The Shape of Revolutions: How Things Change

Ellen Friedman, PhD
13 June 2017
Berlin Buzzwords #bbuzz

Contact Information

Ellen Friedman, PhD

Principal Technologist, MapR Technologies

Committer Apache Drill & Apache Mahout projects

O'Reilly author

Email efriedman@mapr.com ellenf@apache.org

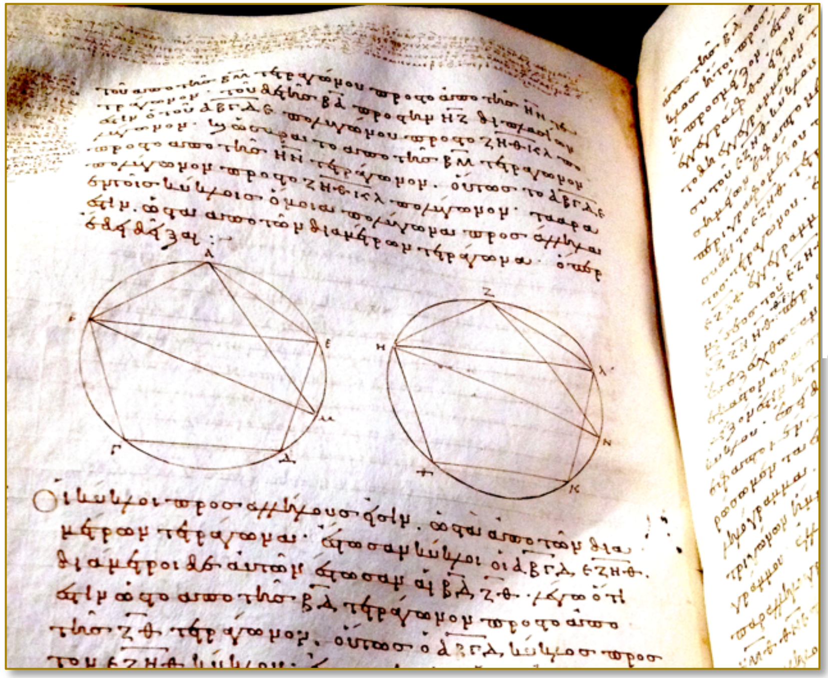
Twitter @Ellen_Friedman

Today: #bbuzz

© 2017 Ellen Friedman

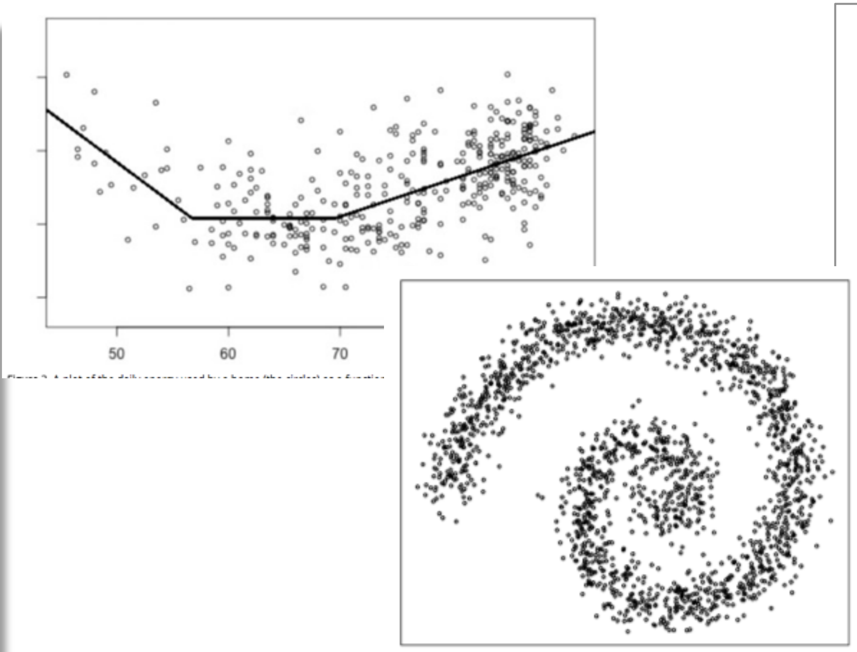
What makes innovation have real impact?

Discovery of new ideas & technology is just part of the picture



Euclid's Geometry, Bodleian Library, Oxford

Image © Ted Dunning 2015 used with permission



Tensors; k-means clustering

Images ©T. Dunning 2017 used with permission

Medical imaging

Image © WesAbrams, used with permission

© 2017 Ellen Friedman

Great Discovery in Genetics: Gregor Mendel

Is blended inheritance correct?

I could test it.

Peas, please….

© 2017 Ellen Friedman

Great Discovery in Genetics: Gregor Mendel

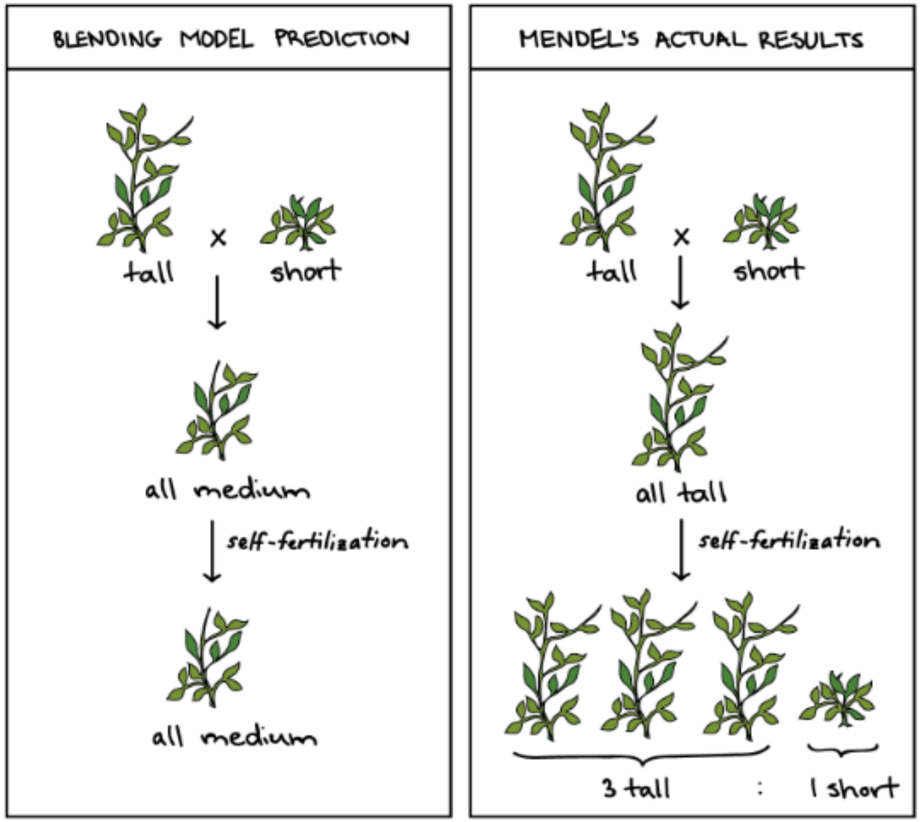
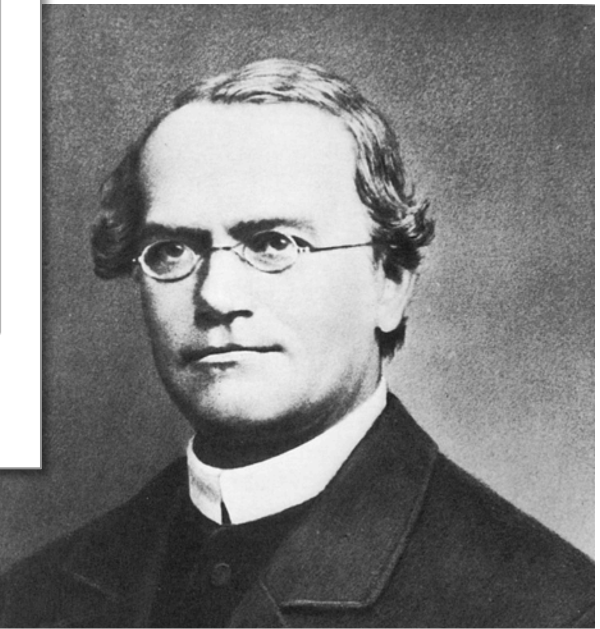


Image modified from "Mendel seven characters," by Mariana Ruiz Villareal (public domain).



