Essential Tools For Your Big Data Arsenal

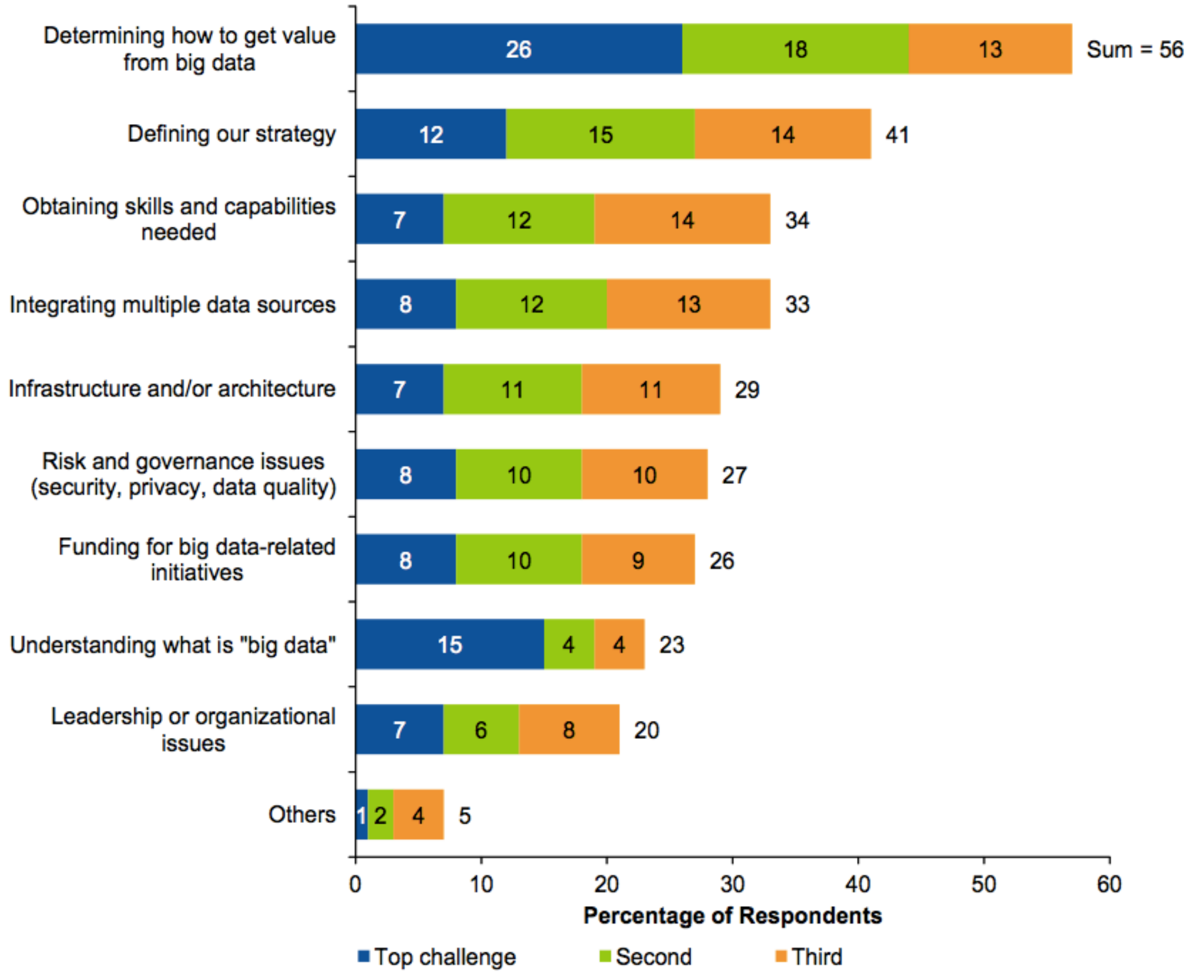
Matt Asay (@mjasay)
VP, Business Development & Strategy, MongoDB

mongoDB

The Big Data Unknown

Top Big Data Challenges?

Translation? Most struggle to know what Big Data is, how to manage it and who can manage it

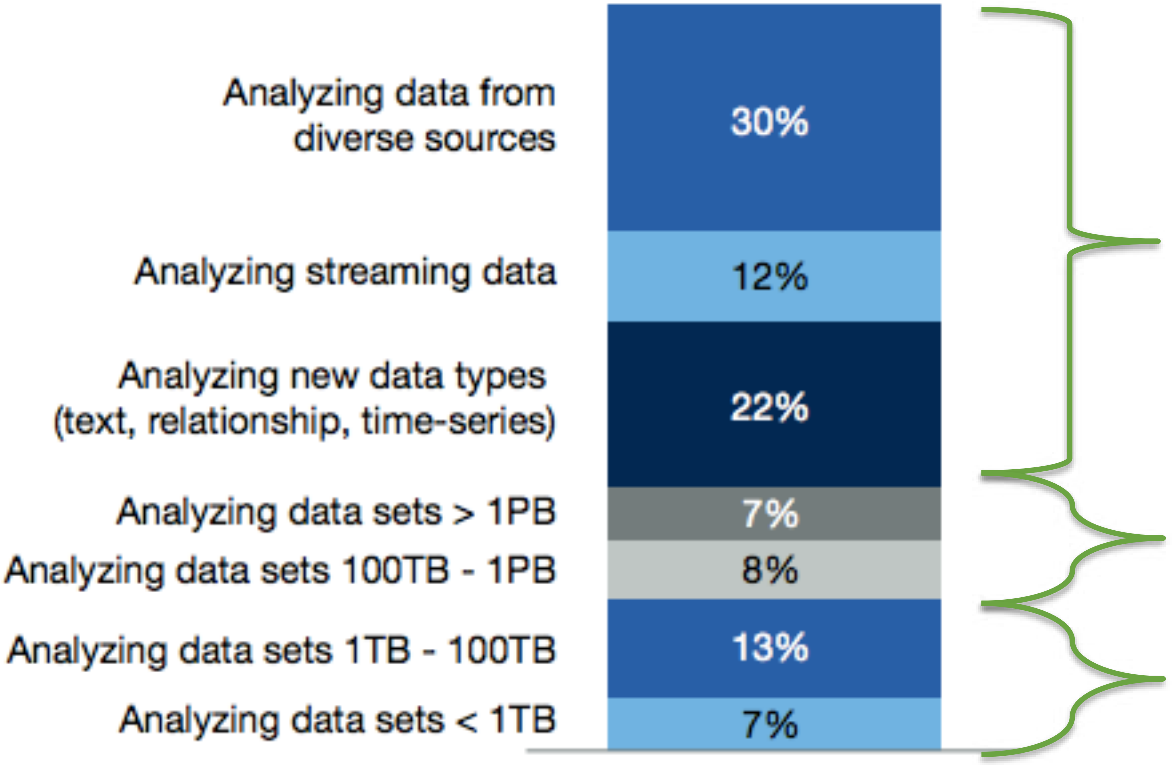


N = 687 (excludes "don't know" responses)

Source: Gartner

mongoDB

Understanding Big Data – It’s Not Very “Big”



