



Turning data into products

Sam Shah

LinkedIn: the professional profile of record

LinkedIn Account Type: Pro 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

www.linkedin.com/in/mosesting/ Contact Info

ACTIVITY

Share an update...

Lauren Bowen via LinkedIn Today

How to Influence Your Company Culture (For the Better)
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.

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
All-Star

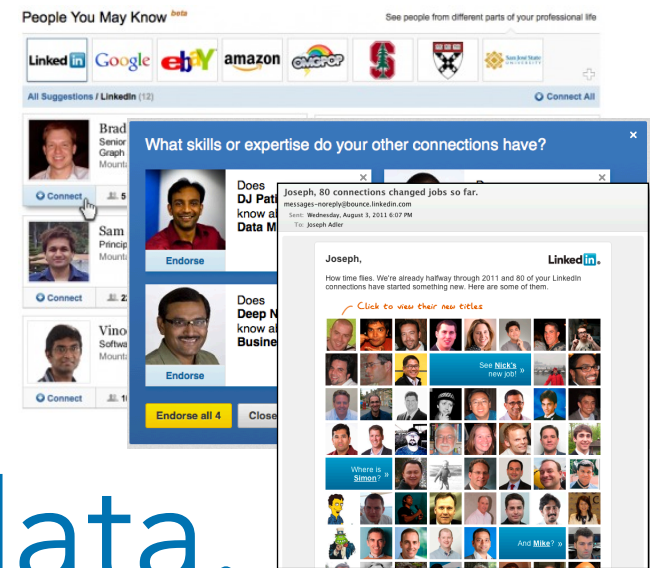
Share your profile »

YOUR NETWORK
Company »

250M Members | **250M Member Profiles**

SUMMARY

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



We have a lot of data.
We want to leverage this data to build products.

Applications

Profile Stats Pro

Last 90 Days December 13, 2011 – March 12, 2012 Settings

Who's Viewed Your Profile

TODAY

Sam Sha

Principal E

San Francis

In Common

Jay Krej

Principal

San Francis

In Common

Roshan S

Senior Sof

San Francis

In Common

YESTERDAY

Matthi

Senior

San Francis

In Common

Lili Wu

Senior Sof

San Francis

In Common

MORE THAN TWO DAYS /

Gordon K

<script>a

San Francis

In Common

Baq Hair

Senior Sof

San Francis

In Common

Mitul Tiw

Senior Sea

San Francis

In Common

Anmol B

Engineerin

San Francis

In Common

William V

Senior Sof

San Francis

In Common

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

Joseph Adler

Joseph,

LinkedIn

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

Views 329

Feb 26

160

Views

22

4

World total 371

Facebook

Microsoft

Amazon

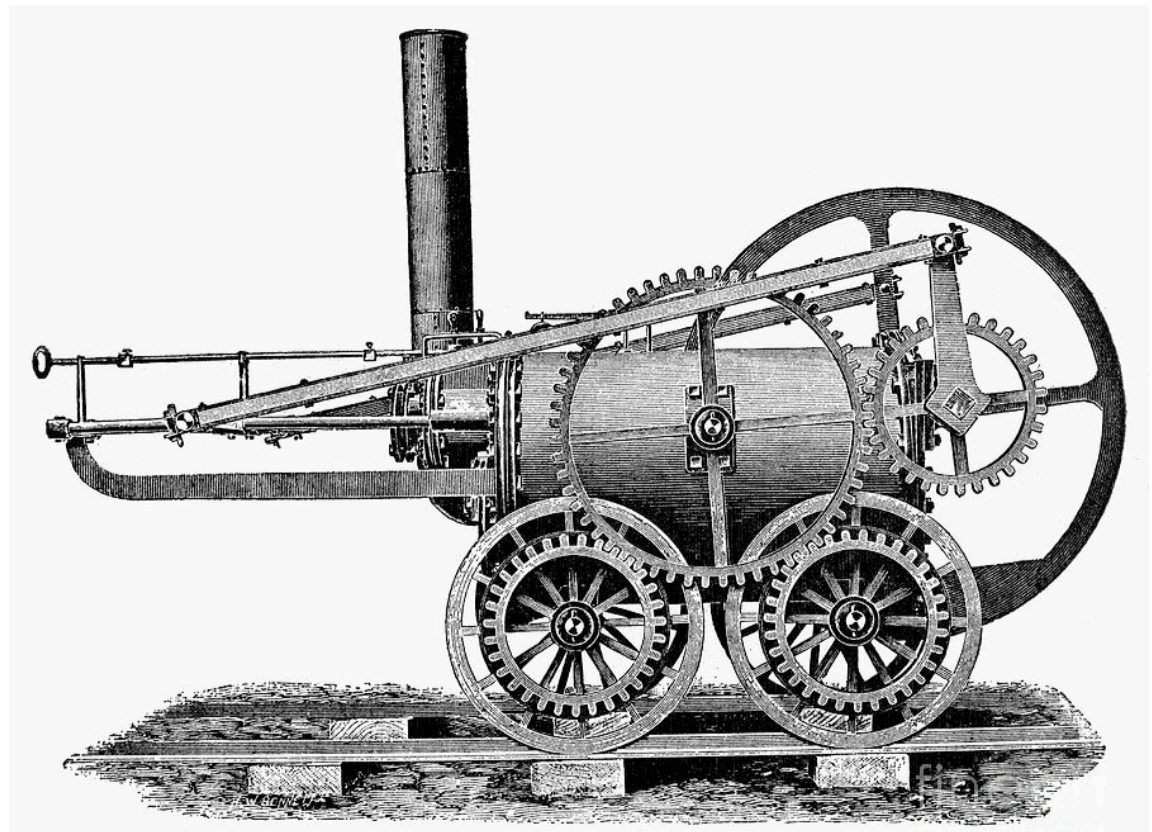
Homepage powered by data

The image shows a screenshot of the LinkedIn homepage. Several sections are highlighted with red dashed boxes, indicating data-driven content:

- Top Navigation:** A red box highlights the top navigation bar, including the LinkedIn logo, navigation links (Home, Profile, Contacts, Groups, Jobs, Inbox, Companies, News, More), a search bar, and user information (Christian Pesse, Add Connections).
- Update Input:** A red box highlights the "Share an update" input area at the top left.
- MIT Executive Education:** A red box highlights a banner for "MIT Executive Education - MIT's Unique Entrepreneurial Education Program, Learn More & Enroll Today."
- LinkedIn Today:** A red box highlights the "LinkedIn Today" section, which features "See all Top Headlines for You" and includes articles from Wikipedia, HBR, and Yahoo.
- All Updates:** A red box highlights the "All Updates" section, which displays a list of recent updates from connections, including profile changes and shared content.
- Who's Viewed Your Profile?:** A red box highlights the "Who's Viewed Your Profile?" section, which shows a circular chart with 515 connections and 100+ profile viewers.
- Jobs You May Be Interested In:** A red box highlights the "Jobs You May Be Interested In" section, which lists job opportunities such as "Bioinformatics/Senior Bioinformatics," "Senior Staff Systems Engineer (Data)," and "Sr. Applied Scientist."
- Groups You May Like:** A red box highlights the "Groups You May Like" section, which lists groups such as "AnalyticBridge," "Bayesian Belief Networks with Bayesialab," and "Sentiment Analysis Symposium."
- Companies You May Want To Follow:** A red box highlights the "Companies You May Want To Follow" section, which lists companies such as "ness," "Linguistics," "ontotext," and "digital travel."

