



**Turning data into products**

Sam Shah

# LinkedIn: the professional profile of record

LinkedIn Account Type: Pro 7

Sydney Shoup Add Connections

Home Profile Contacts Groups Jobs Inbox Companies News More People Search... Advanced

High Yield Savings - Great Rate. No Monthly Fees. Open an Online Account at CIT Bank Today.

**Lauren Bowen**  
Design is my lifelong craft.  
San Francisco, California | Internet  
Current LinkedIn, Napkin Sketches  
Previous LinkedIn, Advent Software, Tamale Software  
Education Georgia Institute of Technology

Improve your profile Edit 500+ connections Contact Info

ACTIVITY Share an update...

Lauren Bowen via LinkedIn Today

How to Influence Your Company Culture (For the Better)  
Inc.com • Benefits, perks and compensation may paint a rosy picture for recruitment, but it's these intangible elements that make for a happy and healthy company culture.

250M Members

BACKGROUND

 SUMMARY

My passion to create and build has served as the pivotal foundation in my career as an engineer and

PEOPLE YOU MAY KNOW

 Juliana Williams 2° Director of Design Connect

Ads by LinkedIn Members

 Learn to design w/Agile Full-day UX design workshop w/Anders Ramsay, Toronto: Nov 30. Learn More »

 Dreamers, Pirates & You Come join a growing group who are playing bigger in the Bay Area Learn More »

PROFILE STRENGTH 2 All-Star

250M Member Profiles

Share your profile

YOUR NETWORK Company 4

# We have a lot of data.

We want to leverage this data to build products.



# Applications

**Who's Viewed Your Profile**

**TODAY**

- Samir Sha Principal San Francisco In Common
- Jay Krep Principal San Francisco In Common

**Viewers of Your Profile**

**S P**

- SP

**YESTERDAY**

- Matthi Senior Software Engineer San Francisco In Common
- Lili Wu Senior Software Engineer San Francisco In Common

**A R**

- AR

**MORE THAN TWO DAYS AGO**

- JP Gordon K <script>a</script> San Francisco
- Baq Hai Senior Software Engineer San Francisco
- Mitul Tiw Senior Sales San Francisco
- Anmol Bh Engineering San Francisco
- William V Senior Software Engineer San Francisco

**Evion H. I**  
Software Engineer at LinkedIn  
San Francisco Bay Area | Computer Software  
In Common: > 3 shared connections > 2 shared groups

Joseph, 80 connections changed jobs so far.

messages-noreply@bounce.linkedin.com

Wednesday, August 3, 2011 6:07 PM

To: Joseph Adler

Joseph,



How time flies. We're already halfway through 2011 and 80 of your LinkedIn connections have started something new. Here are some of them.

Click to view their new titles



See Nick's new job! »

Where is Simon? »

And Mike? »

Have you started something new in 2011?

[Let your connections know.](#)

You are receiving LinkedIn Marketing emails. [Unsubscribe](#)

© 2011, LinkedIn Corporation, 2029 Stierlin Ct, Mountain View, CA 94043

World Total

371

Views 329

Feb 26



News

22

4

371

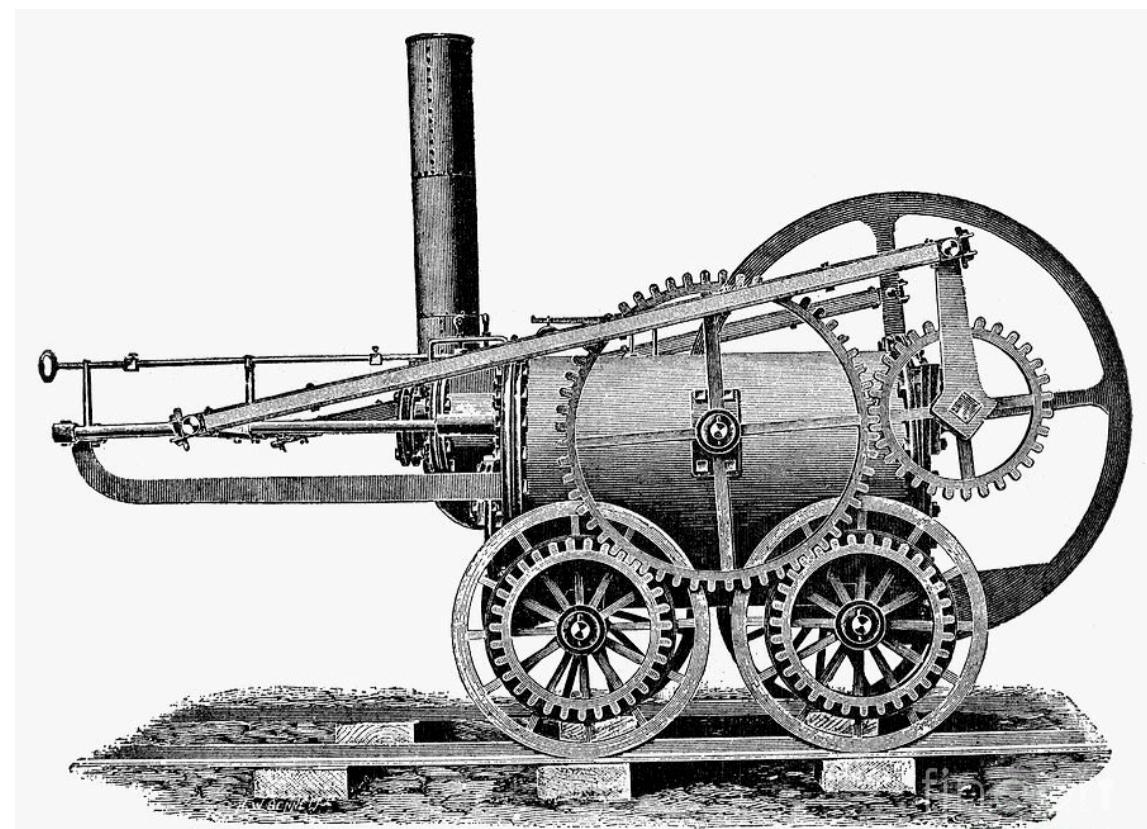
# Homepage powered by data

The screenshot shows a LinkedIn homepage with several data-powered modules:

- Top Navigation:** Home, Profile, Contacts, Groups, Jobs, Inbox, Companies, News, More, People, Search, Advanced.
- Header:** MIT Executive Education - MIT's Unique Entrepreneurial Education Program, Learn More & Enroll Today.
- Share an update:** A placeholder for sharing content.
- LinkedIn Today:** Headlines include "Wikipedia Blackout: Survive with These 12 Alternatives", "'No' is the New 'Yes': Four Practices to Rebrandize Your Company", and "Yahoo Co-Founder Jerry Yang Resigns".
- All Updates:** A feed of recent connections and activity, such as "Chad Whitney is now connected to Benjamin Menasha, Consulting Engineer at ITA Software and Eel-Bart Parco, Telecom Technician at LinkedIn".
- Who's Viewed Your Profile?** Shows 515 connections and 100+ profile viewers in the last 3 months, with 2 new viewers since 01/01/13.
- Jobs You May Be Interested In:** Listings for KDN Bioinformatics/Senior Bioinformatics, Senior Staff Systems Engineer (Data...), and Sr. Applied Scientist.
- Groups You May Like:** AnalyticsBridge, Bayesian Belief Networks with BayesiaLab, and Sentiment Analysis Symposium.
- Companies You May Want To Follow:** ness, Linguamatics, SAS, and ontotext.
- People You May Know:** Recommendations for Bhupesh Heda, Shomit Ghose, and Alay Mathew.
- Advertisement:** Get over 30 premium benefits and powerful spend capacity, with a LEARN MORE button.

## We're in the “pre-industrial age” of Big Data

- Need “bridges & railways”



# Algorithms

# Year in Review

Joseph, 80 connections changed jobs so far.  
messages-noreply@bounce.linkedin.com  
Sent: Wednesday, August 3, 2011 6:07 PM  
To: Joseph Adler

**Joseph,**

How time flies. We're already halfway through 2011 and 80 of your LinkedIn connections have started something new. Here are some of them.