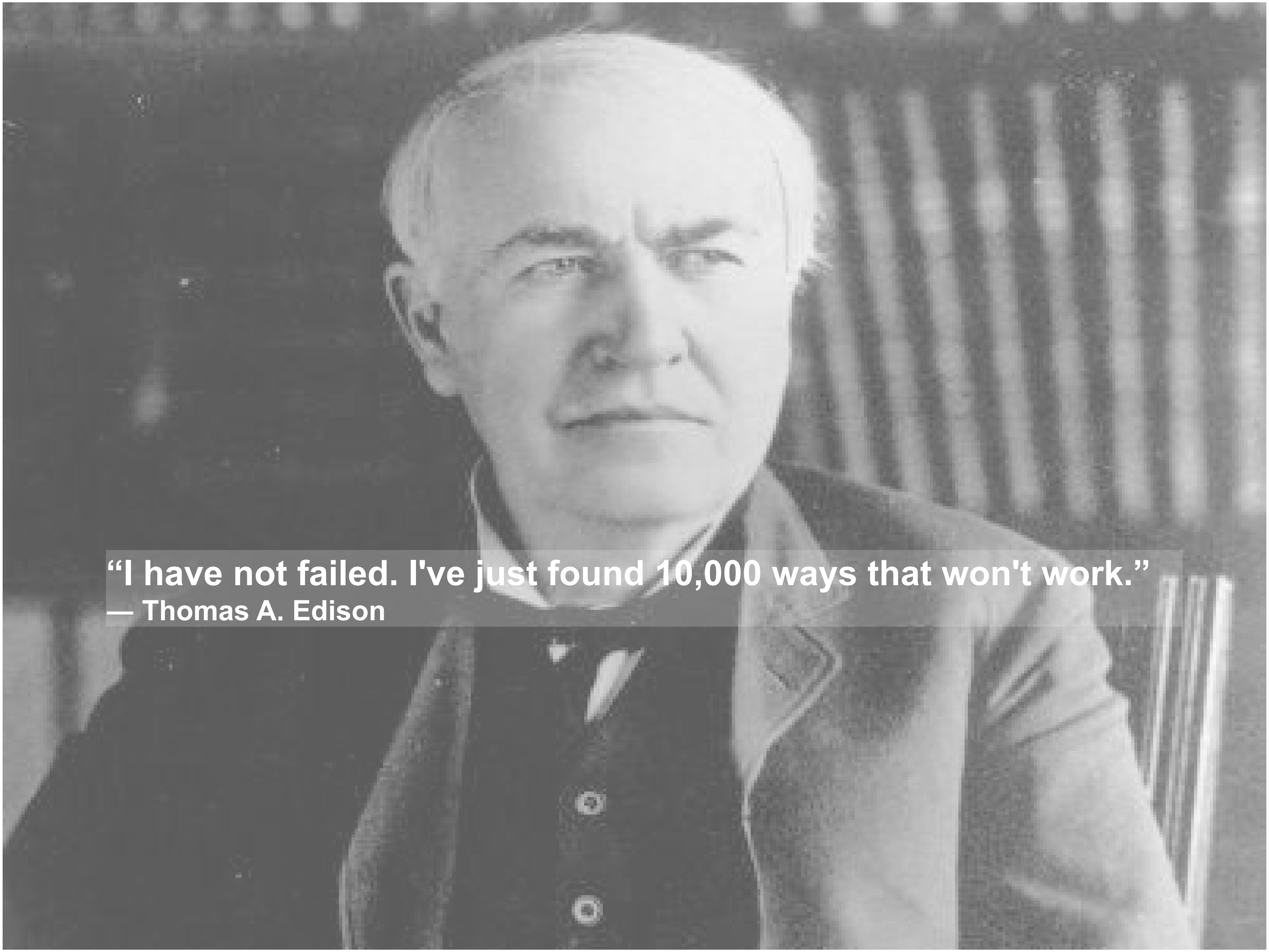
64% - Ingest diverse, new data in real-time

15% - More than 100TB of data

20% - Less than 100TB (average of all? <20TB)

from Big Data Executive Summary – 50+ top executives from Government and F500 firms

Innovation As Iteration



“I have not failed. I've just found 10,000 ways that won't work.”
— Thomas A. Edison

Back in 1970…Cars Were Great!

The New American Car.

This is the American Motors Gremlin.

It is the kind of car this country has needed for a long, long time.

It is designed to give the American motorist a car that is easy to buy, easy to handle, easy to take care of, and, at the same time, fun to drive.

The Gremlin is the smallest production car made in America.

It is 161 inches long, just 2½ inches longer than the Volkswagen.

Yet its turning circle, at 32 feet, 8 inches, is about 3 feet less than VW's.

Which makes the Gremlin about the easiest car in the world to park and handle.

The Gremlin gets the best gas mileage of any car made in America. It goes about 500 miles without stopping for gas.

This is great gas mileage, when you consider that the Gremlin has a bigger standard engine than any car near its size and price. 128 hp to VW's 57.

This engine gets from 0 to 60 in 15.3 seconds, the pickup you need on expressways.

And nobody's going to push you around in a Gremlin. It is 10 inches wider, 7 inches lower and 765 pounds heavier than a VW.

Which gives you about the smoothest, most stable ride possible in a car this size.

The Gremlin is remarkably easy to service and maintain.

Its normal oil change interval is 6 months or 6,000 miles; lubrication is normally needed only every 24,000 miles.

There are two basic Gremlin models.

A two-passenger, with storage area in the rear.

A four-passenger with fold-down rear seats for extra storage and flip-up rear window for easy access.

Both models cost about what you'd pay for an imported economy car.

The four-passenger lists for $1,959.*

The lowest list price of any car made in America.

Except for the two-passenger Gremlin. It lists for $1,879.*

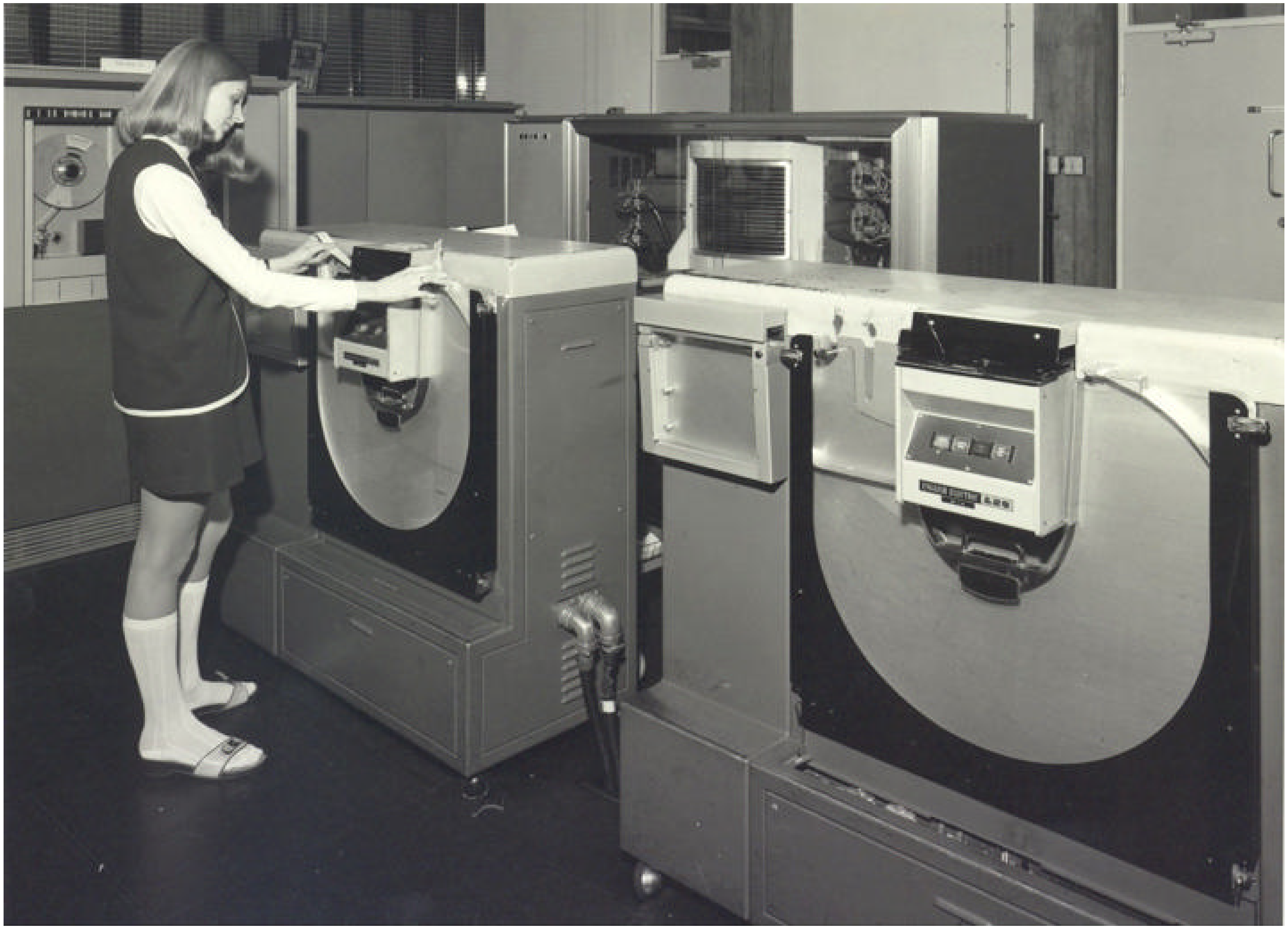
Which is quite a bargain, when you consider what you get for your money.

The new American car.

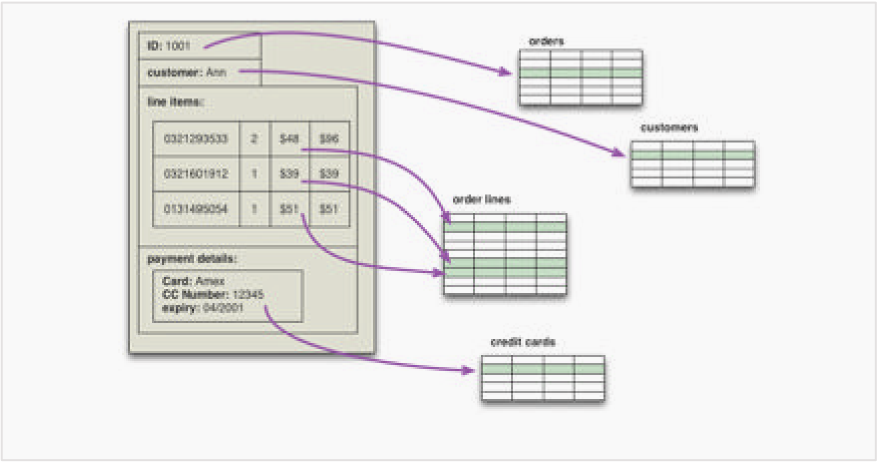
American Motors
Gremlin

$1,879* 2-Passenger **$1,959*** 4-Passenger

So Were Computers!



