O'REILLY®
Strata
CONFERENCE
Making Data Work

11–13 Nov 2013
LONDON, ENGLAND

#strataconf
strataconf.com/london

# Where Polyglot Persistence meets the Lambda Architecture


Michael Hausenblas, Chief Data Engineer EMEA
**MapR Technologies**

# Polyglot Persistence

```
$ tail -f some.log
```

```
$ ls -al
```

```
$ nc localhost 80
```

```
awk 'BEGIN { FS = "," }
     /2013-[[:digit:]]+-[[:digit:]]+/ { print $3 }'
     sample.csv
```

tool box


one-size-fits-all

# Polyglot Persistence—Backdrop

- Michael Stonebraker and Ugur Çetintemel—2005

  **"One Size Fits All": An Idea Whose Time Has Come and Gone**

- Martin Fowler—2011

  **Polyglot Persistence[1]**

- Eric Brewer—2012

  **Ricon Keynote—Advancing Distributed Systems[2]**

1) http://martinfowler.com/bliki/PolyglotPersistence.html

2) https://speakerdeck.com/eric_brewer/ricon-2012-keynote

# Polyglot Persistence—Key Points

- Use different datastores for different needs
- Can apply within an application or cross-enterprise
- Encapsulating data access yields loosely coupled components
- Find sweet spot between dev/op complexity and flexibility

# Lambda Architecture

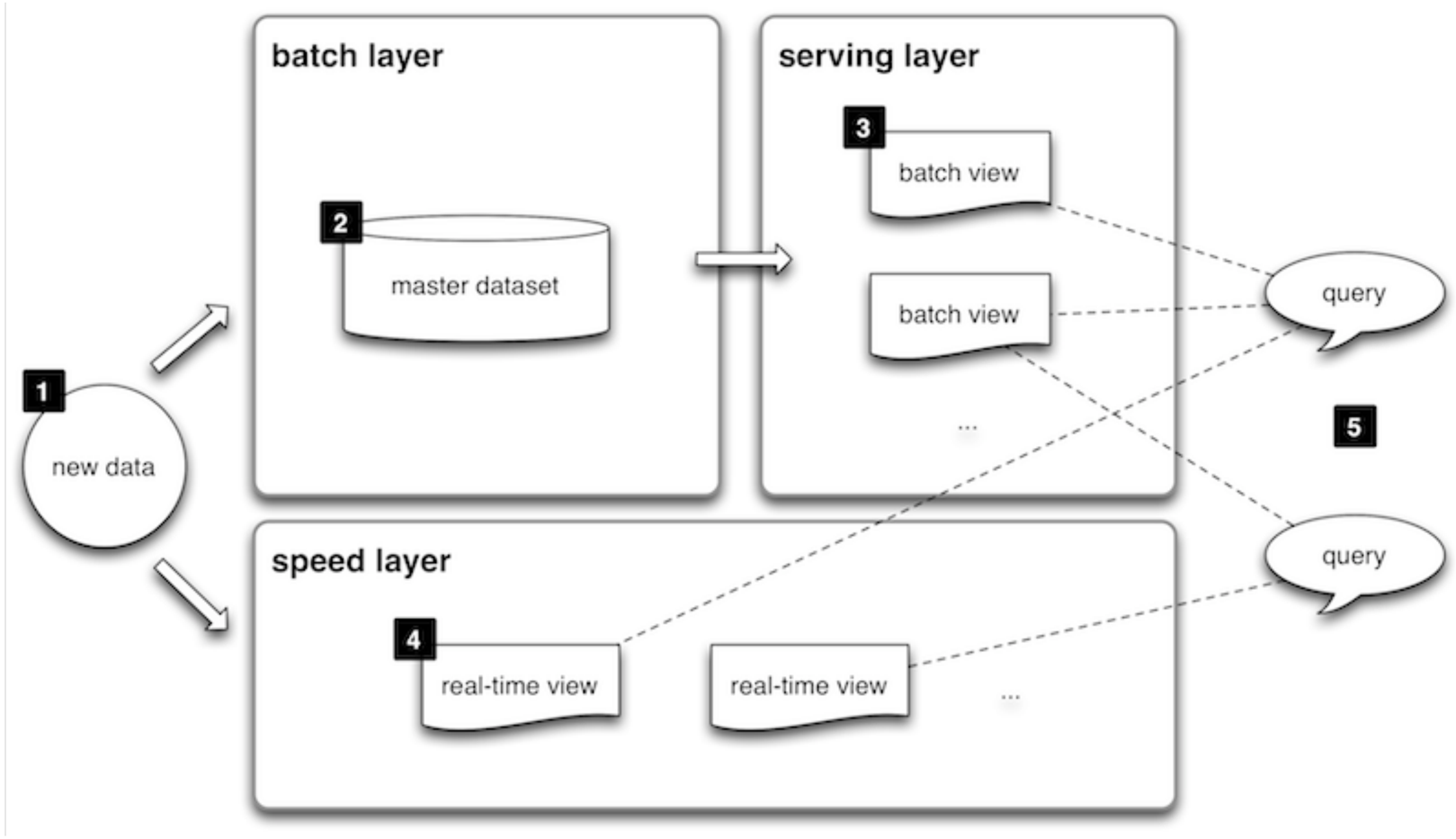
# Lambda Architecture—Backdrop

- Nathan Marz (Backtype, Twitter)