*Click to view their new titles*

See [Nick's](#) new job! »

Where is [Simon](#)? »

And [Mike](#)? »

Have you started something new in 2011?  
[Let your connections know.](#)

You are receiving LinkedIn Marketing emails. [Unsubscribe](#)  
© 2011, LinkedIn Corporation, 2029 Stierlin Ct, Mountain View, CA 94043

## ■ Steps to make the email

- Collect job changers
- Figure out who is connected to them
- Rank job changes

# Example: Year in Review

```

memberPosition = LOAD '$latest_positions' USING BinaryJSON;
memberWithPositionsChangedLastYear = FOREACH (
    FILTER memberPosition BY ((start_date >= $start_date_low ) AND
        (start_date <= $start_date_high))
) GENERATE member_id, start_date, end_date;

allConnections = LOAD '$latest_bidirectional_connections' USING
BinaryJSON;

allConnectionsWithChange_nondistinct = FOREACH (
    JOIN memberWithPositionsChangedLastYear BY member_id,
    allConnections BY dest
) GENERATE allConnections::source AS source,
allConnections::dest AS dest;

allConnectionsWithChange = DISTINCT
    allConnectionsWithChange_nondistinct;

memberinfowpics = LOAD '$latest_memberinfowpics' USING
BinaryJSON;
pictures = FOREACH ( FILTER memberinfowpics BY
    ((cropped_picture_id is not null) AND
    (member_picture_privacy == 'N') OR
    (member_picture_privacy == 'E')))
) GENERATE member_id, cropped_picture_id, first_name as
dest_first_name, last_name as dest_last_name;

resultPic = JOIN allConnectionsWithChange BY dest, pictures
BY member_id;
connectionsWithChangeWithPic = FOREACH resultPic GENERATE
    allConnectionsWithChange::source AS source_id,
    allConnectionsWithChange::dest AS member_id,
    pictures::cropped_picture_id AS pic_id,
    pictures::dest_first_name AS dest_first_name,
    pictures::dest_last_name AS dest_last_name;

```

```

joinResult = JOIN connectionsWithChangeWithPic BY source_id,
    memberinfowpics BY member_id;
withName = FOREACH joinResult GENERATE
    connectionsWithChangeWithPic::source_id AS source_id,
    connectionsWithChangeWithPic::member_id AS member_id,
    connectionsWithChangeWithPic::dest_first_name as first_name,
    connectionsWithChangeWithPic::dest_last_name as last_name,
    connectionsWithChangeWithPic::pic_id AS pic_id,
    memberinfowpics::first_name AS firstName,
    memberinfowpics::last_name AS lastName,
    memberinfowpics::gmt_offset as gmt_offset,
    memberinfowpics::email_locale as email_locale,
    memberinfowpics::email_address as email_address;

resultGroup0 = GROUP withName BY (source_id, firstName,
lastName, email_address, email_locale, gmt_offset);

-- get the count of results per recipient
resultGroupCount = FOREACH resultGroup0 GENERATE group,
withName as tomany, COUNT_STAR(withName) as num_results;
resultGroupPre = filter resultGroupCount by num_results > 2;
resultGroup = FOREACH resultGroupPre {
    withName = LIMIT tomany 64;
    GENERATE group, withName, num_results;
}

x_in_review_pre_out = FOREACH resultGroup GENERATE
    FLATTEN(group) as (source_id, firstName, lastName,
email_address, email_locale, gmt_offset),
    withName.(member_id, pic_id, first_name, last_name) as
jobChanger, '2011' as changeYear:chararray,
    num_results as num_results;

x_in_review = FOREACH x_in_review_pre_out GENERATE
    source_id as recipientID, gmt_offset as gmtOffset,
    firstName as first_name, lastName as last_name, email_address,
    email_locale,
    TOTUPLE( changeYear, source_id,firstName, lastName,
    num_results, jobChanger) as body;

rmf $xir;
STORE x_in_review INTO '$xir' USING BinaryJSON('recipientID');

```



# People You May Know

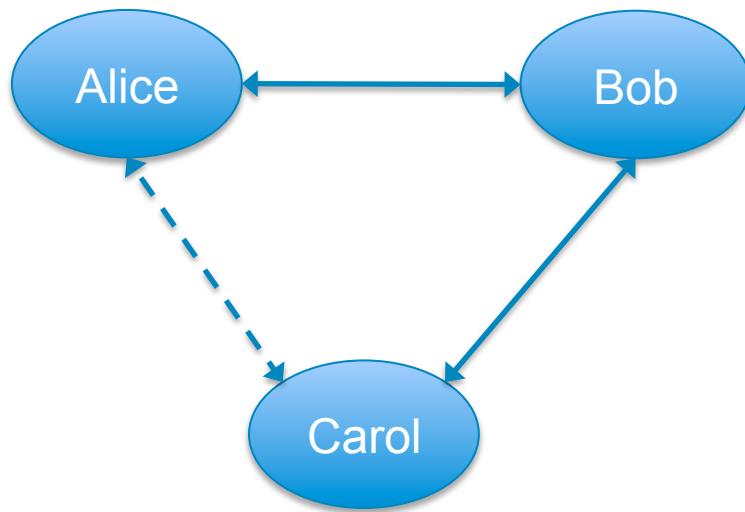
People You May Know beta

See people from different parts of your professional life

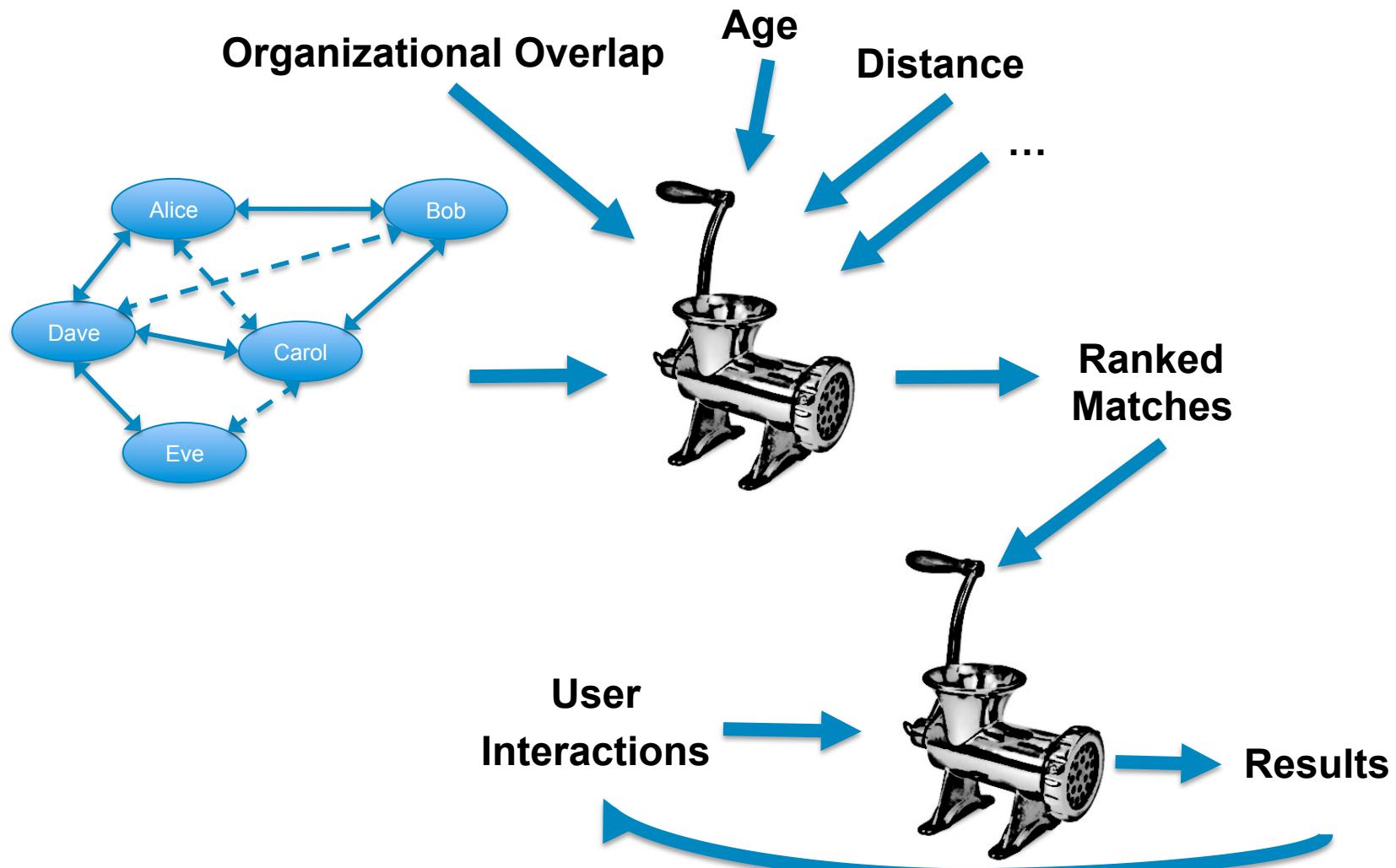
All Suggestions / LinkedIn (12)  Connect All

