the lookout for good resources. The topic is covered all across the web, but I found many of them focused on the code, ...

A graphical introduction to dynamic programming

medium.com

47 Likes • 14 Comments

Like Comment Share

Madhumita Mantri, **CHI-YI Kuan** and 12 others follow **Talend**

Talend
23,918 followers

If you're eager to see what machine learning can do, we've got just what you need to build smarter data pipelines. <https://bit.ly/2oB5n1c>

What people are talking about now

- Boeing 737 MAX may soon fly again
17h ago • 3,829 readers
- Salesforce.com buys Salesforce.org
8h ago • 15,769 readers
- TikTok snaps up ex-Snap employees
2h ago • 4,210 readers
- The most stressed people at work
8h ago • 74,441 readers
- Time lists 'most influential' people
2h ago • 12,760 readers

Show more

Promoted

- Companies apply to you. Tell us what you love to do and have companies compete over you.
- Top-Ranked Jack Welch MBA
Most MBA programs study great leaders. Ours is led by one. Apply by 07/01.
- UC Berkeley Masters
Earn your master's in cybersecurity online in 20 months. GRE/GMAT required.

LEARNING

Stay sharp on your current skills

- Finding similar values with LIKE

From Analyzing Big Data with Hive

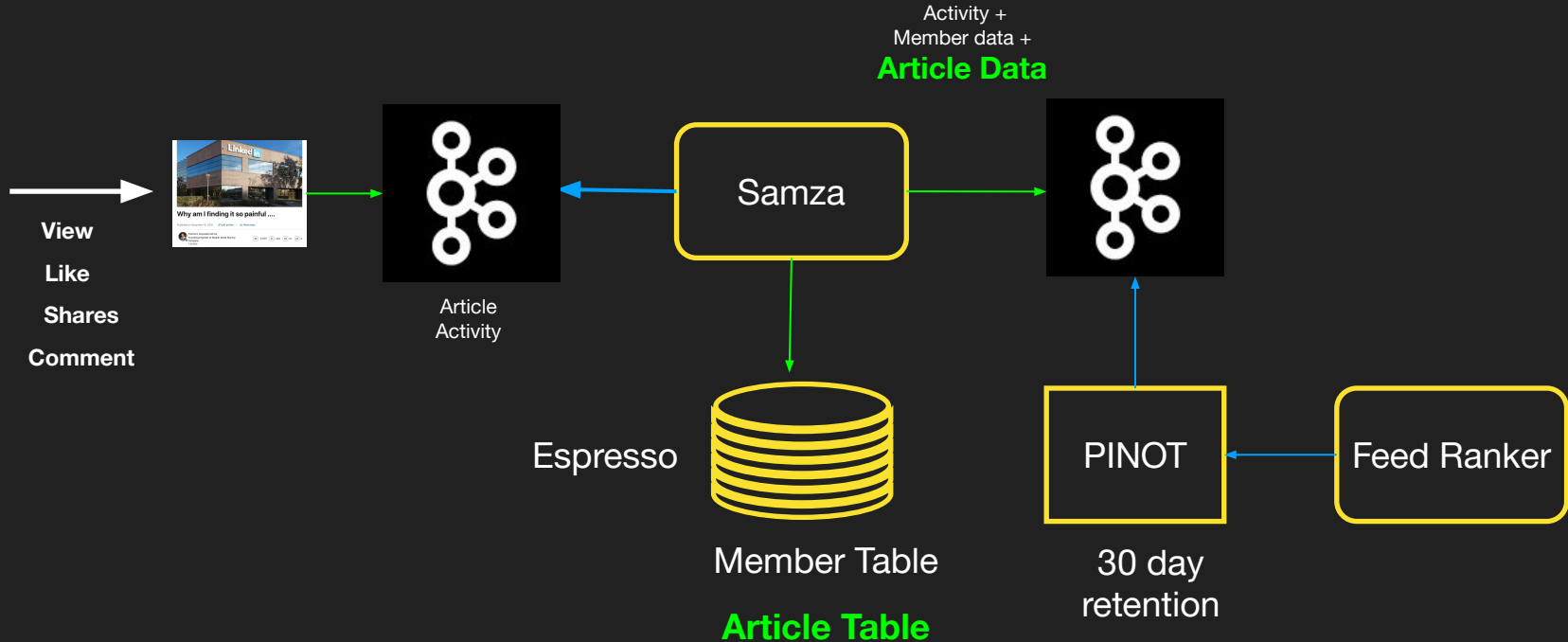
About Help Center Privacy & Terms

Advertising Business Services

Get the LinkedIn app More

LinkedIn LinkedIn Corporation © 2019

Feed Ranking Architecture



Feed Ranking Perf Numbers

Companies apply to you... Tell us what you love to do and have companies compete over you. Ad ...

kishore gopalakrishna
Founding Engineer at Stealth Mode Startup Company

Who's viewed your profile 905
Views of your post 8,730

Recent

- Distributed Systems D...
- SQL on Hadoop
- LinkedIn Alumni Netw...
- The LI Massage Team
- Real Time Analytics

Groups

- Distributed Systems D...
- SQL on Hadoop
- LinkedIn Alumni Netw...

Followed Hashtags

Discover more

Start a post

Write an article on LinkedIn

Sort by: Recent

Rupesh Dabhir likes this

Punit Singh Soni • 2nd
CEO - Suki, Ex-Flipkart, Motorola, Google

Instead of stable truth, I chose unstable possibilities - Murakami

I like this a lot. It's all about embracing those unstable elements that surround us, enriching our life story. And nothing is more unstable than the life of a startup entrepreneur.

