# Real World Use Cases: Hadoop & NoSQL in Production

Ted Dunning and Ellen Friedman

Strata Hadoop World Conference 19 February 2015

# Contact Information

Ted Dunning

- Chief Applications Architect, MapR Technologies
- Committer & PMC for Apache’s Drill, Zookeeper & Mahout
- Mentor for Myriad & Apache’s Storm, Flink, Datafu, Optiq, Drill

Email tdunning@apache.org tdunning@maprtech.com

Twitter @ted_dunning

Hashtag today: #StrataHadoop

# Contact Information

Ellen Friedman

Solutions Consultant and Commentator

Apache Mahout committer, Apache Drill contributor

| | | |
|---|---|---|
| Email | ellenf@apache.org | efriedman@maprtech.com |
| Twitter | @Ellen_Friedman | @ApacheDrill |
| | Hashtag today: | #StrataHadoop |

# What can you do with Hadoop?

# How can you succeed?

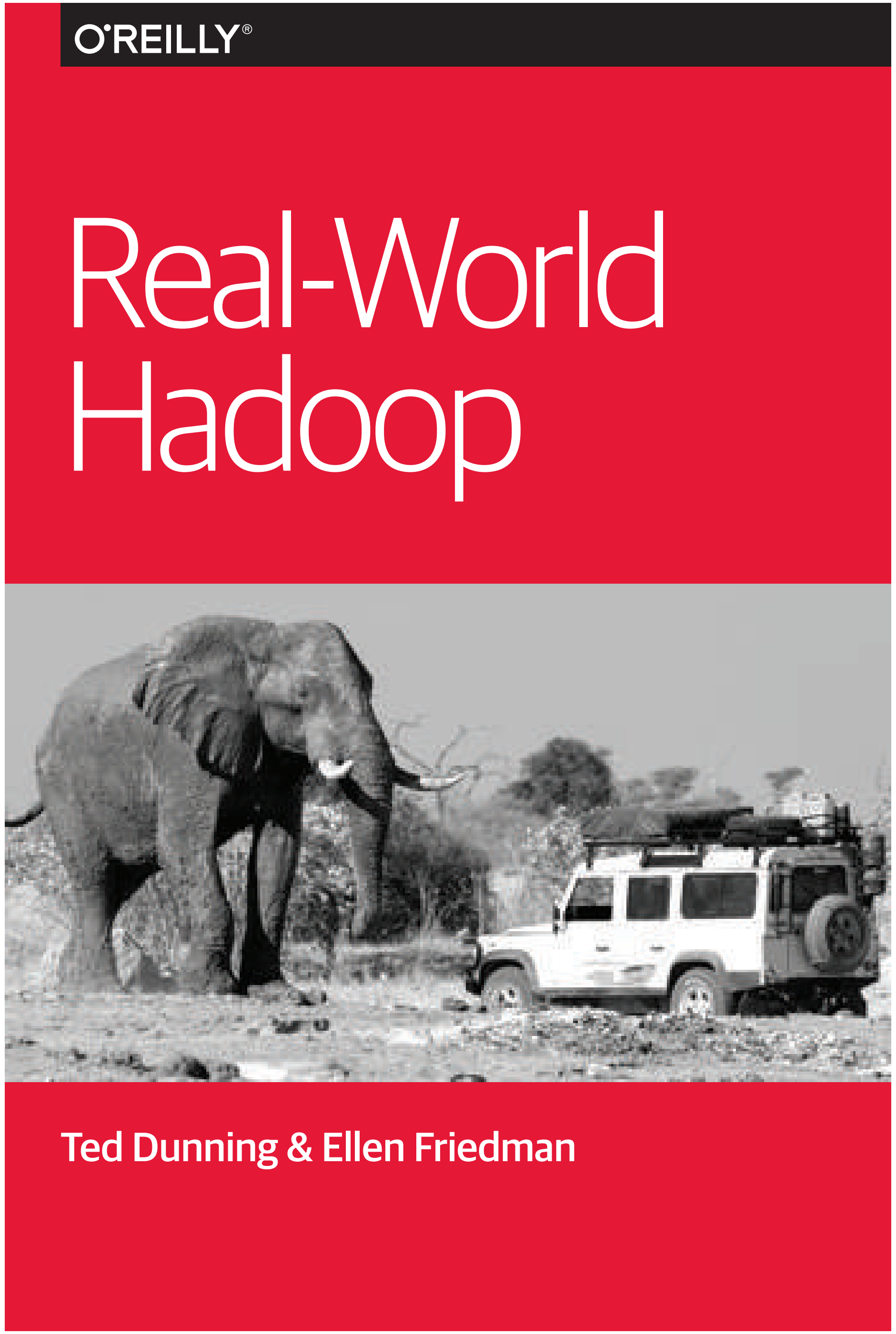
One good way is to see what others are doing

- Look at use cases in your own vertical
- What about use cases in other verticals?
  - They may look different but have the same basic issues and yield to the same basic solutions
  - Look for common design patterns that cut across verticals
- Shows you how things work in practice, not in theory