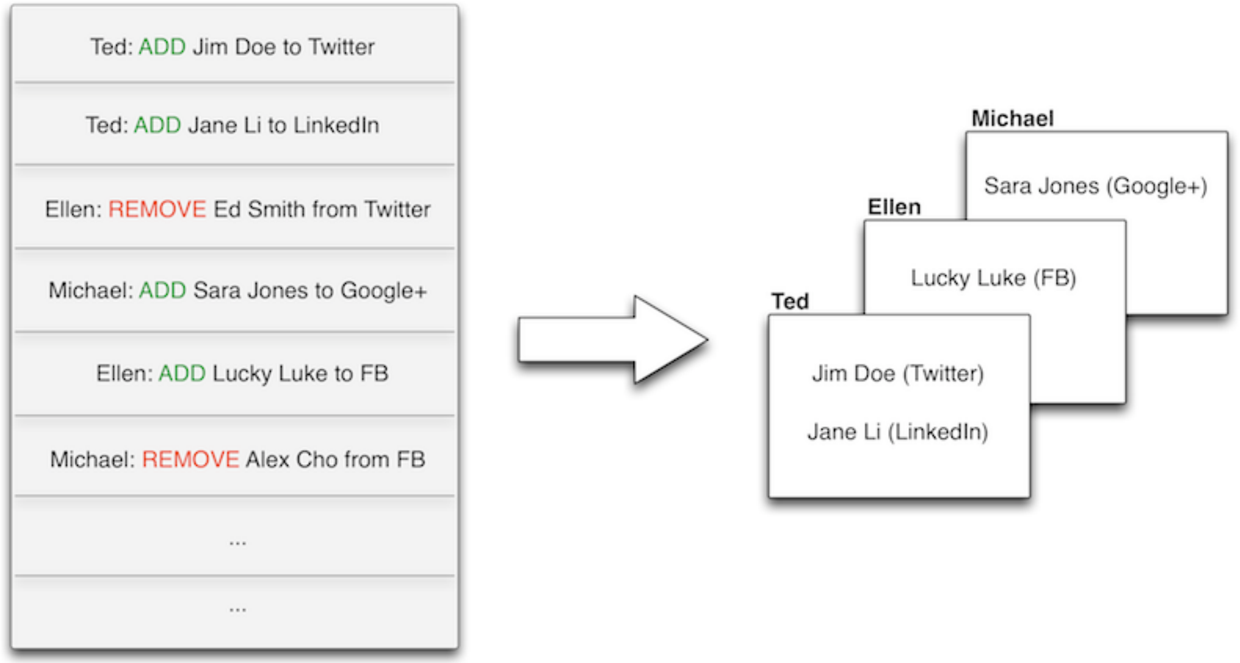
- Creator of …
  - Storm
  - Cascalog
  - ElephantDB


http://manning.com/marz/

# Lambda Architecture—Overview



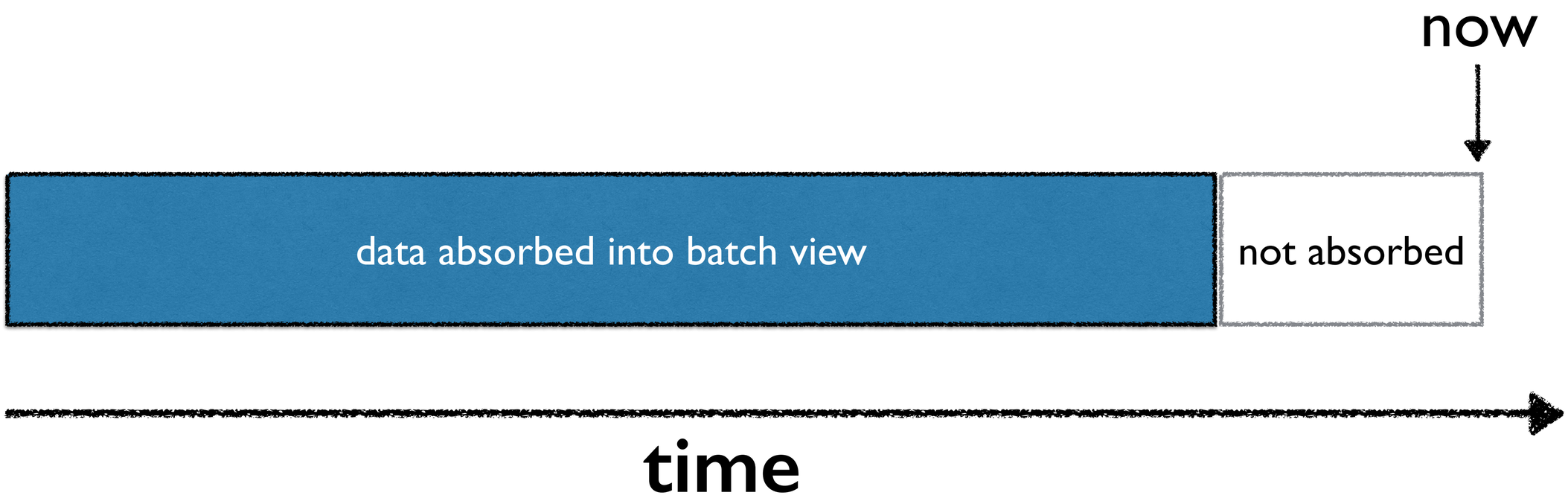
# Lambda Architecture—Immutable Data + Views