Mendel’s Discovery

- Experimental data did not support blended inheritance;
  Supported *discontinuous inheritance*

- Defined terms “dominant” and “recessive” for inheritance

- Basis for modern genetics!!

- Presented results in1865 talks & 1866 paper
  “Experiments on Plant Hybridization”

© 2017 Ellen Friedman

Huge discovery, with strong evidence to support it

Initial impact:
none

Re-Discovery of Mendel’s Work

- 40 scientists had failed to see the significance of Mendel’s work…

Until…

- William Bateson (Cambridge University botanic garden) work on discontinuous inheritance 1894 – 1900 credited Mendel + “rediscovery” of Mendel paper by DeVries & Correns

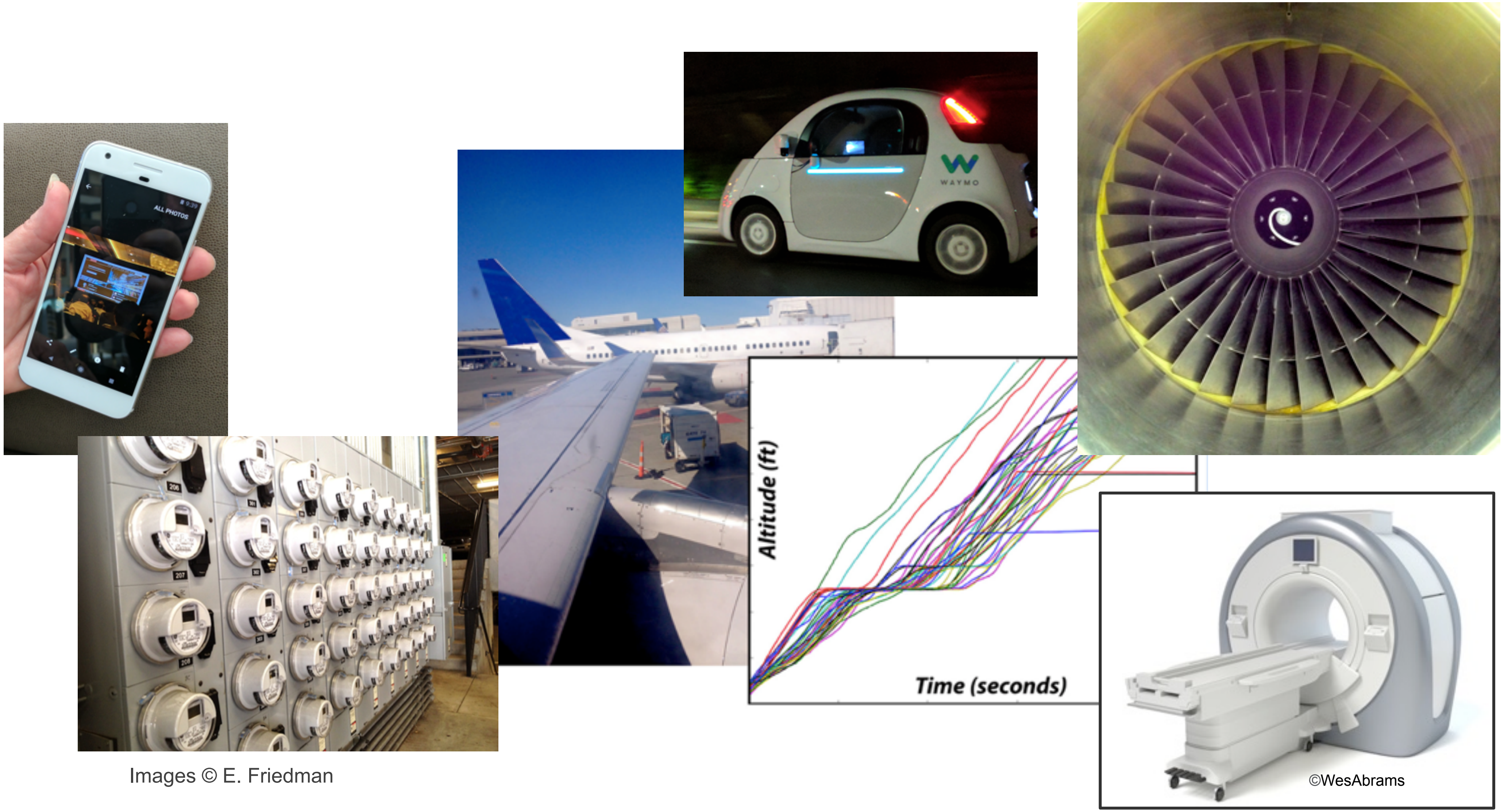
- Huge amount of work followed → modern genetics

For innovation to have impact, the *users* must have vision

© 2017 Ellen Friedman

Why stream?

