Harnessing the Internet of Things with NoSQL

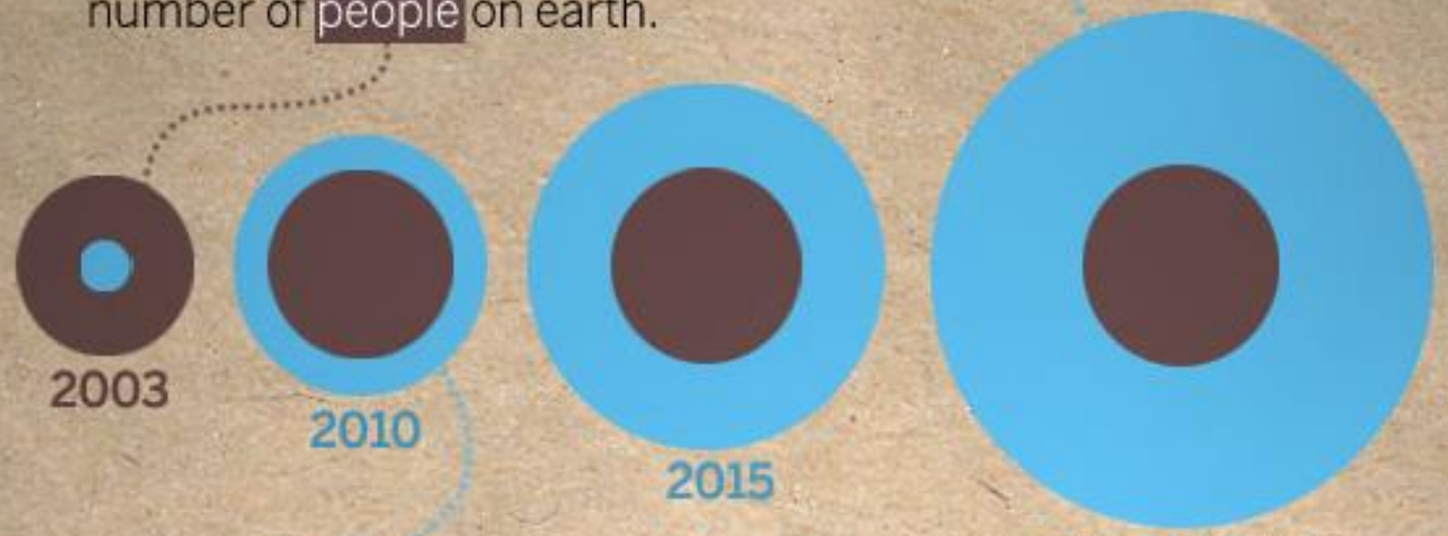
NoSQL matters, 2013-11-30, Barcelona, Spain

Michael Hausenblas
Chief Data Engineer, MapR Technologies

The INTERNET of THINGS

We already have cameras and computers that are one cubic millimeter. You could fit 150 of them in this icon.

During 2008, the number of things connected to the Internet exceeded the number of people on earth.

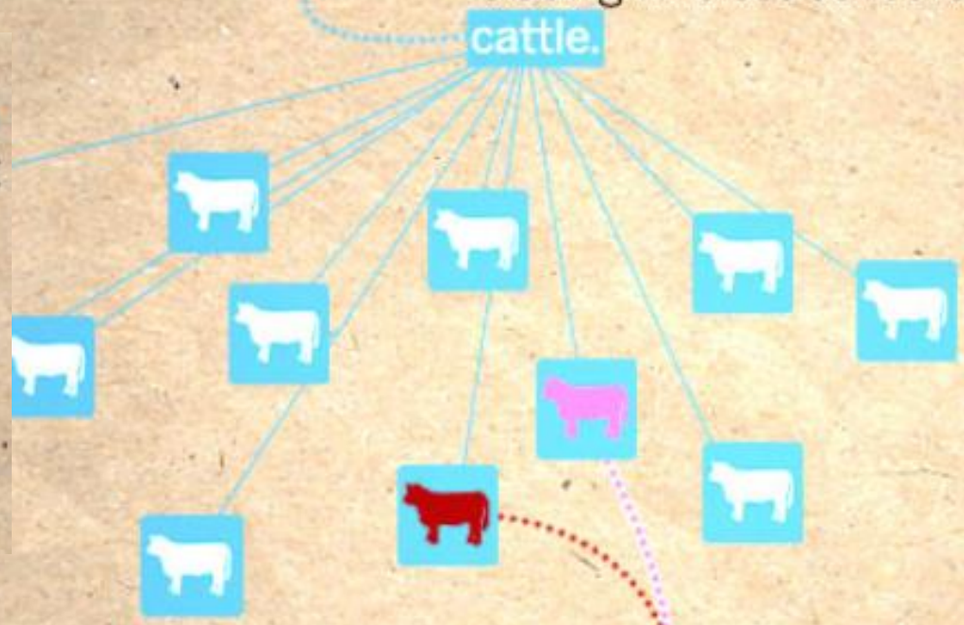


By 2020 there will be 50 billion.