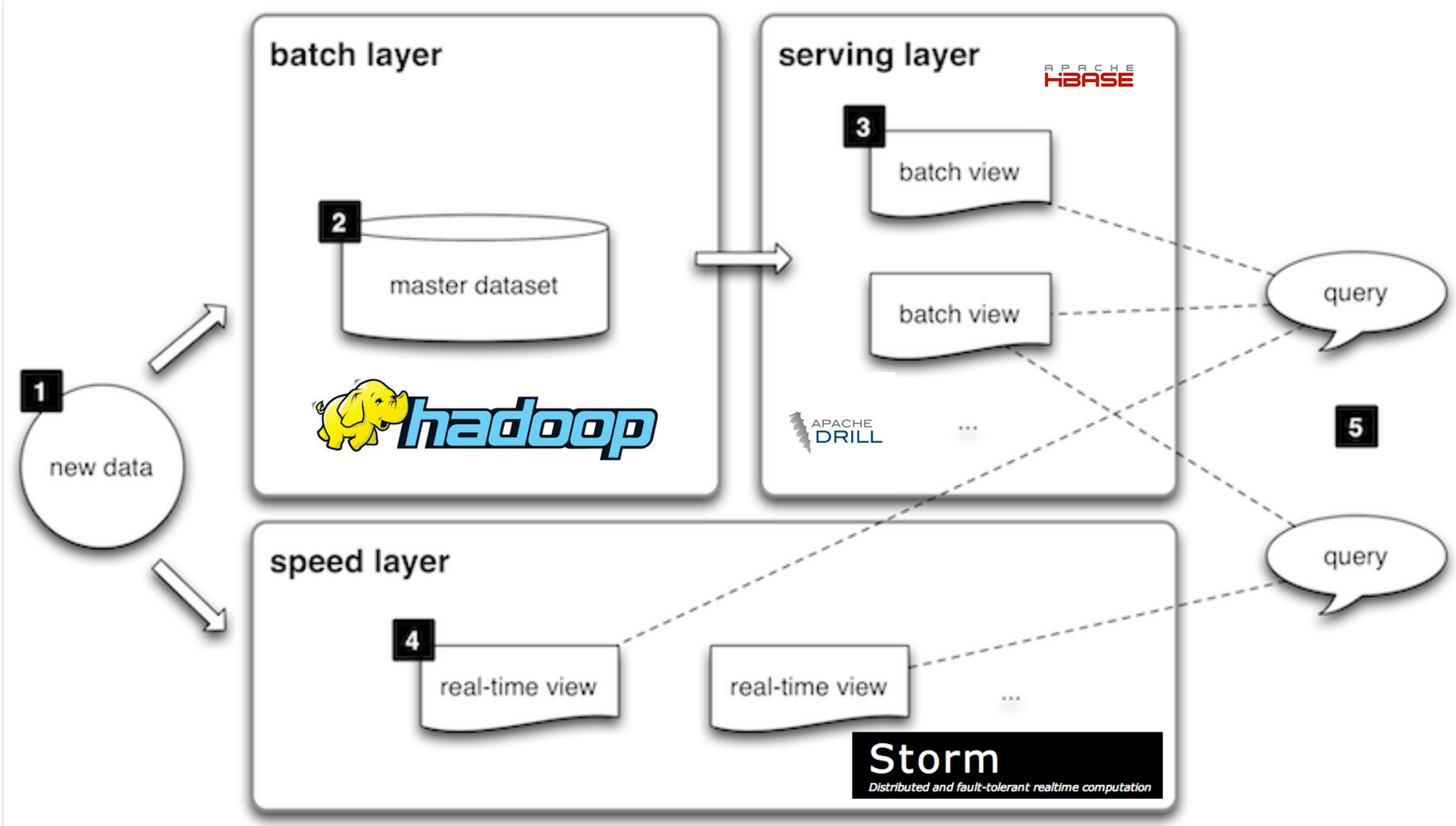

master dataset

batch view

# Lambda Architecture—Compensate Batch

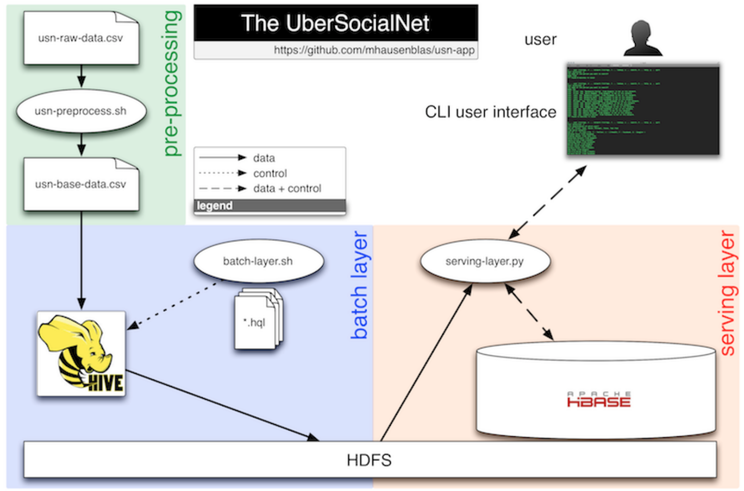


# Lambda Architecture—Technologies



# Lambda Architecture—Resources

- http://www.slideshare.net/nathanmarz/runaway-complexity-in-big-data-and-a-plan-to-stop-it
- http://www.slideshare.net/nathan_gs/a-real-time-architecture-using-hadoop-and-storm
- http://www.drdobbs.com/database/applying-the-big-data-lambda-architectur/240162604