# Is Hadoop ready for production?

yes

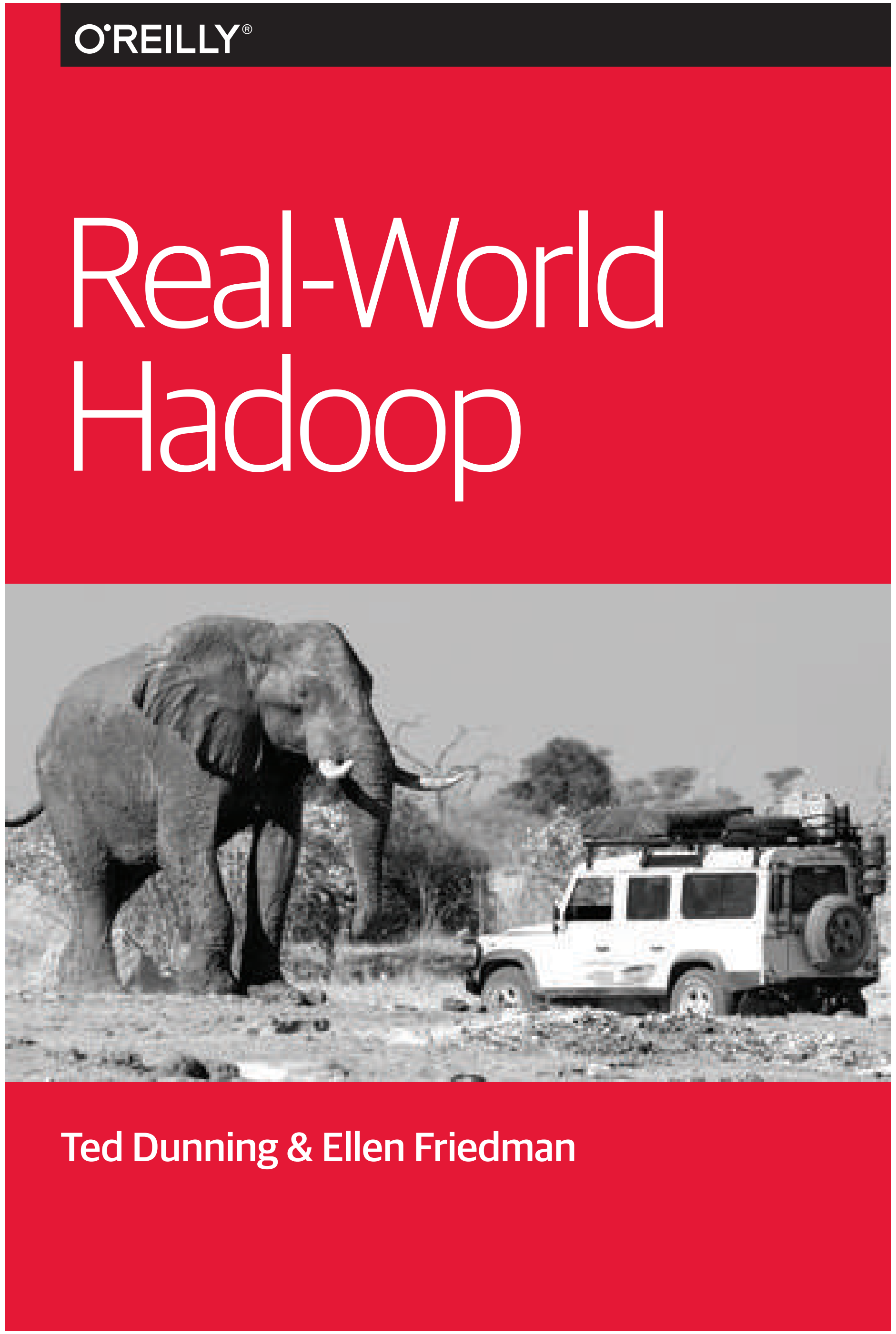
# Evidence:
# So many people are using Hadoop and NoSQL successfully in production already



O'REILLY®

# Real-World Hadoop

Ted Dunning & Ellen Friedman



O'REILLY®

# Real-World Hadoop

Ted Dunning & Ellen Friedman

How MapR customers are using Apache Hadoop and NoSQL

# What is MapR?

MAPR®

MapR is Hadoop and more…

# MapR is a distribution for Apache Hadoop, but…

- It is API compatible with Apache Hadoop (no vendor lock in)
- Has it's own distributed file system: MapR-FS
  - MapR-FS is a **real time, fully read/write file system**
  - Supports NFS/POSIX
- You can use Hadoop commands but also non-Hadoop commands
  - Also use Linux, Python, JAVA, etc.
- **MapR cluster is not isolated: Use it like any file system**

# MapR's real file system has advantages

- Snapshots are consistent
- Mirroring is fast, efficient and reliable
  - Secondary data center for disaster recovery *much* easier to set up
- You can use legacy code and applications directly
  - Don't have to copy everything in and out for use

# MapR has no NameNode

- Extremely reliable
- High availability
- Good performance; less traffic problems

# MapR-FS includes a NoSQL db: MapR-DB

- It is API compatible with Apache Hbase
- MapR-DB does not have delays due to compactions
    - Makes it very highly available
- More column families; great performance

# If you're new to Hadoop…

# Pick one thing and get started

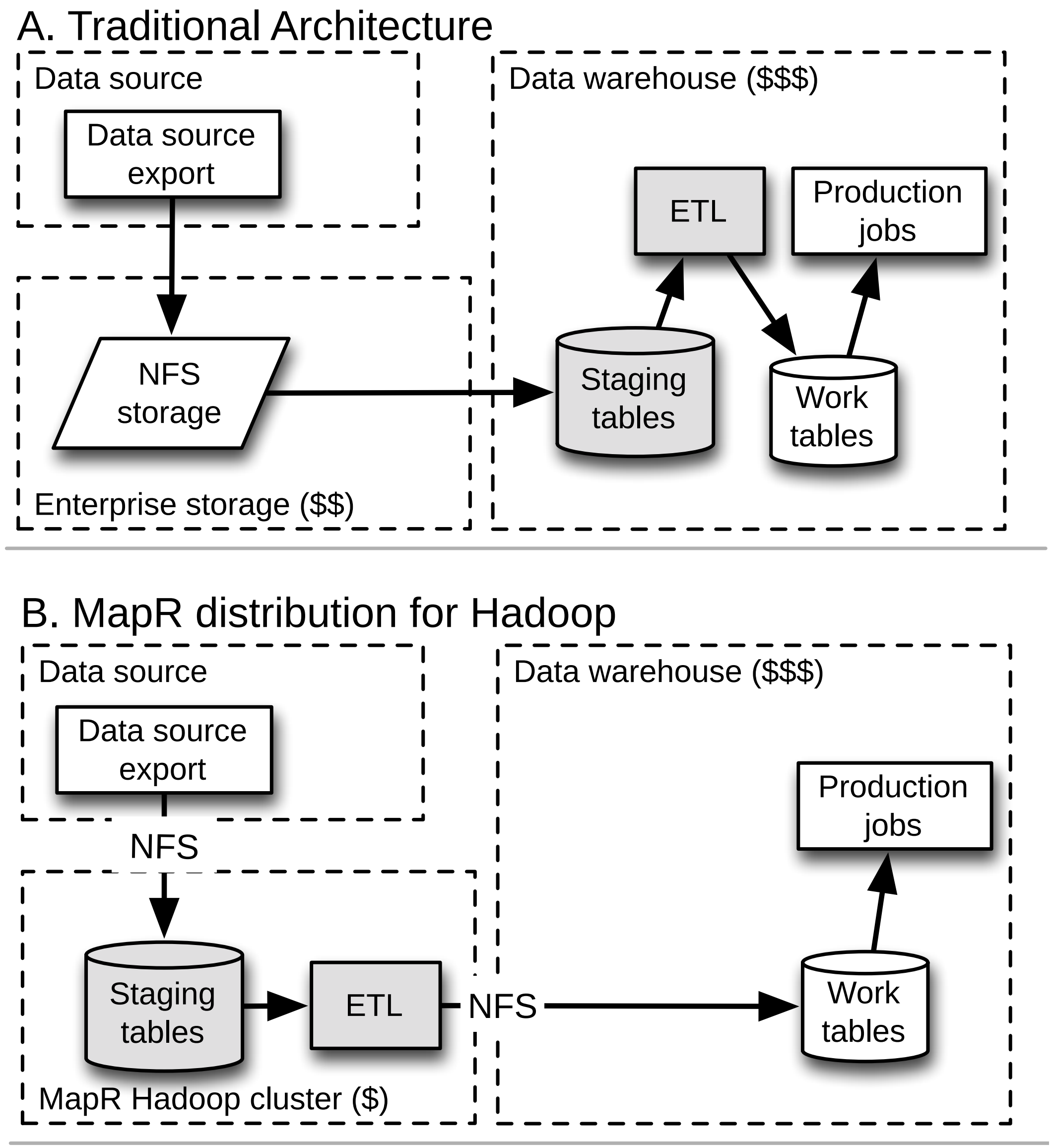
- Don’t have to decide all-at-once all the ways you may use Hadoop