507 Likes · 23 Comments

Like Comment Share

Rupesh Dabhir likes this

Avik Das • 2nd
Software Engineer at LinkedIn

I've been helping a friend understand dynamic programming (DP for short), so I've been on the lookout for good resources. The topic is covered all across the web, but I found many of them focused on the code, c...see more

A graphical introduction to dynamic programming

medium.com

47 Likes · 14 Comments

Like Comment Share

Madhumita Mantri, **Chi-Yi Kuan** and 12 others follow **Talend**

Talend
25,918 followers

If you're eager to see what machine learning can do, we've got just what you need to build smarter data pipelines. <https://tldr.info/8F9n>

What people are talking about now

- Boeing 737 MAX may soon fly again 17h ago • 3,609 readers
- Salesforce.com buys Salesforce.org 8h ago • 15,709 readers
- TikTok snaps up ex-Snap employees 2h ago • 4,210 readers
- The most stressed people at work 8h ago • 74,441 readers
- Time lists 'most influential' people 2h ago • 12,760 readers

Promoted

- Companies apply to you. Tell us what you love to do and have companies compete over you.
- Top-Ranked Jack Welch MBA Most MBA programs study great leaders. Ours is led by one. Apply by 07/01.
- UC Berkeley Masters' Earn your master's in cybersecurity online in 20 months. GRE/GMAT required.

LEARNING

Stay sharp on your current skills

- Finding similar values with LIKE

From Analyzing Big Data with Hive

About Help Center Privacy & Terms Advertising Business Services Get the LinkedIn app More

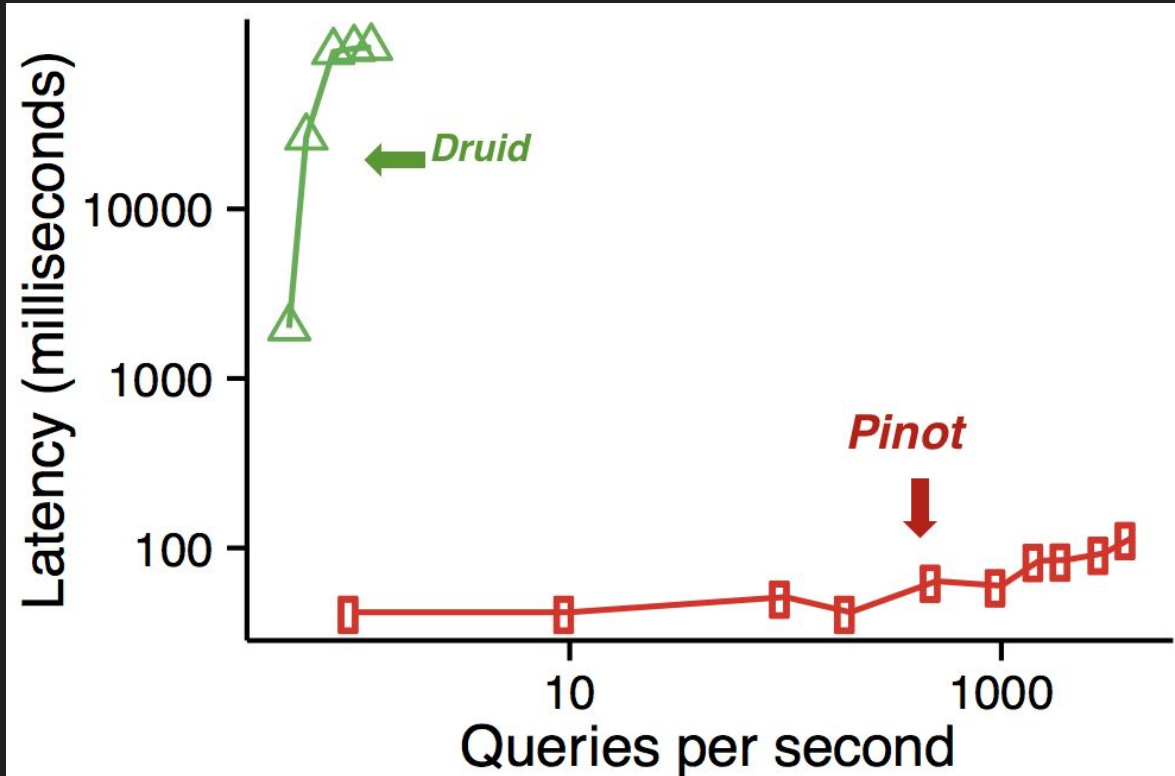
LinkedIn LinkedIn Corporation © 2019

SELECT sum(count) from T
WHERE memberId = <>
AND article in (list of 1500 items)
AND time >= (now - 14 days)
GROUP BY action, item, position, time