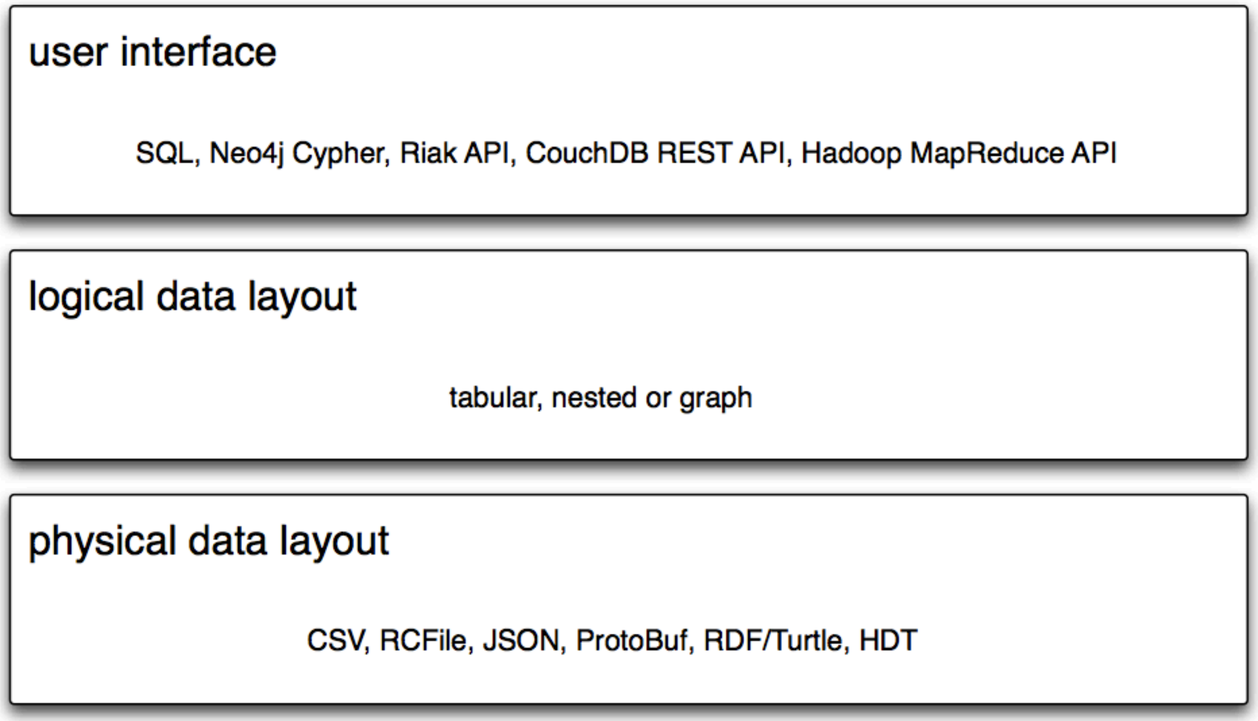

# How are Polyglot Persistence and Lambda Architecture related?

# Polyglot Persistence & Lambda Architecture

MAPR TECHNOLOGIES

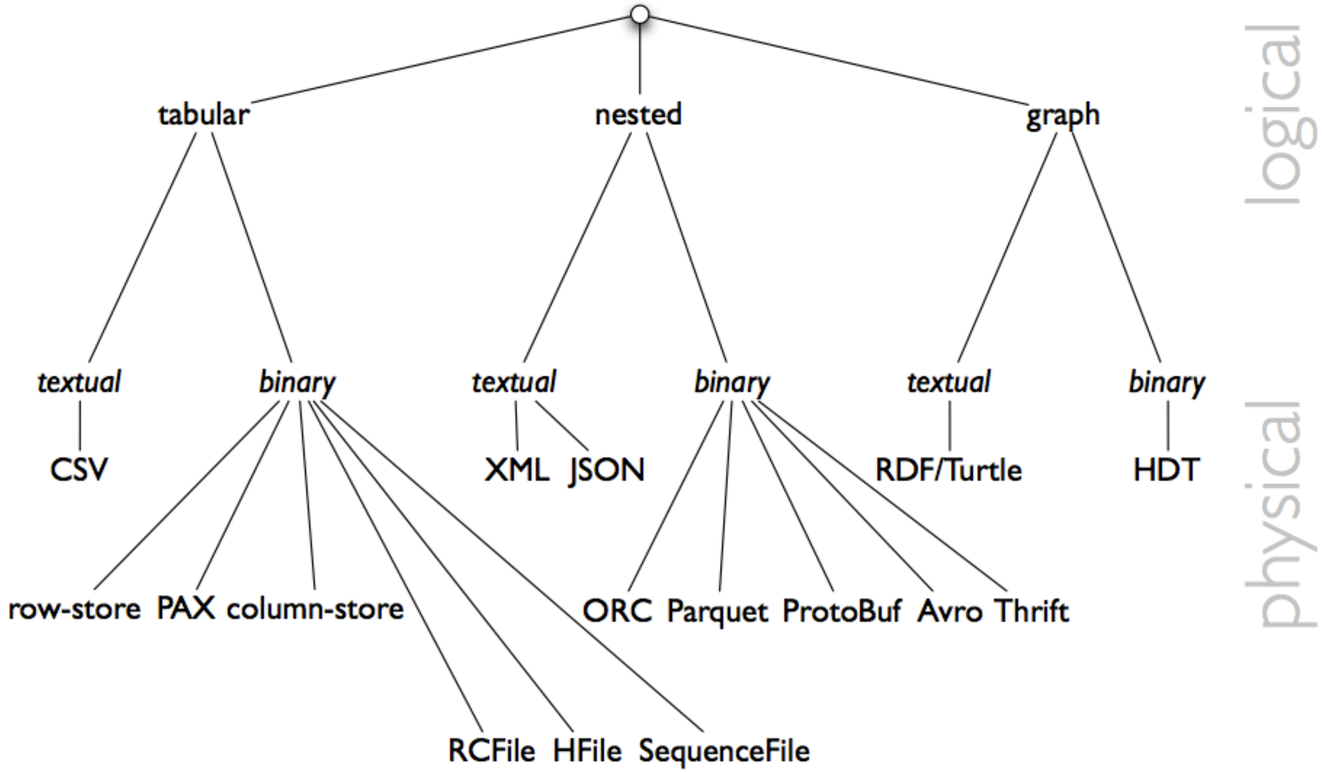
- Both are conceptually ‘old’
- Strongly related
- LA implies PP but not necessarily the other way round
- **Physical and logical data layouts …**