- Future-proof your organization: Build experience
  - You won’t be a Hadoop pioneer, but there’s still an early mover advantage

- Lose your fear of failure (plan for a few false starts)

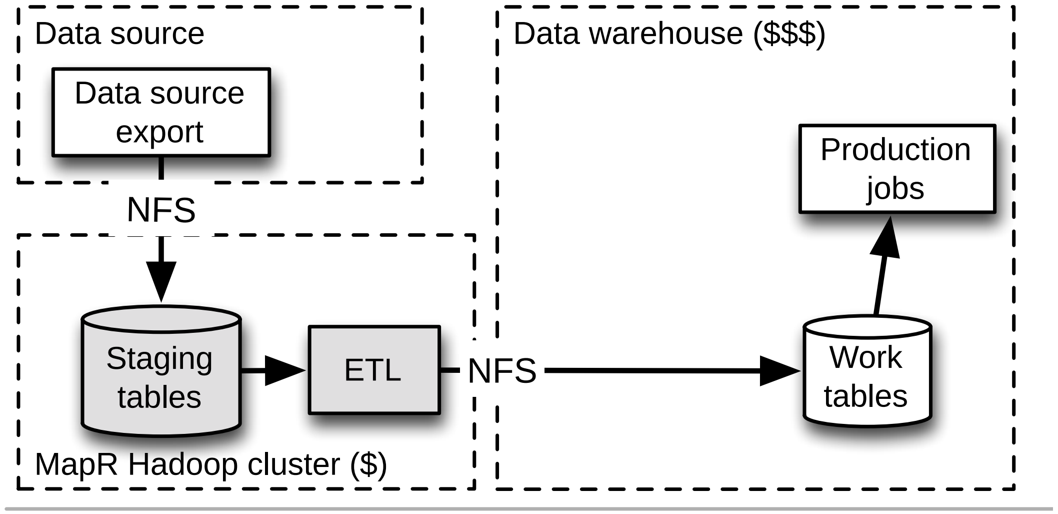
- Start conservatively and plan to expand

# Good 1st use case: Data Warehouse Optimization

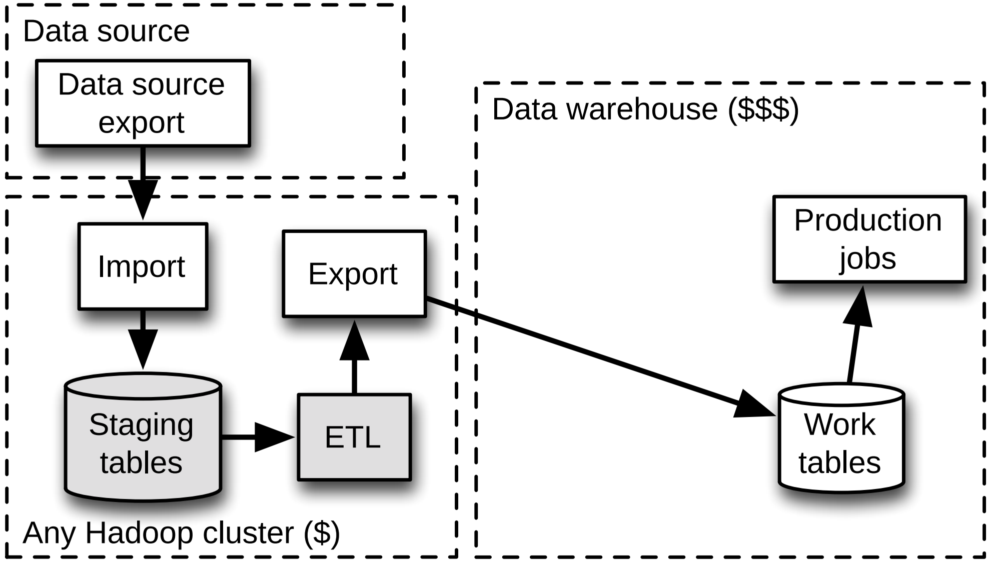


# Good 1st use case: Data Warehouse Optimization

B. MapR distribution for Hadoop



C. HDFS-based distribution for Hadoop



# Benefits of DW Optimization

- Reduce strain on DW and save money
- Keep using traditional systems for what they do best
- Additional benefit: Option for further explore the original data
  - Feasible to have saved it thanks to the cost-effective nature of Hadoop

If you’re experienced Hadoop user…

# Plan across entire organization

- Expand cluster as you identify new use cases of interest

- Build a centralized data hub:
  - break silos, provide access to same data by multiple groups

- Propagate knowledge of Hadoop & NoSQL to other groups

- Continue to give your teams time to explore & experiment

- Plan co-existence of traditional, legacy & new applications (MapR makes this easier to do)

# Additional tips

- Be realistic about SLA’s (example: some projects need 24/7 availability or very fast response times)

- Be flexible: Shake off old assumptions and look for opportunities to build new insights