IoT Data: Sensors & Smart Parts

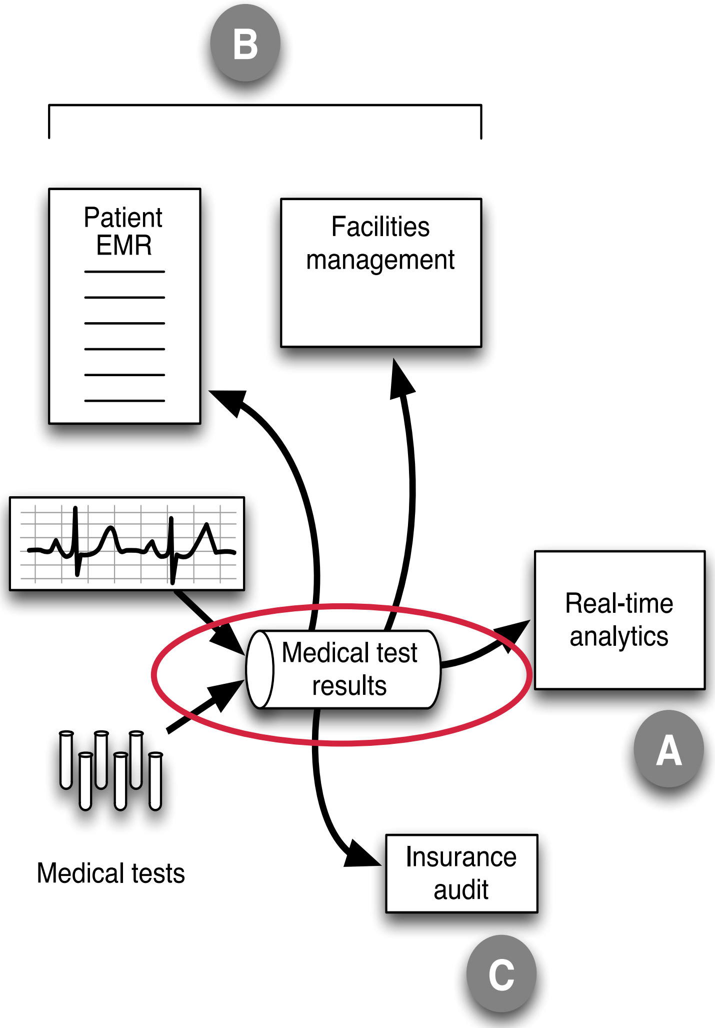


Images © E. Friedman

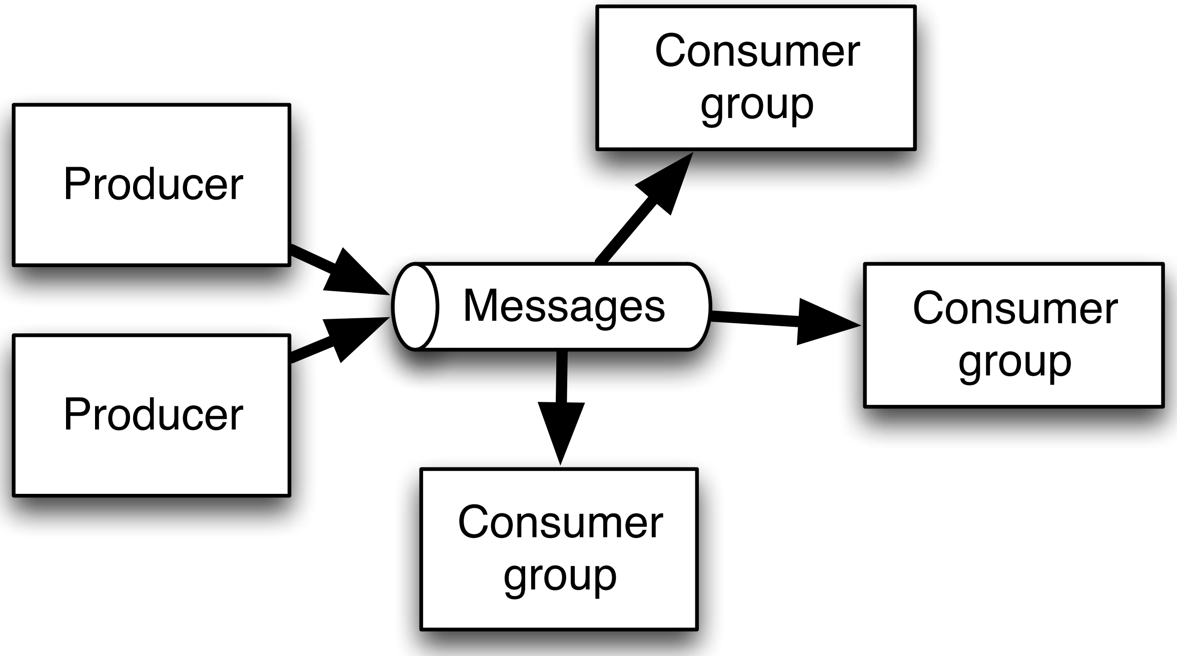
©WesAbrams

At the Heart: Message Transport

With the right messaging tool at the heart of stream-1st architecture you support other classes of use cases (B & C)



Message Transport Technology: Apache Kafka & MapR Streams



Key capabilities

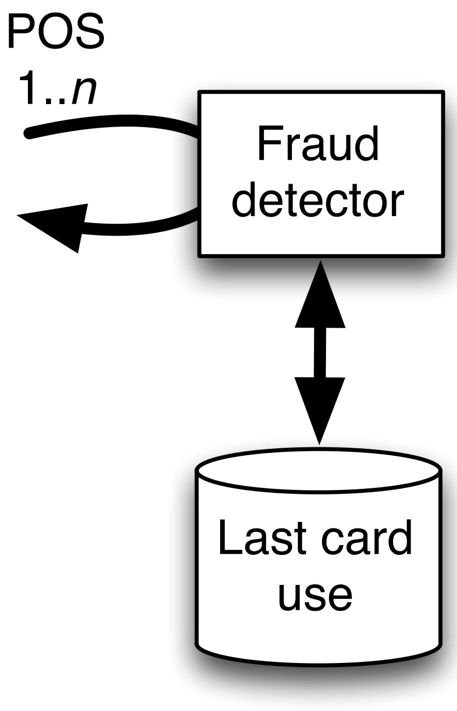
- Highly scalable
- High throughput, low latency
- Multiple producers & consumers: decoupled
- Durable messages
- Geo-distributed replication preserves offsets, high topic cardinality (unique to MapR Streams)

© 2017 Ellen Friedman

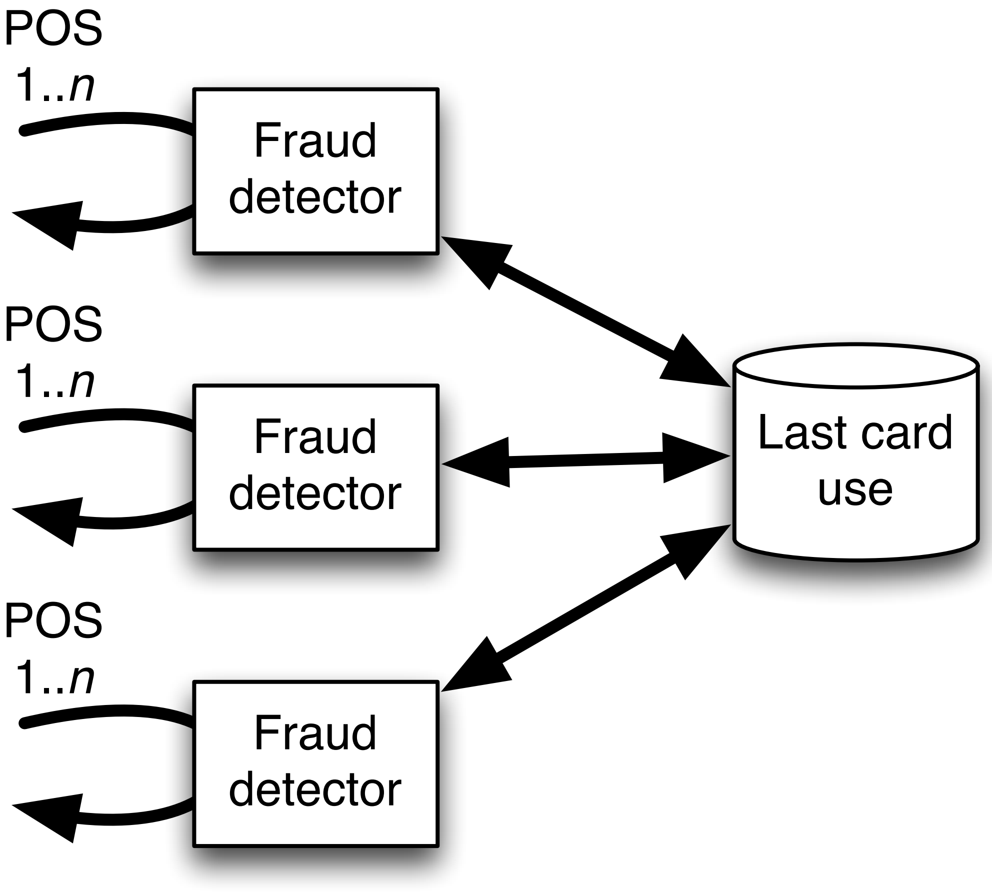
Stream transport gives you the flexibility of micro services