These things are not just smartphones and tablets.

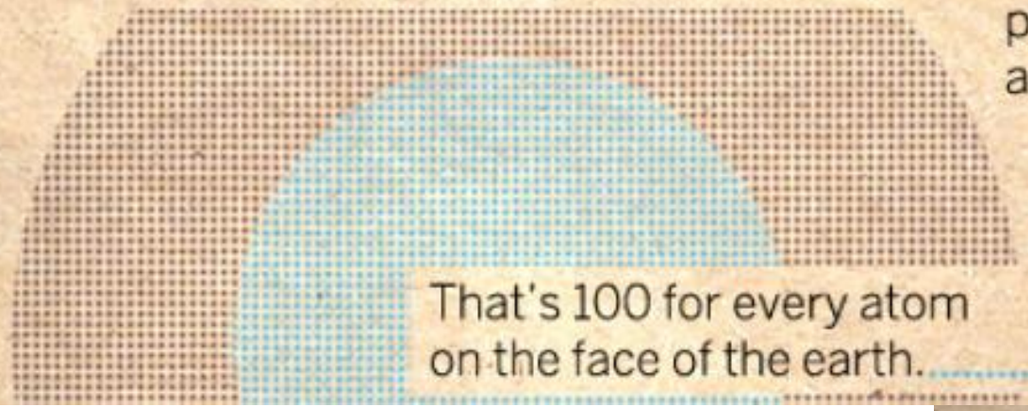
They're everything.

A Dutch startup, **Sparked**, is using wireless sensors on cattle.



So that when one is sick or pregnant, it sends a message to the farmer.
Each cow transmits 200 mb of data per year.

With the IPv6 protocol, we will have
340,282,366,920,938,463,463,374,607,431,768,211,456
possible Internet addresses.



That's 100 for every atom on the face of the earth.

http://blogs.cisco.com/news/the-internet-of-things-infographic/

By 2020 there will be …

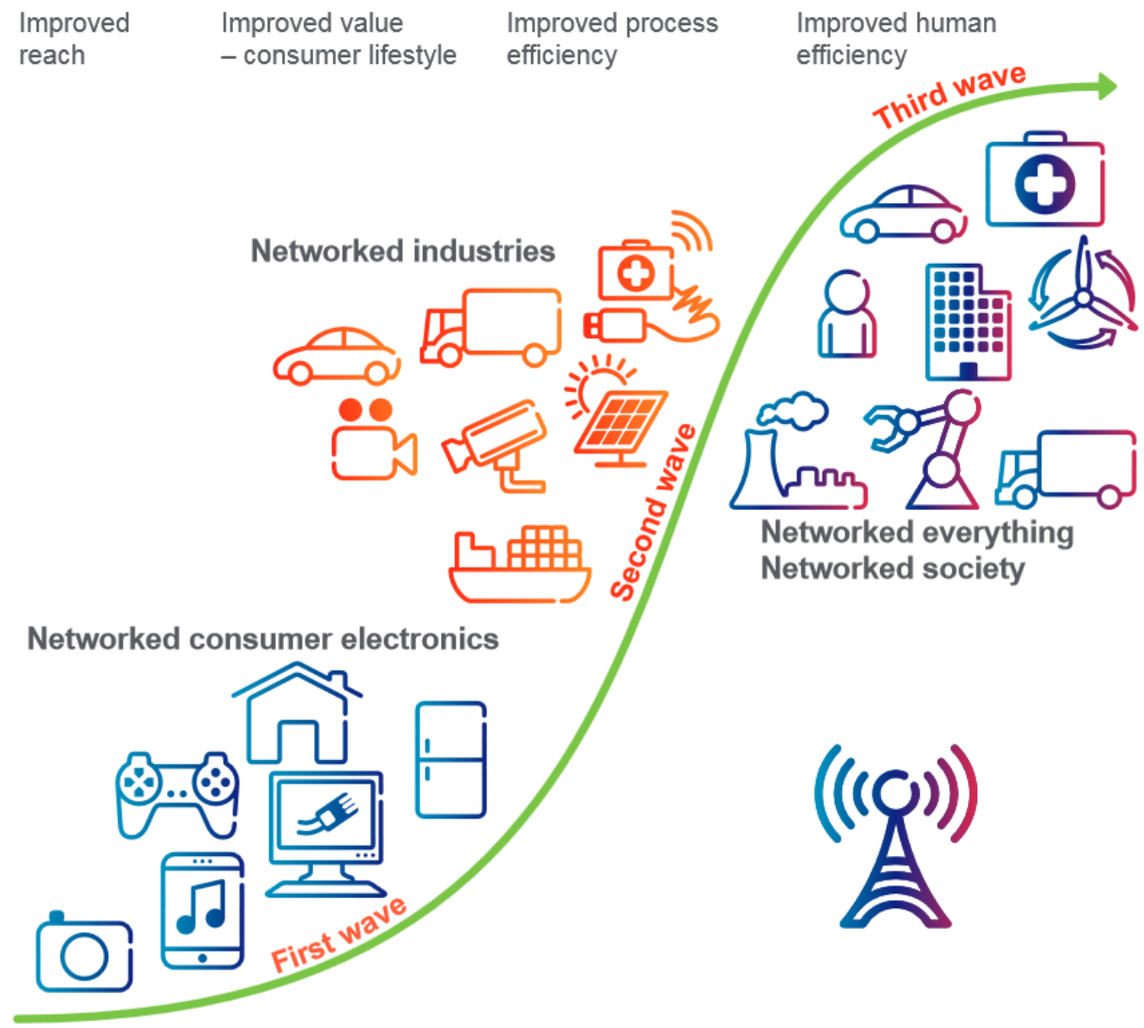
- 3 billion subscribers with sufficient means to buy information on a 24/7 basis