QPS	p50	p90	p99	p99.9
6400	5ms	25ms	45ms	100ms

Significant increase in engagement

Site Facing use case: Pinot vs Druid

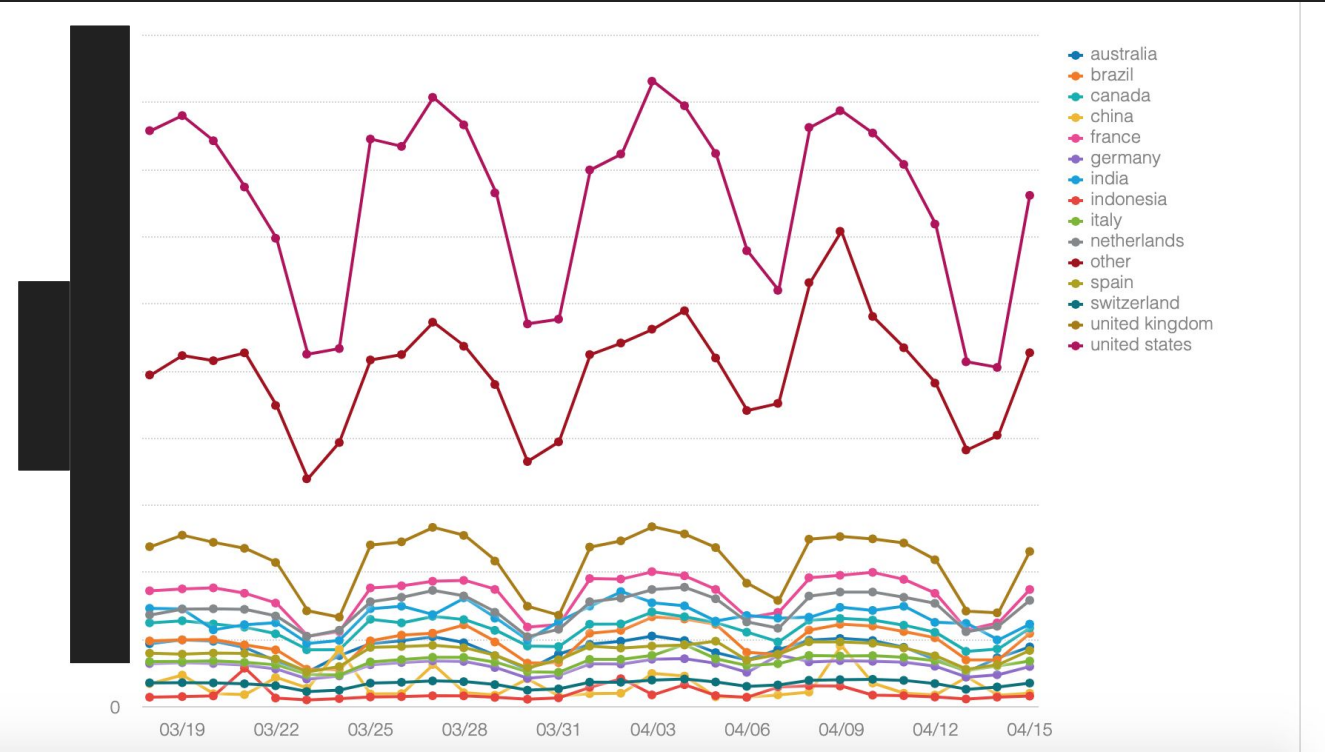


- Sorted Index
- Per query optimizer
- Optional indexing



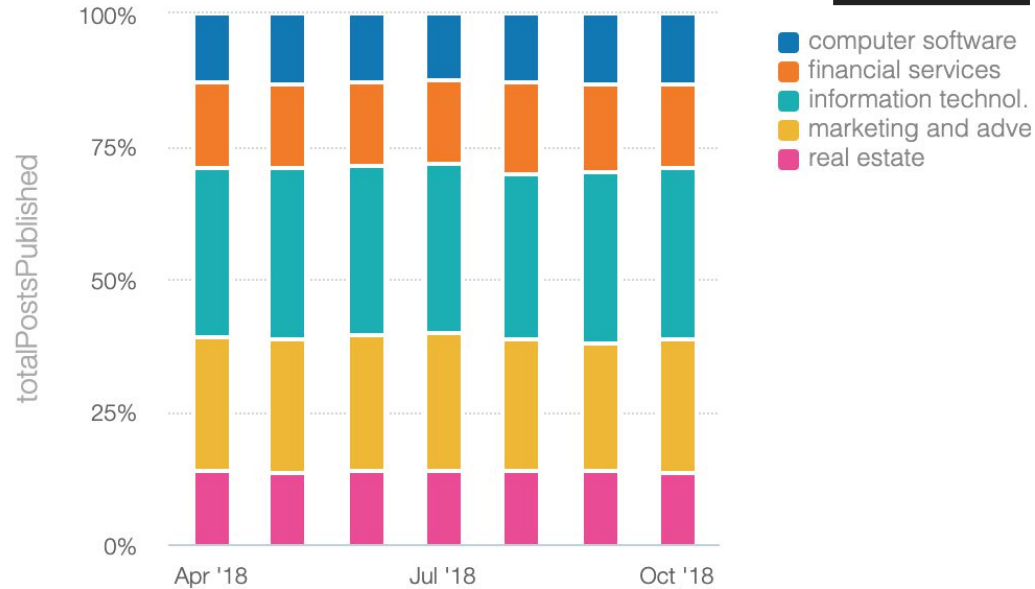
What Business Insights can we generate from this data?