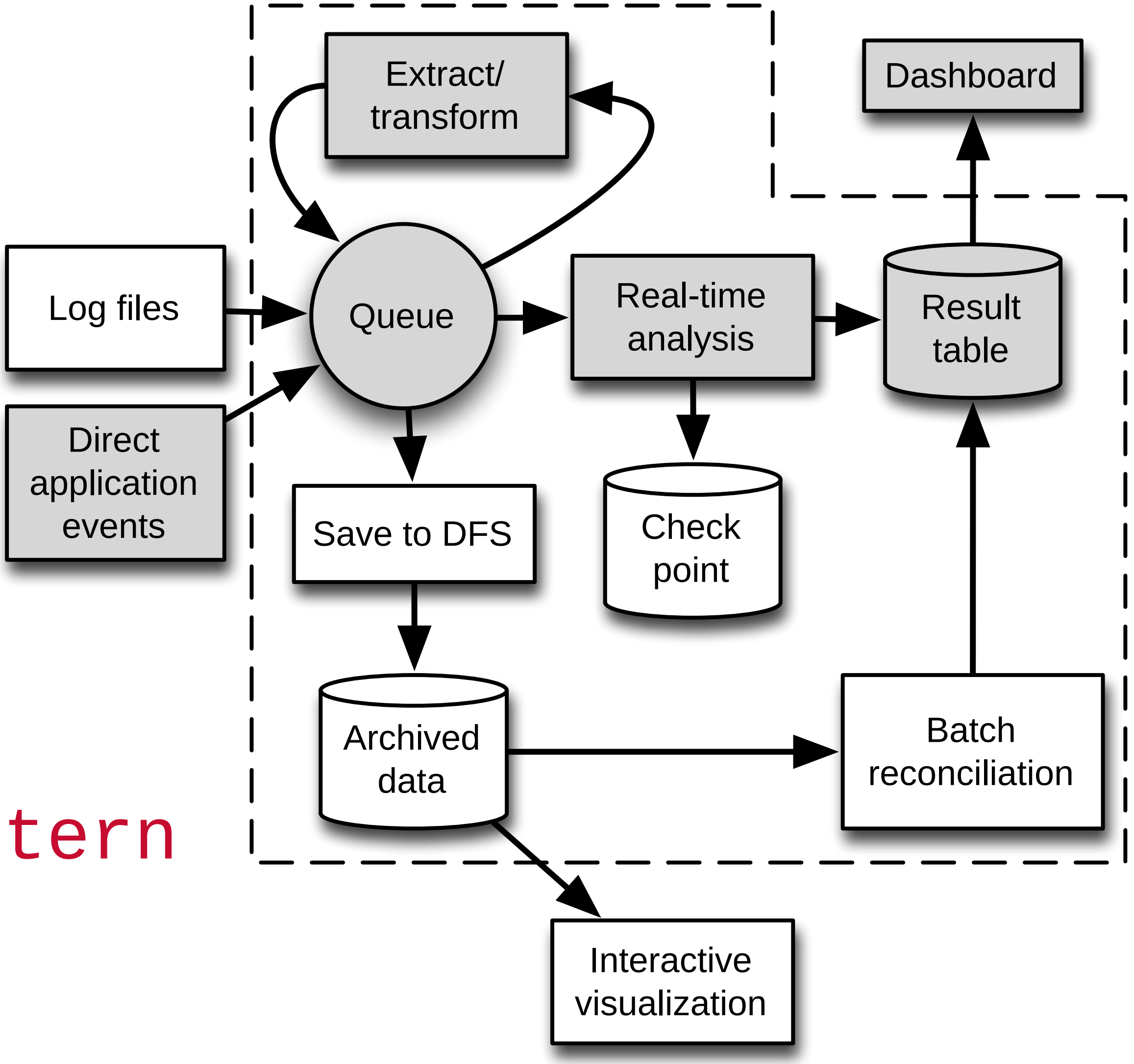
# Another use case...

# Streaming Log Analysis: Business Goals

- Customer may be trying to track down a security breach
- Customer may be interested in identifying anomalous behaviors or other patterns clickstream data from user interactions on a website
- Customer may want to supply data to a real-time dashboard