- In mature markets, these customers will typically possess between 5-10 connected devices each

- 1.5 billion vehicles globally, not counting trams and railways

- 3 billion utility meters, like electricity, water and gas

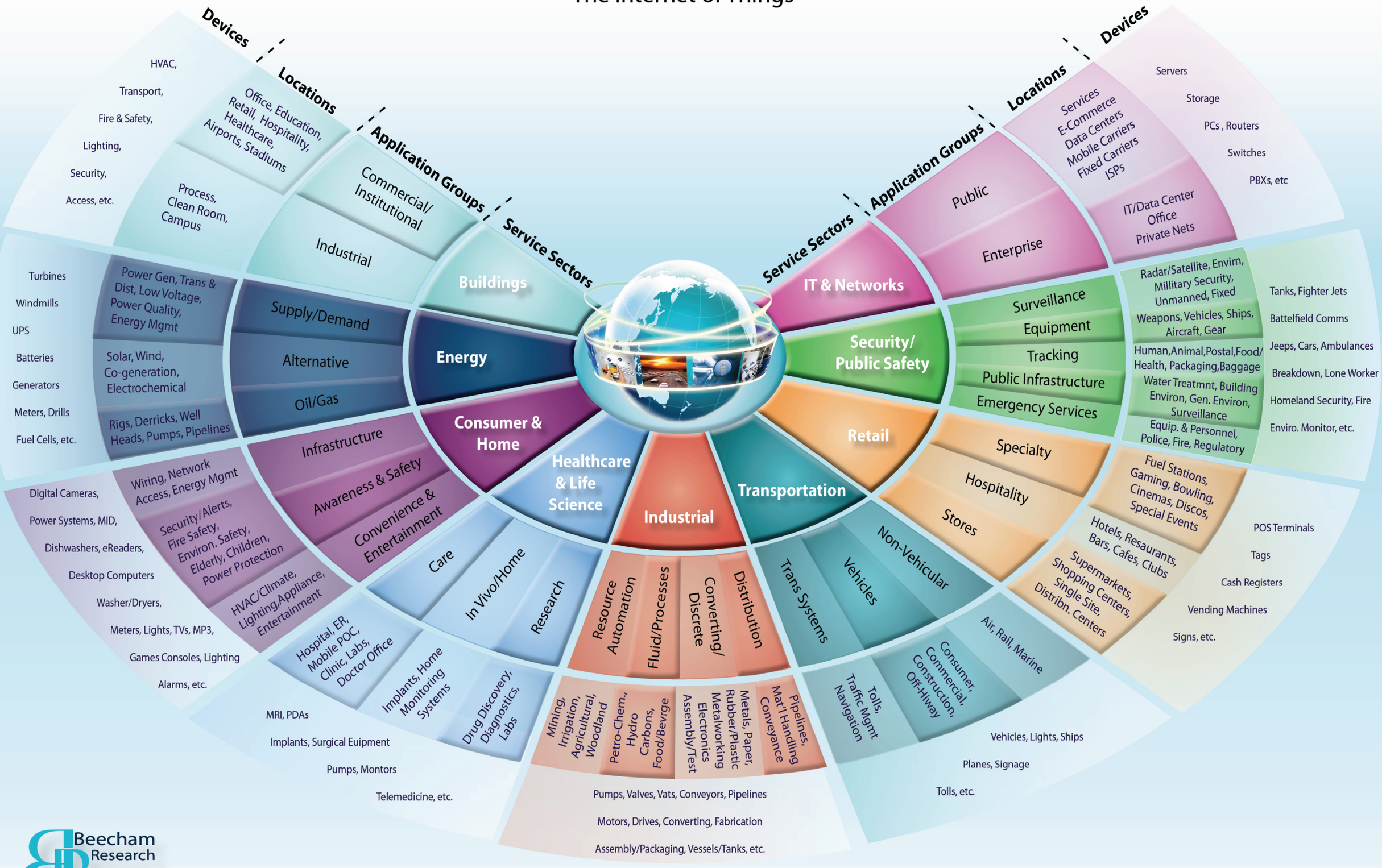
- A cumulative 100 billion processors shipped, each capable of processing information and communicating



http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf

M2M World of Connected Services

The Internet of Things



Beecham Research

Boston | London

info@beechamresearch.com +44 (0)845 533 1758 www.beechamresearch.com

Application: personalised ads in AR environments

Application: supply chain management for retailers

Application: pro-active servicing

Application: ETA of planes

Application: patient monitoring

Application: optimisation in logistics

Application: smart city

http://www.wired.com/gadgetlab/2013/05/internet-of-things/all/

Application: increasing operation efficiency

What have all these apps in common?