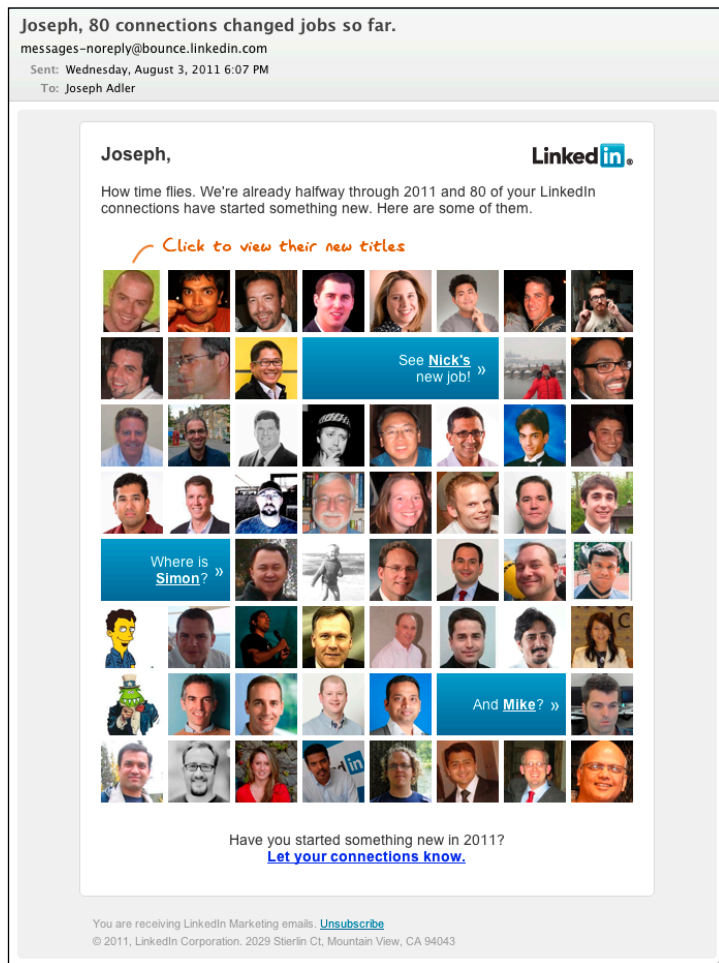
We're in the "pre-industrial age" of Big Data

- Need "bridges & railways"



Algorithms

Year in Review



- 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 toomany, COUNT_STAR(withName) as num_results;
resultGroupPre = filter resultGroupCount by num_results > 2;
resultGroup = FOREACH resultGroupPre {
    withName = LIMIT toomany 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

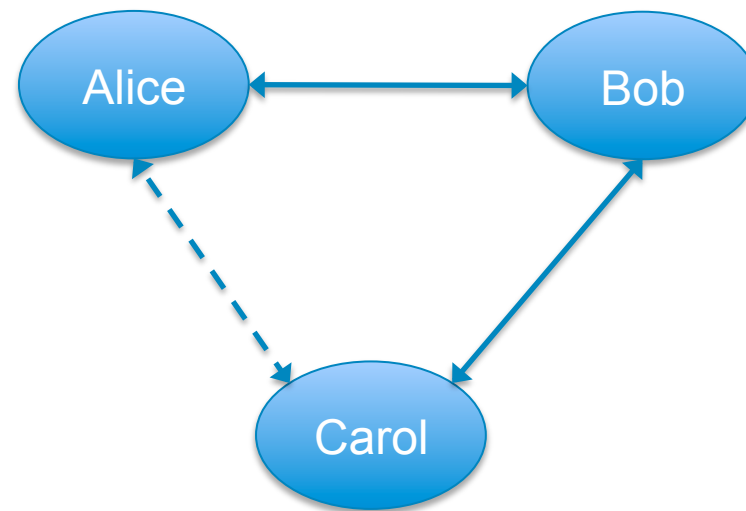
People You May Know beta See people from different parts of your professional life



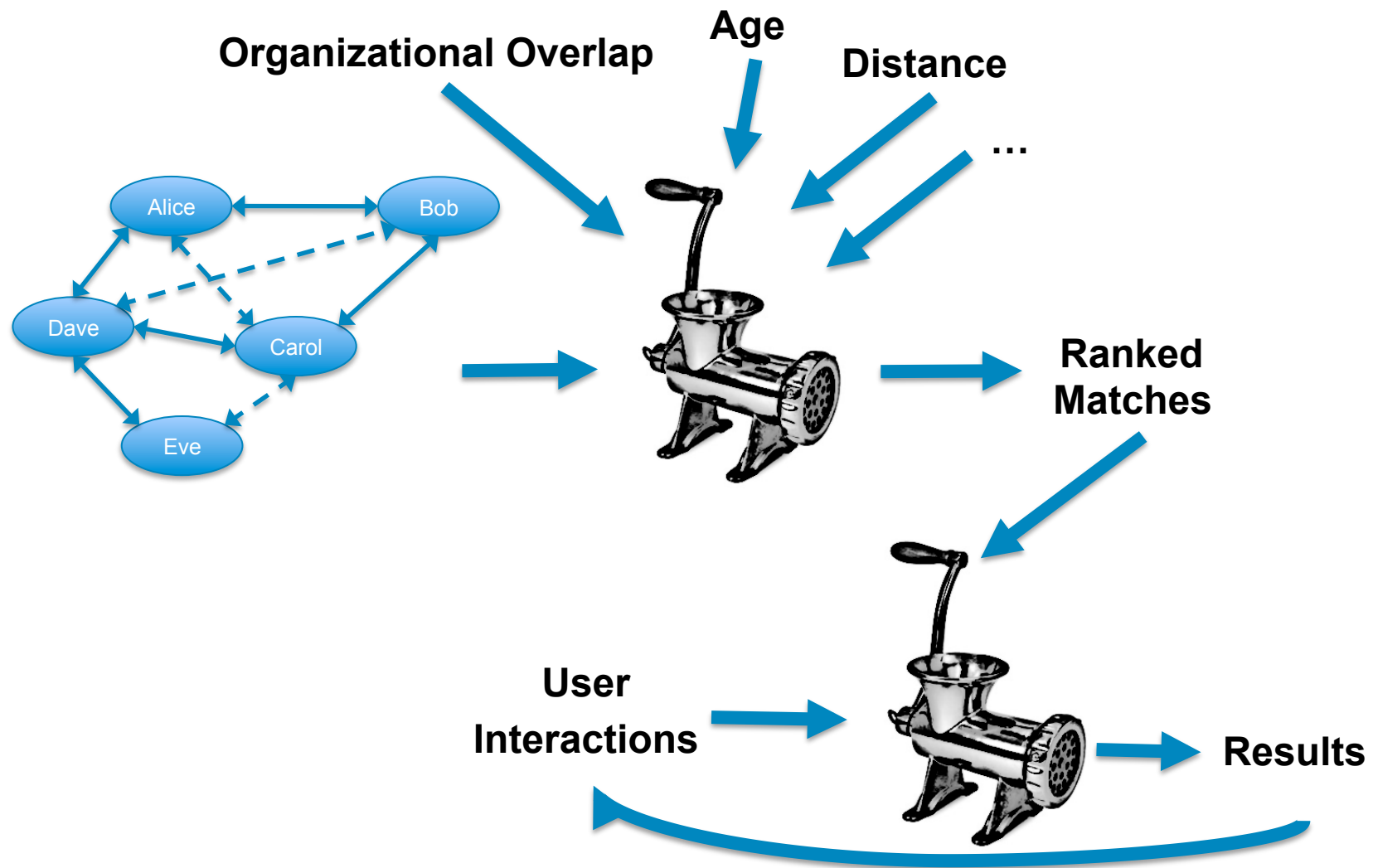
All Suggestions / LinkedIn (12) [Connect All](#)

| | |
|--|--|
|  <p>Brad Mauney <small>2nd</small>
Senior Product Manager, Search & Social Graph at LinkedIn
Mountain View, California</p> <p>Connect  5 shared connections</p> |  <p>Albert Wang <small>2nd</small>
Senior User Experience Designer at LinkedIn
Mountain View, California</p> <p>Connect  127 shared connections</p> |
|  <p>Sam Shah <small>2nd</small>
Principal Engineer at LinkedIn
Mountain View, California</p> <p>Connect  22 shared connections</p> |  <p>Tan Nhu <small>2nd</small>
Senior Web Developer at LinkedIn
Mountain View, California</p> <p>Connect  16 shared connections</p> |
|  <p>Vinodh Jayaram <small>2nd</small>
Software Engineering Manager at LinkedIn
Mountain View, California</p> <p>Connect  10 shared connections</p> |  <p>Andy Chen <small>2nd</small>
Software Engineer at LinkedIn
Mountain View, California</p> <p>Connect  78 shared connections</p> |