 LinkedIn	 Google	 eBay	 amazon	 GIGAOM	 STANFORD UNIVERSITY	 San Jose State UNIVERSITY	
<a href="#">All Suggestions / LinkedIn (12)</a> <input checked="" type="radio"/> Connect All <input type="radio"/>							
 Brad Mauney <small>2nd</small> Senior Product Manager, Search & Social Graph at LinkedIn Mountain View, California	 Albert Wang <small>2nd</small> Senior User Experience Designer at LinkedIn Mountain View, California						
<a href="#">Connect</a>  5 shared connections	<a href="#">Connect</a> 127 shared connections						
 Sam Shah <small>2nd</small> Principal Engineer at LinkedIn Mountain View, California	 Tan Nhu <small>2nd</small> Senior Web Developer at LinkedIn Mountain View, California						
<a href="#">Connect</a> 22 shared connections	<a href="#">Connect</a> 16 shared connections						
 Vinodh Jayaram <small>2nd</small> Software Engineering Manager at LinkedIn Mountain View, California	 Andy Chen <small>2nd</small> Software Engineer at LinkedIn Mountain View, California						
<a href="#">Connect</a> 10 shared connections	<a href="#">Connect</a> 78 shared connections						

## People You May Know

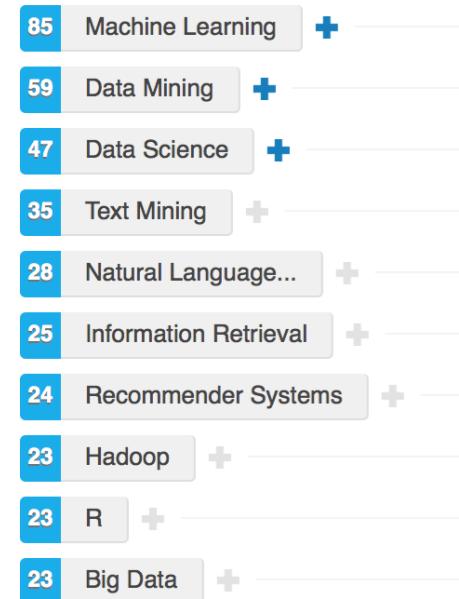
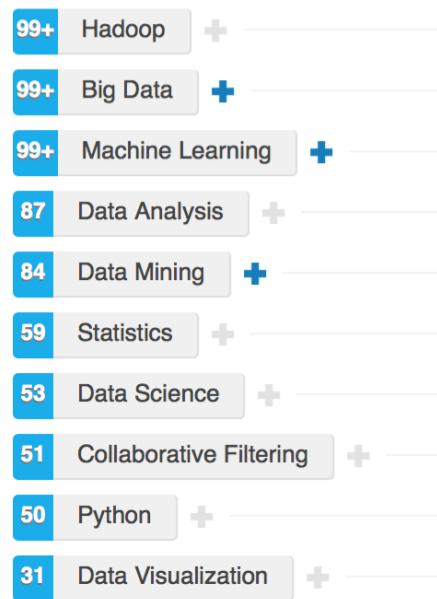
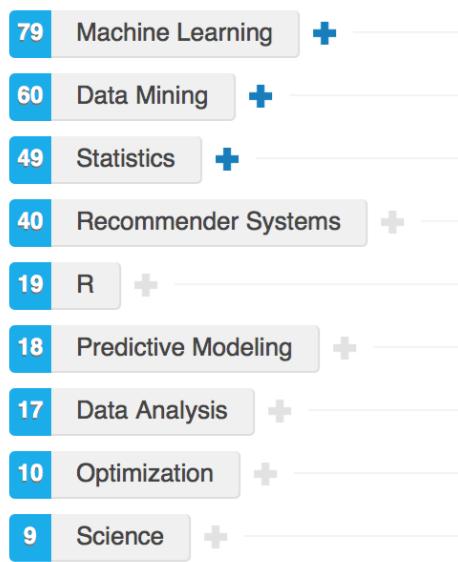


# People You May Know



# Infrastructure

# Skill sets



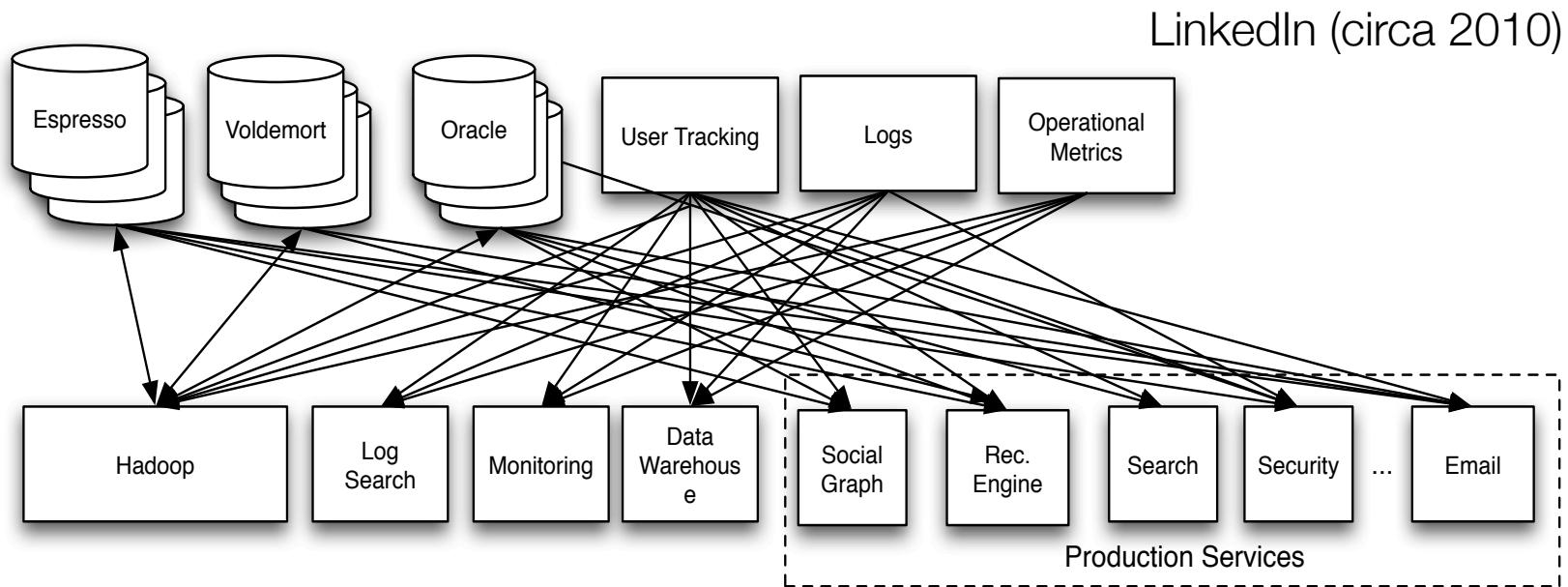
## Top Complaints from Data Scientists

- **Discovery**: where is the data?
- **Wrangling**: can I make sense of the data?
- **Verifying**: is the data correct?
- **Scaling**: how can I scale my computation?
- **Workflow**: how can I operate my processing?
- **Publishing**: how can I get my results into production?

## Top Complaints from Data Scientists

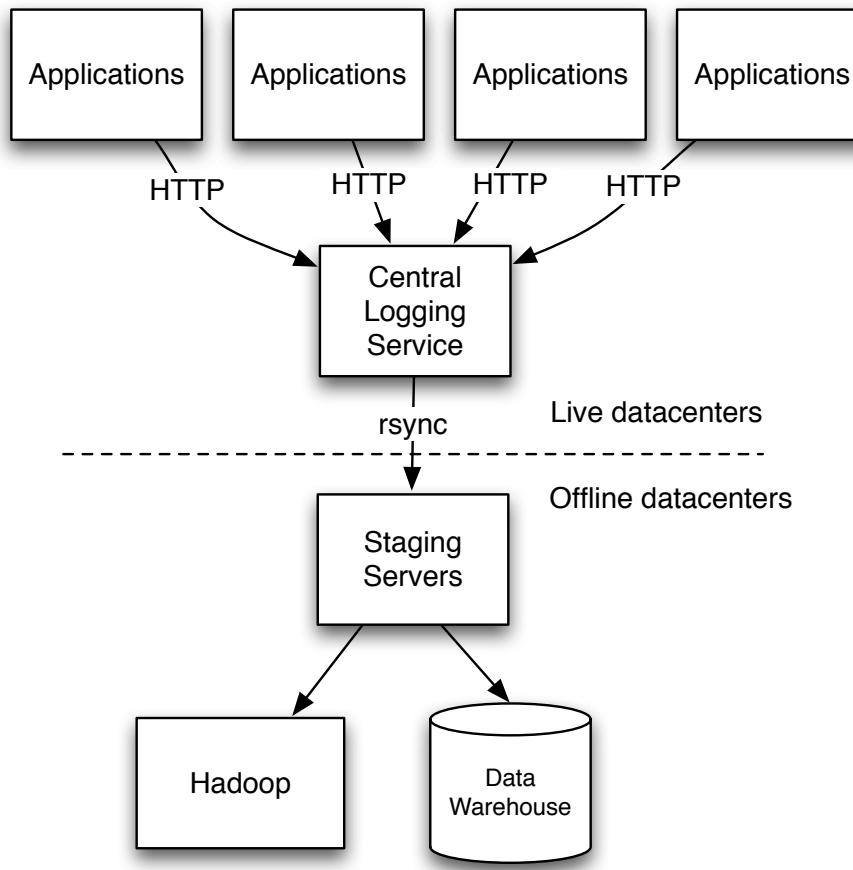
- **Discovery:** where is the data?
- **Wrangling:** can I make sense of the data?
- **Verifying:** is the data correct?
- **Scaling:** how can I scale my computation?
- **Workflow:** how can I operate my processing?
- **Publishing:** how can I get my results into production?