# Levels of representation and interaction




http://arxiv.org/abs/1305.6506

# Taxonomy for logical & physical data layouts

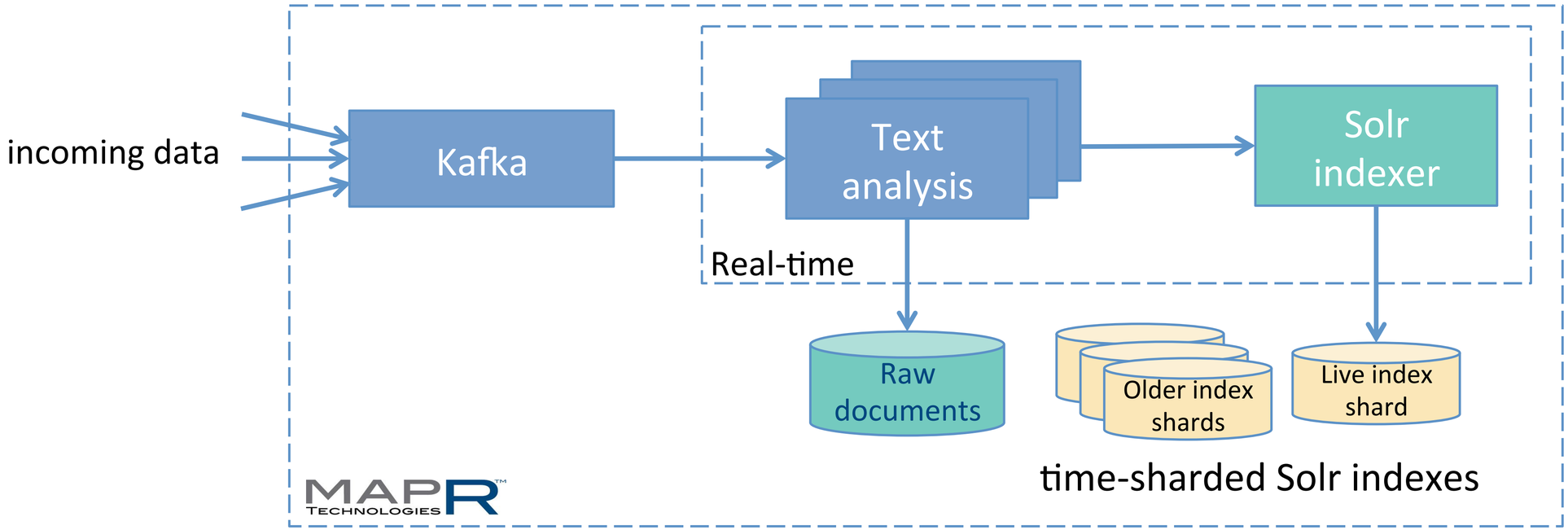



http://arxiv.org/abs/1305.6506