People You May Know

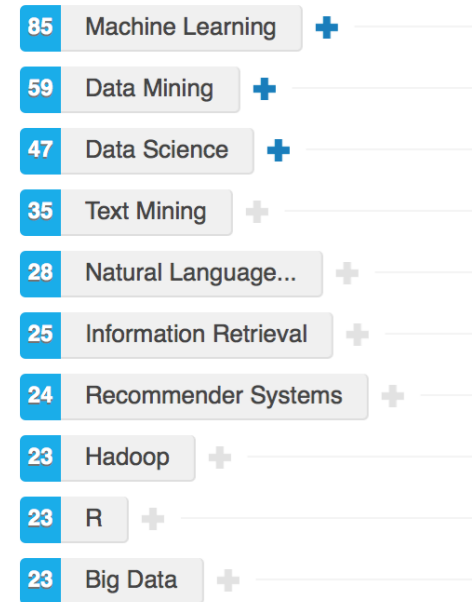
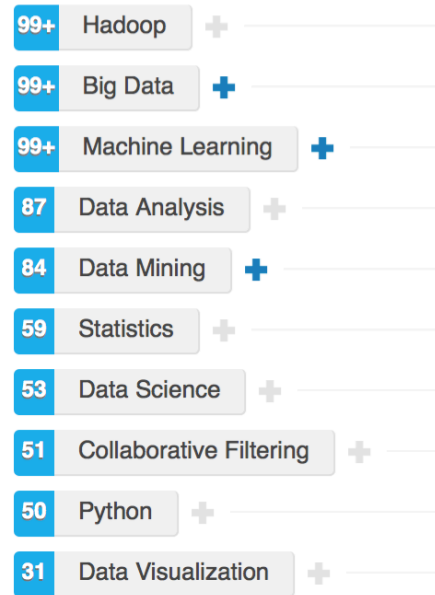
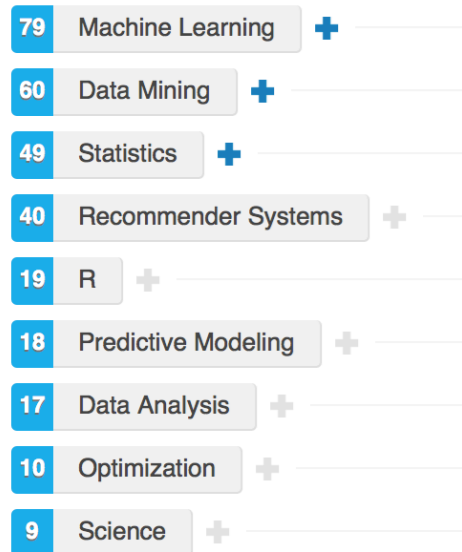


People You May Know



Infrastructure

Skill sets



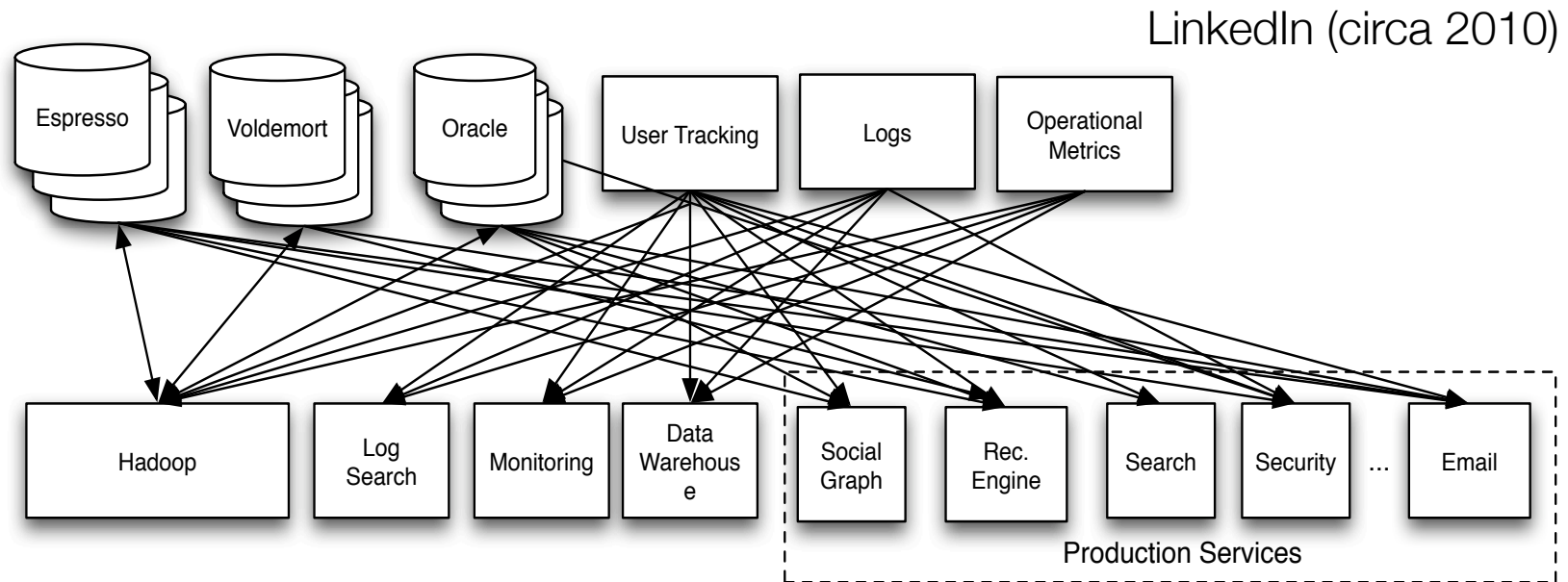
Top Complaints from Data Scientists

- **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

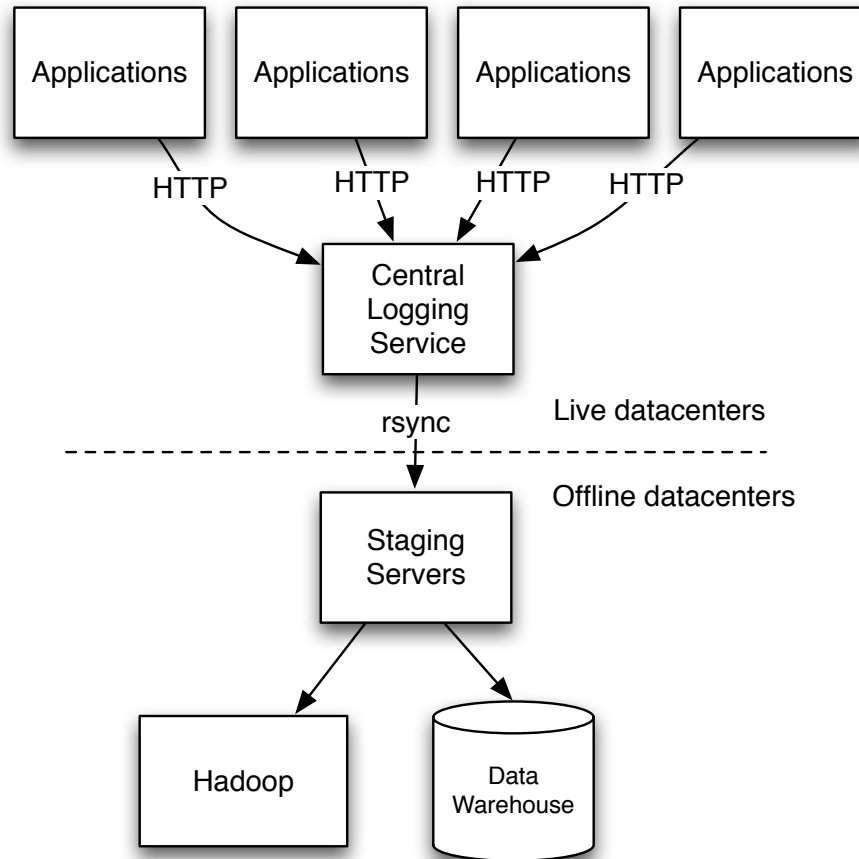
- **Discovery**: where is the data?
- **Wrangling**: can I make sense of the data?
- **Verifying**: is the data correct?
- **Scaling**: how can I scale my computation?
- **Workflow**: how can I operate my processing?
- **Publishing**: how can I get my results into production?

Discovery: where is the data?