Lots of Great Innovations Since 1970



ID: 1001

customer: Ann

line items:

0321293533	2	$48	$96
0321601912	1	$39	$39
0131495054	1	$51	$51

payment details:

Card: Amex
CC Number: 12345
expiry: 04/2001

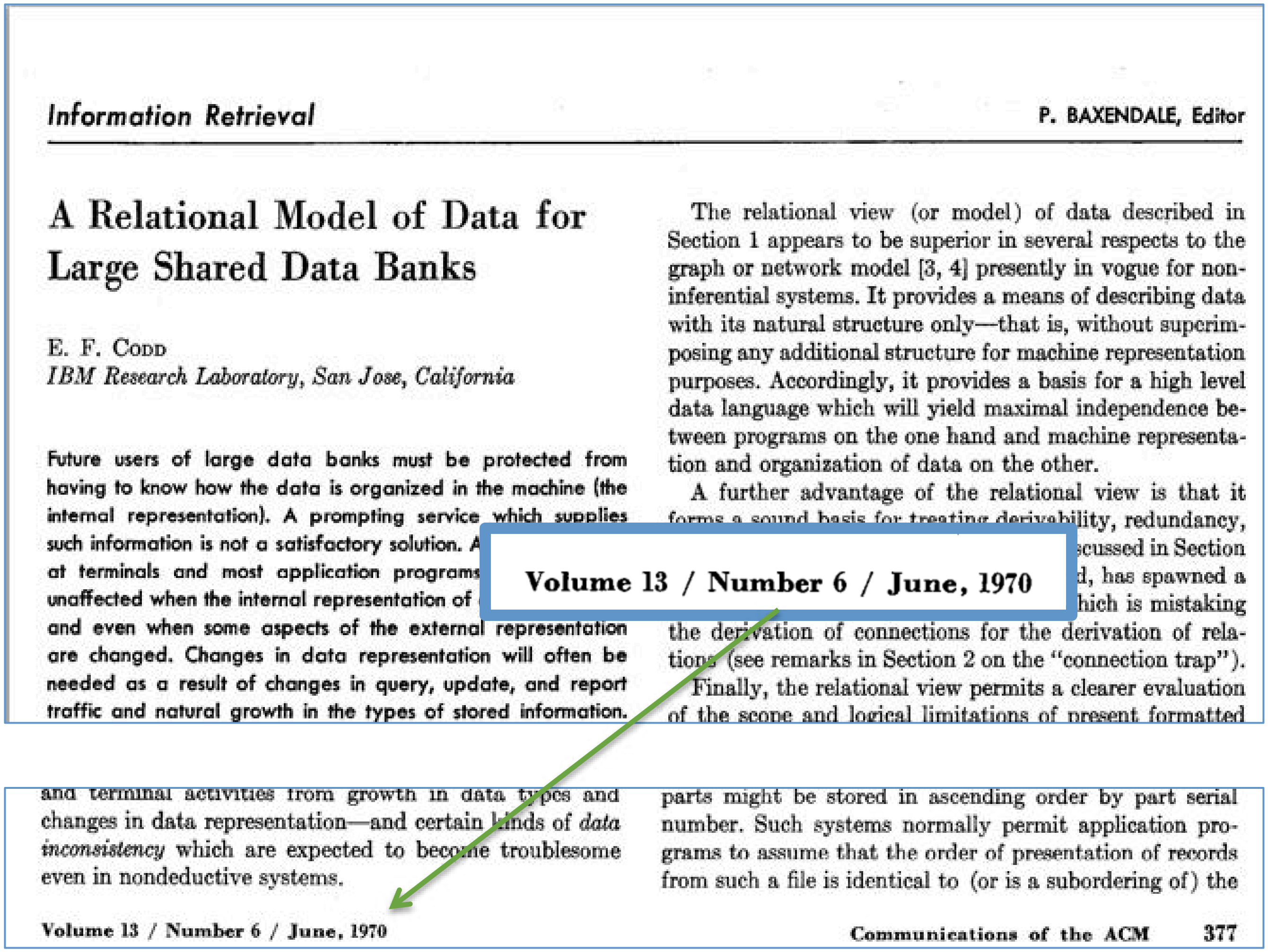
orders

customers

order lines

credit cards

Including the Relational Database



Information Retrieval

P. BAXENDALE, Editor

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. A at terminals and most application programs unaffected when the internal representation of and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, cussed in Section d, has spawned a hich is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the "connection trap").

Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted

Volume 13 / Number 6 / June, 1970

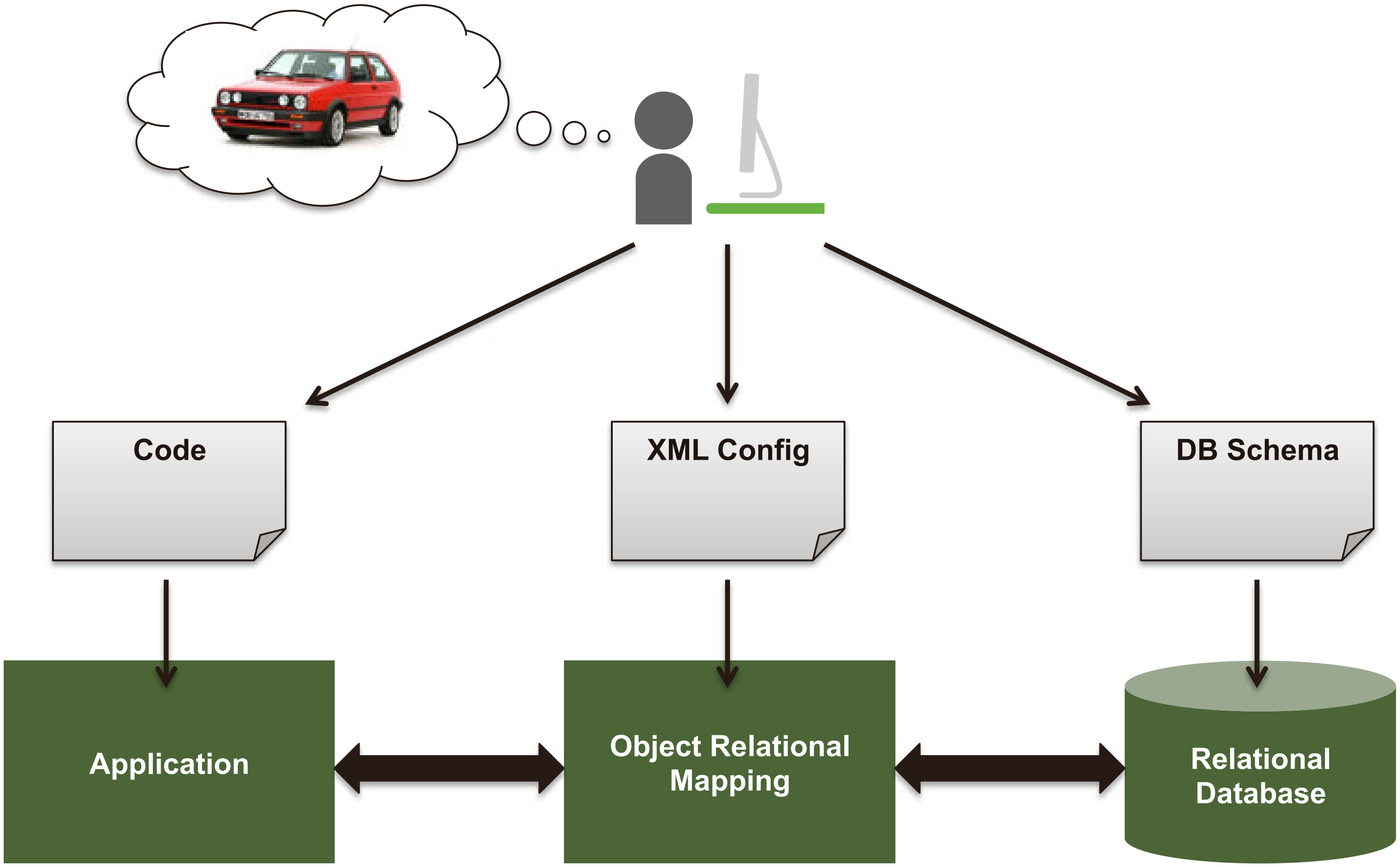
and terminal activities from growth in data types and changes in data representation—and certain kinds of *data inconsistency* which are expected to become troublesome even in nondeductive systems.

parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from such a file is identical to (or is a subordering of) the