# Polyglot Persistence and Lambda Architecture in the Wild

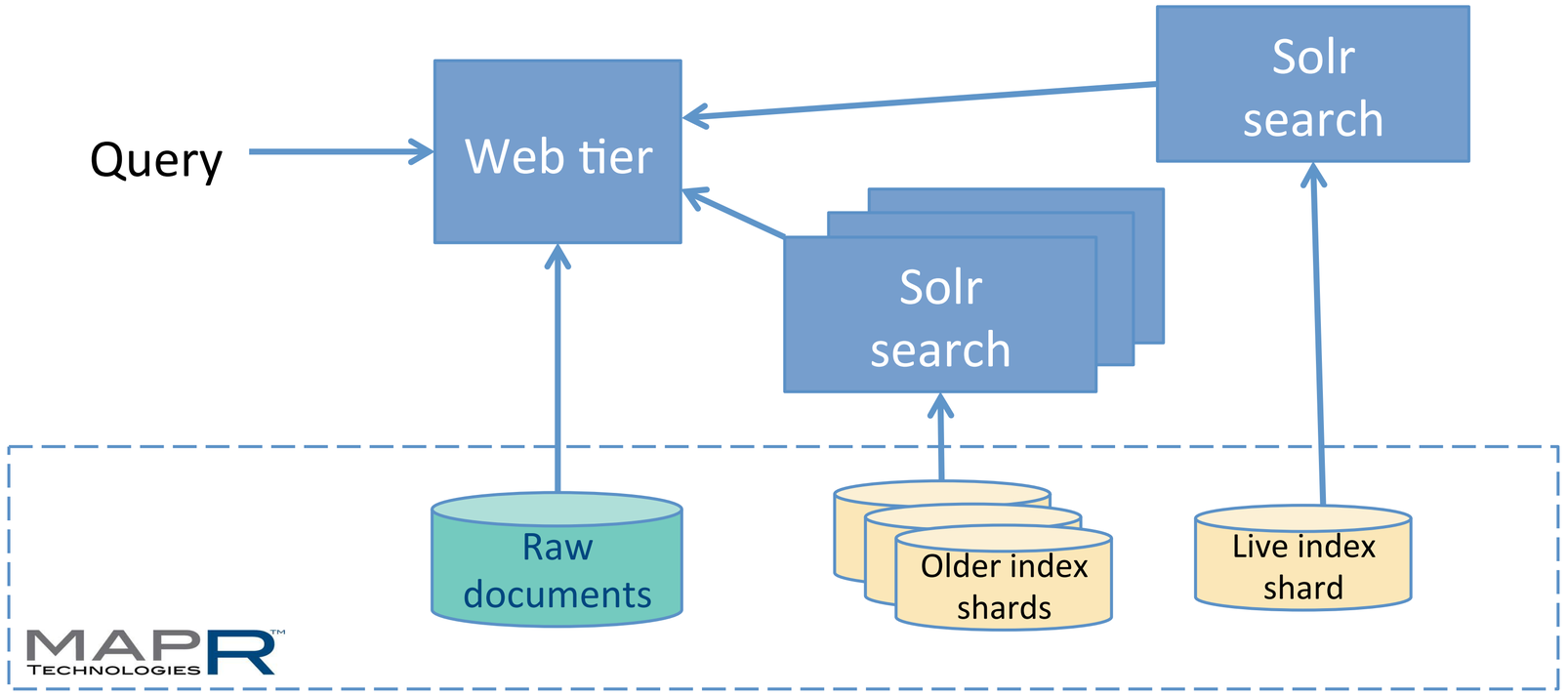
# Case Study:
# Log Analysis

# Log Analysis—Architecture




**Data Ingestion and Indexing**

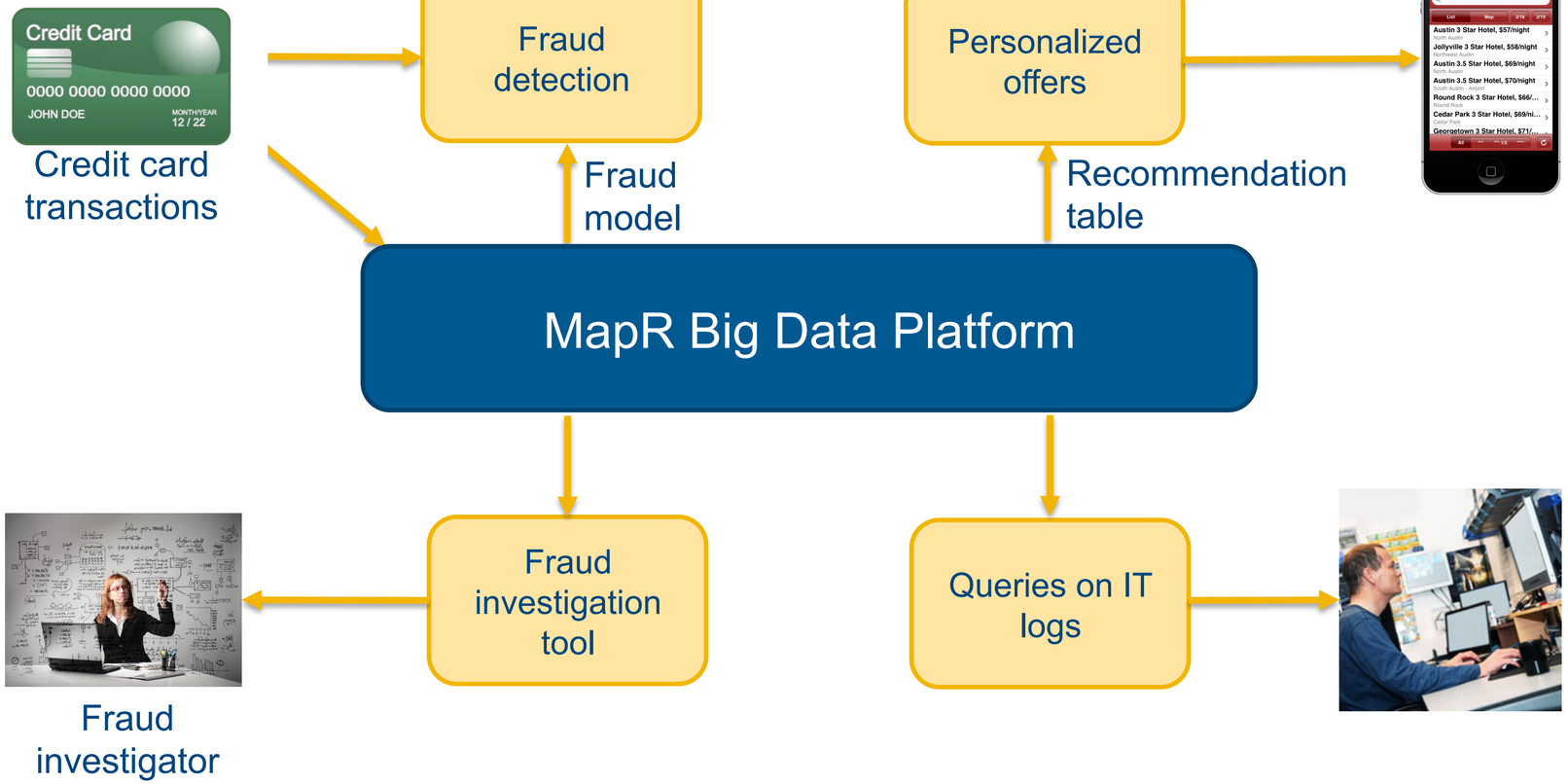