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

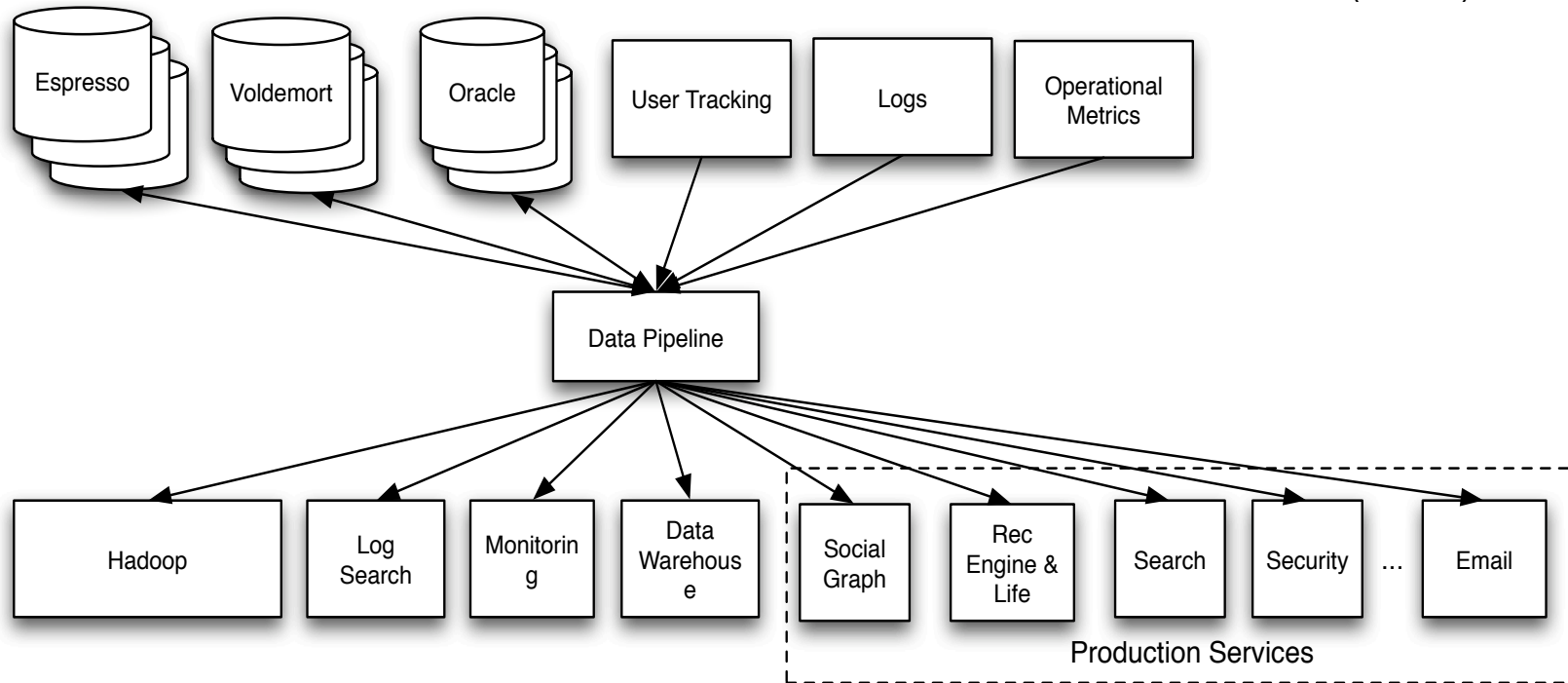
Infrastructure fragility



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

Ingress - $O(n)$ data integration

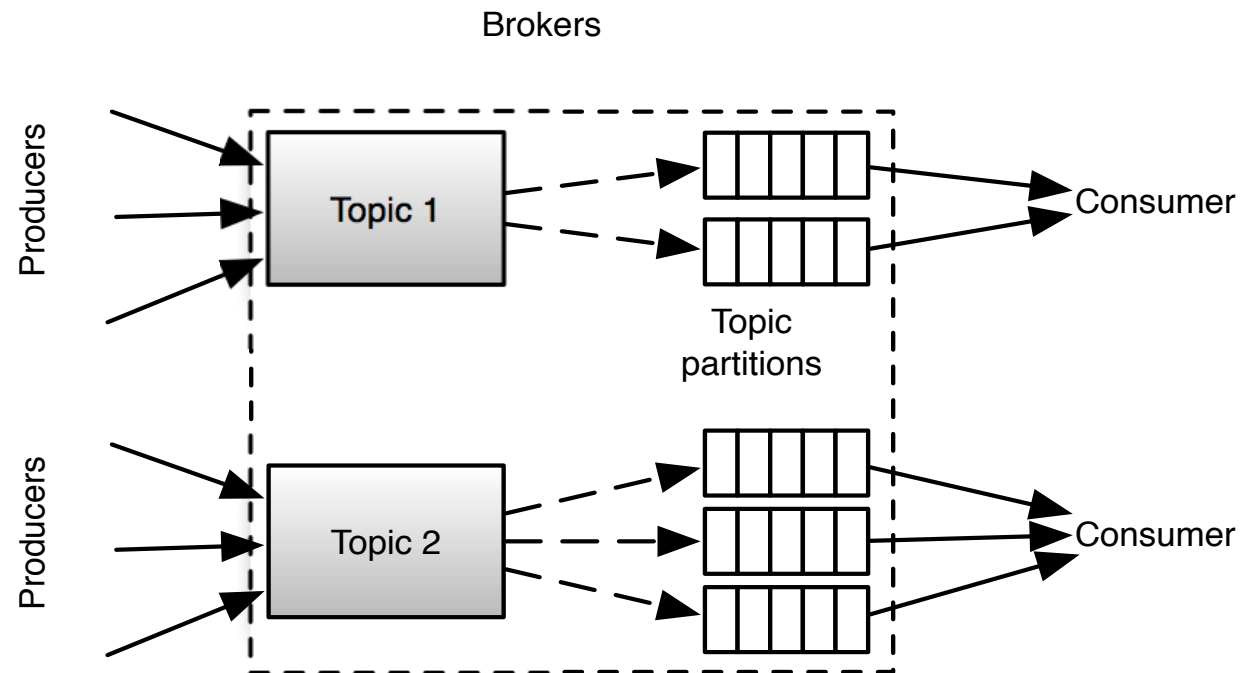
LinkedIn (2013)



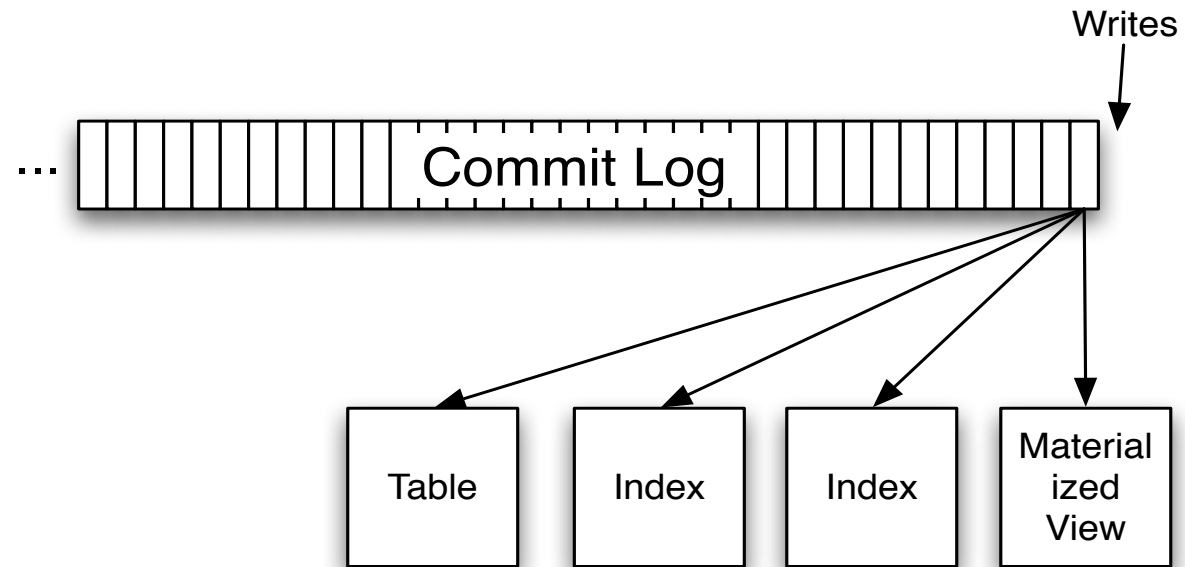
Ingress – Apache Kafka



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



What is a commit log?