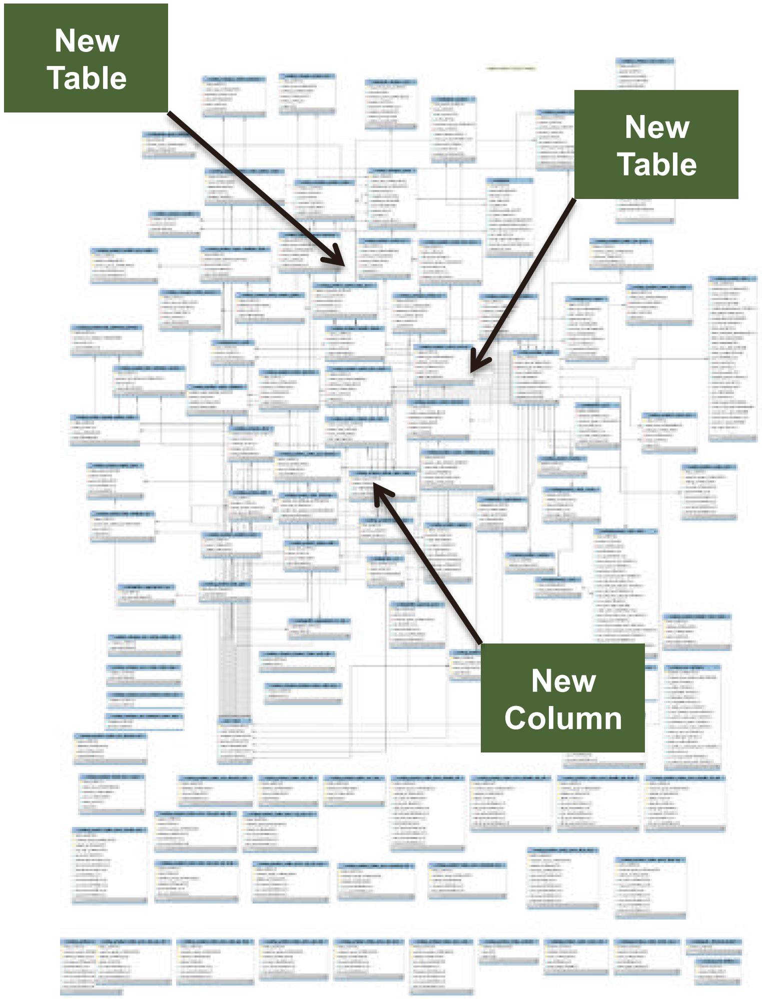
Volume 13 / Number 6 / June, 1970

Communications of the ACM 377

RDBMS Makes Development Hard



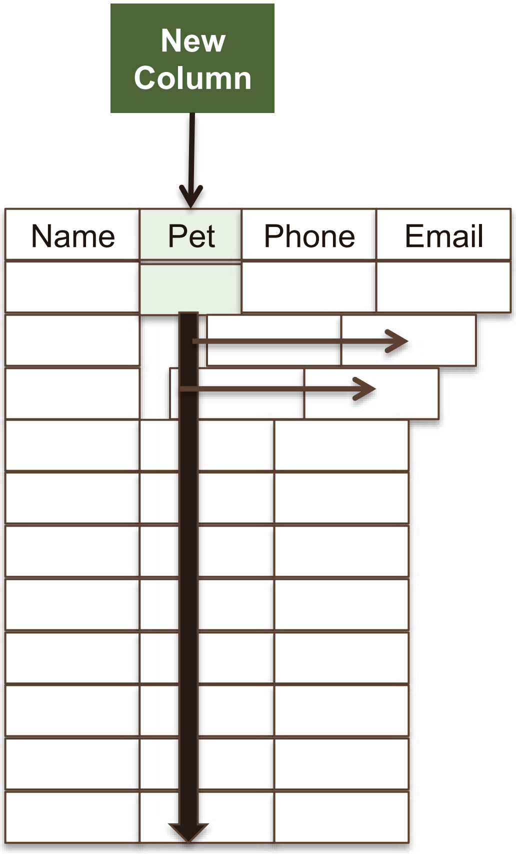
And Even Harder To Iterate



New Table

New Table

New Column



New Column

Name

Pet

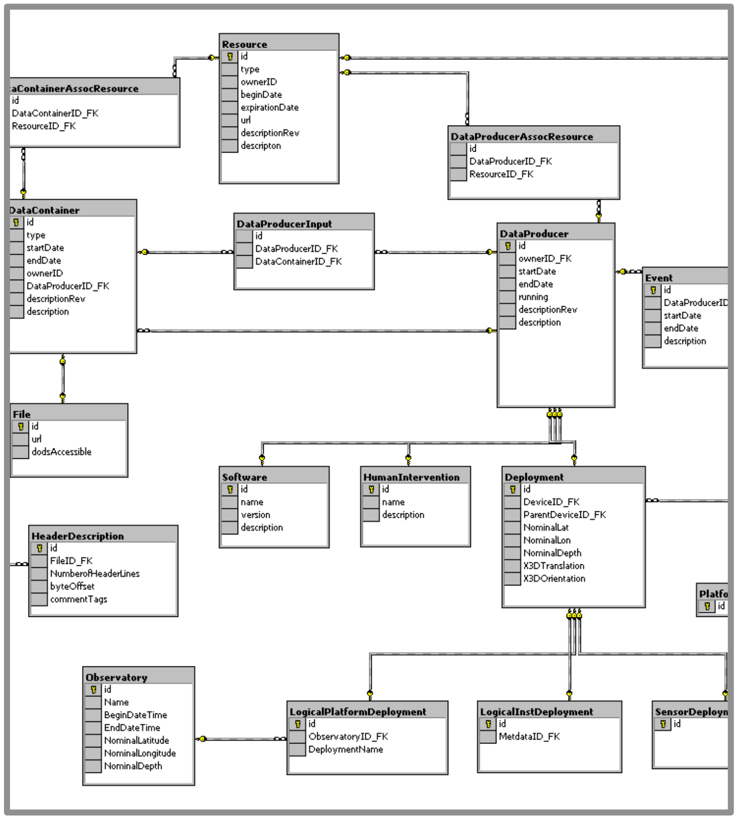
Phone

Email

3 months later…

From Complexity to Simplicity

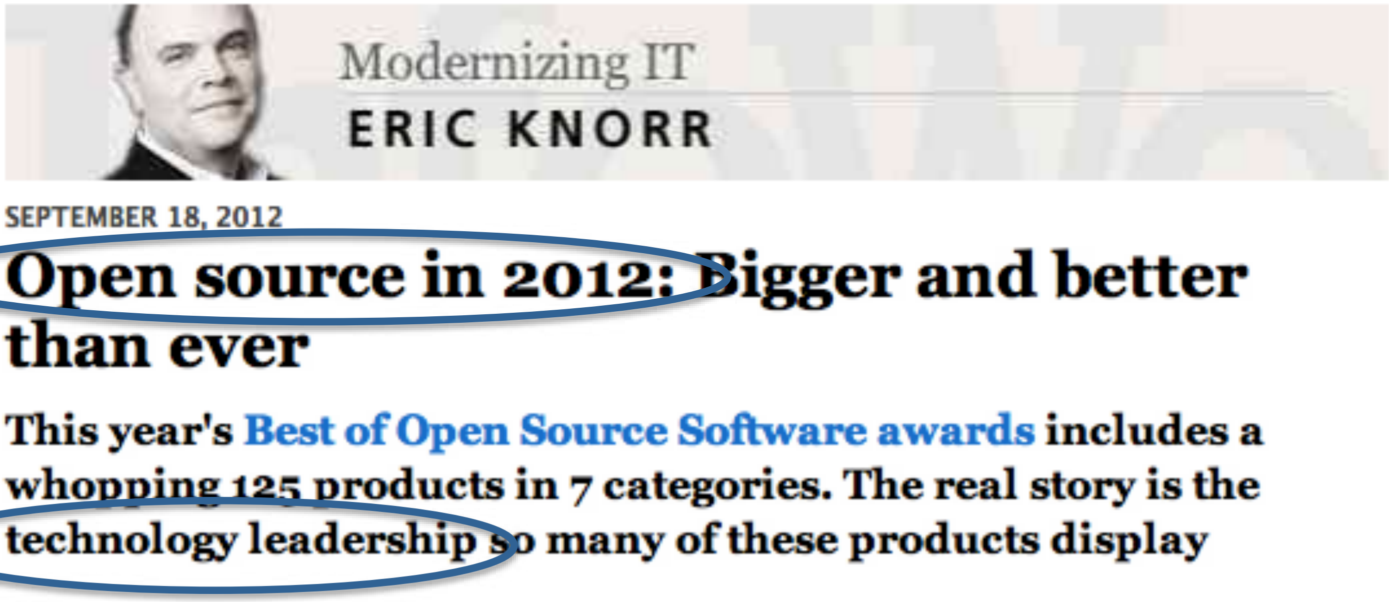
RDBMS



MongoDB

```
{
    _id : ObjectId("4c4ba5e5e8aabf3"),
    employee_name: "Dunham, Justin",
    department : "Marketing",
    title : "Product Manager, Web",
    report_up: "Neray, Graham",
    pay_band: “C”,
    benefits : [
           {  type :  "Health",
              plan : "PPO Plus" },
           {  type :   "Dental",
              plan : "Standard" }
               ]
}
```