Posts Published: Breakdown By Country

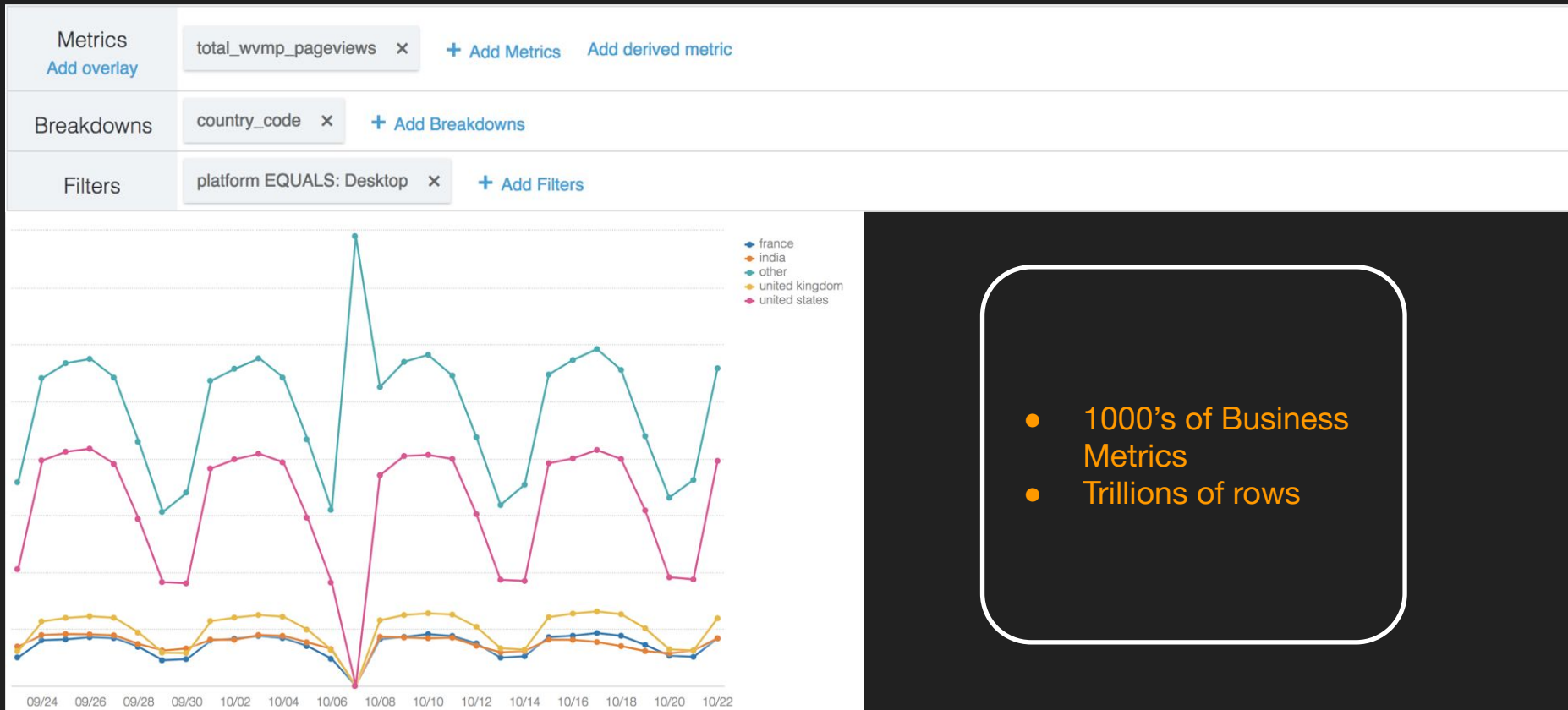


Distribution: By Industry

Monthly Posts Published by Industry

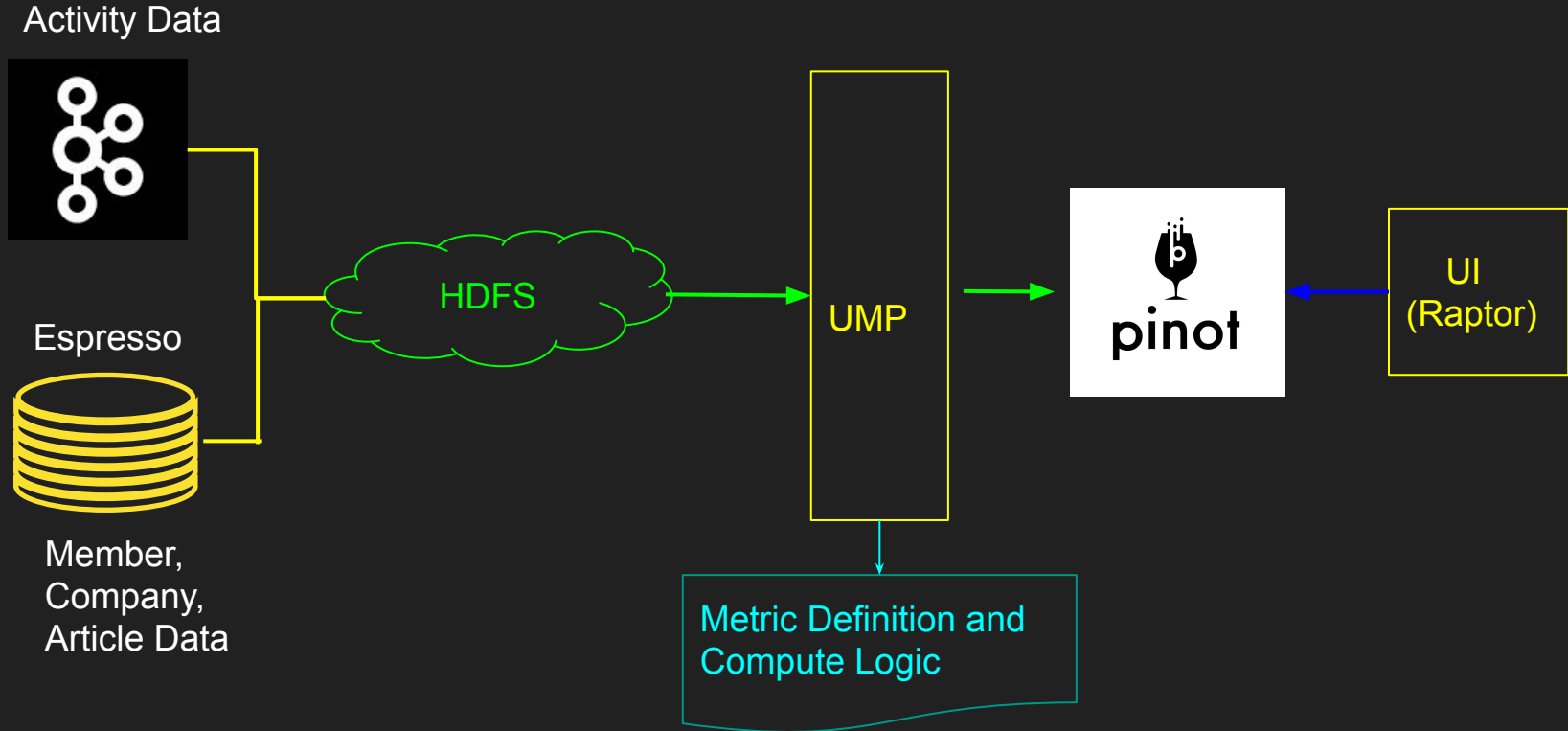


Slice and Dice UI

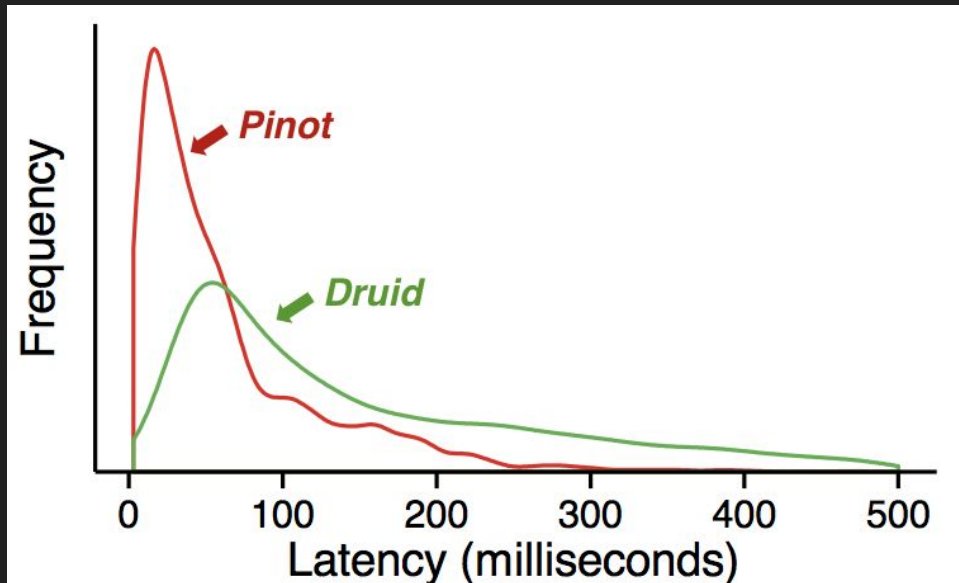


- 1000's of Business Metrics
- Trillions of rows

Dashboard Pipeline Architecture



Dashboard use case: Pinot vs Druid