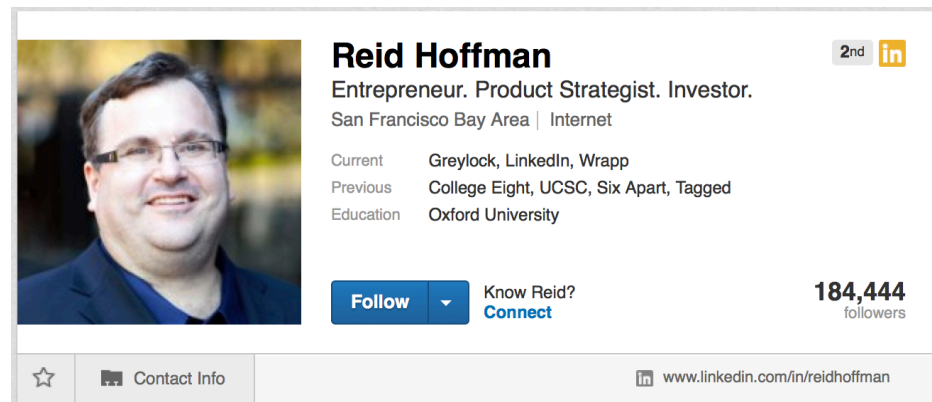
Top Complaints from Data Scientists

- **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

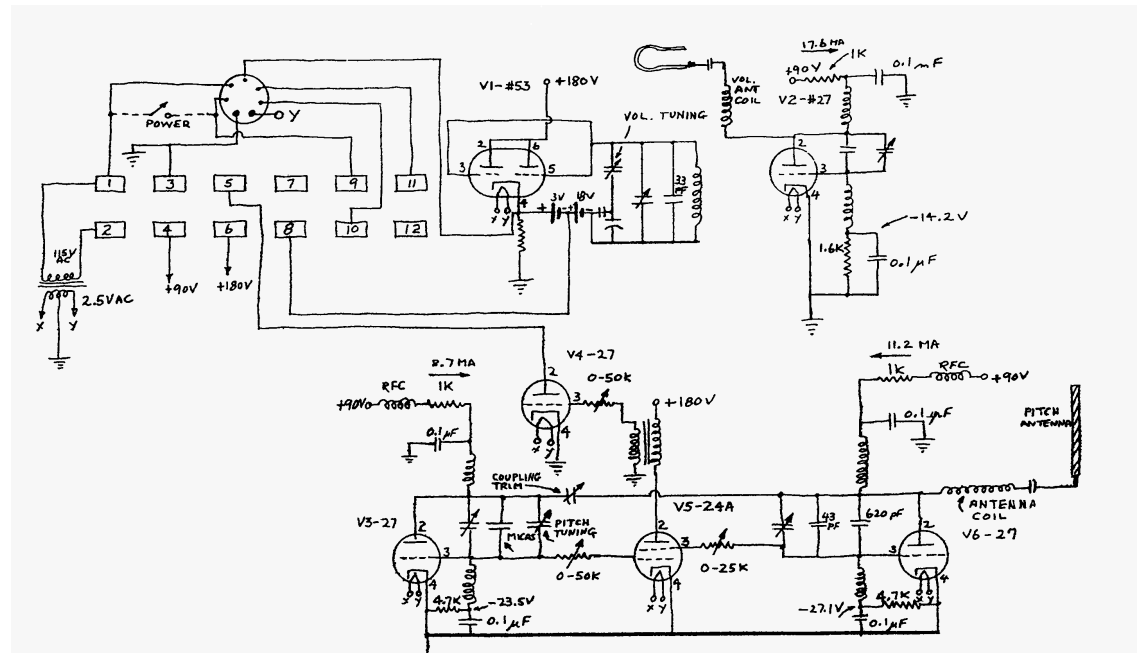
```
{
  tracking_code=null,
  session_id=42,
  tracking_time=Tue Jul 31 07:27:25 PDT 2010,
  error_key=null,
  locale=en_us,
  browser_id=ddc61a81-5311-4859-be42-ca7dc7b941e3,
  member_id=1214,
  page_key=profile,
  tracking_info=Viewee=1213,lnl=f,nd=1,o=1214,^SP=pId-'pro_stars',rslvd=t,vs=v
,vid=1214,ps=EDU|EXP|SKIL|,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

LinkedIn (circa 2010)

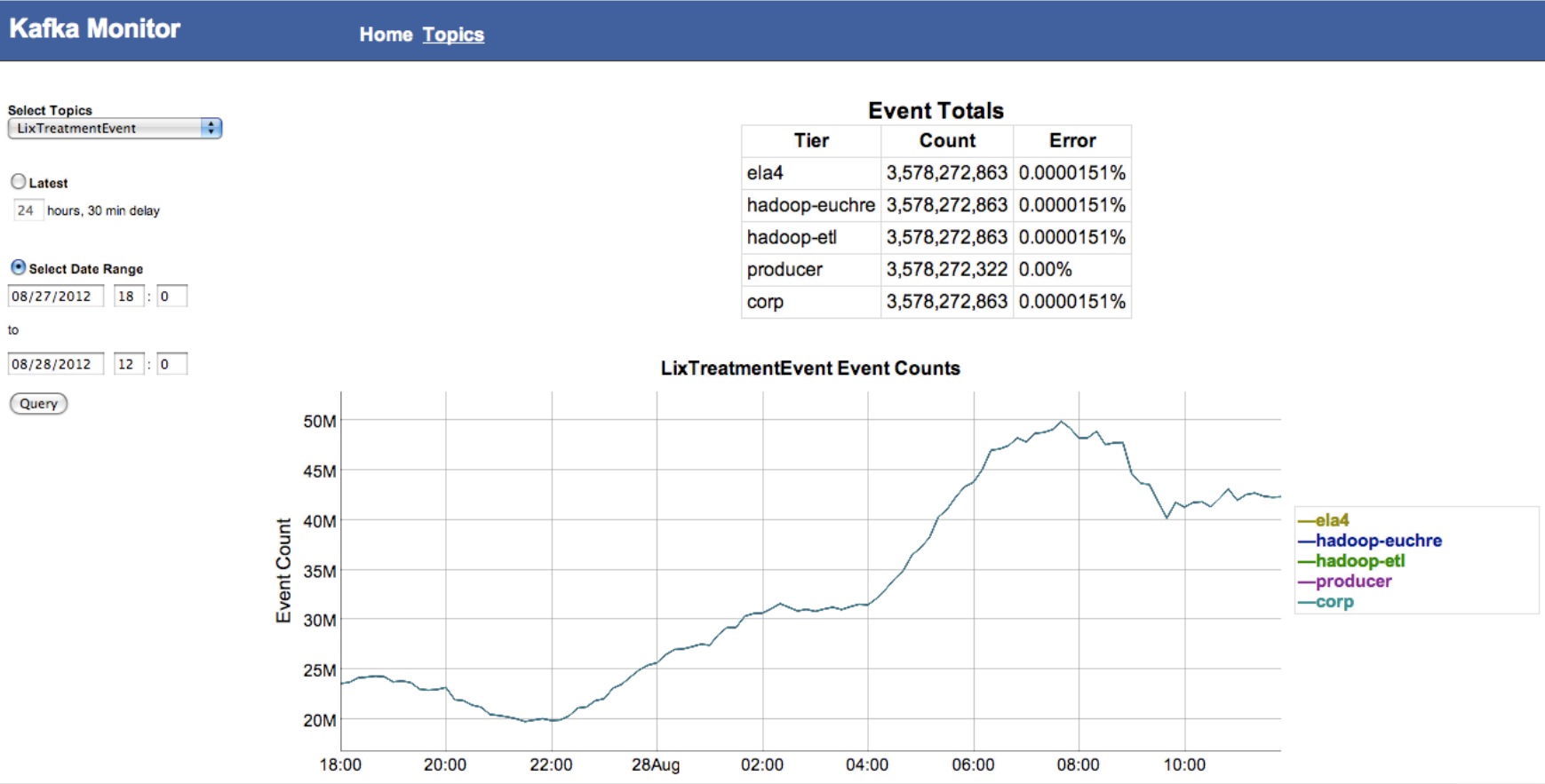


Schemas

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



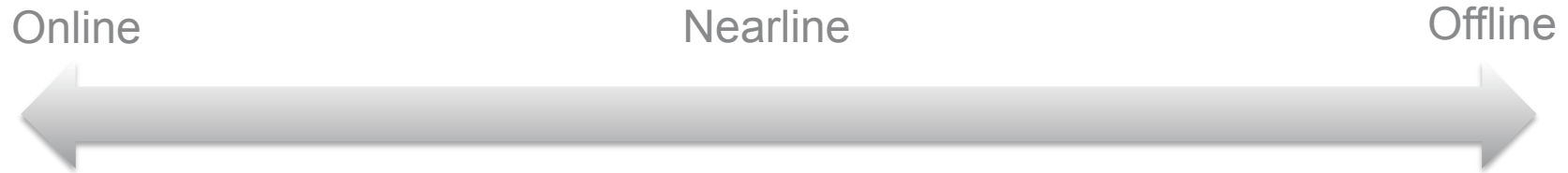
Audit trail



Top Complaints from Data Scientists

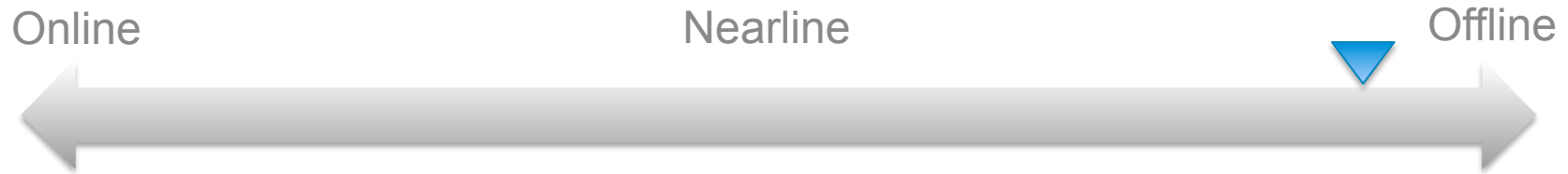
- **Discovery**: where is the data?
- **Wrangling**: can I make sense of the data?
- **Verifying**: is the data correct?
- **Scaling**: how can I scale my computation?