So…Use Open Source



Modernizing IT

ERIC KNORR

SEPTEMBER 18, 2012

Open source in 2012: Bigger and better than ever

This year's Best of Open Source Software awards includes a whopping 125 products in 7 categories. The real story is the technology leadership so many of these products display

Big Data != Big Upfront Payment

RDBMS Is Expensive To Scale

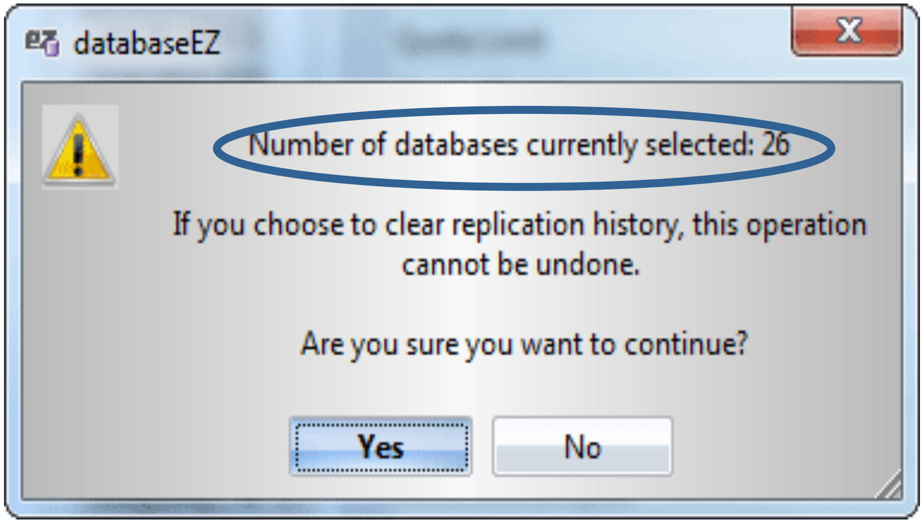
“Clients can also opt to run zEC12 without a raised datacenter floor -- a first for high-end IBM mainframes.”

IBM Press Release 28 Aug, 2012

Spoiled for choice

DB-Engines.com Database Ranking

1	Oracle	Relational DBMS	1583.84	54.23
2	MySQL	Relational DBMS	1331.34	25.58
3	Microsoft SQL Server	Relational DBMS	1207	-106.78
4	PostgreSQL	Relational DBMS	177.01	-5.22
5	DB2	Relational DBMS	175.83	3.58
6	MongoDB	NoSQL Document Store	149.48	-2.71
7	Microsoft Access	Relational DBMS	142.49	-4.21
8	SQLite	Relational DBMS	77.88	-4.9
9	Sybase	Relational DBMS	73.66	-1.68
10	Teradata	Relational DBMS	54.41	3.32

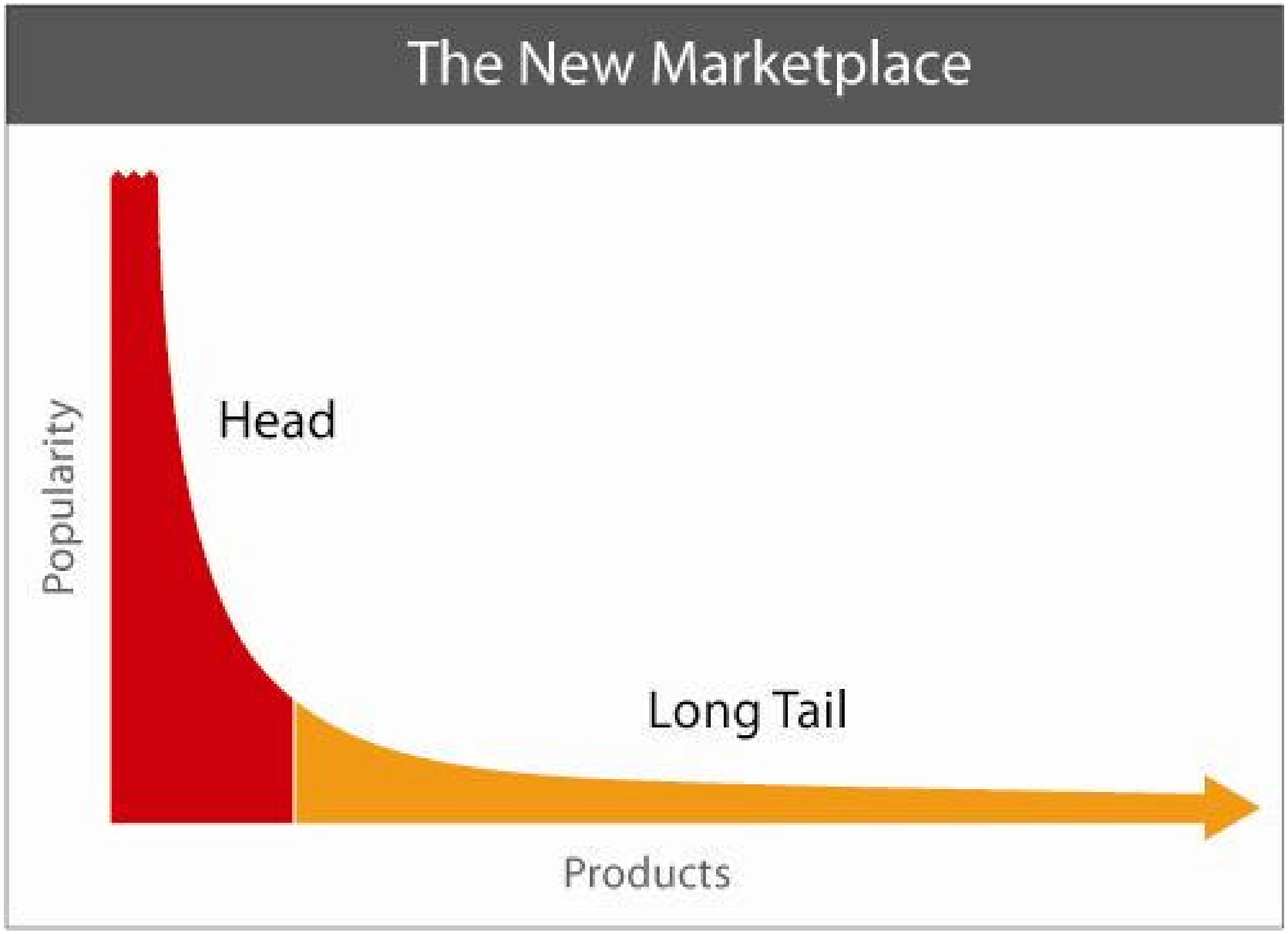


Number of databases currently selected: 26

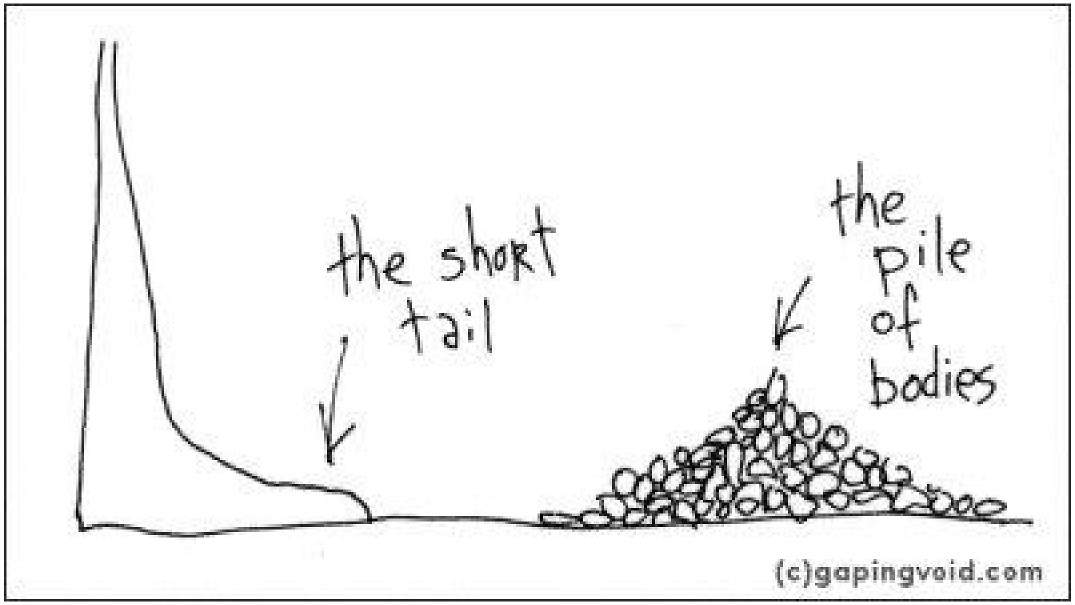
If you choose to clear replication history, this operation cannot be undone.

Are you sure you want to continue?