# Discovery: where is the data?



- $O(n^2)$  point-to-point data integration complexity

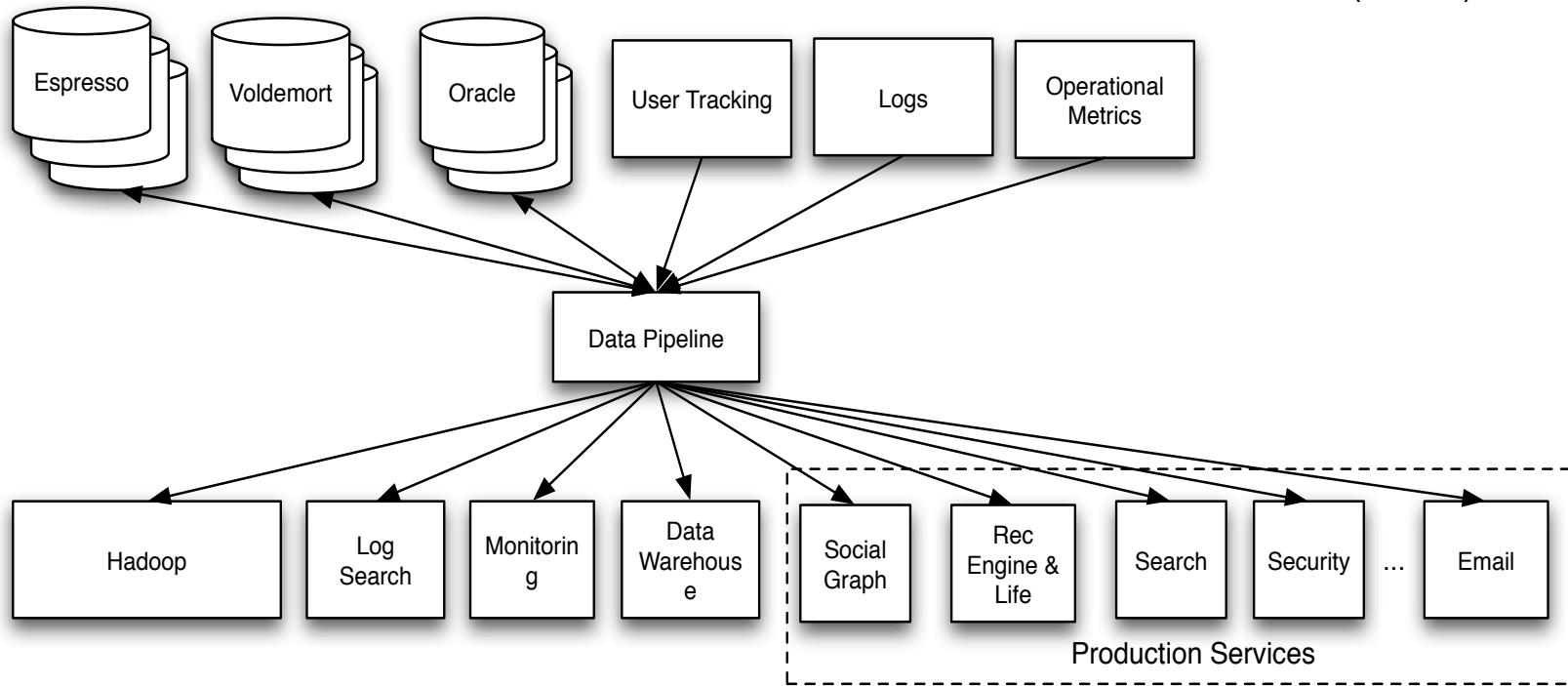
# Infrastructure fragility



- Can't get all data
- Hard to operate
- Multi-hour delay
- Labor intensive
- Slow
- Does it work?

# Ingress - O(n) data integration

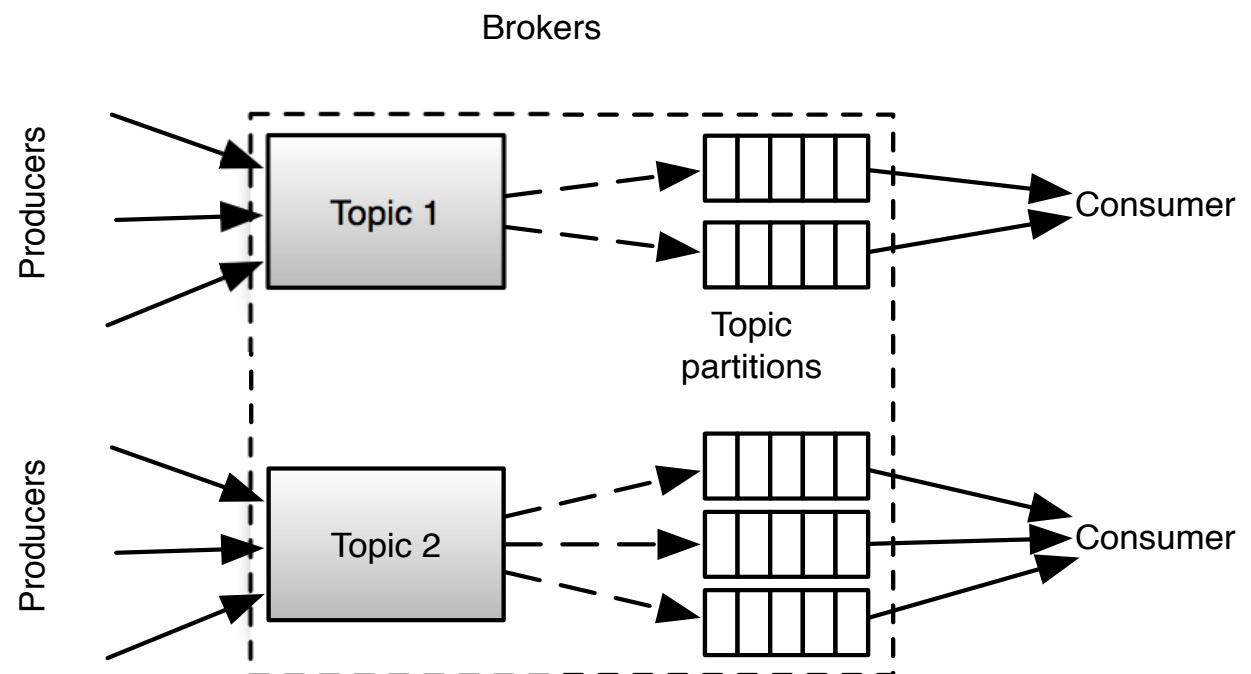
LinkedIn (2013)



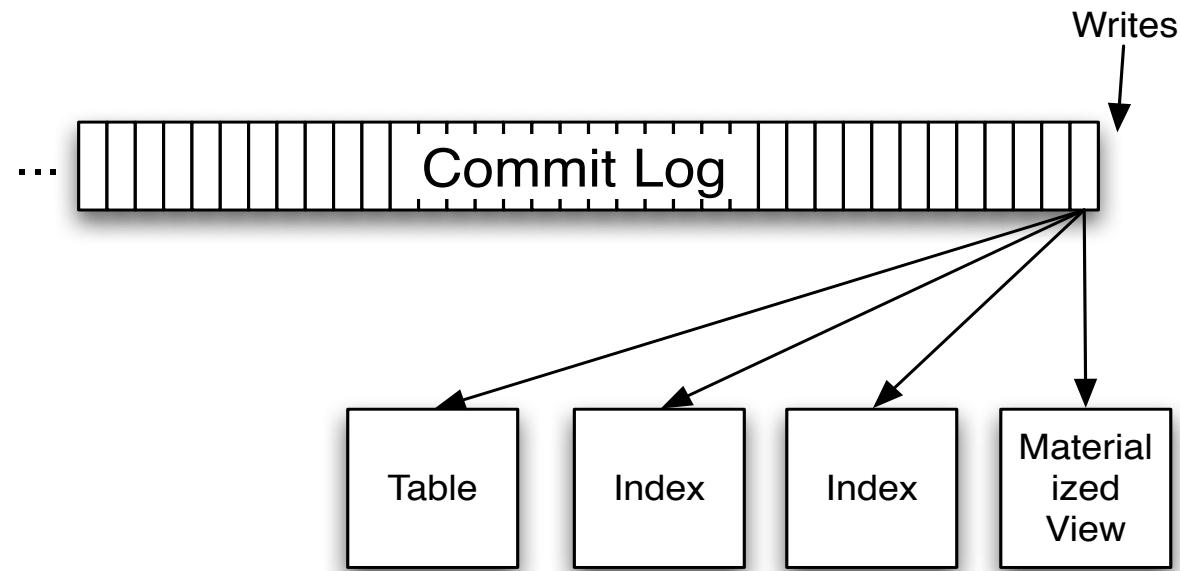
# Ingress – Apache Kafka



- Multi-broker publish/subscribe system
- Categorized topics
  - “PeopleYouMayKnowTopic”
  - “ConnectionUpdateTopic”



# What is a commit log?



## Top Complaints from Data Scientists

- **Discovery**: where is the data?
- **Wrangling**: can I make sense of the data?
- **Verifying**: is the data correct?
- **Scaling**: how can I scale my computation?
- **Workflow**: how can I operate my processing?
- **Publishing**: how can I get my results into production?