- **Workflow**: how can I operate my processing?
- **Publishing**: how can I get my results into production?

Models of computation



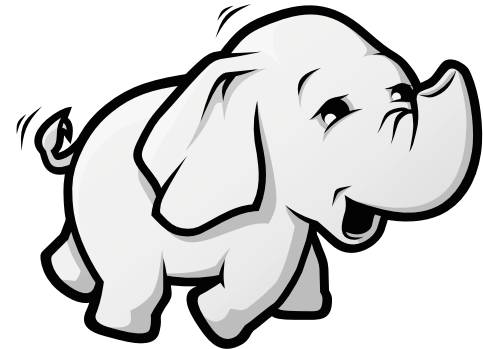
- Sub-second processing
- Harder to scale
- Must handle failures gracefully
- Computationally intensive
- Easier to scale
- Easier to tolerate failures
- Faster iteration

Models of computation



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

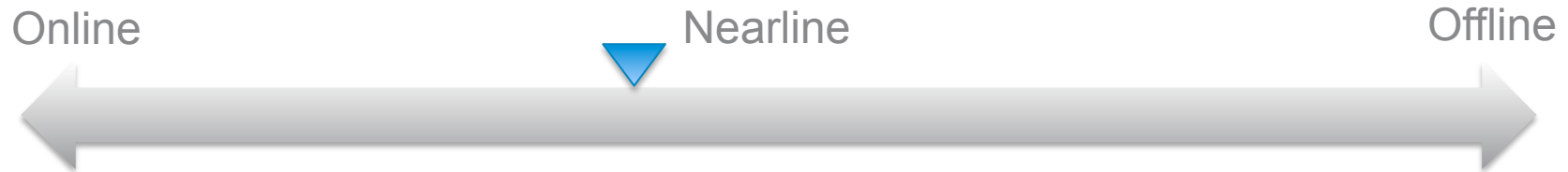
Hadoop



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

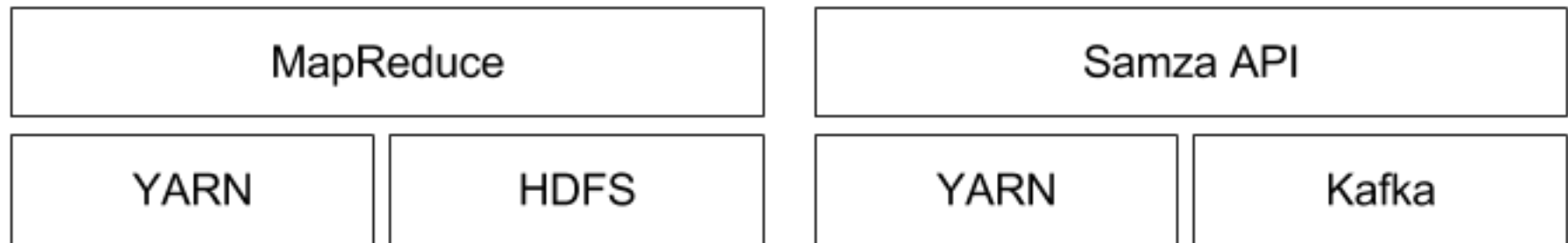
Models of computation



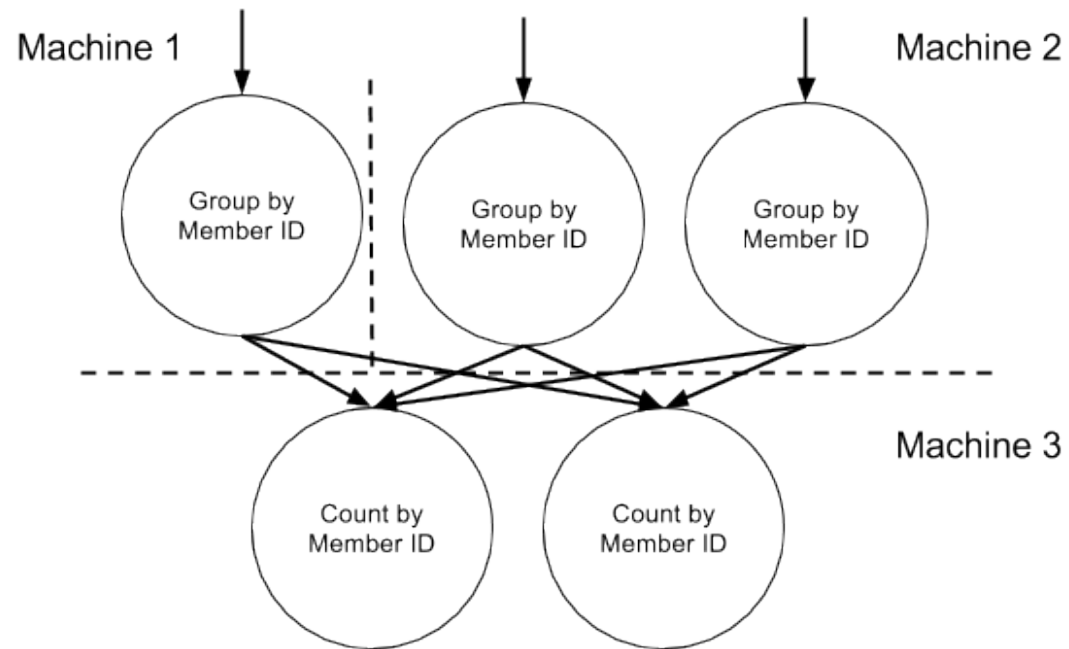
- Sub-second processing
- Harder to scale
- Must handle failures gracefully
- Computationally intensive
- Easier to scale
- Easier to tolerate failures
- Faster iteration

Apache Samza – “MapReduce for streams”

samza

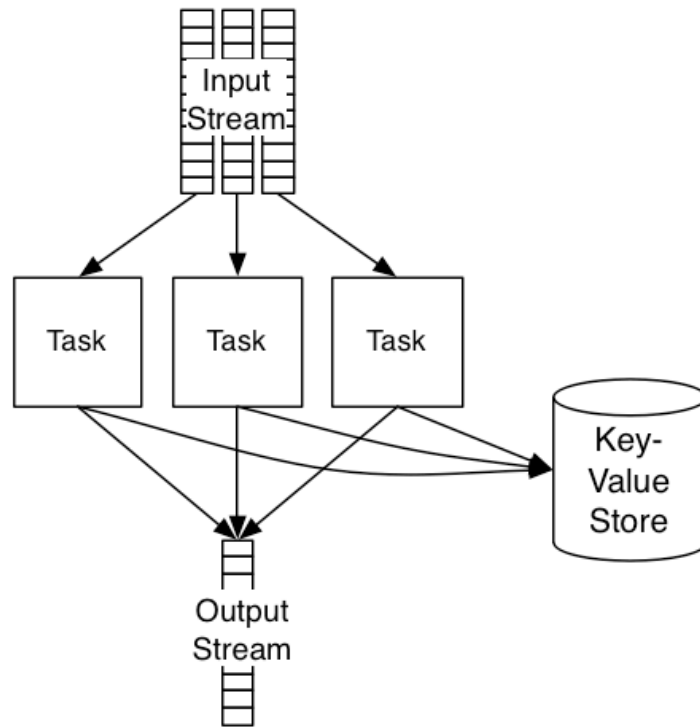


Samza

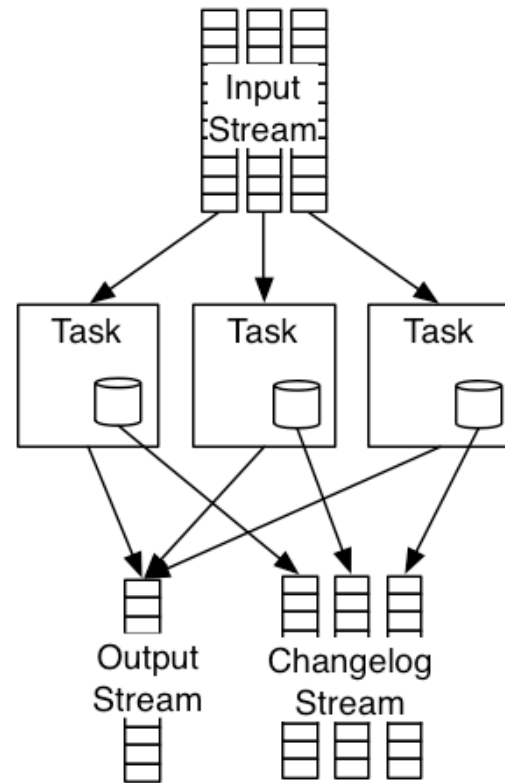
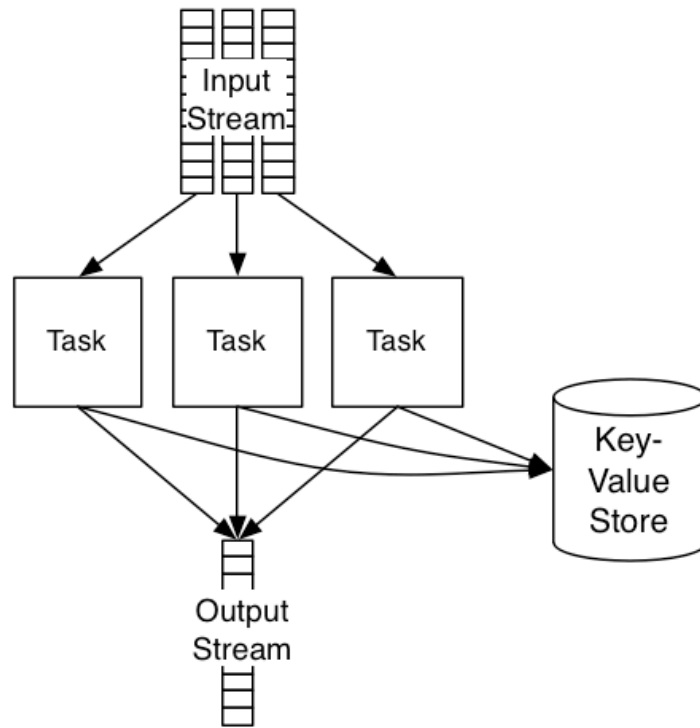