- ~ 5000 random queries of the form
 - `select sum(views), time from T where country = us, browser = chrome,...`
`group by Date`
- run sequentially one after the other

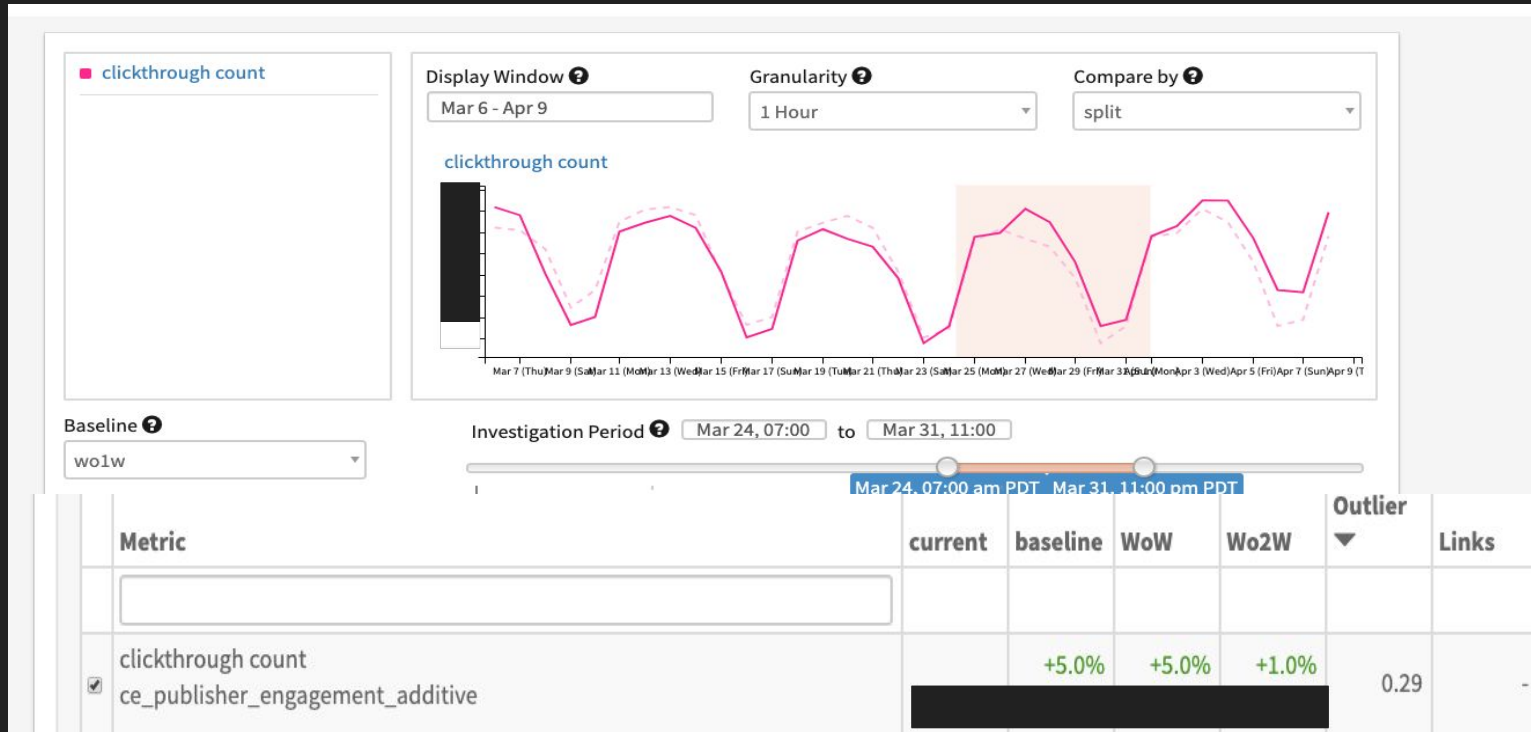
	Pinot	Druid
Total time	11 minutes	24 minutes
p50	84 ms	136 ms
p90	206 ms	667 ms

Anomaly Detection



Why don't we monitor these metric and alert?