# Data model

```
{  
    tracking_code=null,  
    session_id=42,  
    tracking_time=Tue Jul 31 07:27:25 PDT 2010,  
    error_key=null,  
    locale=en_us,  
    browser_id=ddc61a81-5311-4859-be42-ca7dc7b941e3,  
    member_id=1214,  
    page_key=profile,  
tracking_info=Viewee=1213,lnl=f,nd=1,o=1214,^SP=pId-'pro_stars',rslvd=t,vs=v,  
,vid=1214,ps=EDU|EXP|SKIL|,  
    error_id=null,  
    page_type=FULL_PAGE,  
    request_path=view  
    ...  
}
```

LinkedIn (circa 2010)



**Reid Hoffman**  
Entrepreneur. Product Strategist. Investor.  
San Francisco Bay Area | Internet

Current Greylock, LinkedIn, Wrapp  
Previous College Eight, UCSC, Six Apart, Tagged  
Education Oxford University

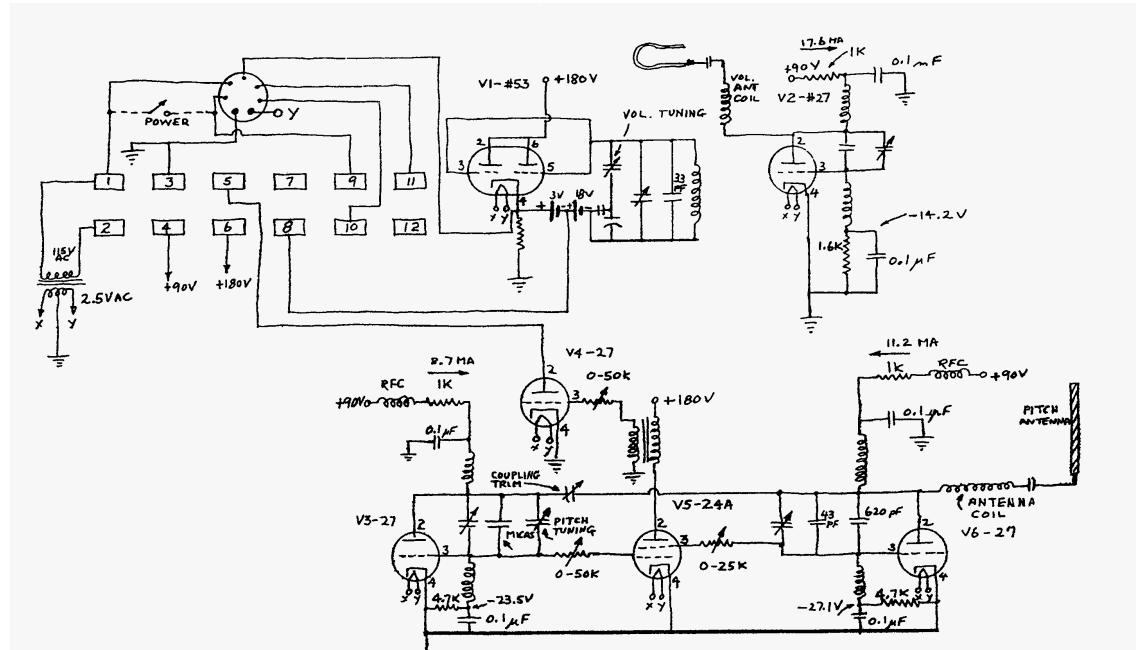
Follow ▾ Know Reid?  
Connect

**184,444**  
followers

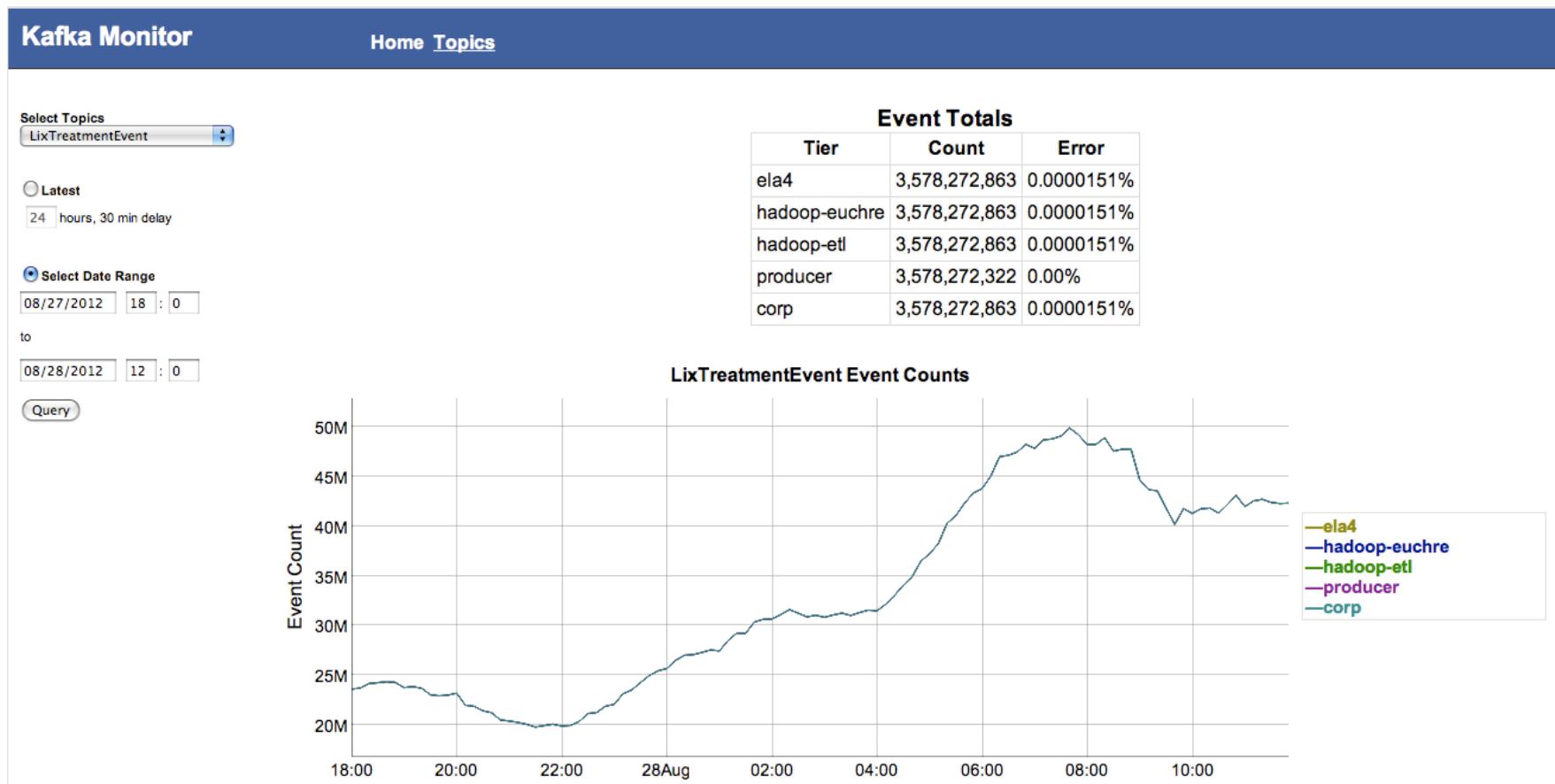
[www.linkedin.com/in/reidhoffman](http://www.linkedin.com/in/reidhoffman)

# Schemas

- Schemas are the contract
  - DDL for data definition and schema
- Central versioned registry of all schemas
- Schema evolution with programmatic checks