```
SELECT COUNT(*) FROM PageViewEvent GROUP BY member_id
```

Samza: State Management



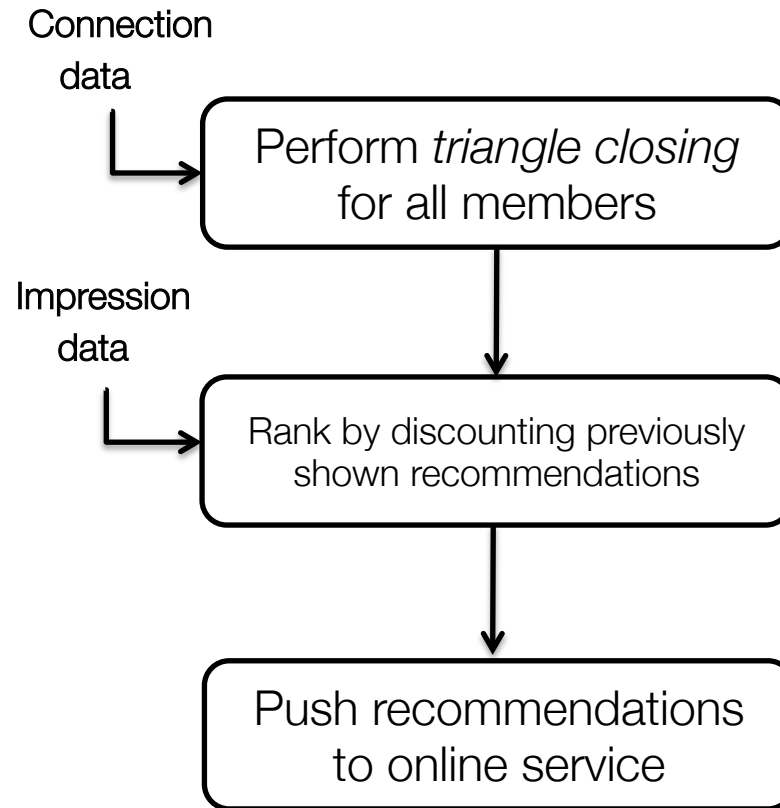
Samza: State Management



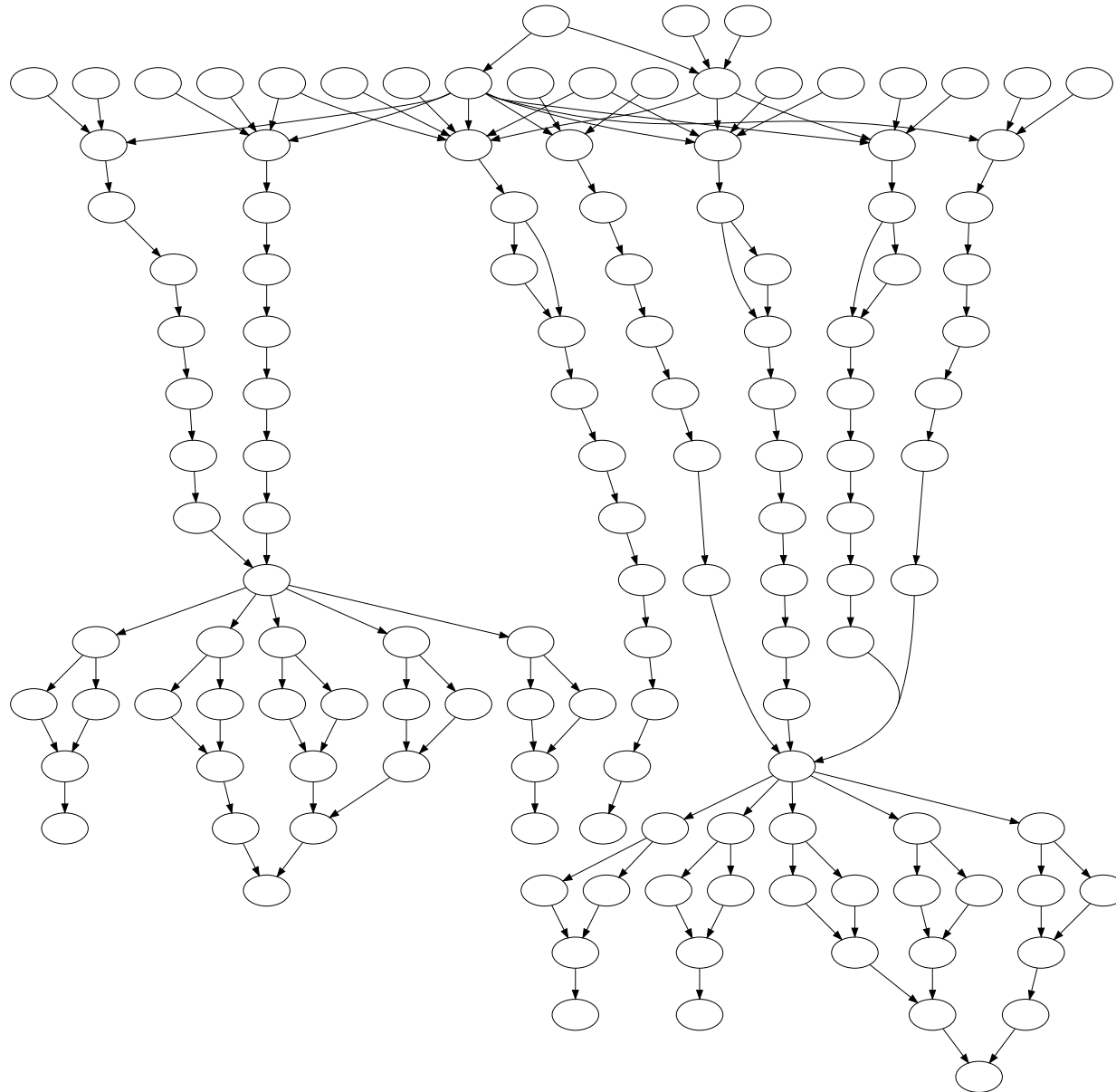
Top Complaints from Data Scientists

- **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



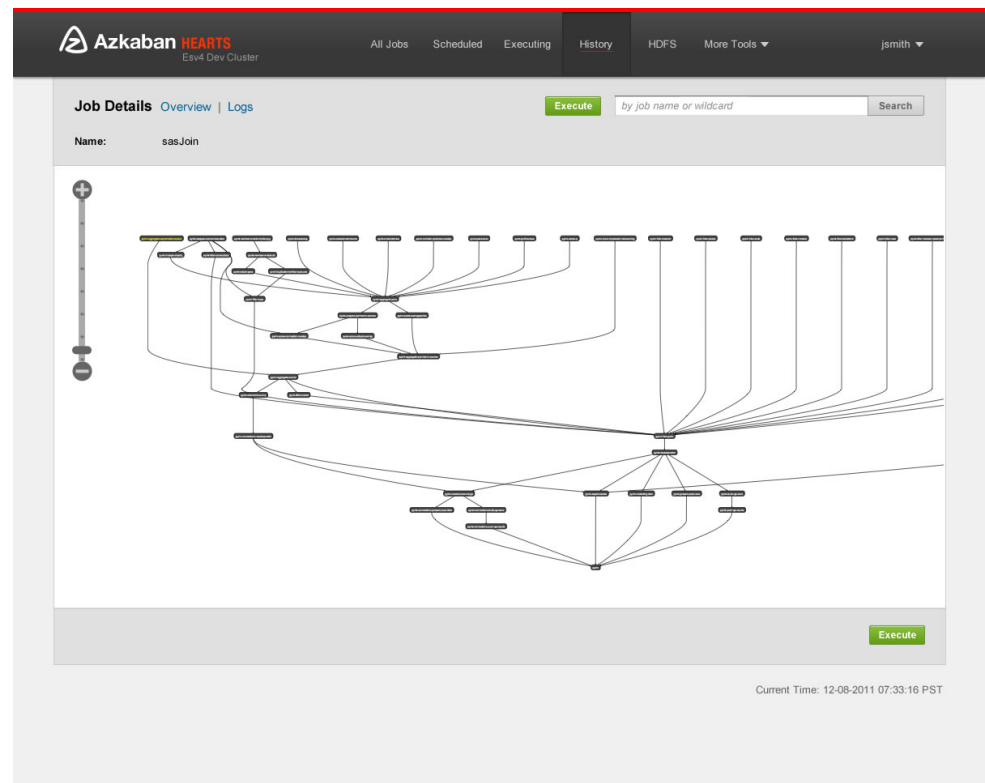
People You May Know – Workflow (in reality)



Workflow Management - Azkaban