- lots of **things** (devices + humans)
- **location**
- sensor data is **messy**
- sensor data is **incomplete**
- **streams** of data

Requirements

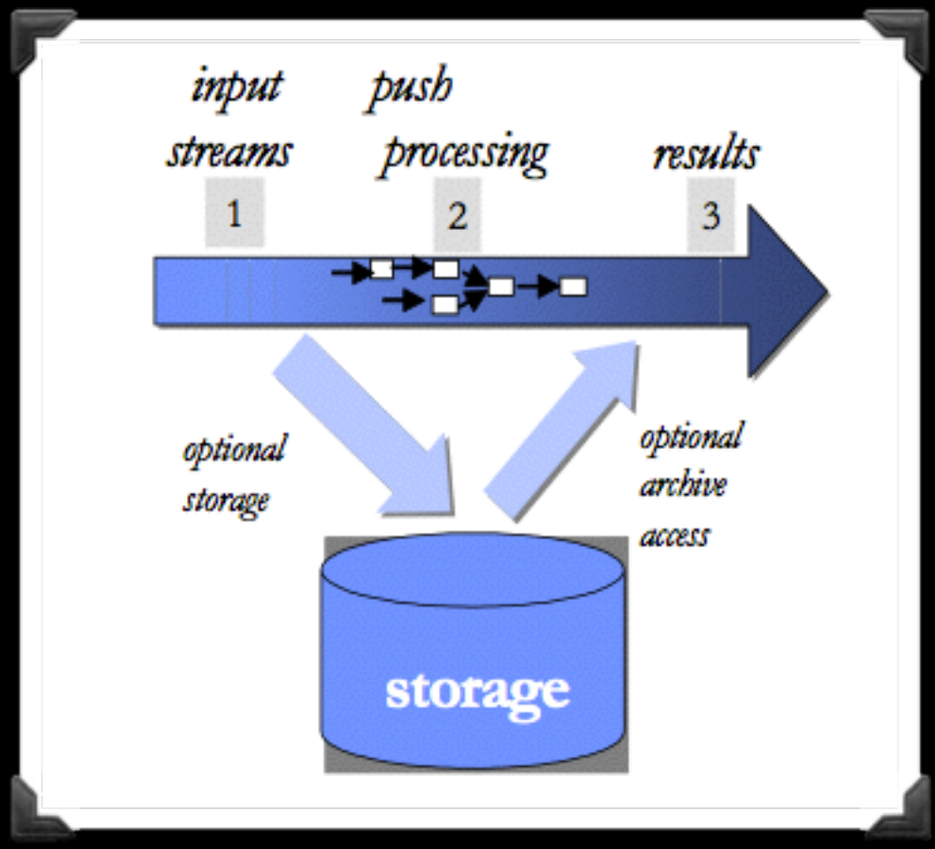
- Be able to **capture, process and store all** the sensor data
- Can **combine historical** data with new, **incoming** data from sensors

How NOT to do it

- Oh, I'm gonna use my good old RDBMS
- Stonebraker 2005

“One Size Fits All”: An Idea Whose Time Has Come and Gone

In summary, there may be a substantial number of domain-specific database engines with differing capabilities off into the future. We are reminded of the curse “may you live in interesting times”. We believe that the DBMS market is entering a period of very interesting times. There are a variety of existing and newly-emerging applications that can benefit from data management and processing principles and techniques. At the same time, these applications are very much different from business data processing and from each other — there seems to be no obvious way to support them with a single code line. The “one size fits all” theme is unlikely to successfully continue under these circumstances.



OK, so what else could I do?

Commoditisation

ØMQ

samza

fünf
Open Sensing Framework

Storm
Distributed and fault-tolerant realtime computation

the thing system

Apache Kafka
A high-throughput distributed messaging system.