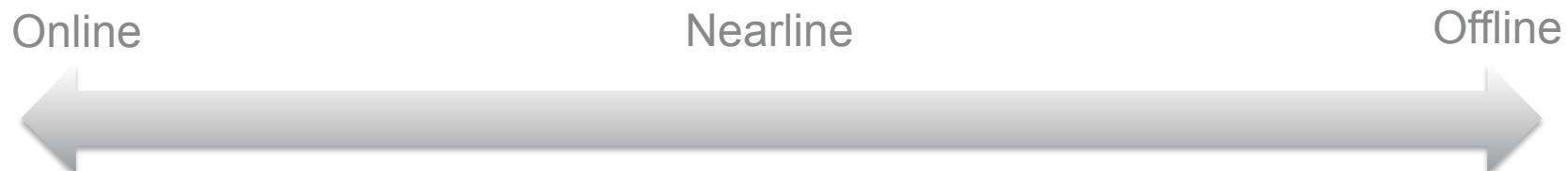
# Audit trail



## Top Complaints from Data Scientists

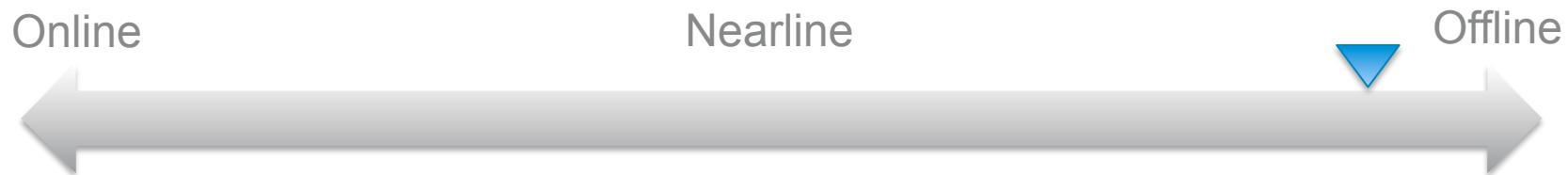
- **Discovery**: where is the data?
- **Wrangling**: can I make sense of the data?
- **Verifying**: is the data correct?
- **Scaling**: how can I scale my computation?
- **Workflow**: how can I operate my processing?
- **Publishing**: how can I get my results into production?

# Models of computation



- Sub-second processing
  - Harder to scale
  - Must handle failures gracefully
  - Computationally intensive
  - Easier to scale
  - Easier to tolerate failures
  - Faster iteration

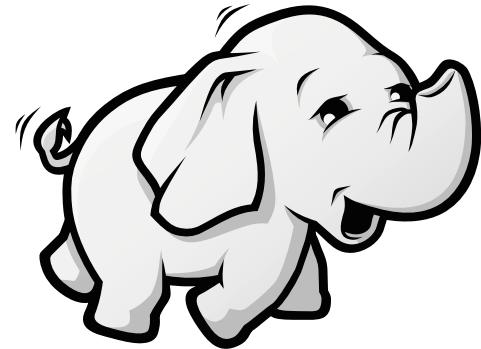
# Models of computation



- Sub-second processing
- Harder to scale
- Must handle failures gracefully
- Computationally intensive
- Easier to scale
- Easier to tolerate failures
- Faster iteration

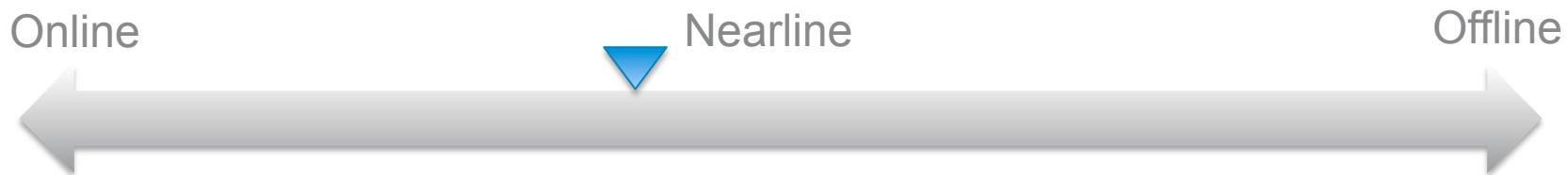
# Hadoop

## Why we use Hadoop



- Simple programmatic model
- Rich developer ecosystem
  - Languages: Pig, Hive, Crunch, Cascading, ...
  - Libraries: Mahout, DataFu, ElephantBird, ...
- DataFu
  - Large-scale machine learning and statistical operations
- Horizontal scalability, fault tolerance, multi-tenancy
  - Reliably process multiple TB of data

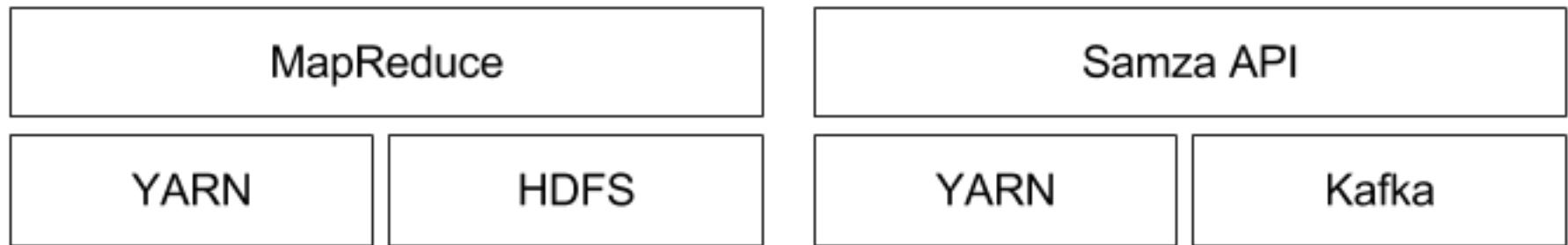
# Models of computation



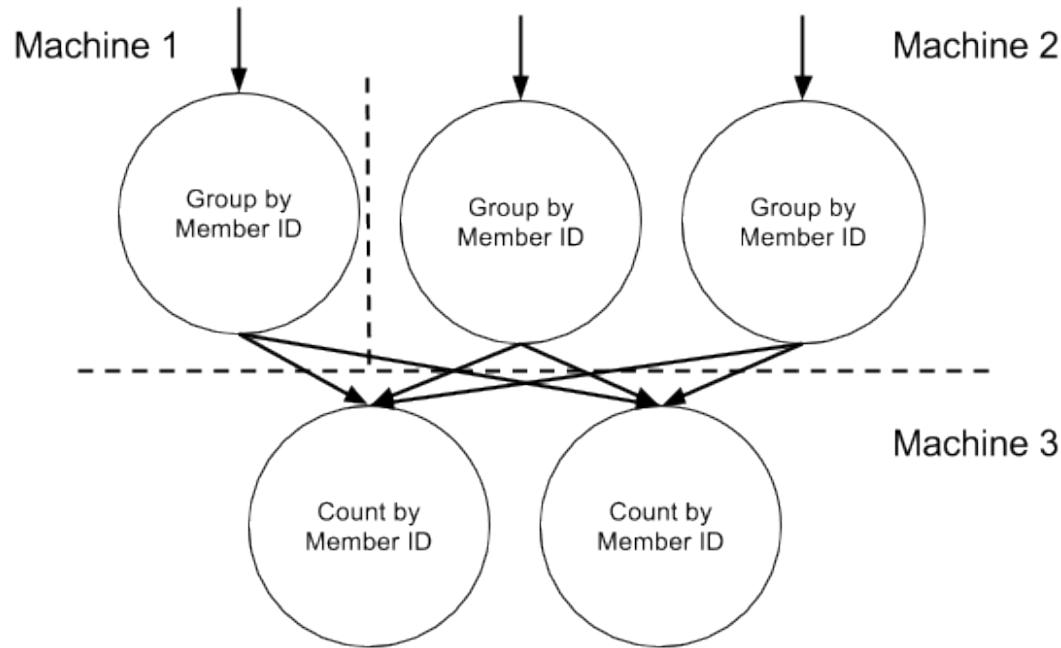
- Sub-second processing
- Harder to scale
- Must handle failures gracefully
- Computationally intensive
- Easier to scale
- Easier to tolerate failures
- Faster iteration