ThirdEye: Anomaly Detection



ThirdEye:Root Cause Analysis

18
Dimensions

action_type	viewLink (18.74)	viewArticle (18.77)				
Domain	Other Site (-0.6)	LinkedIn (-0.53)	null (3.84)	Busi...	OTHER	
article_type	3rd Party (-3.69)	1st Party (-0.15)	null (3.84)			
author_type	Member (-0.66)	OTHER				
connection_count	90-499 (-0.86)	500-999 (0.5)	1000-1999 (0.72)	2000+ (0.83)	30-89 (-0.01)	OTHER
country_code	united states (-0.38)	other (0.03)	united king... france (... netherl... india (... canada... brazil (0... spain (... germ... australia... italy (... chin...	OTHER		
author_type	other (-0.14)	information... financ...	computer s... banking... oil a... auto... pharbumsaacco... edu... marketing... hospita... inter... reta... high... real... law...	OTHER		
#connections	not_tagged (-0.93)	OTHER				
industry	fourbyfour (0.68)	onebythree (-0.35)	OTHER			
....	home-feed:phone (0.78)	home-feed:desktop (... feed-item... conten...	home_f... feed-item... OTHER			
location	NULL (-1.56)	linkedin:share (1.54)				
verb_type	NULL (2.27)	desktop (-0.1)	api (-0.1)	phone app... OTHER		
	NULL (2.27)	member (-0.97)	company (-0.1)			
	VOYAGER (1.12)	LINKEDIN (-0.64)	OT...			
	share (-0.01)	viral_update (0.17)	content_po... recs (0.0... OTHER			
	NULL (0)					
	phone app (0.74)	desktop (-0.63)	phone br... OTHER			
	linkedin:share (-0.59)	linkedin:like (0.53)	linkedin... linkedi... OTHER			

Interactive

break down

sub-second

Multiple queries

Anomaly Detection

```
SELECT sum(view), time  
FROM PostView  
GROUP BY time
```

TOP LEVEL

```
for d1 in [us, ca, ... ]  
for d2 in [chrome, ie, ... ]  
...  
SELECT sum(view), time  
FROM PostView  
WHERE  
    country = d1  
    AND browser = d2  
    AND ...  
GROUP BY time
```

MULTI DIMENSIONAL

Multi-dimensional anomaly detection challenges

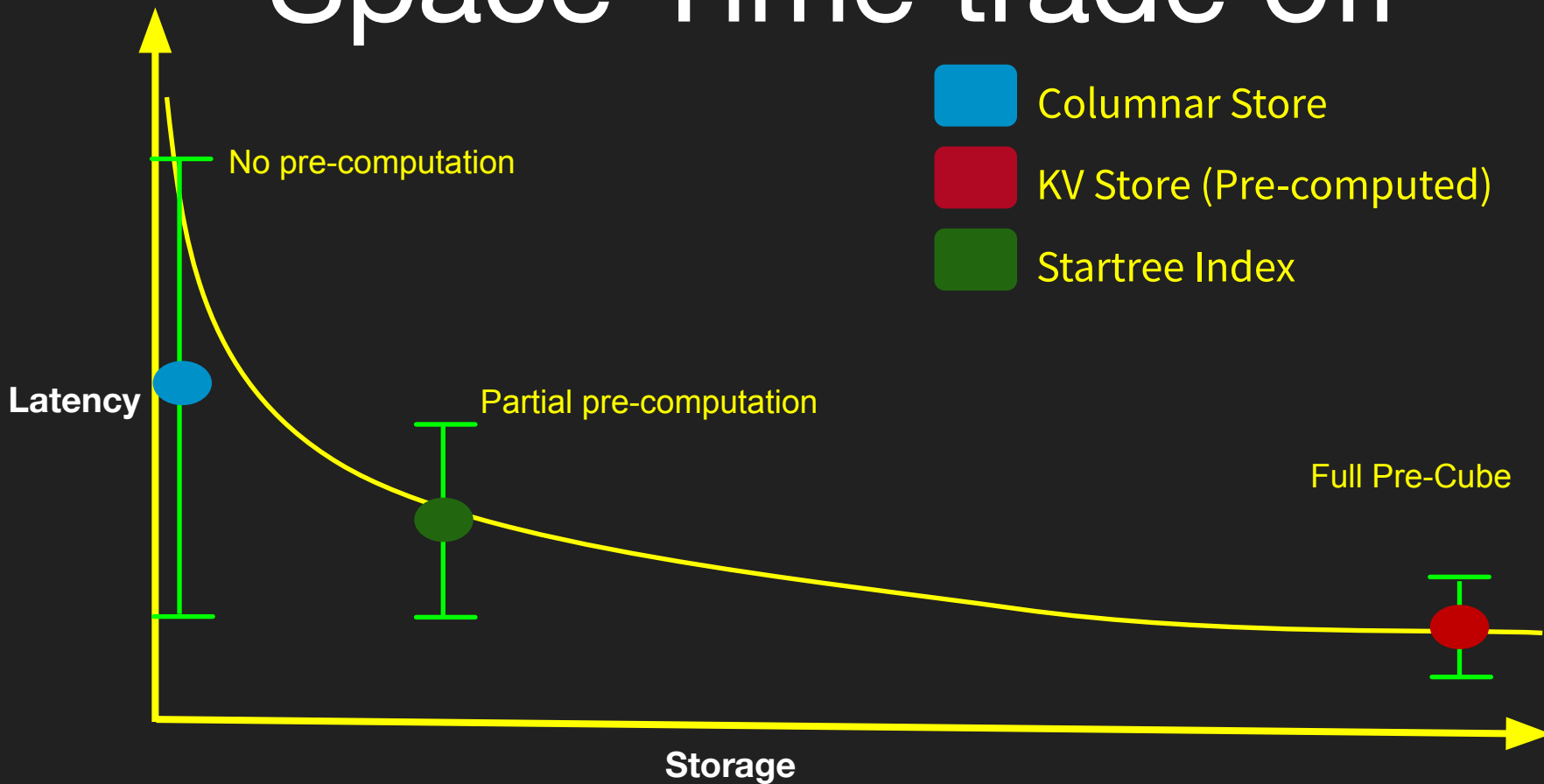
```
for d1 in [us, ca, ... ]
for d2 in [chrome, ie, ... ]
...
SELECT sum(view), time
FROM PostView
WHERE
    country = d1
    AND browser = d2
    AND ...
GROUP BY time
```