Polyglot Persistence

Lambda Architecture

Polyglot Persistence

Lambda Architecture


```
$ 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) http://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

Polyglot Persistence: Example

Polyglot Persistence

Lambda Architecture

Lambda Architecture: Backdrop

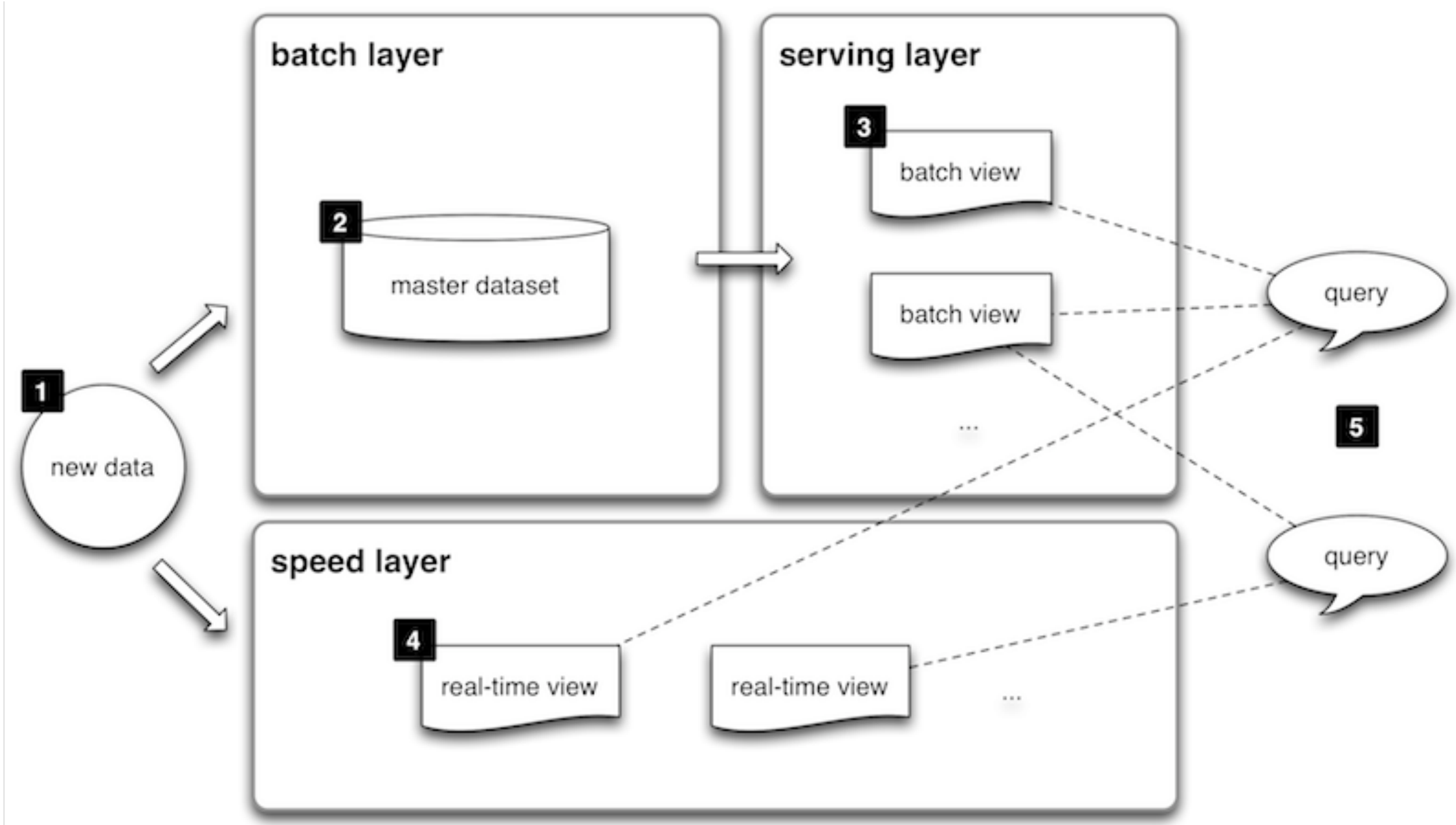
- Nathan Marz (Backtype, Twitter, stealth startup)

- Creator of …
  - Storm
  - Cascalog
  - ElephantDB

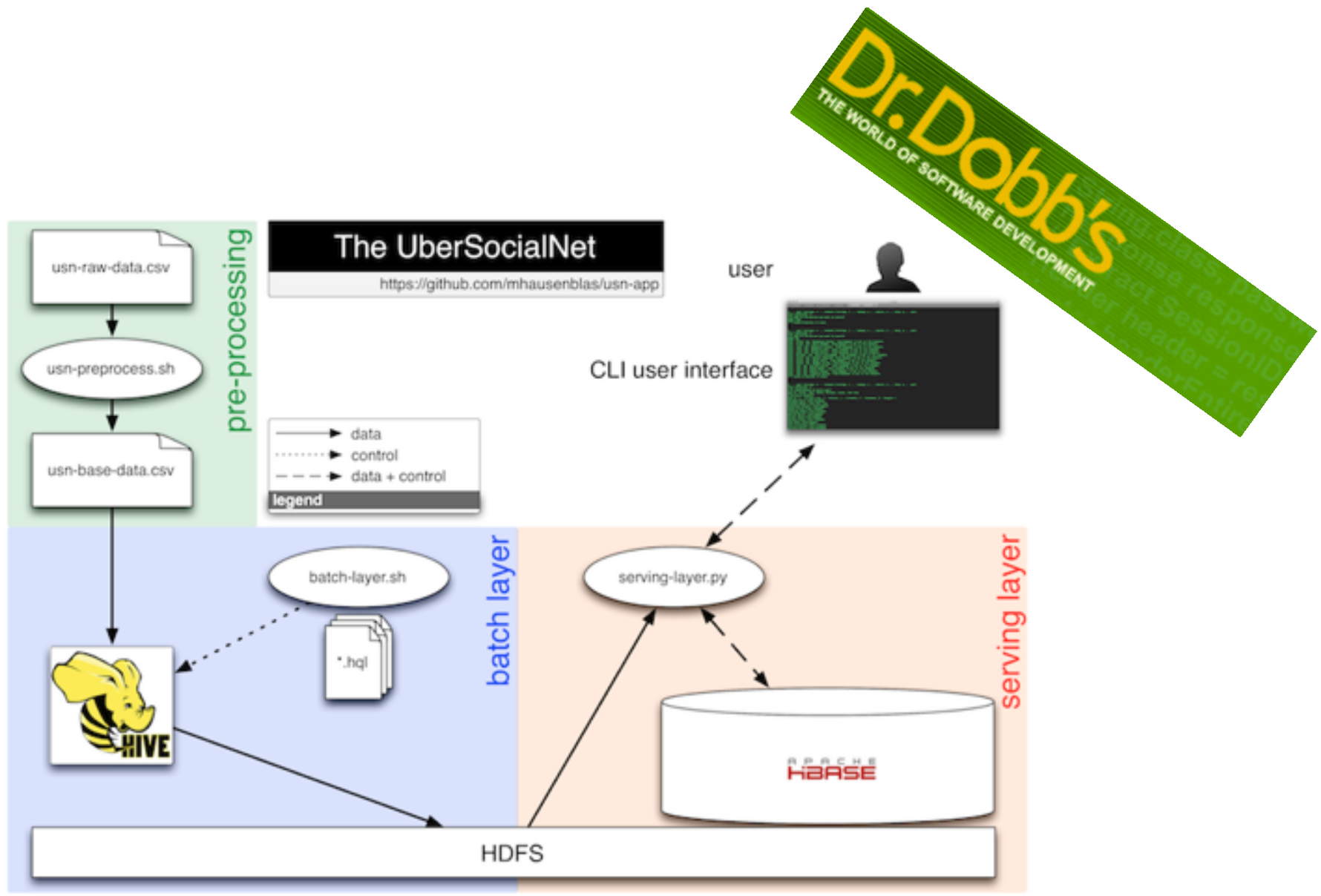