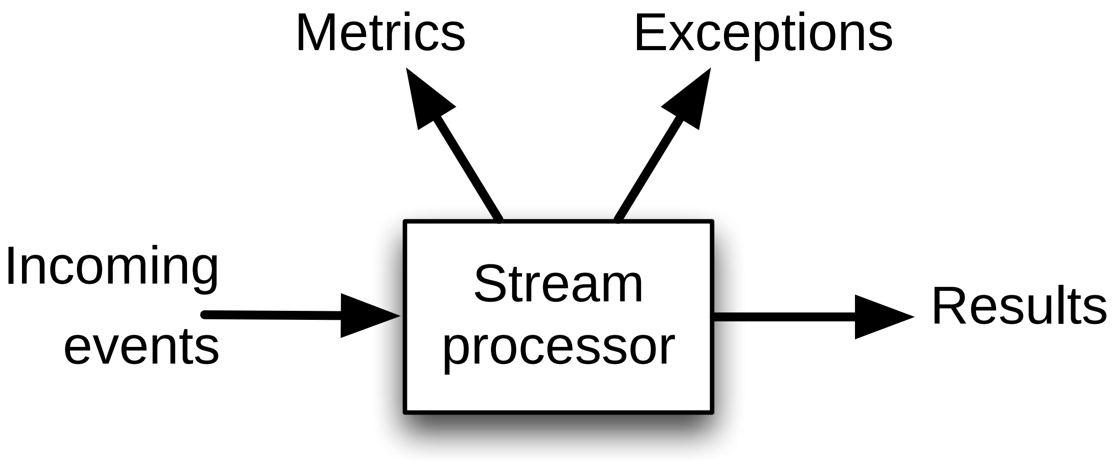
# Persistent queuing is key

# Streaming log analysis

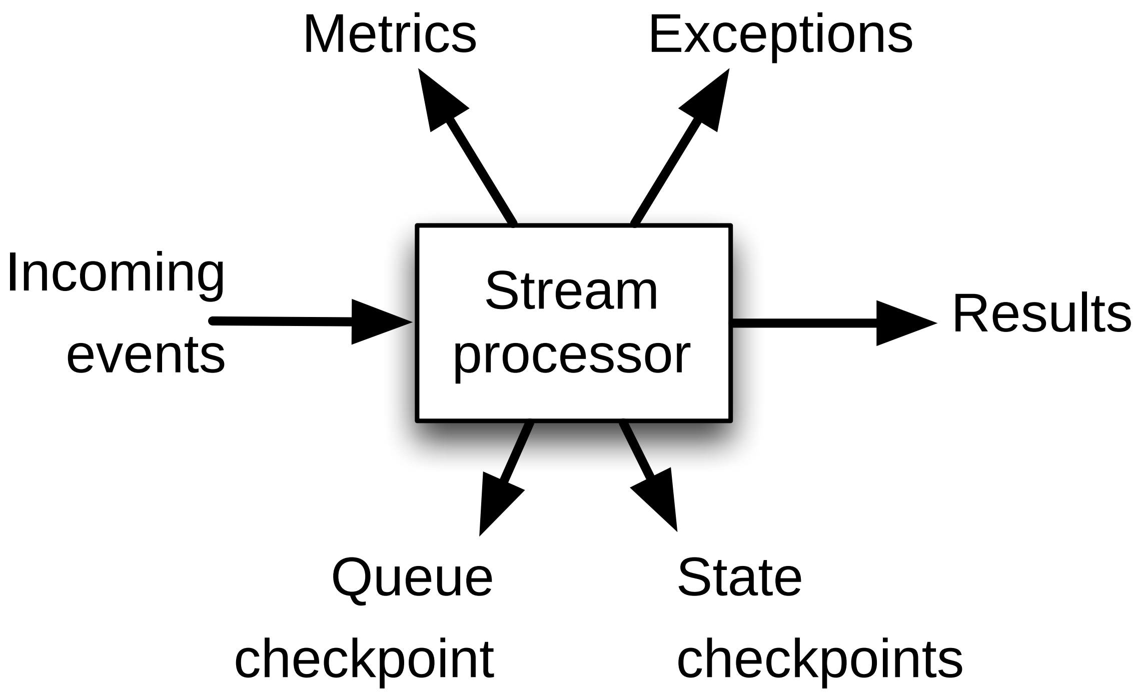


Persistent queuing is
Key architectural pattern

# Universal Architectural Pattern



# Stateful Reliable Processing



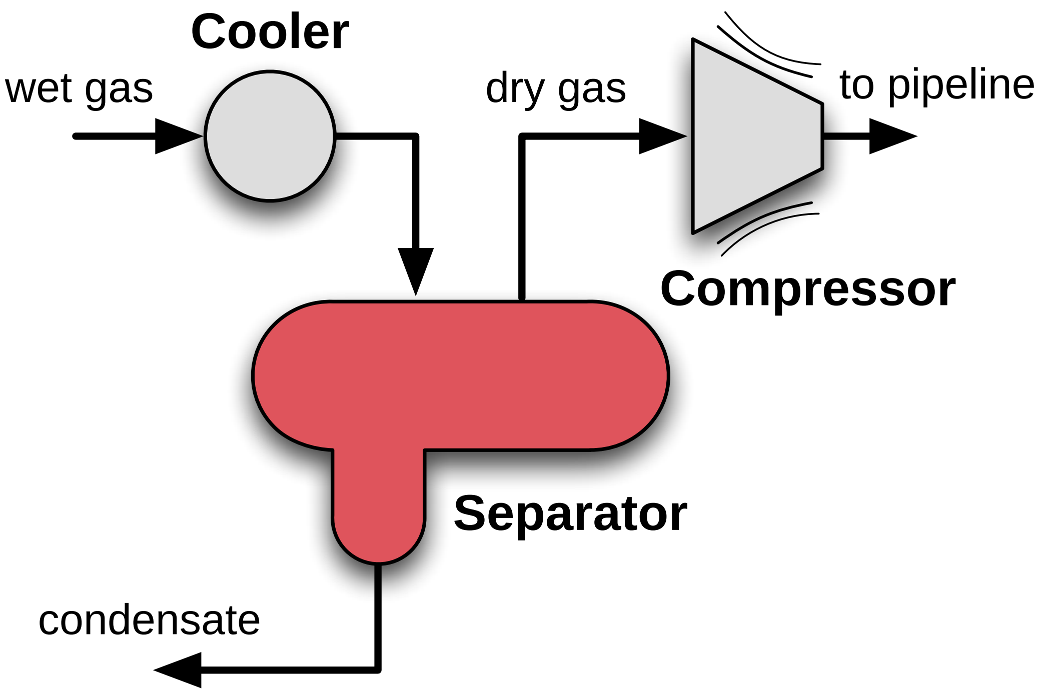
# Keys to Queue Architectures

- Standardize on record formats
  - More than one may be needed
  - Parquet (sadly) doesn’t like record by record
  - Simple Binary Encoding has very fast record codecs
  - Low latency and mechanical sympathy communities are good resources

- Standardize on component shapes
  - Goes-ins and goes-outs first
  - Metrics and exception channels are required
  - Checkpoint to files, push checkpoint record to queue

# Another use case...

# Predictive Maintenance



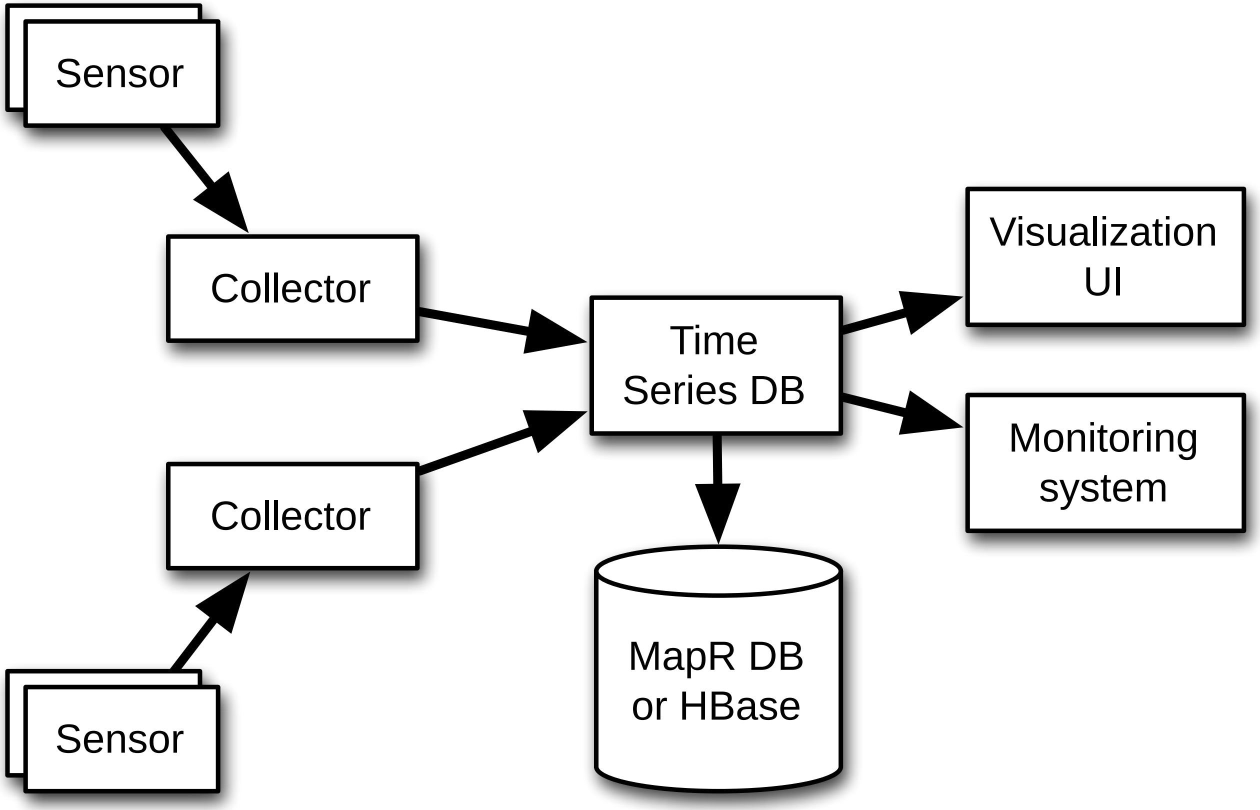
Degradation Detected
PREDICTIVE ALERT

Images courtesy MTell

# Time Series Data: Predictive Maintenance

- Streaming sensor data for variety of measurements made at multiple times
- Keep a long term maintenance history (part #, location, when serviced; when failed)
- Use machine learning techniques to identify indicators of a potential near-term failure and send alert

# Time Series Data from Sensors



# Time Series Notes

- Sustained load is what people worry about
  - Look for secondary loading effects like compactions
  - Consider pre-compaction in memory