© 2017 Ellen Friedman

Traditional Solution – Use a Profile Database

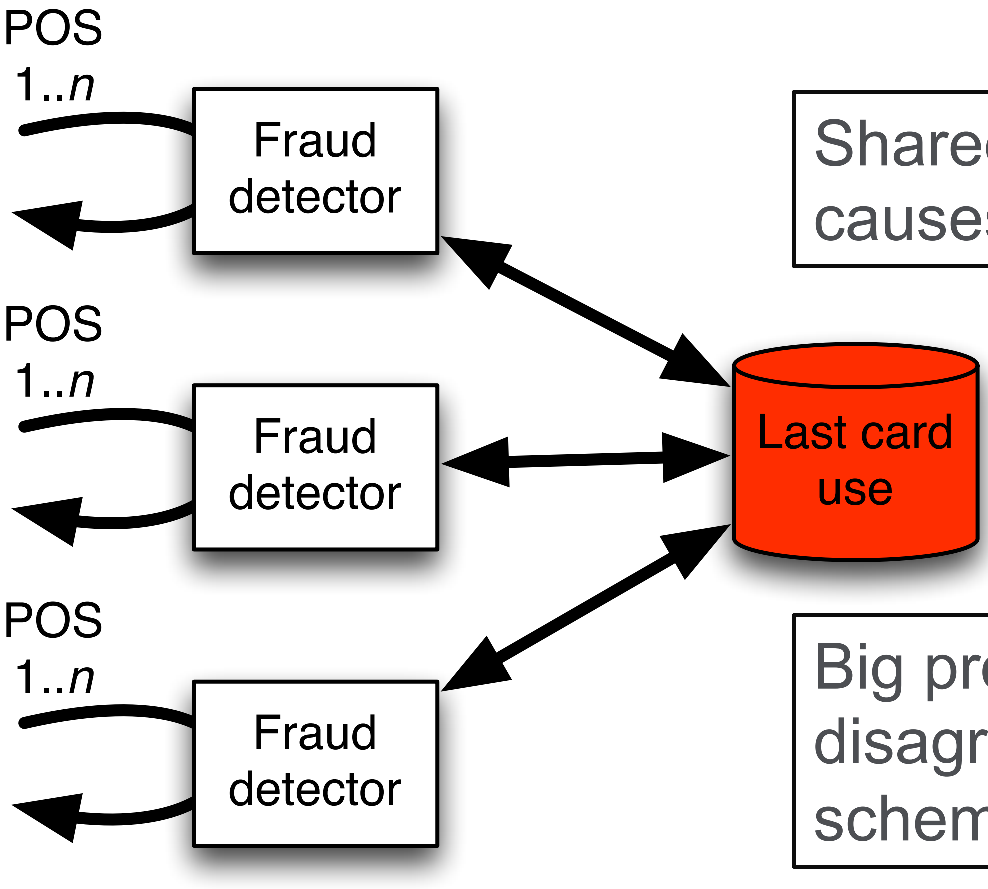


What Happens Next?



© 2017 Ellen Friedman

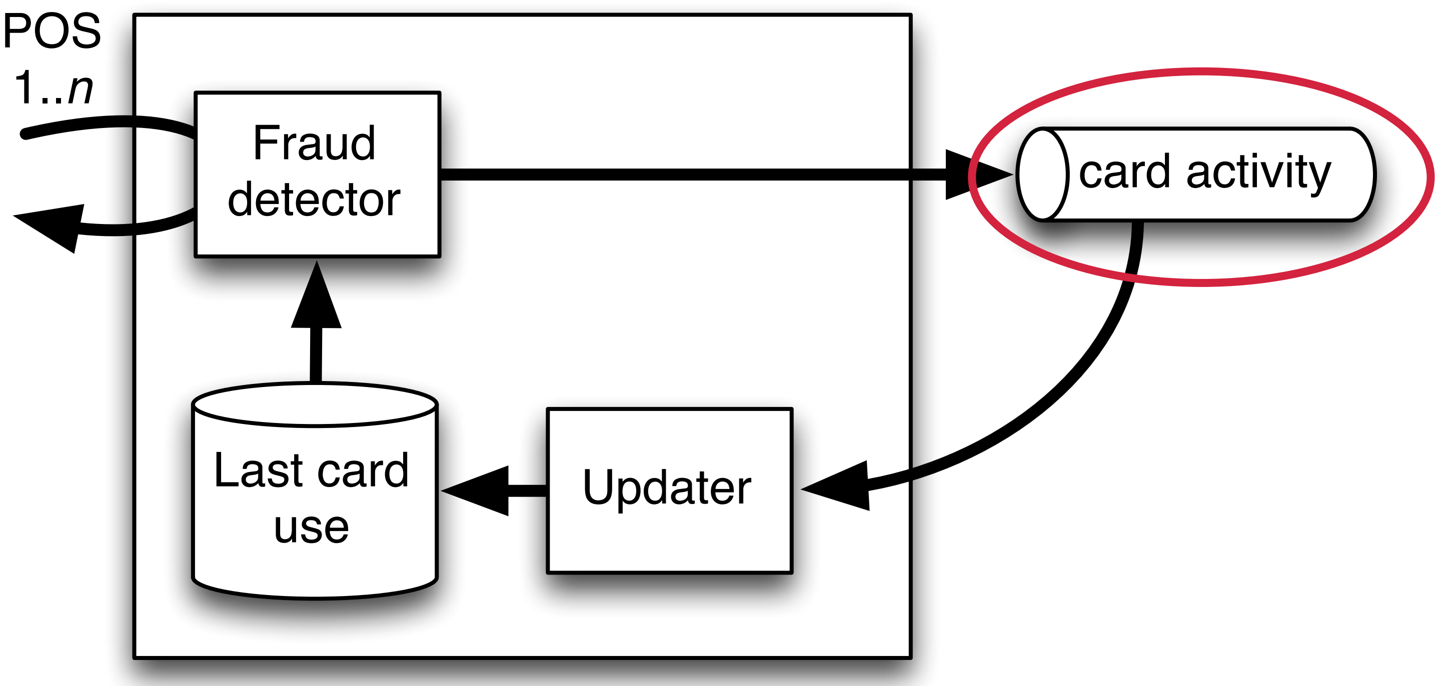
What Happens Next?



Shared database causes problems

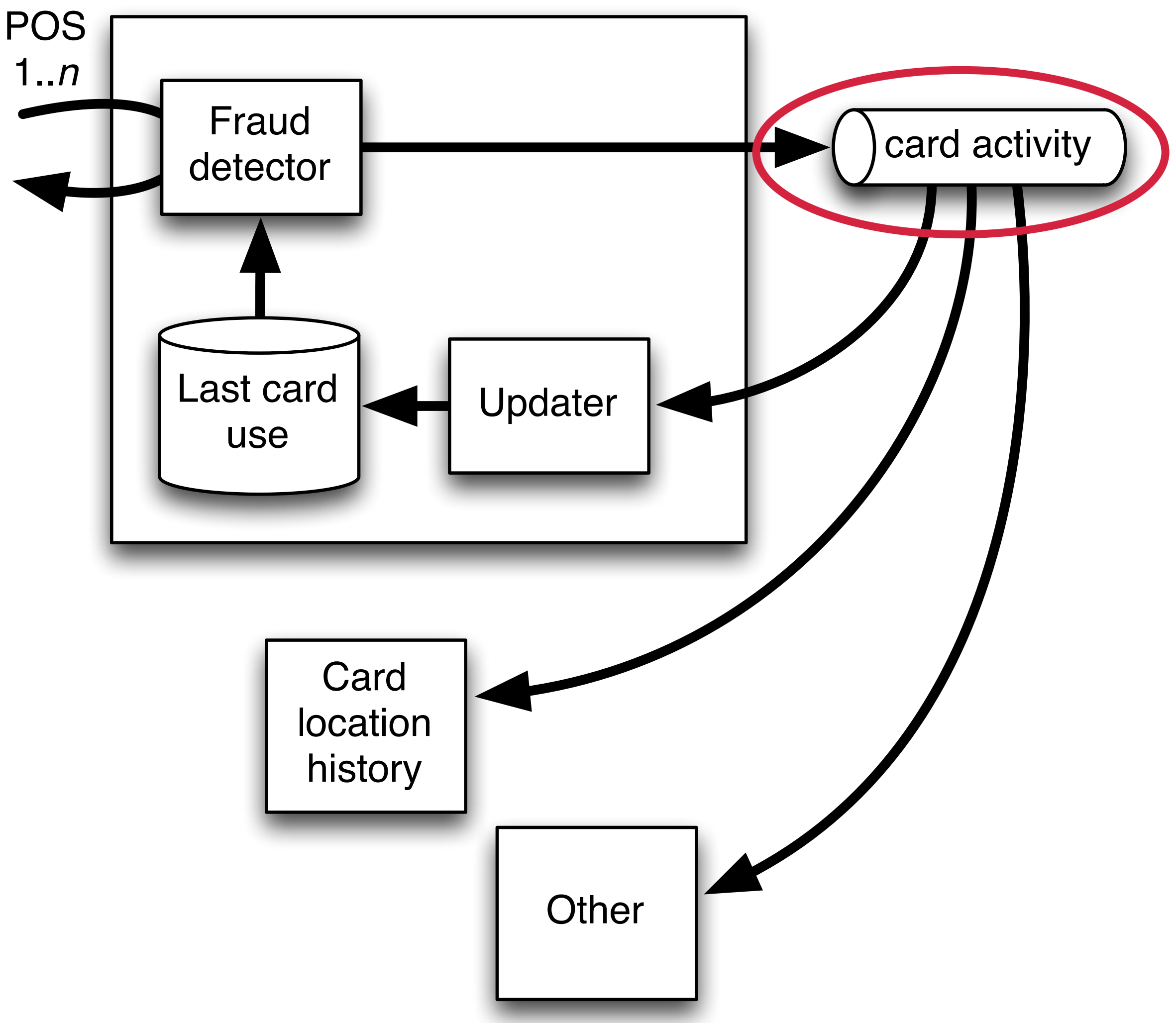
Big problem is disagreement about schema and indexing