Lambda Architecture: Backdrop

http://manning.com/marz/

Lambda Architecture: Overview



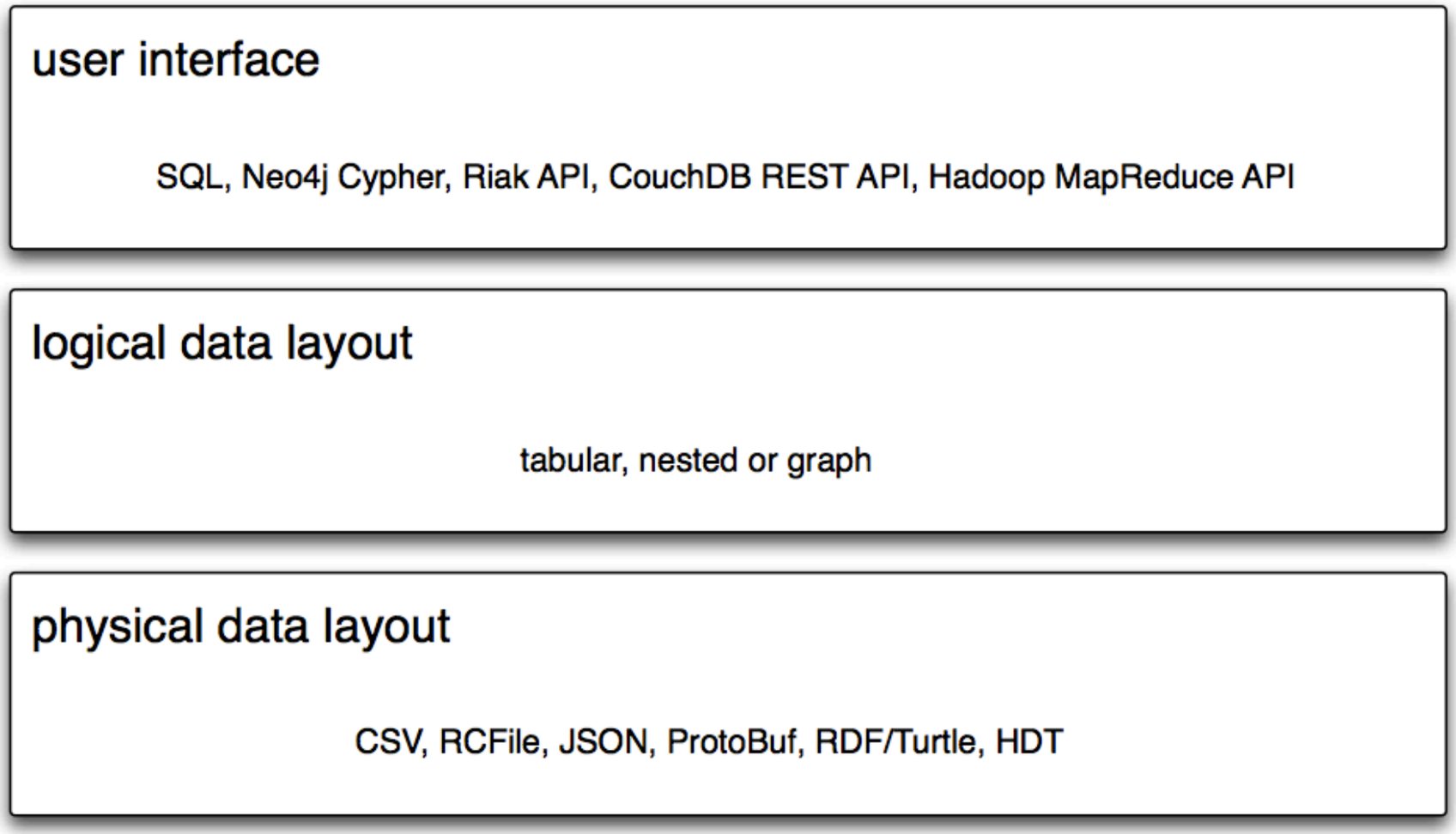
Lambda Architecture: Try it out …



http://www.drdobbs.com/database/applying-the-big-data-lambda-architectur/240162604

Ah, and one more thing …

Levels of representation and interaction



http://arxiv.org/abs/1305.6506

Right. That sounds all well, but also tough to realise …

… can I have this out-of-the-box?

MapR Platform

use cases

log file analysis