MULTI DIMENSIONAL

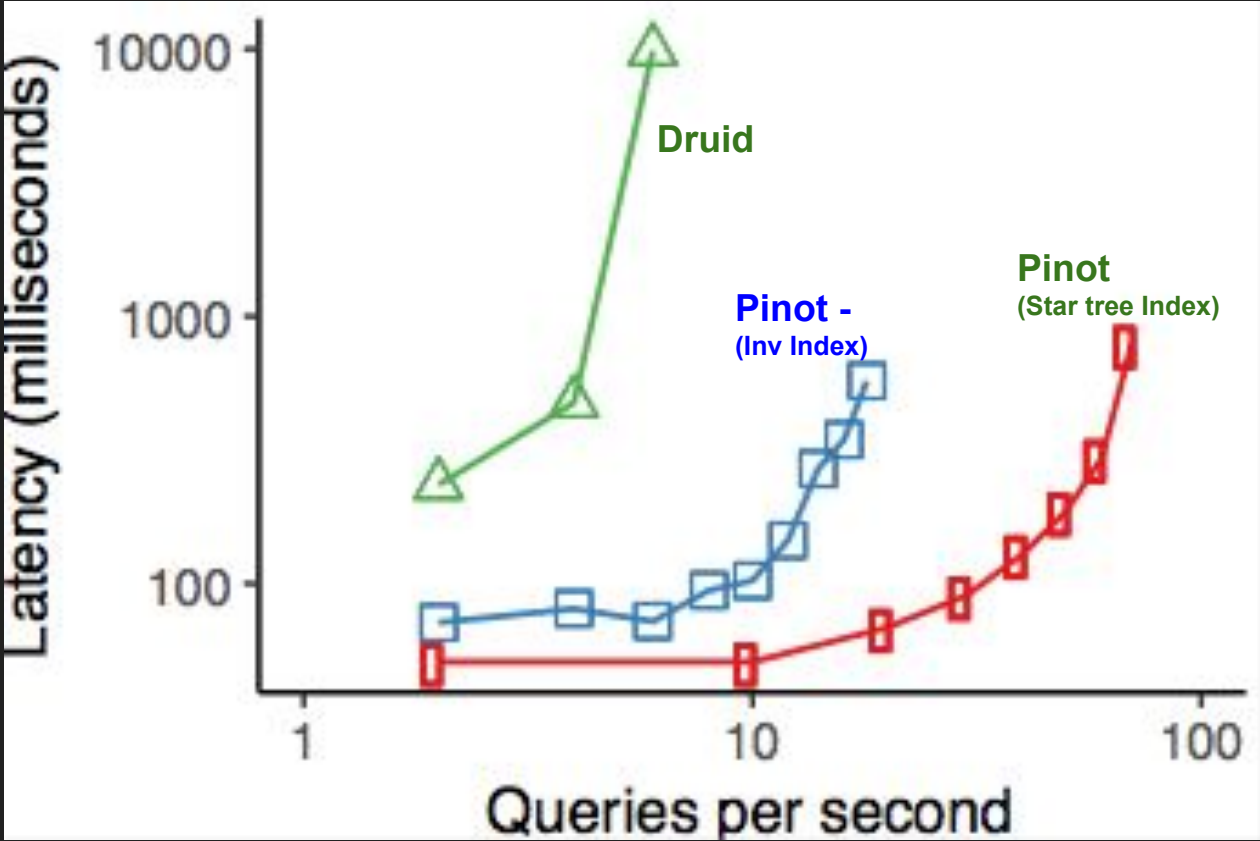
1. *Identifying issues requires monitoring all possible combinations*
2. *No Id column (ArticleId, Member Id)*
3. *Latency is unpredictable even with Inverted Index*

select sum(view) where country="us'	scan 60-70% of the rows	Slow
country="ireland'	scan <1% of the rows	Fast