Use a Stream Isolate Services



© 2017 Ellen Friedman

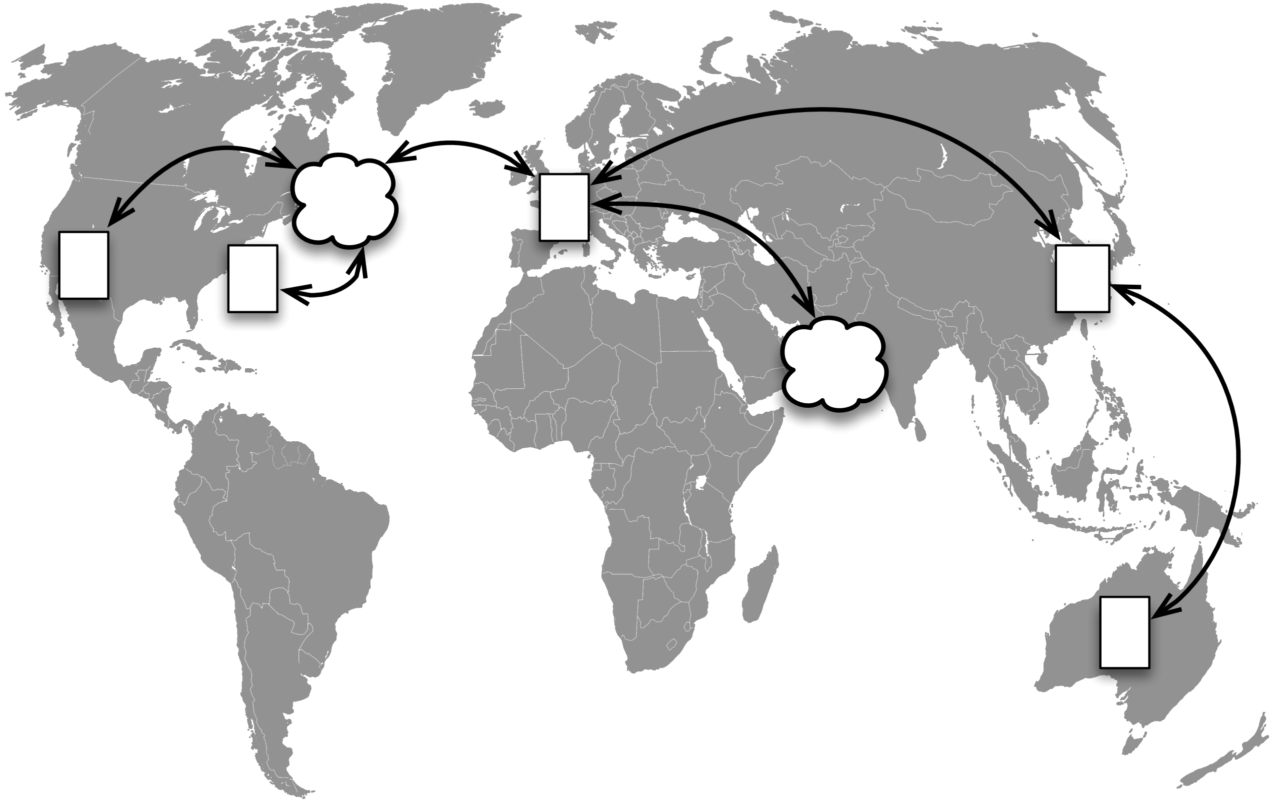
Add New Services via the Stream



Act locally, learn globally

© 2017 Ellen Friedman

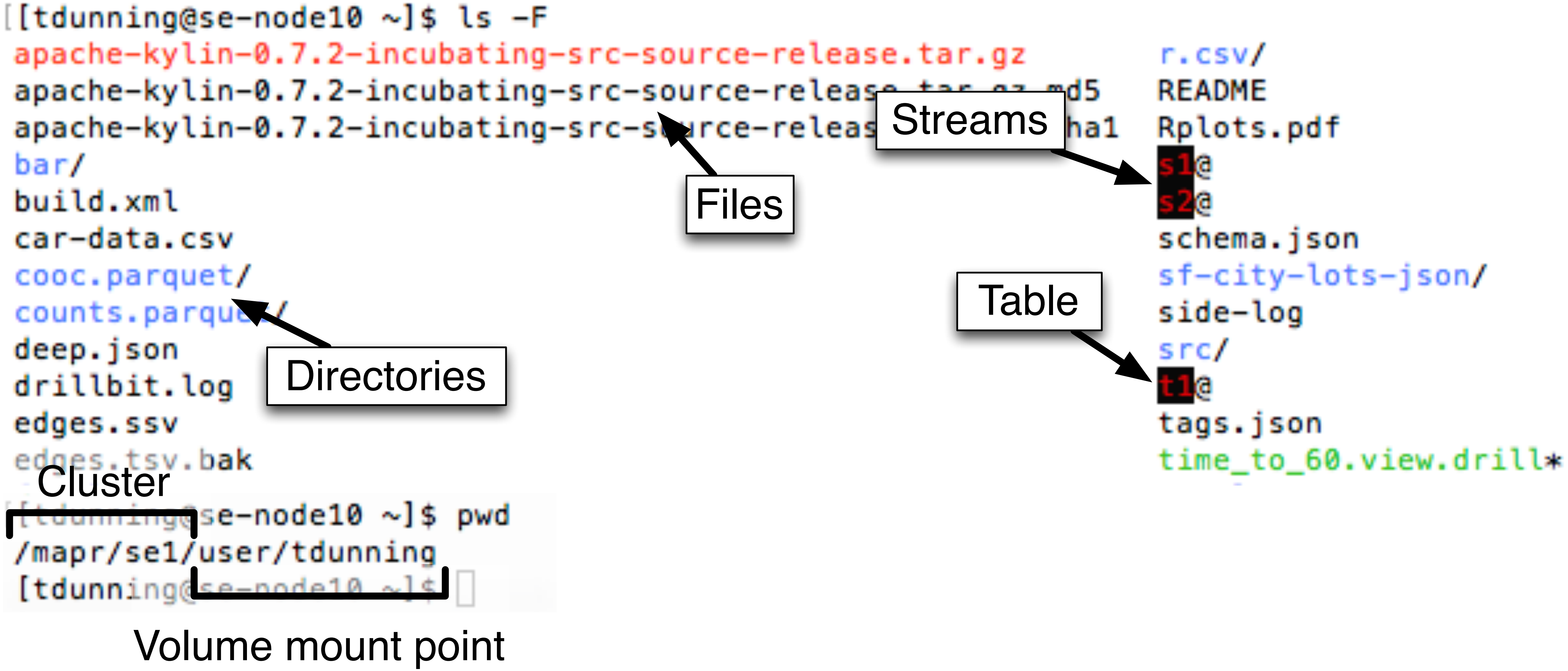
Global Data Fabric



Collect, access & analyze big data where ever you need it, in a seamless system under the same security & administration.