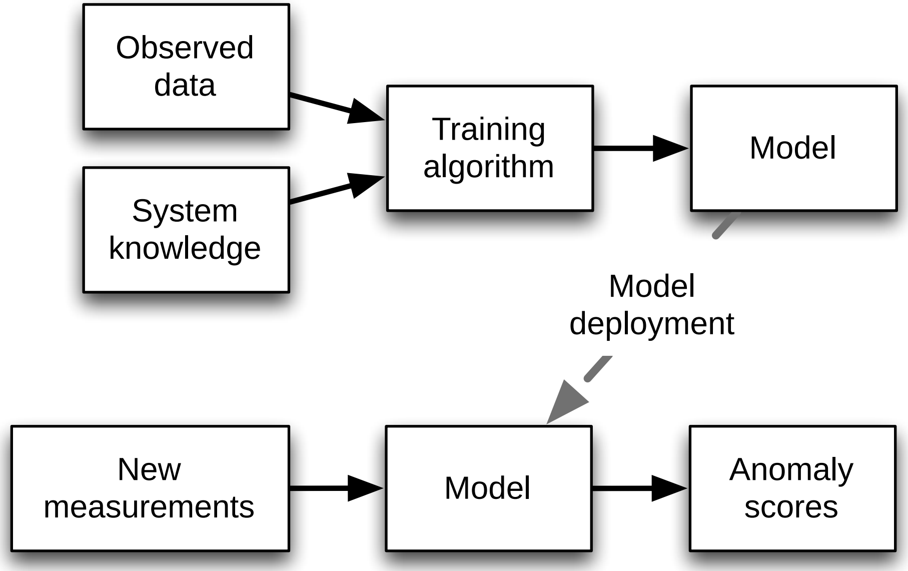
- Backfill is actually the hardest part technically
  - (1000x higher data rate)
  - See our time series book for 200 M points / sec

# Another use case...

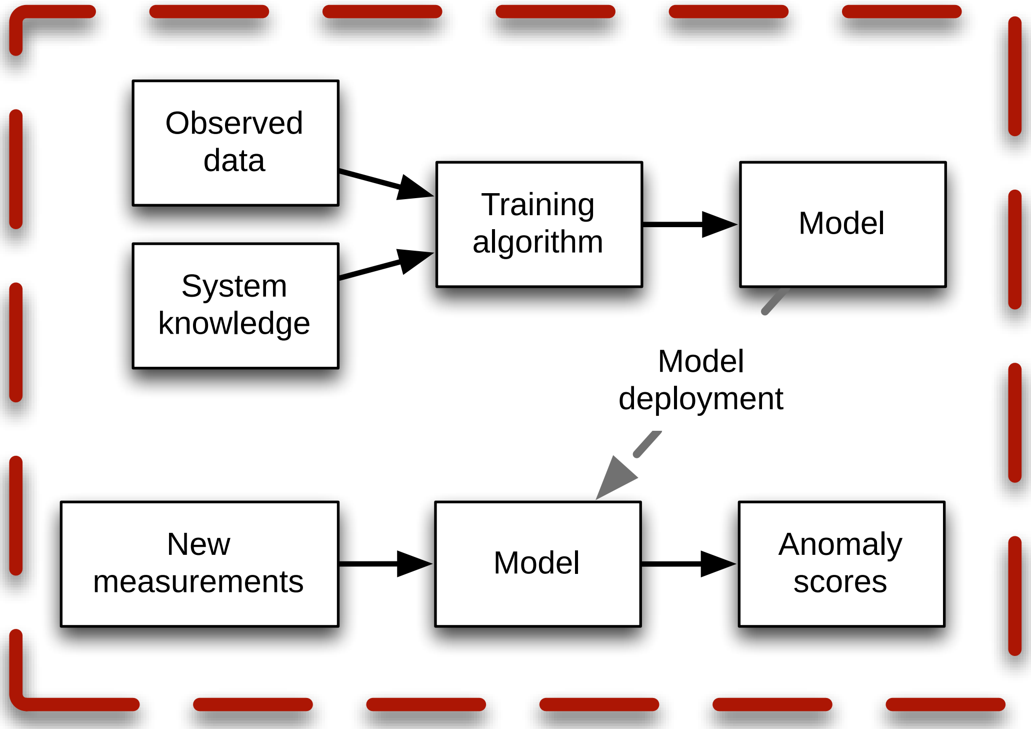
# Anomaly Detection and Fraud Analytics

- Customer wants to identify zero-day attacks
- And advanced persistent threats
- By sophisticated adversaries who don’t use known vectors
- Must keep logs and other data secret
  - But must also collaborate on detection algorithms