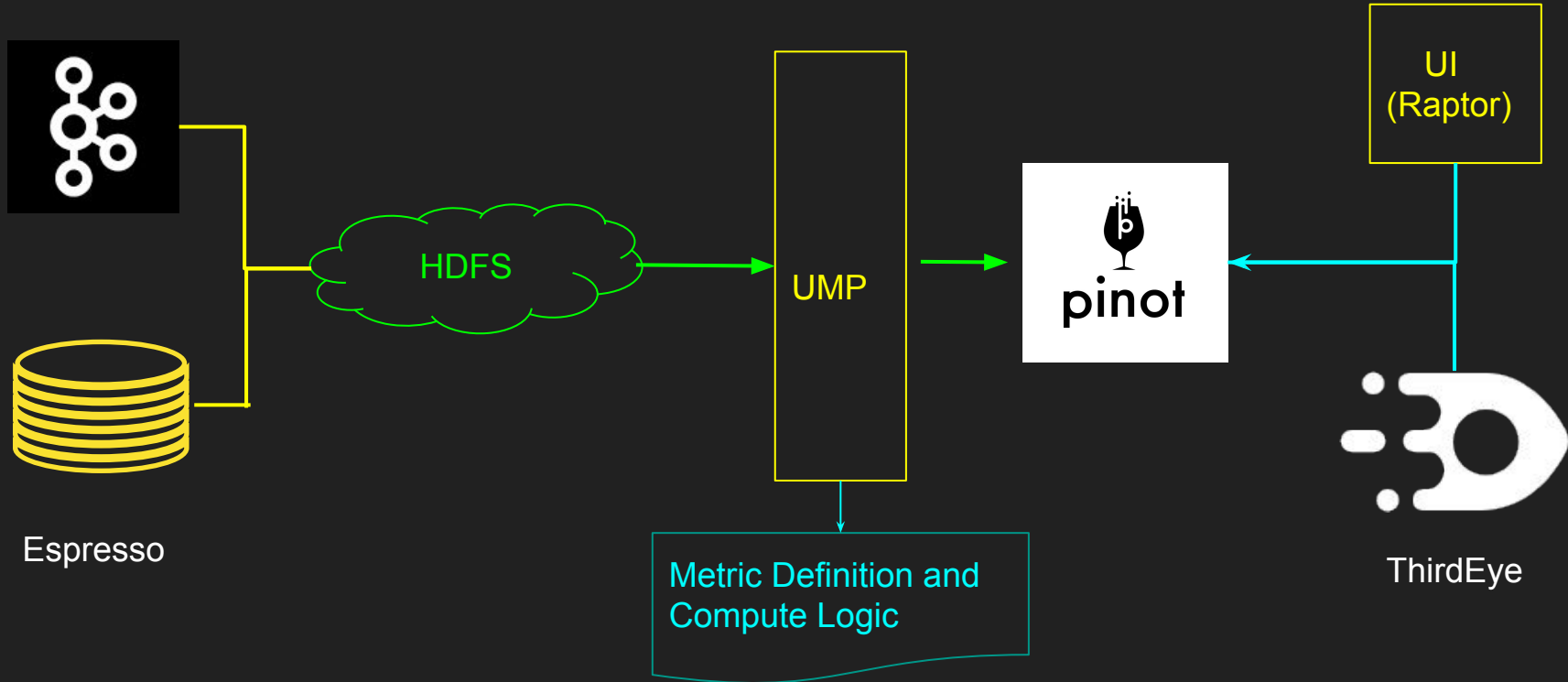
Space-Time trade off



Anomaly Detection: Druid vs Pinot



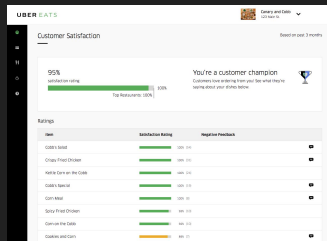
Anomaly Detection Architecture



Pinot usage



50TB
1000 qps



✓ UberEATS

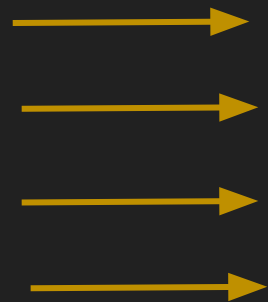
- ✓ MarketPlace
- ✓ UberPool
- ✓ UberFreight
- ✓ Jump



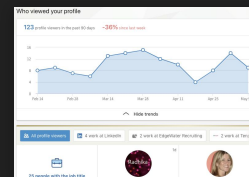
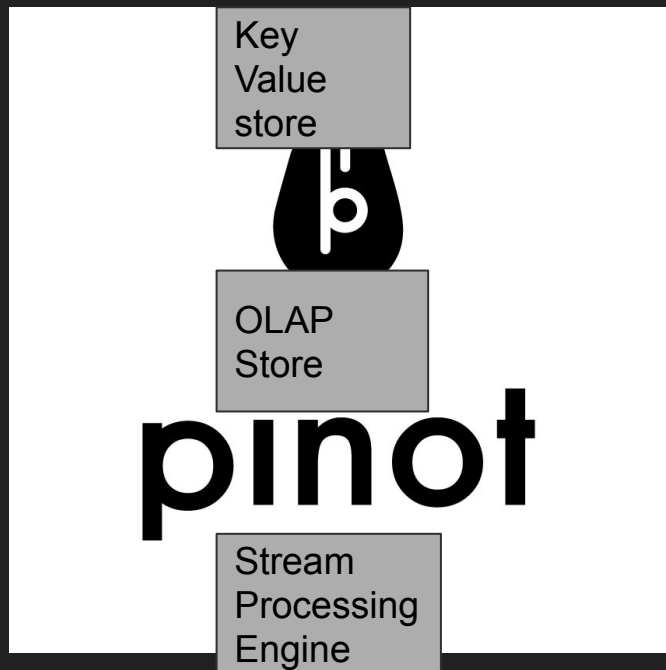
Hi Just want to let you know the power of Pinot, the same data & usecase I have to use 45 node each 256gb ram machines to index it in druid. I'm using 18 node 122gb machine now for Pinot, but that itself is over provisioned!! 😊



Conclusion



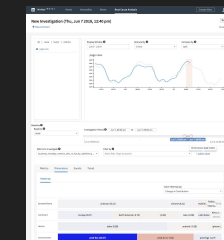
Activity Data



Site Facing Applications



Dashboard: Business Analytics



Anomaly Detection

Questions



Website	http://pinot.apache.org
Slack	apache-pinot.slack.com
Twitter Handle	@apachepinot, @kishoreBytes