© 2017 Ellen Friedman

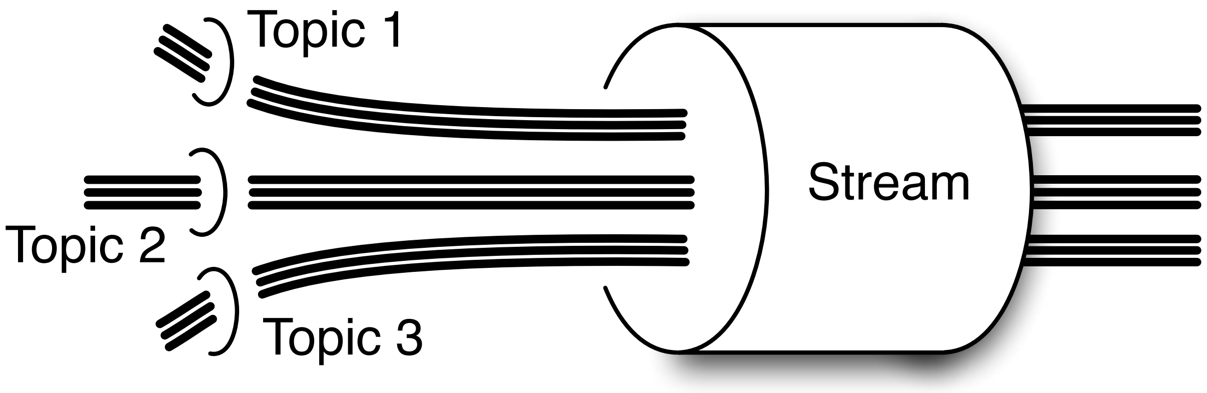
Example: MapR Converged Data Platform



```
[[tdunning@se-node10 ~]$ ls -F
apache-kylin-0.7.2-incubating-src-source-release.tar.gz       r.csv/
apache-kylin-0.7.2-incubating-src-source-release.tar.gz.md5   README
apache-kylin-0.7.2-incubating-src-source-release.tar.gz.sha1  Rplots.pdf
bar/                                                          s1@
build.xml                                                     s2@
car-data.csv                                                  schema.json
cooc.parquet/                                                 sf-city-lots-json/
counts.parquet/                                               side-log
deep.json                                                     src/
drillbit.log                                                  t1@
edges.ssv                                                     tags.json
edges.tsv.bak                                                 time_to_60.view.drill*
[tdunning@se-node10 ~]$ pwd
/mapr/se1/user/tdunning
[tdunning@se-node10 ~]$
```

Unique to MapR: Manage Topics at Stream Level

- *Many* more topics on MapR cluster
- Topics are grouped together in Stream (different from Kafka)



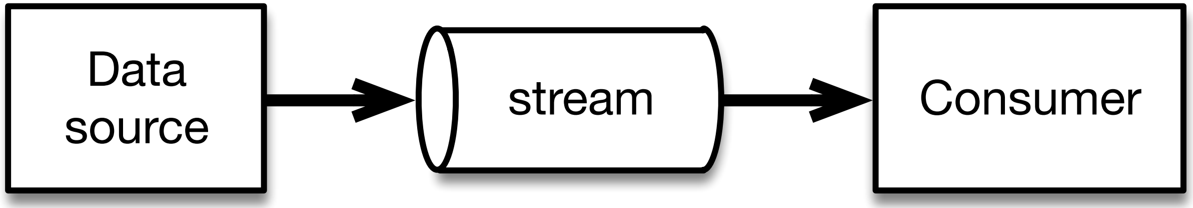
- Policies set at Stream level: time-to-live, ACEs, geo-distributed stream replication (different from Kafka)

© 2017 Ellen Friedman

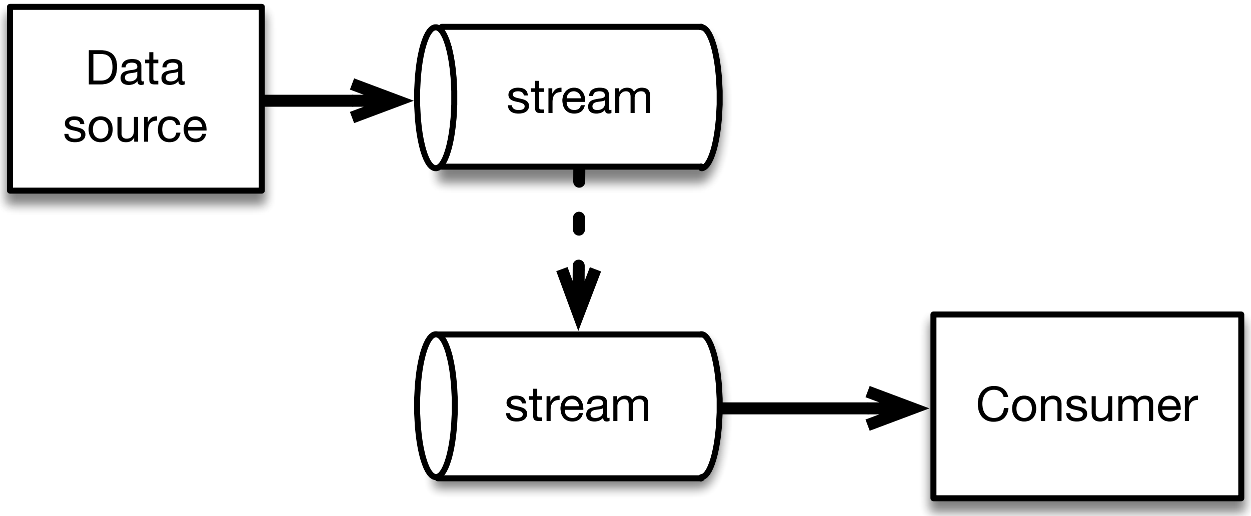
MapR has multi-master, bi-directional table & stream replication

© 2017 Ellen Friedman

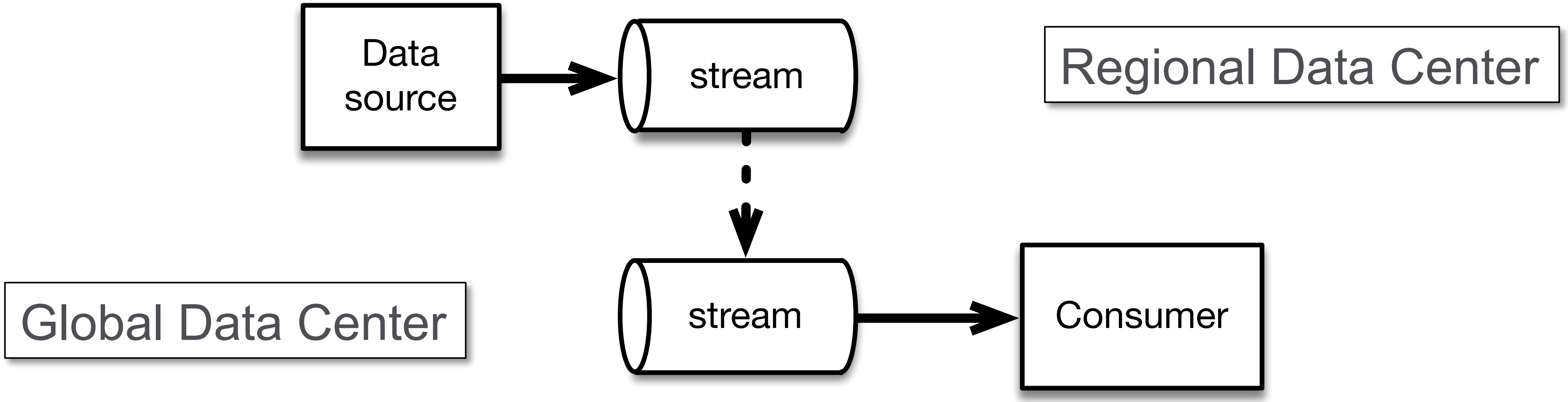
With MapR, Geo-distributed Data Appears Local