# Secure Development is Hard



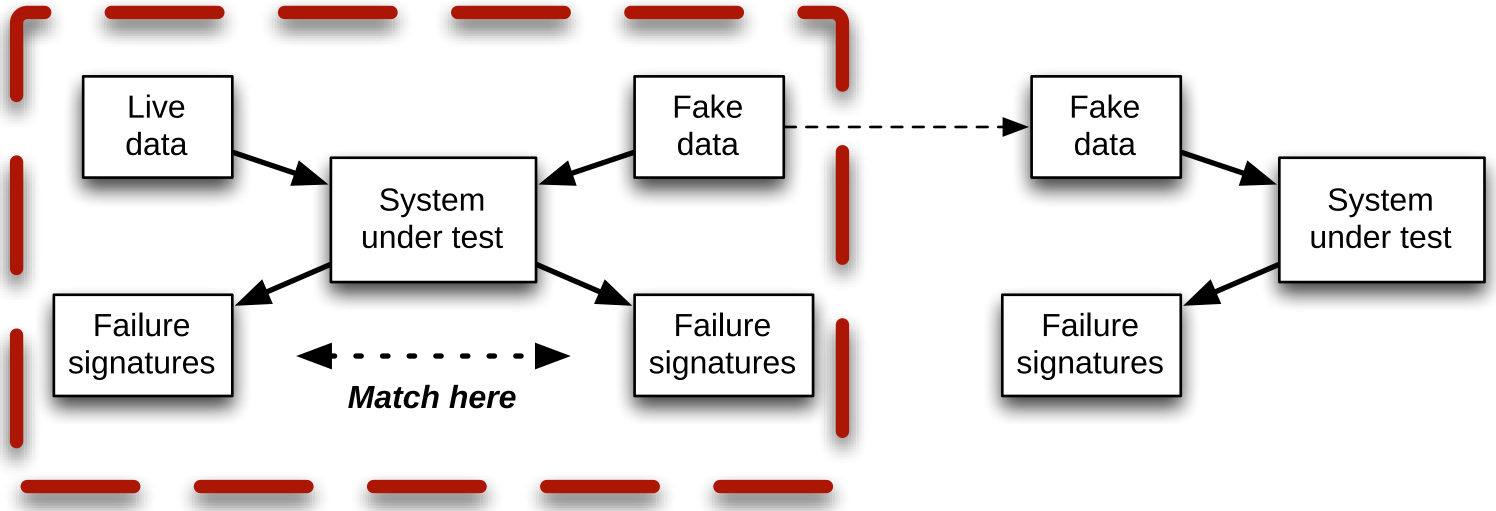
# Secure Development is Hard



Outside collaborators are outside the security perimeter

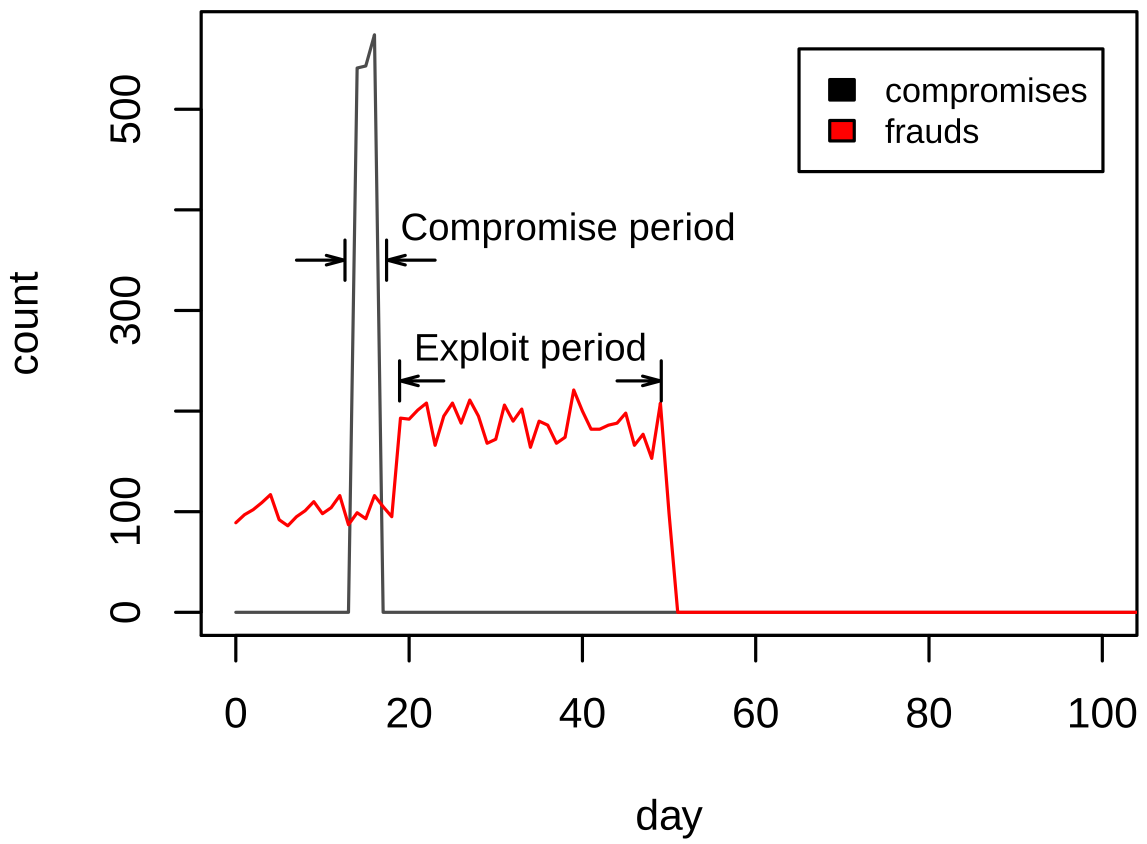
# Parametric Simulation

Parametric matching of failure signatures allows emulation of complex data properties