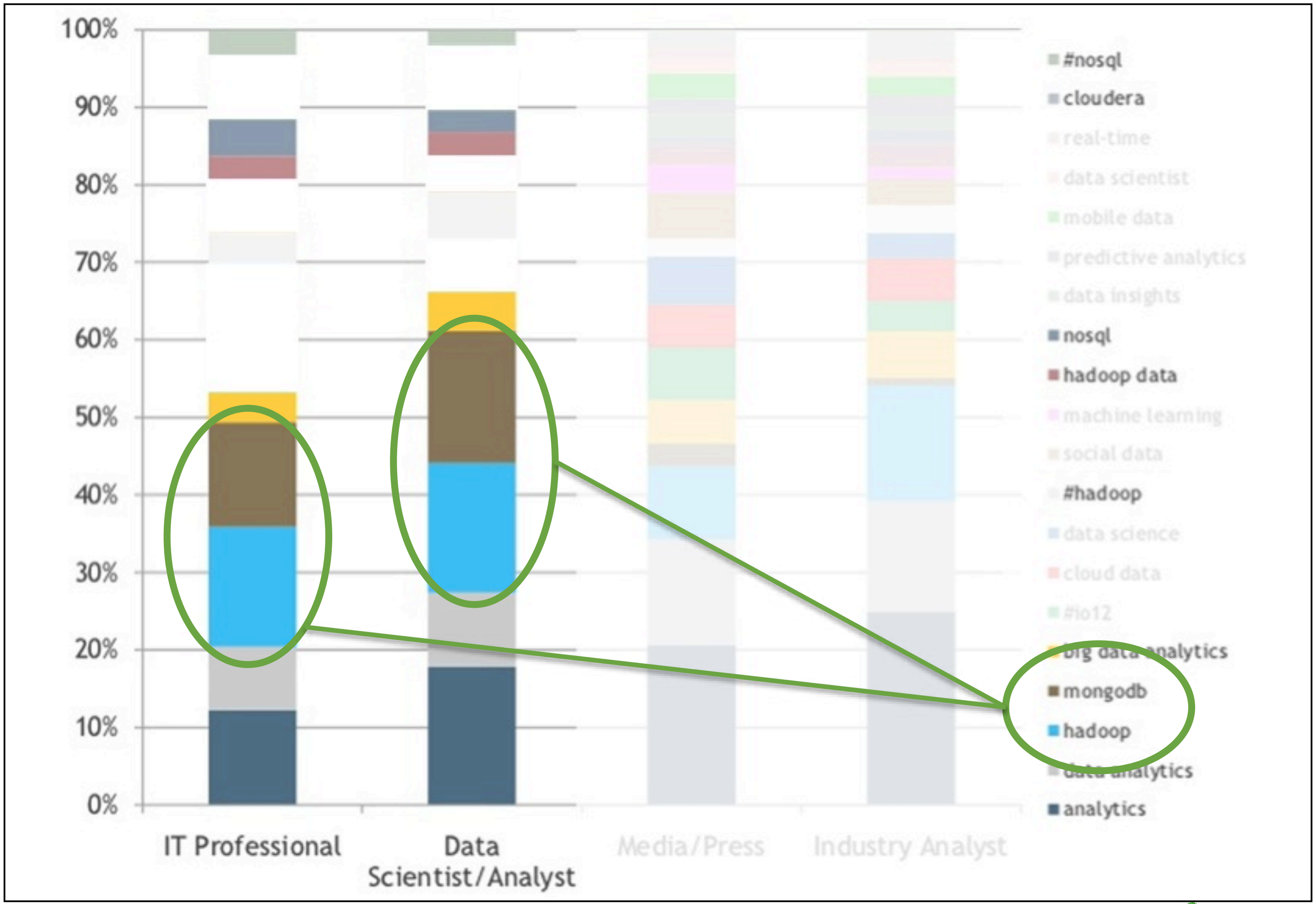
Remember the Long Tail?



It Didn’t Work Out So Well



Use Popular, Well-Known Technologies



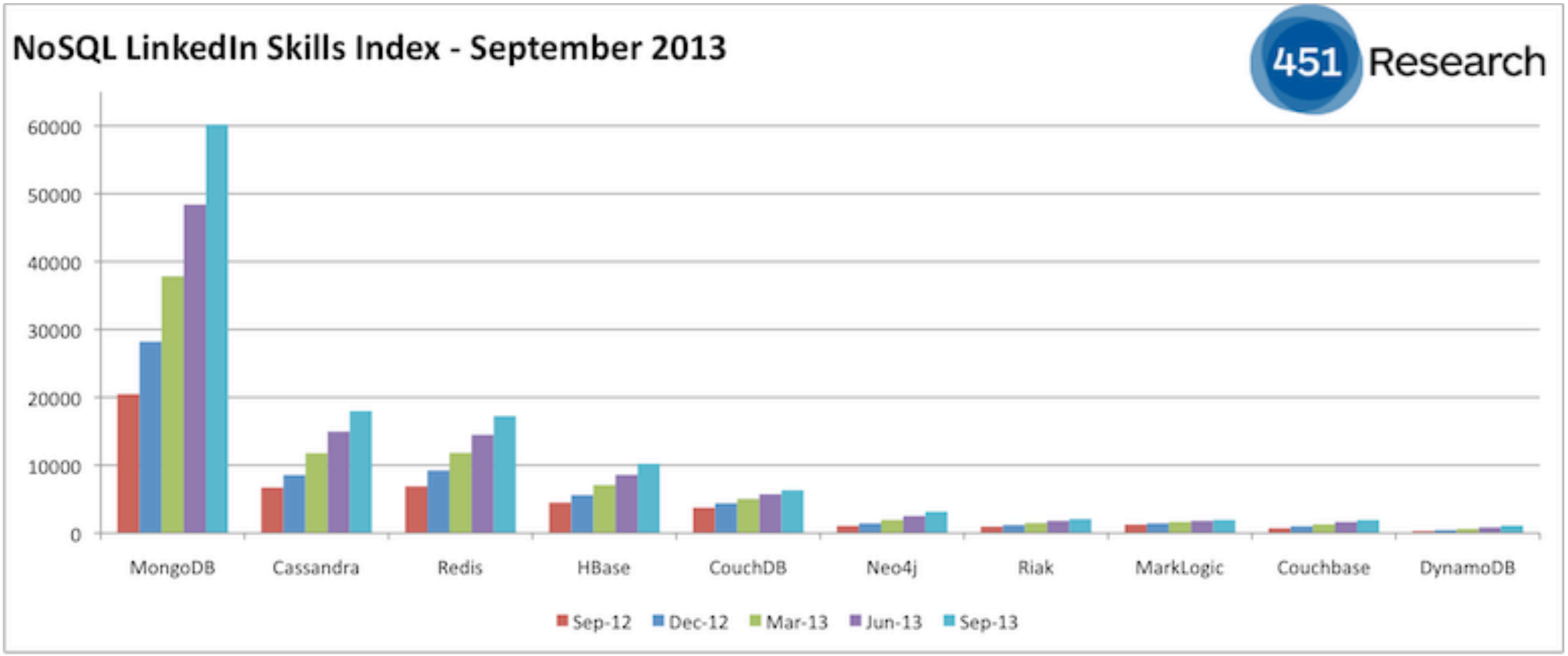
Source: Silicon Angle, 2012

Ask the Right Questions…

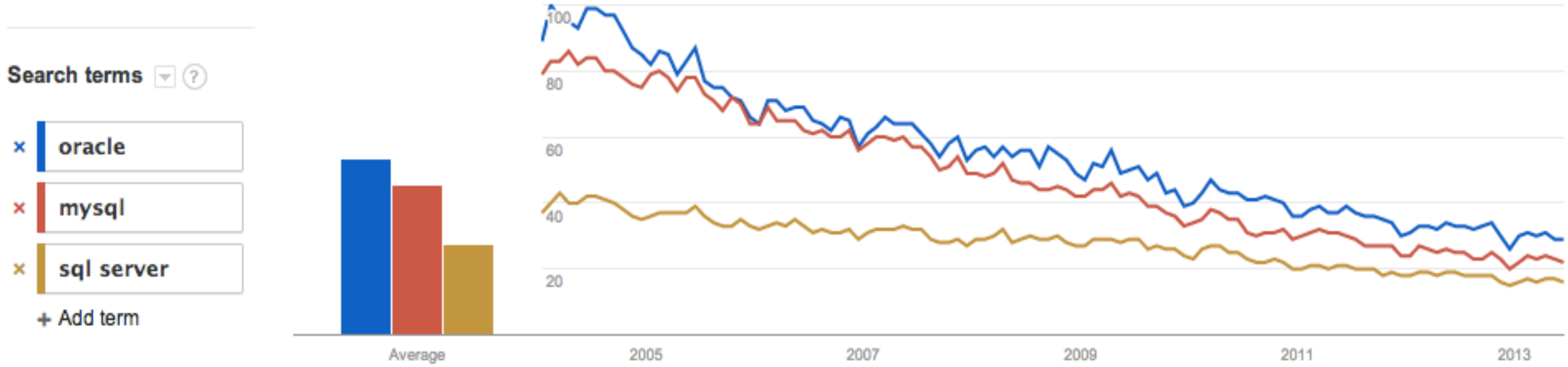
“Organizations already have people who know their own data better than mystical data scientists….Learning Hadoop [or MongoDB] is easier than learning the company’s business.”