With MapR, Geo-distributed Data Appears Local



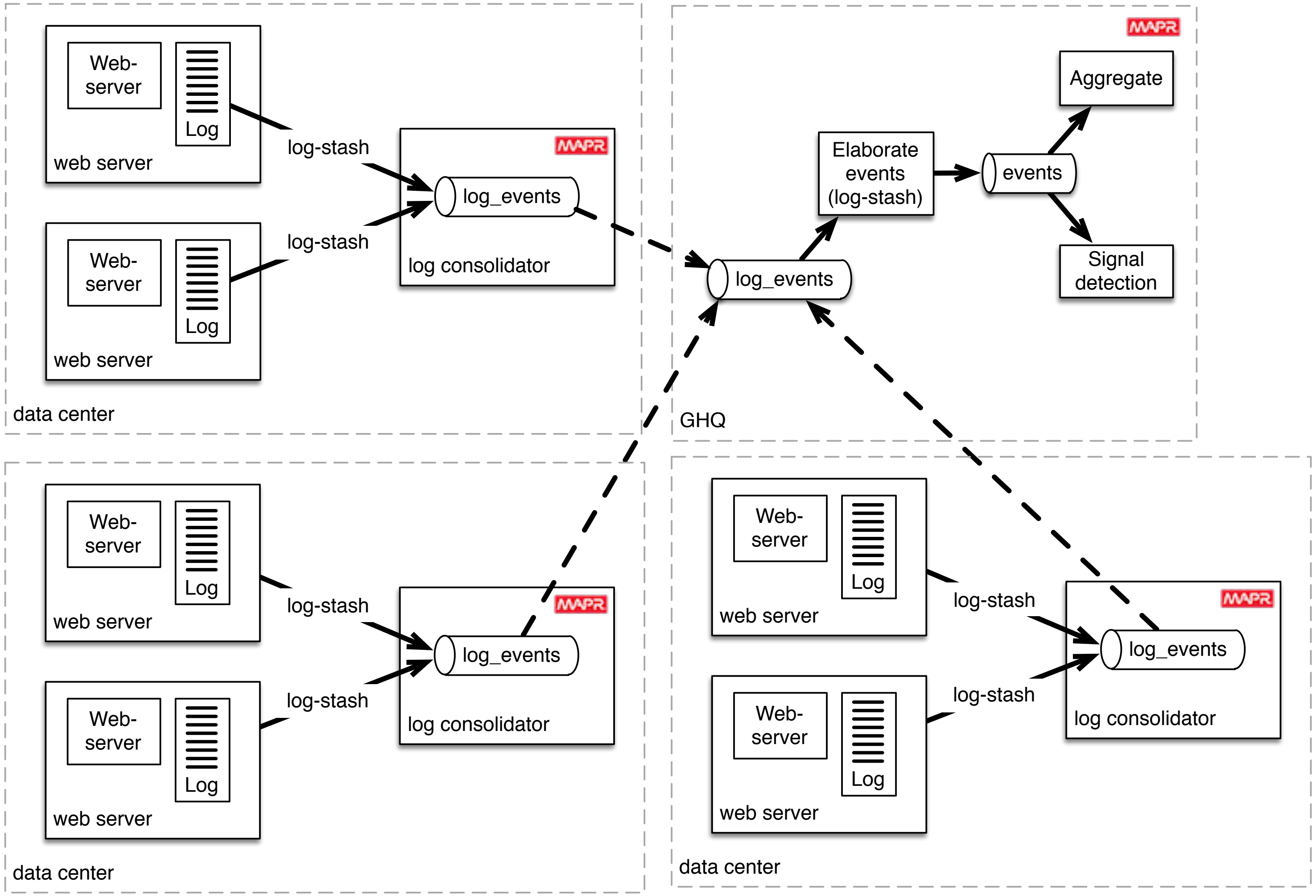
With MapR, Geo-distributed Data Appears Local



Separation of concerns

© 2017 Ellen Friedman

Metrics: Collect Data & Transport to Global Analytics

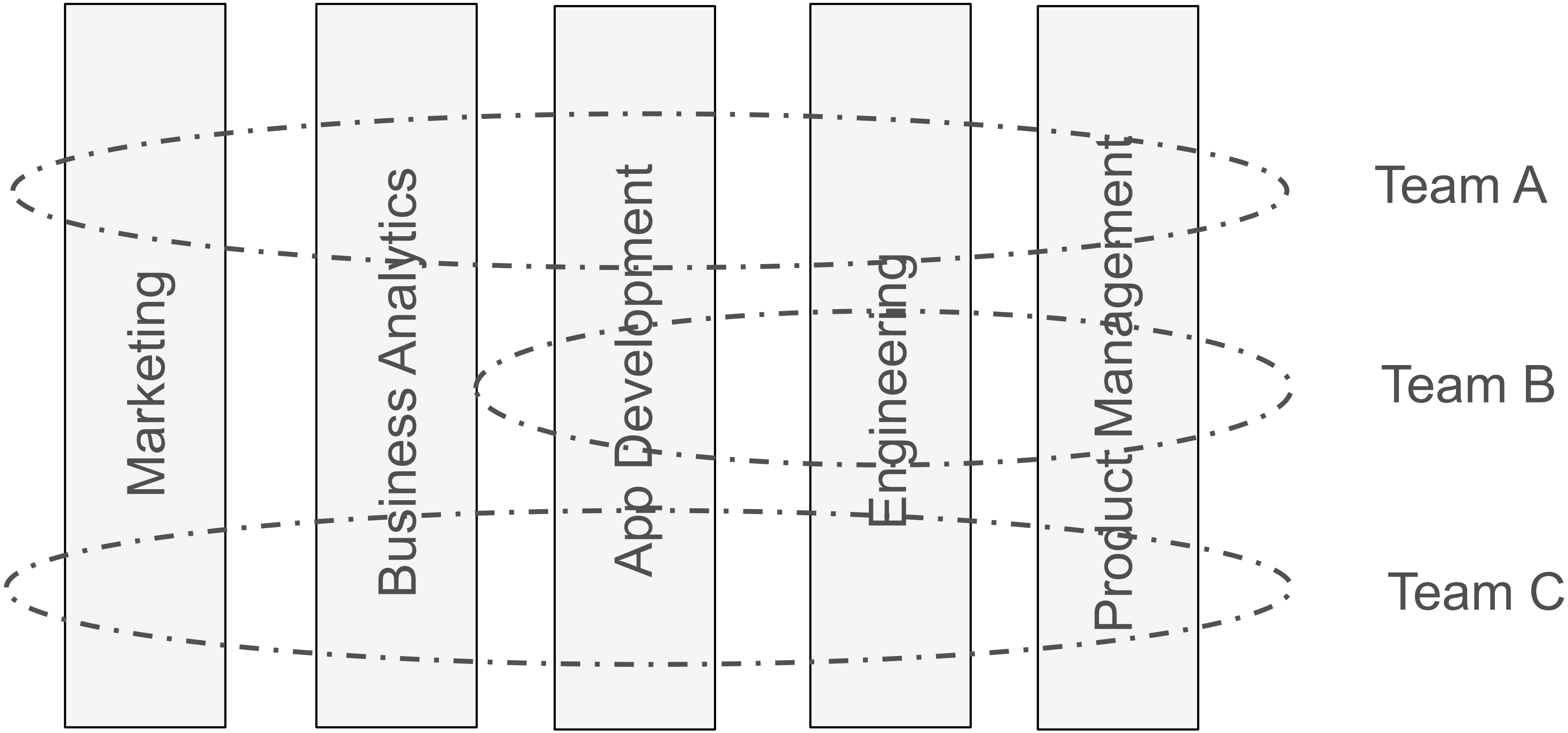


© 2017 Ellen Friedman

“A year of dev time in the bank”

© 2017 Ellen Friedman

Cross-functional teams each with common goal



© 2017 Ellen Friedman

Data is Changing a Society

1.2 B

PEOPLE

Aadhaar Project: Largest Biometric DB in the World

- Unique 12 – digit number for each person in India
- Proof of identity, authenticated anytime, anywhere
- Authentication runs on NoSQL database MapR-DB