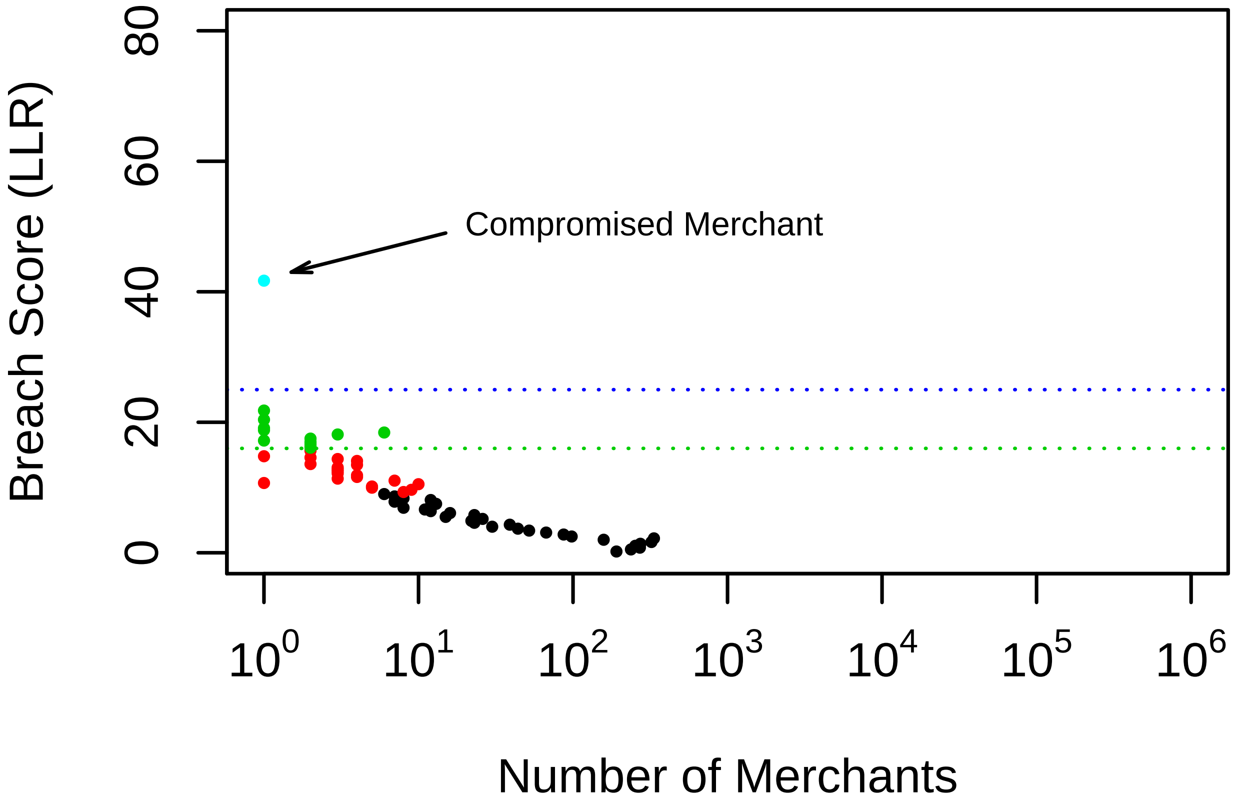


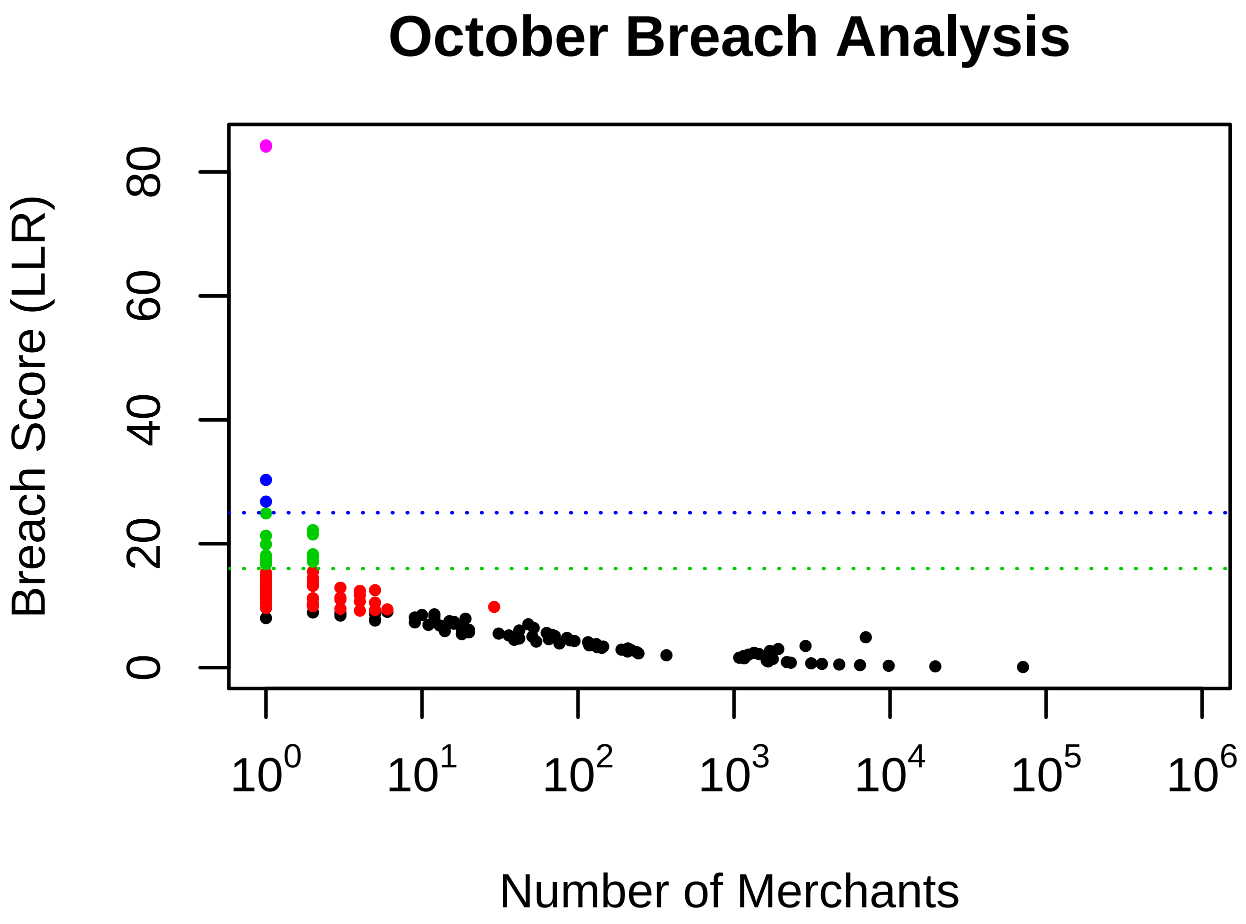
Matching on KPI’s and failure modes guarantees ***practical*** fidelity

# Simulation Setup



# LLR score for simulated merchants




October Breach Analysis
Breach Score (LLR)
0
20
40
60
80
Number of Merchants
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
$10^6$

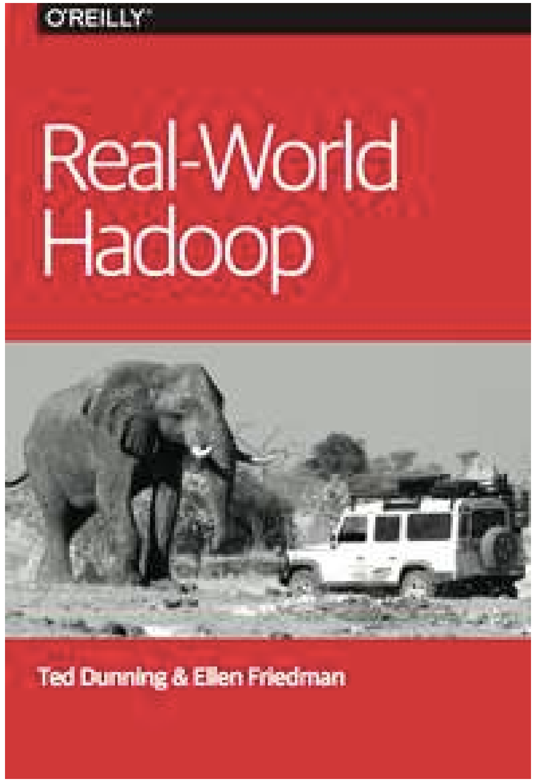
Ask me about Myriad

# Ask me about Myriad

Ask me about Myriad
and about zeta (ζ)

# *Real World Hadoop*

by Ted Dunning and Ellen Friedman © Feb 2015 (published by O'Reilly)

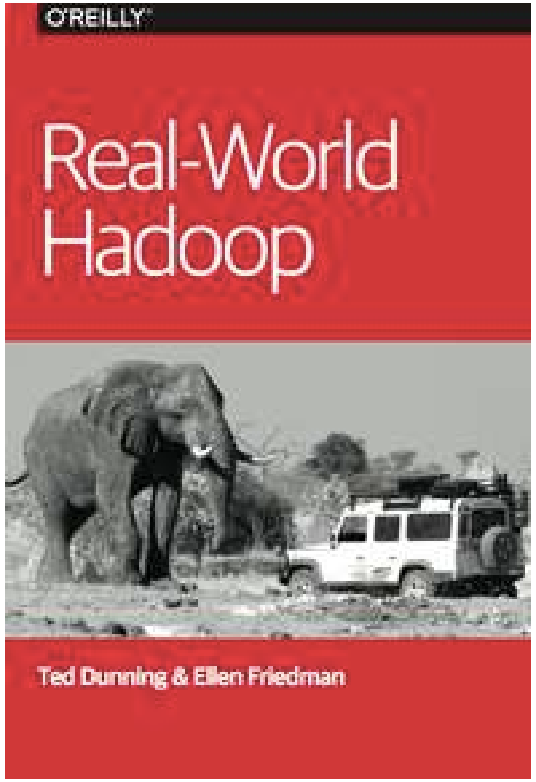


eBook courtesy of MapR:

http://bit.ly/mapr-real-world-hadoop

# *Real World Hadoop*

by Ted Dunning and Ellen Friedman © Feb 2015 (published by O'Reilly)



Free print copy during book signings at MapR booth

| | |
|---|---|
| Thur | 5:30 pm |
| Fri | 10:10 am |

# Related events at Strata this week:

Office Hour  Ellen Friedman  Thur 19 Feb 2015 at 11:30 am

Plus news of Myriad: new OSS collaboration for global resource management:

“YARN vs. Mesos: Can’t We All Just Get Along” Technical talk by Ted Dunning
Fri 20 Feb 2015 at 2:20pm
http://bit.ly/strata2015-myriad

# Thank You!

# Q&A

## Engage with us!

| | | | |
|---|---|---|---|
| @mapr | t | f | maprtech |
| mapr-technologies | in | g+ | MapR |
| tdunning@mapr.tech.com | ✉ | You Tube | maprtech |