(Gartner, 2012)

Leverage Existing Skills



Search as a Sign?



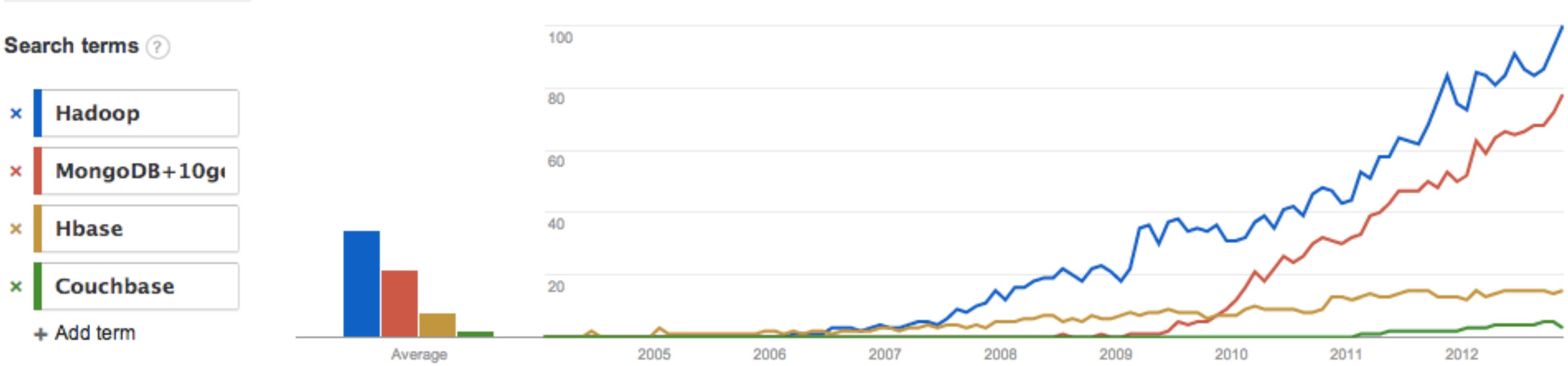
Search terms

× oracle

× mysql

× sql server

+ Add term



Search terms

× Hadoop

× MongoDB+10g

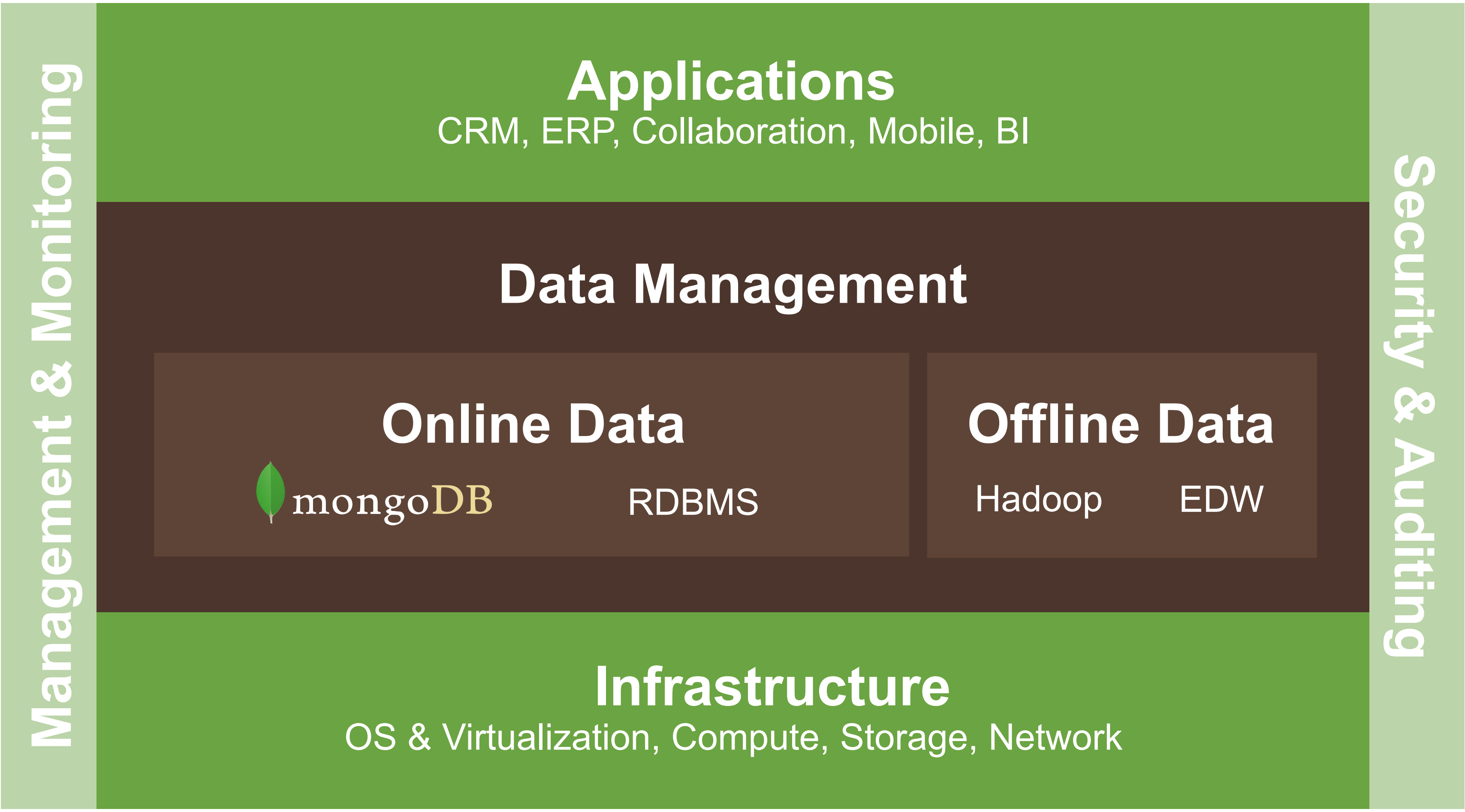
× Hbase

× Couchbase

+ Add term

When To Use Hadoop, NoSQL

Enterprise Big Data Stack



Consideration – Online vs. Offline

Online *vs.* Offline

- Real-time
- Low-latency
- High availability

- Long-running
- High-Latency
- Availability is lower priority

Consideration – Online vs. Offline

Online *vs.* Offline

mongoDB

hadoop

Hadoop Is Good for…

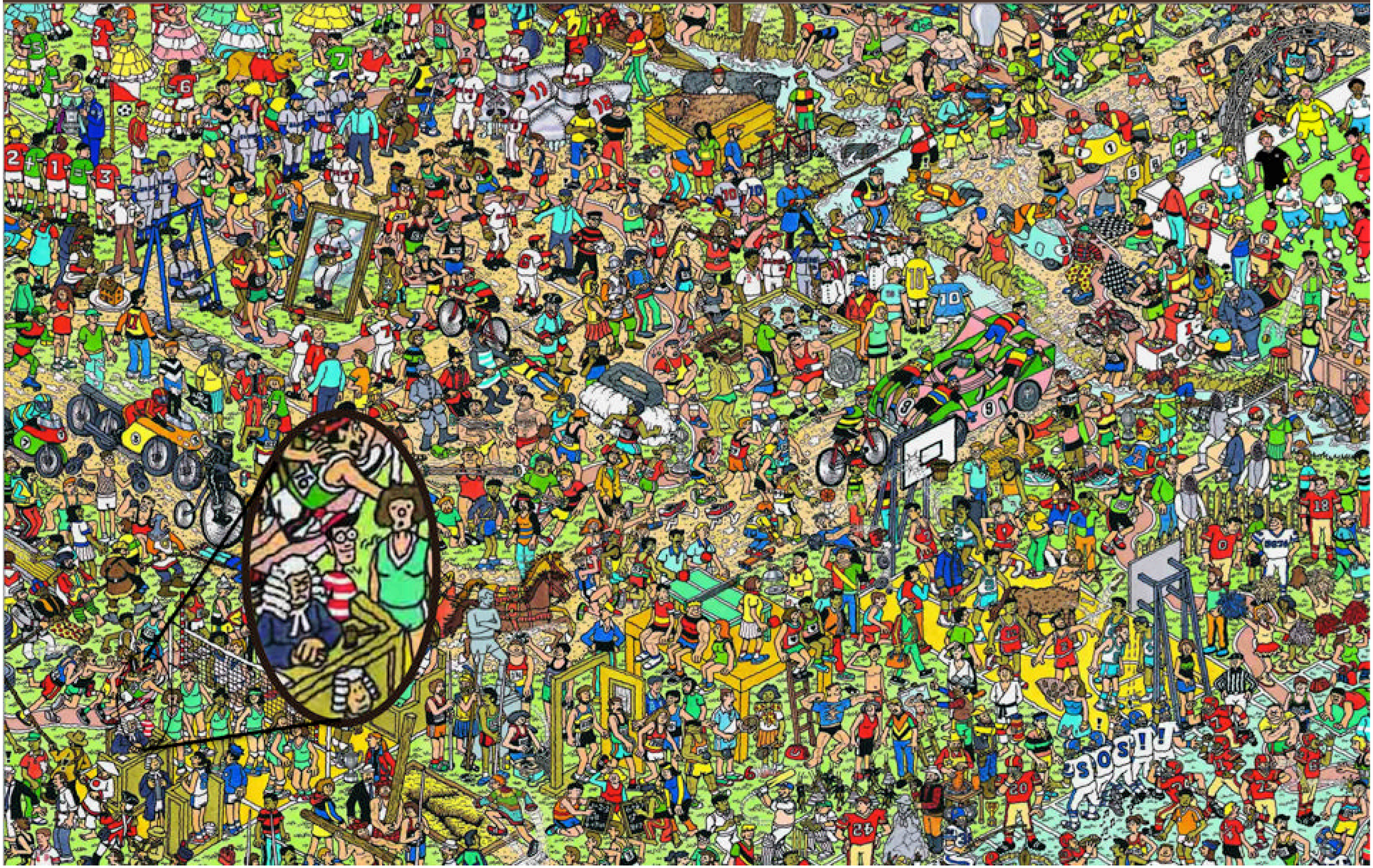
- Risk Modeling
- Churn Analysis
- Recommendation Engine
- Ad Targeting
- Transaction Analysis
- Trade Surveillance
- Network Failure Prediction
- Search Quality
- Data Lake