# Log Analysis—Architecture




**Search**

# Case Study:
# Credit Card Company

# Credit Card Company—Architecture



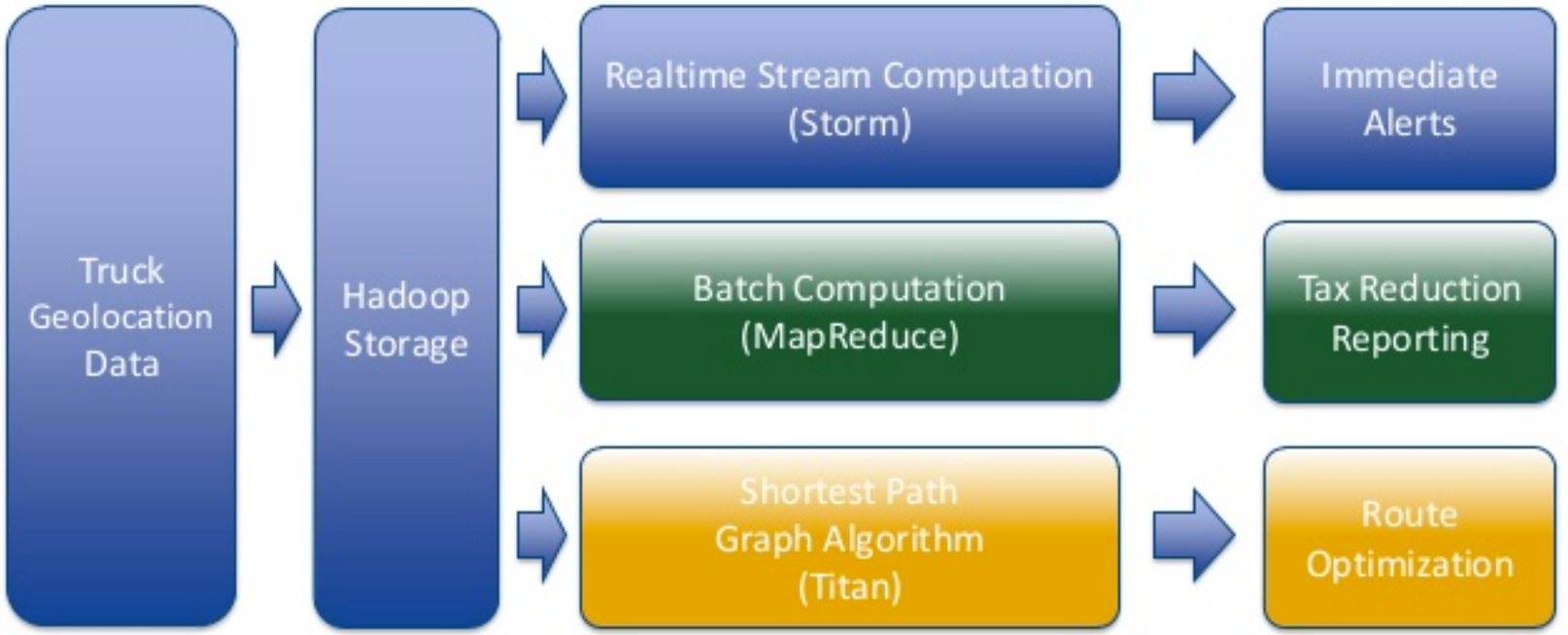
# Case Study:
# Waste & Recycling Leader