# Apache Samza – “MapReduce for streams”

Samza

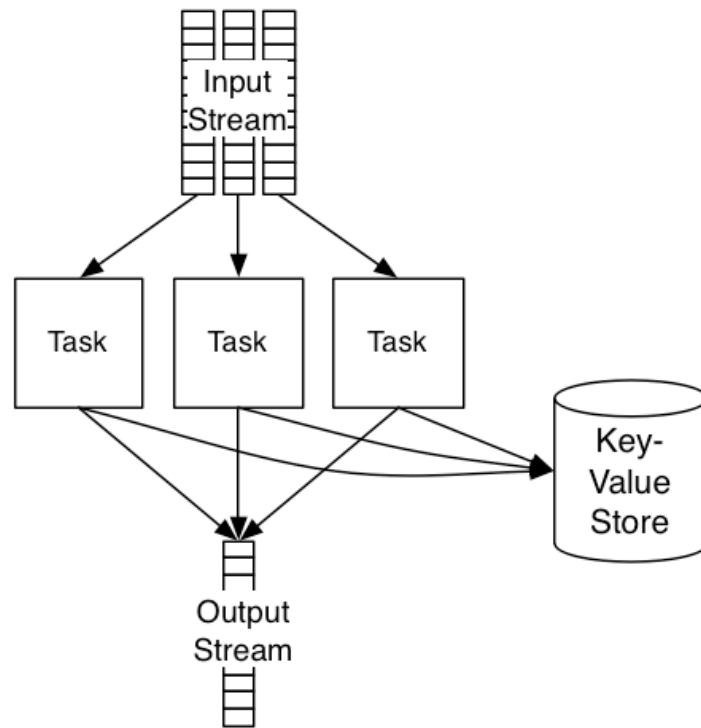


# Samza

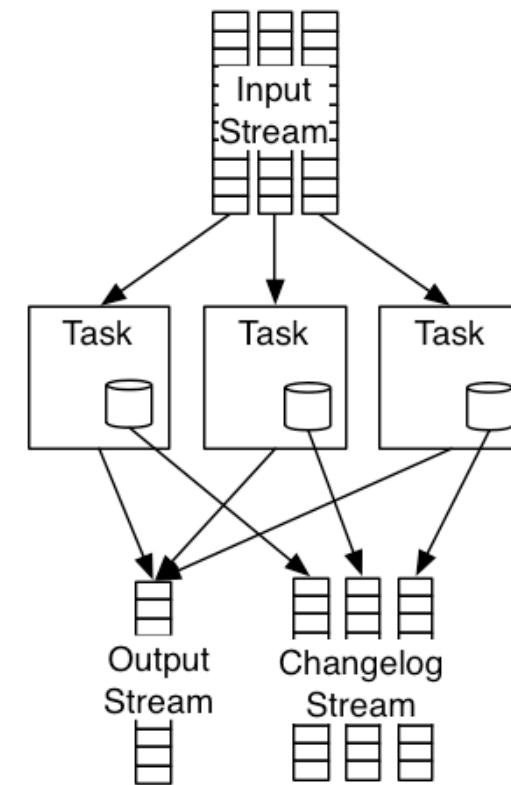
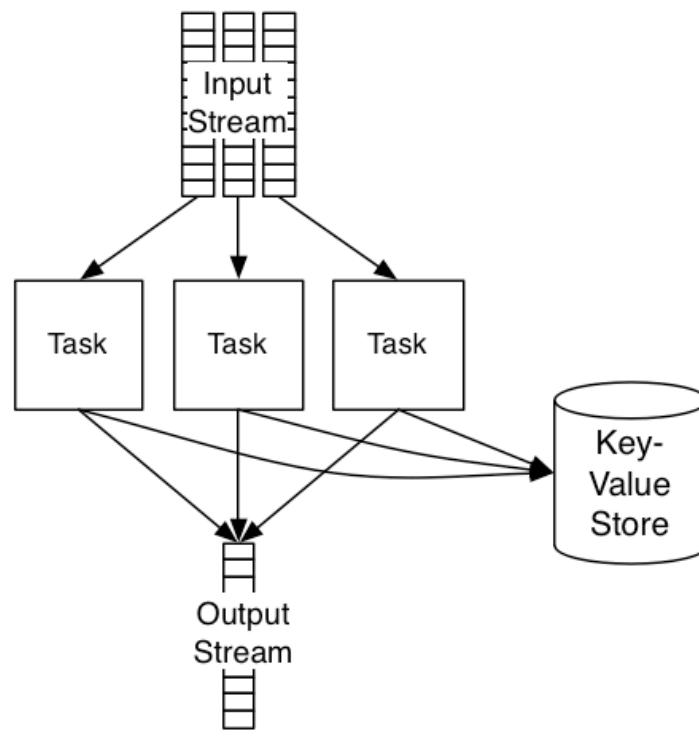


```
SELECT COUNT(*) FROM PageViewEvent GROUP BY member_id
```