AADHAAR

© 2017 Ellen Friedman

Data + Inspiration = Great Things

sarah everts @saraheverts · Apr 22

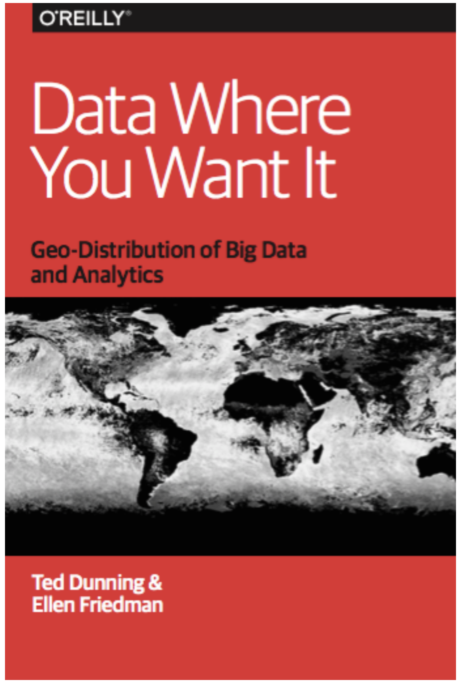
Thousands here at the Brandenburg Gate for Berlin's #sciencemarch #sciencemarchBER #chemistsmarch pic.twitter.com/XXrzCtB4V8

What will you build?

© 2017 Ellen Friedman

Geo-Distribution of Big Data and Analytics

New O’Reilly data report by Ted Dunning & Ellen Friedman © March 2017



O'REILLY

Data Where You Want It

Geo-Distribution of Big Data and Analytics

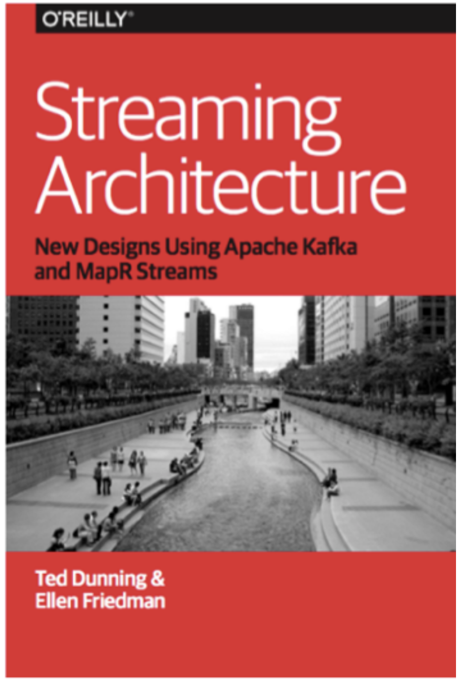
Ted Dunning & Ellen Friedman

Download free pdf courtesy of MapR: http://bit.ly/mapr-geo-distribution-ebook-pdf

Streaming Architecture:

New Designs Using Apache Kafka & MapR Streams

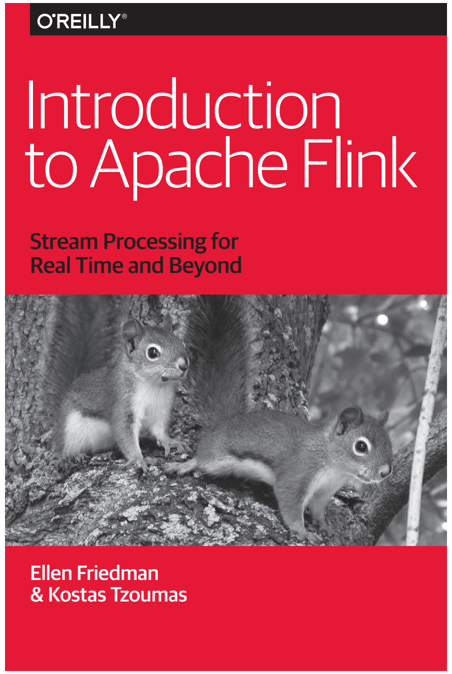
Book by Ted Dunning & Ellen Friedman © March 2017


O'REILLY
Streaming Architecture
New Designs Using Apache Kafka and MapR Streams
Ted Dunning & Ellen Friedman

Free copy online courtesy of MapR:
http://bit.ly/mapr-apache-flink-ch1

Introduction to Apache Flink

Book by Ellen Friedman & Kostas Tzoumas © September 2016


O'REILLY
Introduction to Apache Flink
Stream Processing for Real Time and Beyond
Ellen Friedman & Kostas Tzoumas

Free copy online courtesy of MapR:
http://bit.ly/mapr-streaming-architecture-book

Please support women in tech – help build girls’ dreams of what they can accomplish

#womenintech #datawomen

© Ellen Friedman 2015

Thank you !

© 2017 Ellen Friedman

Contact Information

Ellen Friedman, PhD

Principal Technologist, MapR Technologies

Committer Apache Drill & Apache Mahout projects

O'Reilly author

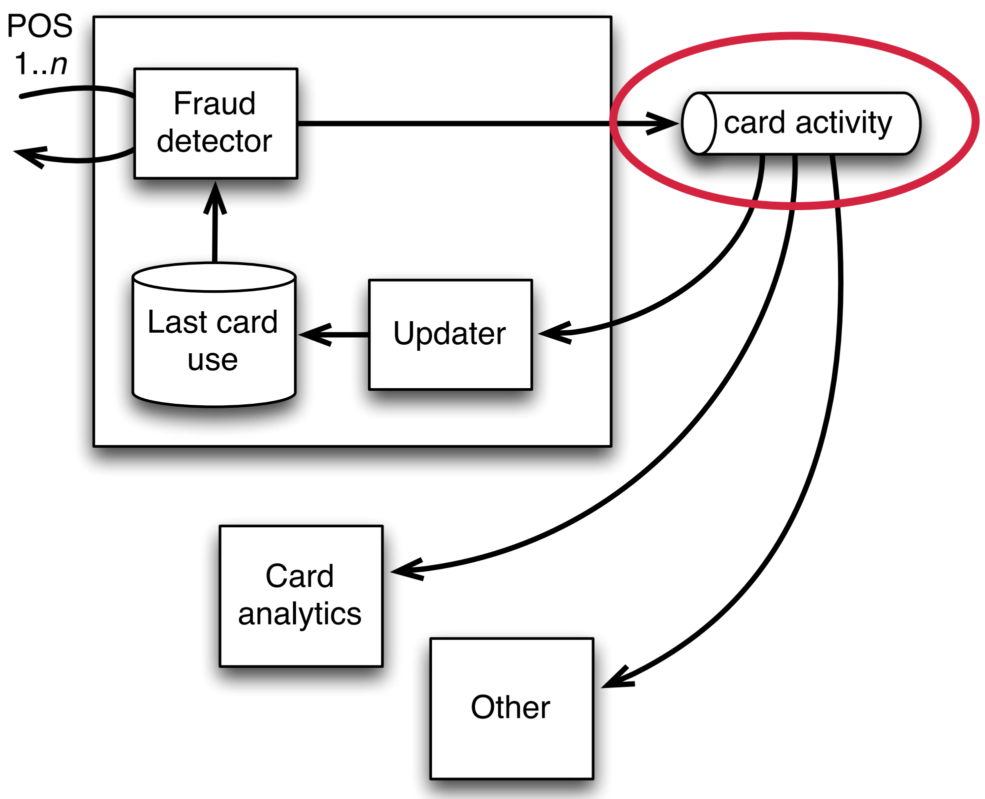
Email efriedman@mapr.com ellenf@apache.org

Twitter @Ellen_Friedman

Today: #bbuzz

© 2017 Ellen Friedman

Stream-1st Architecture: Basis for Micro-Services



Stream instead of database as the shared "truth"

© 2017 Ellen Friedman