# Waste & Recycling Leader

- Data
  - geolocation of 20,000 trucks
  - arriving every 5sec
  - geographic boundaries of landfills

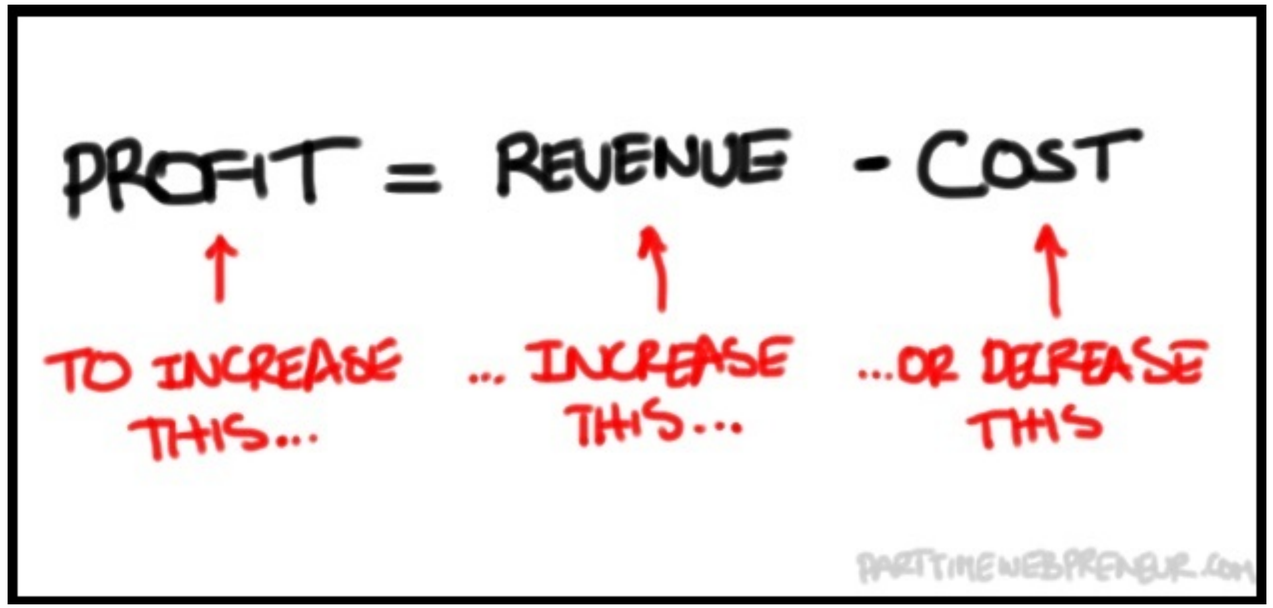
- Goal
  - online alerts
  - tax reduction reporting
  - route optimisation

# Waste & Recycling Leader—Architecture



$$$

# What about the Business Value?



# Return of Investment

- Storage economics ($$$/TB)
- Agile Development (dev/ops)
- Leverage existing knowledge and tools (SQL, anyone?)
- Human fault-tolerance (at scale)

MAPR TECHNOLOGIES

use cases

log file analysis
fraud detection
ETL off-load

customer insights
forensics
drug discovery

supply chain management
logistics
360 social media

**Big Data platform for Hadoop**

workloads

*file-based applications* | *batch processing* | *OLTP* | *interactive query (SQL)* | *stream processing* | *search*

processing

Direct Access NFS™

***MapReduce***
Apache Hive
Apache Pig
Cascading

***Machine Learning***
Apache Mahout
Skytree

Apache HBase
GraphDB Titan

Apache Drill
Impala

Apache Storm

Solr
ElasticSearch

configuration, monitoring
MCS

HA, DR, multi-tenancy

security (PAM/Kerberos)

storage

**MapR Distributed File System**
(structured, semi-structured and unstructured data—POSIX compliant)

nodes

For example:
64GB RAM, 12 cores
10GbE
12x3TB SATA HDD

on-premise and/or cloud

amazon web services | Google | openstack

# Let's discuss!

- Twitter:

  @mhausenblas
  @MapR_EMEA
  @MapR

- Come to my office hour
  Mo 11 Nov 12:30, Table 3



**We're hiring!**