# Samza: State Management



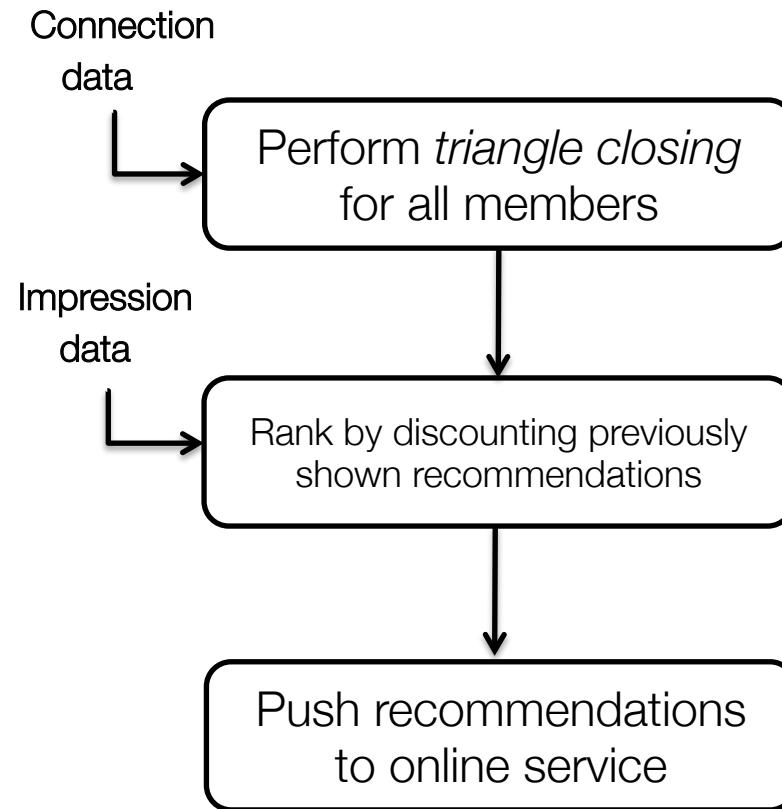
# Samza: State Management



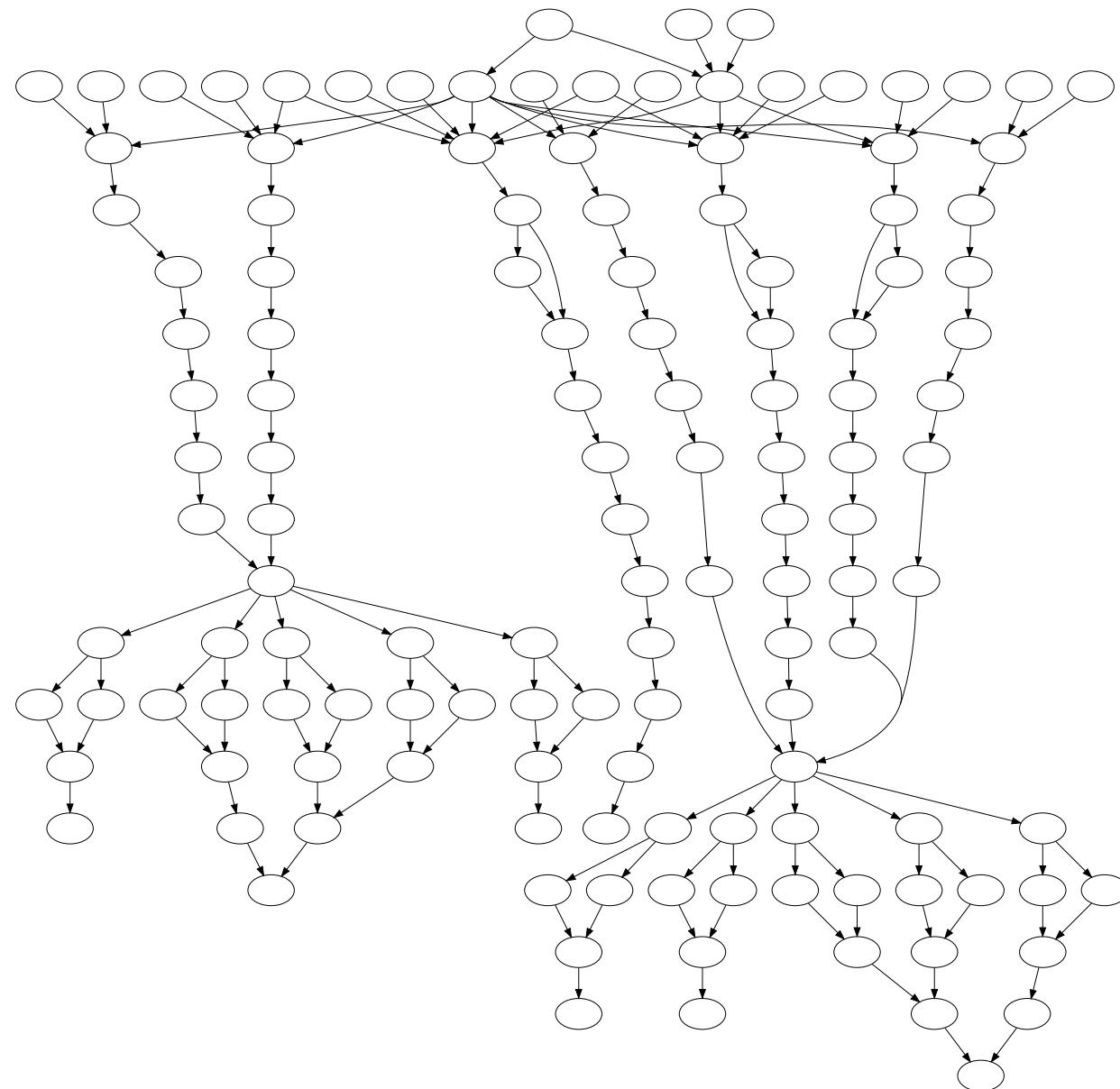
## Top Complaints from Data Scientists

- **Discovery**: where is the data?
- **Wrangling**: can I make sense of the data?
- **Verifying**: is the data correct?
- **Scaling**: how can I scale my computation?
- **Workflow**: how can I operate my processing?
- **Publishing**: how can I get my results into production?

# People You May Know – Workflow



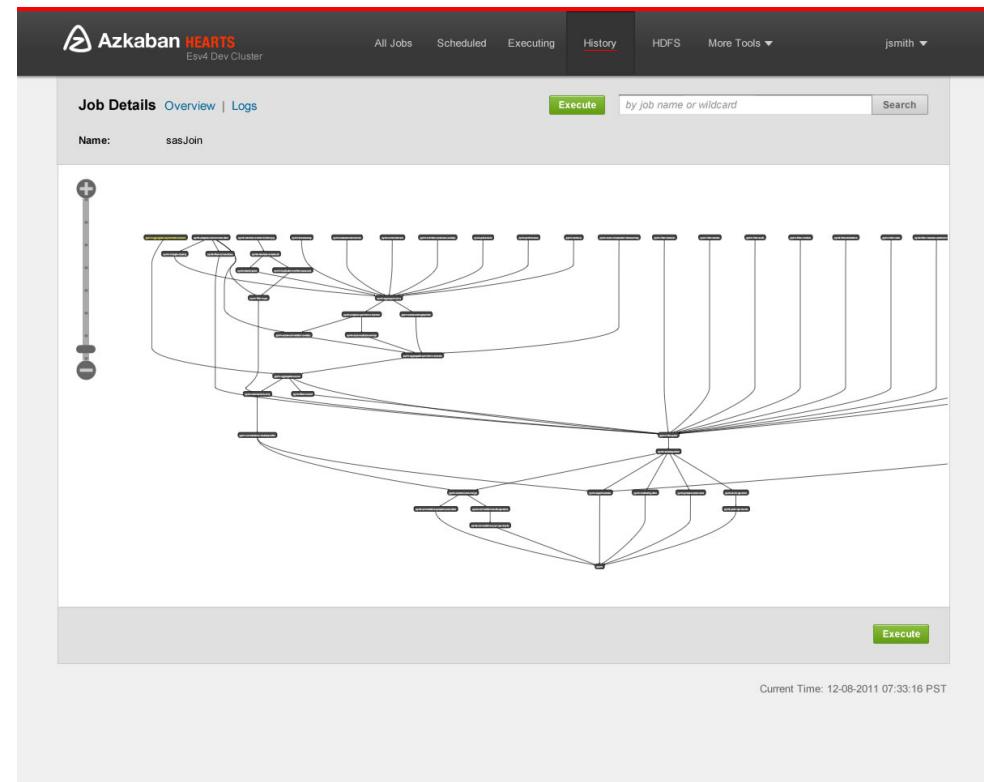
# People You May Know – Workflow (in reality)



# Workflow Management - Azkaban



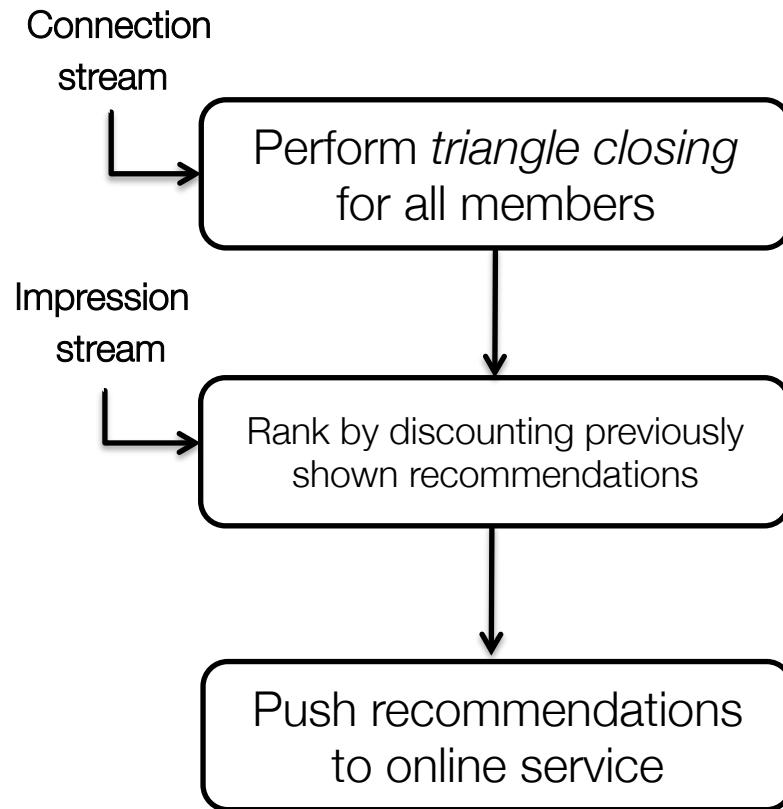
- Dependency management
- Diverse job types
- Scheduling
- Monitoring
- Visualization
- Configuration
- Retry/restart on failure
- Resource locking



## Top Complaints from Data Scientists

- **Discovery**: where is the data?
- **Wrangling**: can I make sense of the data?
- **Verifying**: is the data correct?
- **Scaling**: how can I scale my computation?
- **Workflow**: how can I operate my processing?
- **Publishing**: how can I get my results into production?

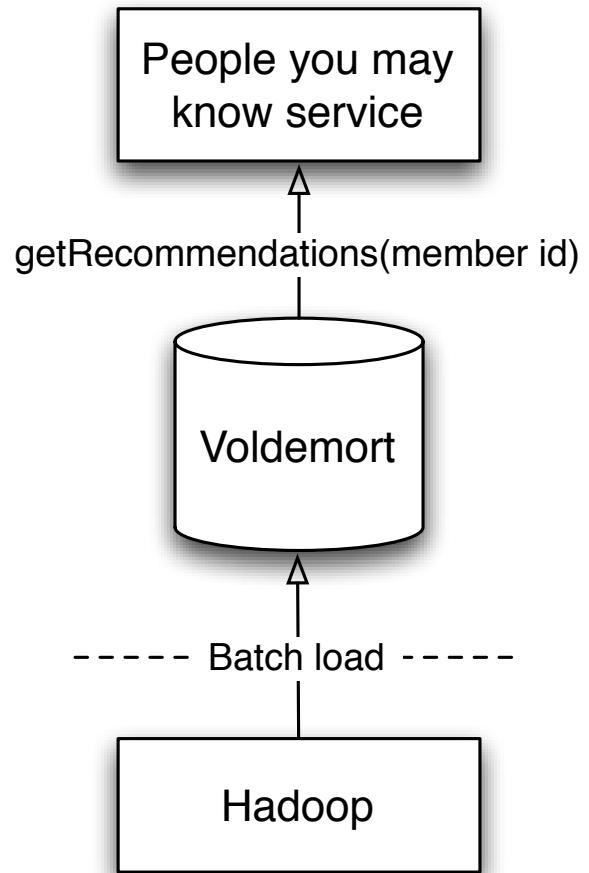
# People You May Know – Workflow



Member Id 1213 =>  
[ Recommended member id 1734,  
Recommended member id 1523  
...  
Recommended member id 6332 ]

# Egress – Key/Value

- Voldemort
  - Based on Amazon's Dynamo
- Distributed and elastic
- Horizontally scalable



## Systems (all open source)

- Apache Kafka: publish/subscribe commit log
- DataFu: Common data routines
- Apache Samza: stream processing framework
- Azkaban: workflow management
- Voldemort: key/value store

**Empowers data scientists and engineers to focus on new product ideas, not infrastructure**

Learning More

[data.linkedin.com](https://data.linkedin.com)