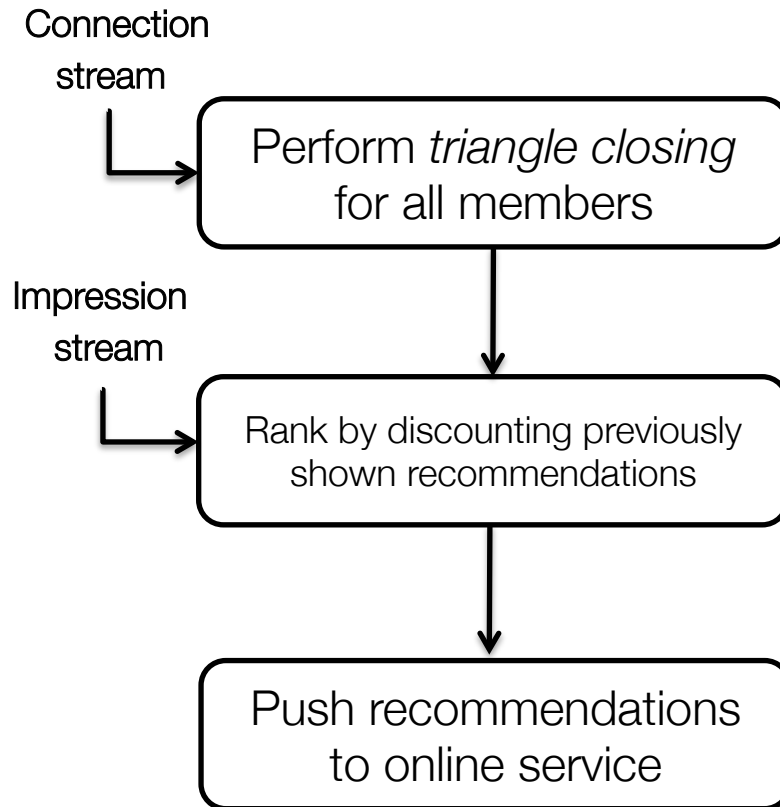
- Dependency management
- Diverse job types
- Scheduling
- Monitoring
- Visualization
- Configuration
- Retry/restart on failure
- Resource locking



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 =>

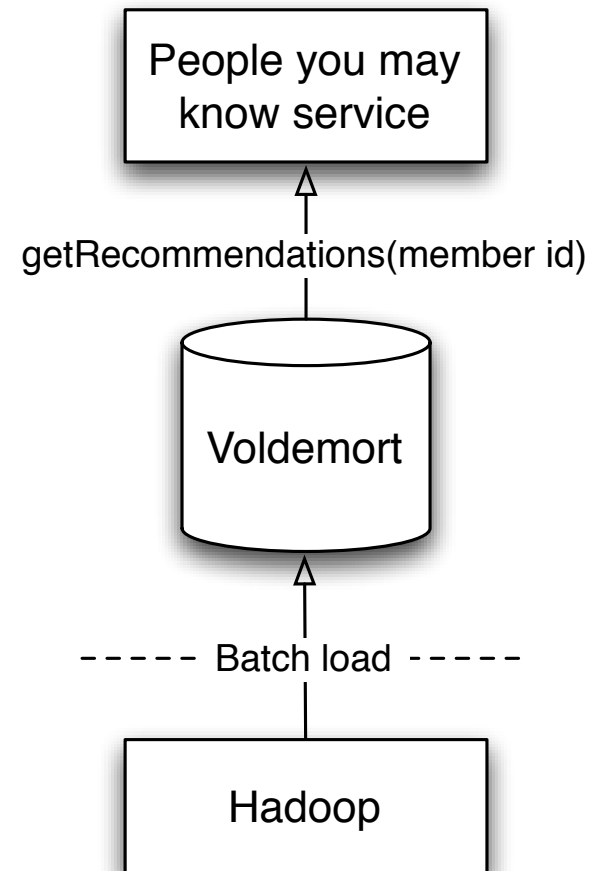
[Recommended member id 1734,
Recommended member id 1523

...

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