fraud detection
ETL off-load

customer insights
forensics
drug discovery

supply chain management
logistics
360 social media

MAPR TECHNOLOGIES™

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

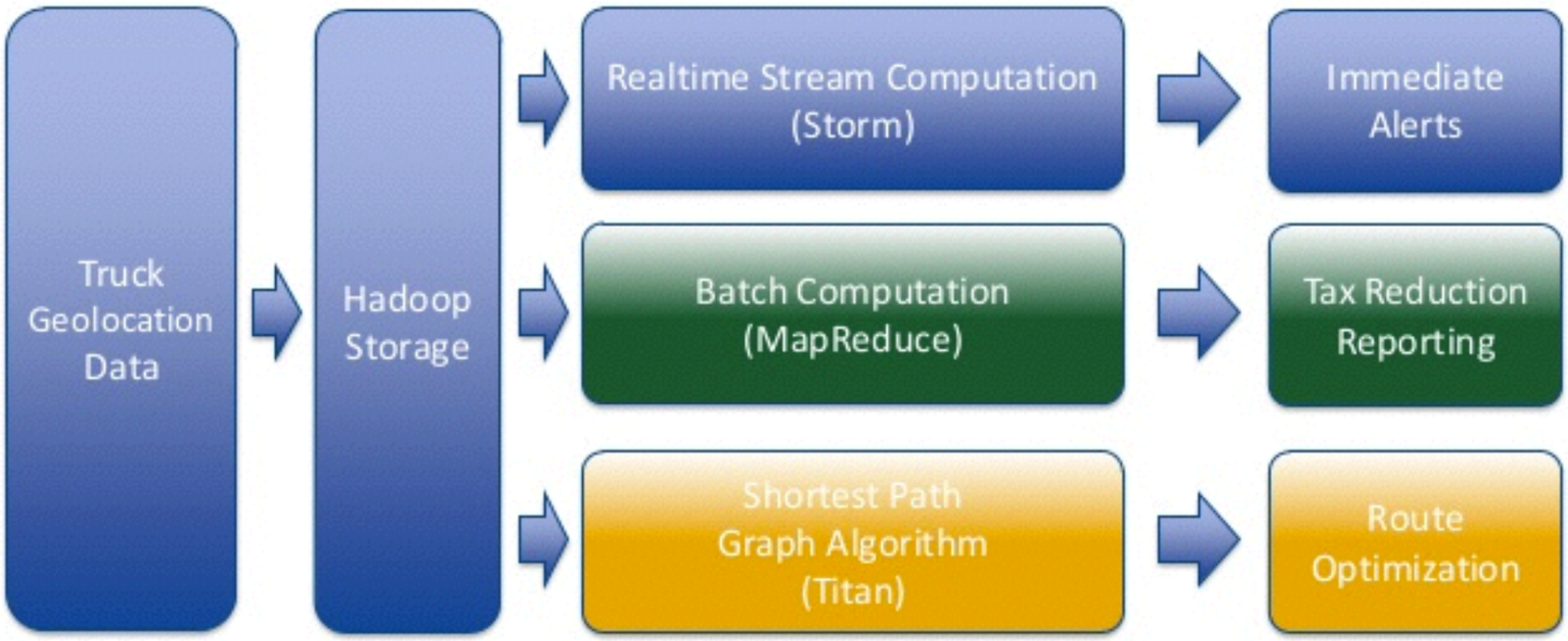
Case Study:
Waste & Recycling Leader

Case Study: Waste & Recycling Leader

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

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

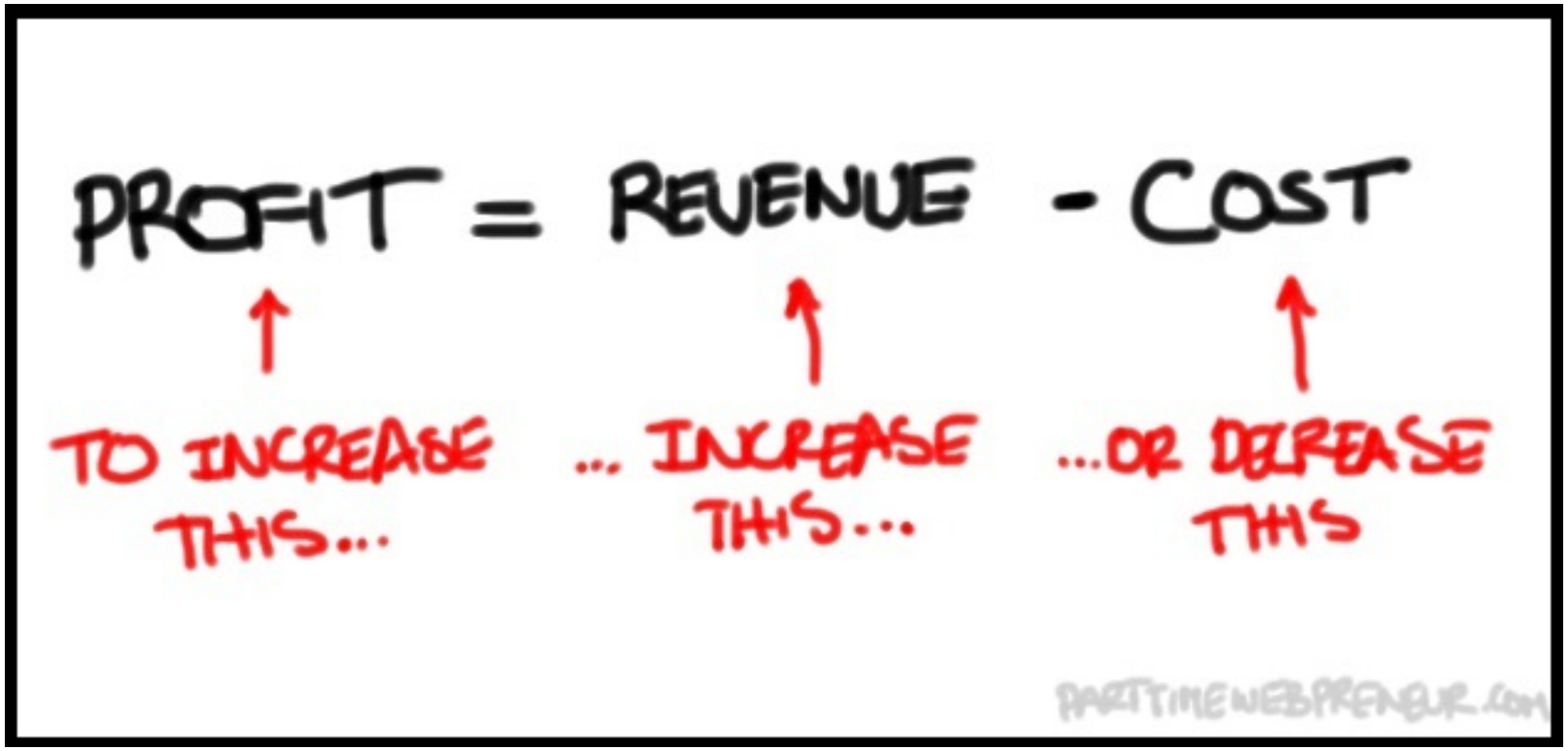
Case Study: Waste & Recycling Leader



Finally. What about the Business Value?

ROI

TCO


PROFIT = REVENUE - COST
TO INCREASE THIS...
... INCREASE THIS...
...OR DECREASE THIS
PARTTIMEWEBPRENEUR.COM

Return of Investment

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

Total Cost of Ownership

- There is nothing like a free lunch
- Open Source is good (but open ≠ free of costs)
- Dev/op knowledge
  - training (in-house? DIY?)
  - outsource

Let's stay in touch …

- @mhausenblas
- @MapR_EMEA
- @MapR