MongoDB/NoSQL Is Good for…

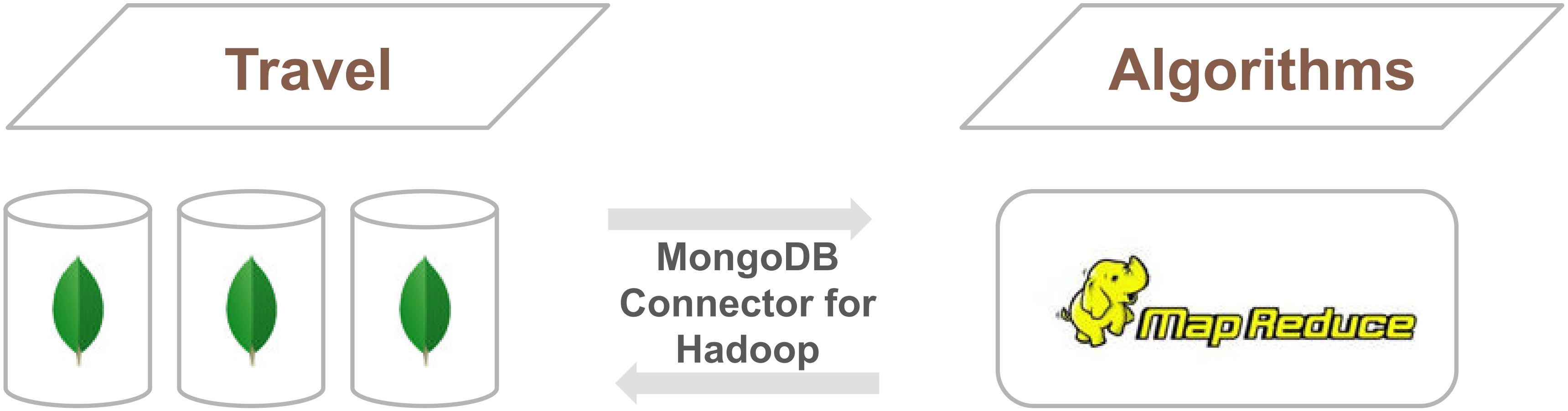
- 360° View of the Customer
- Mobile & Social Apps
- Fraud Detection
- User Data Management
- Content Management & Delivery
- Reference Data
- Product Catalogs
- Machine to Machine Apps
- Data Hub

How To Use The Two Together?

Finding Waldo



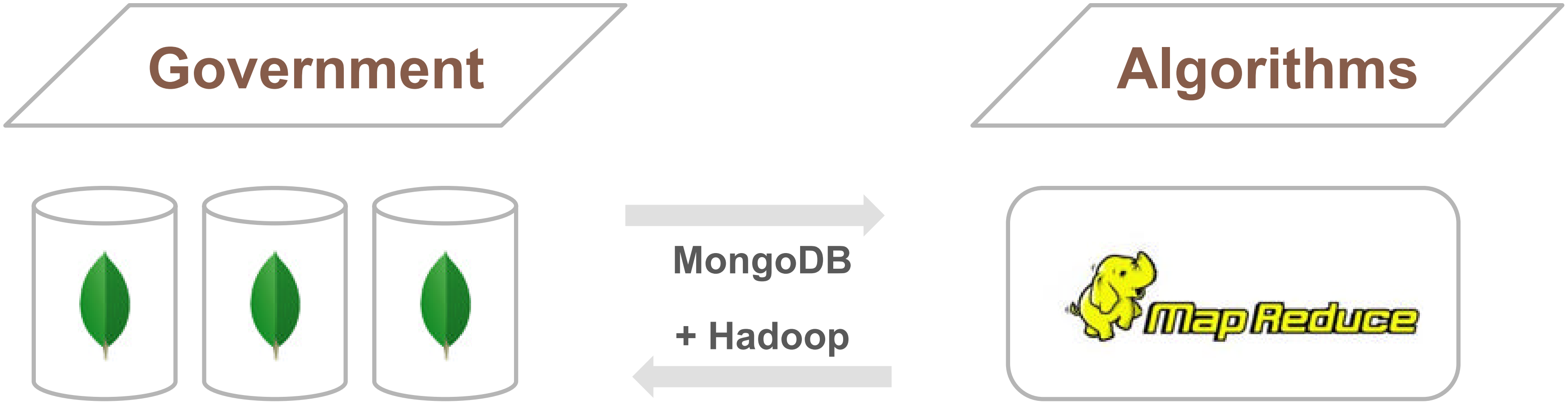
Customer example: Online Travel



- Flights, hotels and cars
- Real-time offers
- User profiles, reviews
- User metadata (previous purchases, clicks, views)

- User segmentation
- Offer recommendation engine
- Ad serving engine
- Bundling engine

Predictive Analytics



Government

MongoDB

+ Hadoop

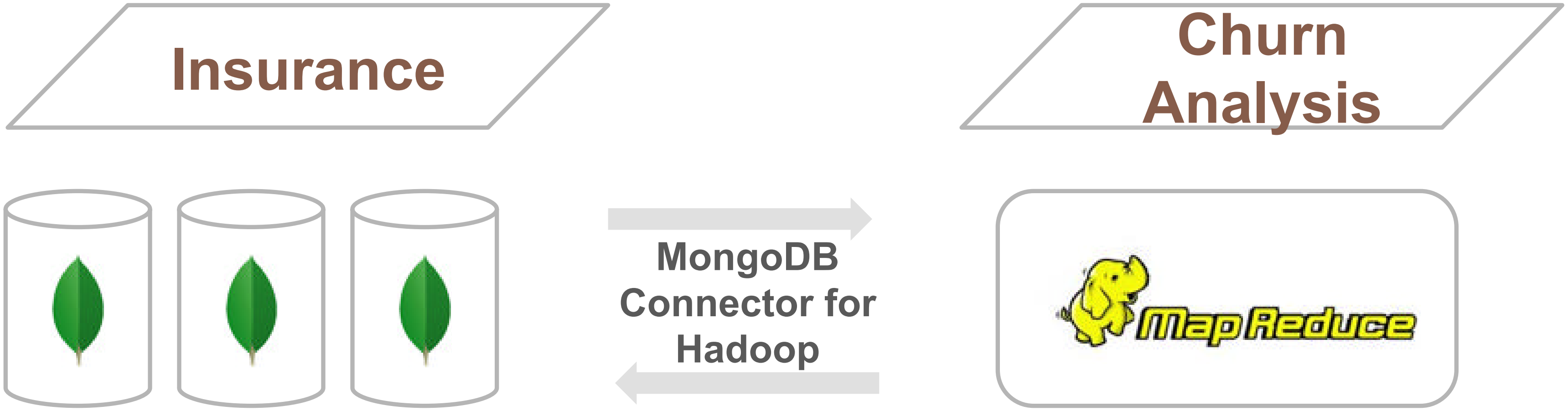
Algorithms

- Predictive analytics system for crime, health issues
- Diverse, unstructured (incl. geospatial) data from 30+ agencies
- Correlate data in real-time

- Long-form trend analysis
- MongoDB data dumped into Hadoop, analyzed, re-inserted into MongoDB for better real-time response

Data Hub

MetLife



Insurance

MongoDB Connector for Hadoop

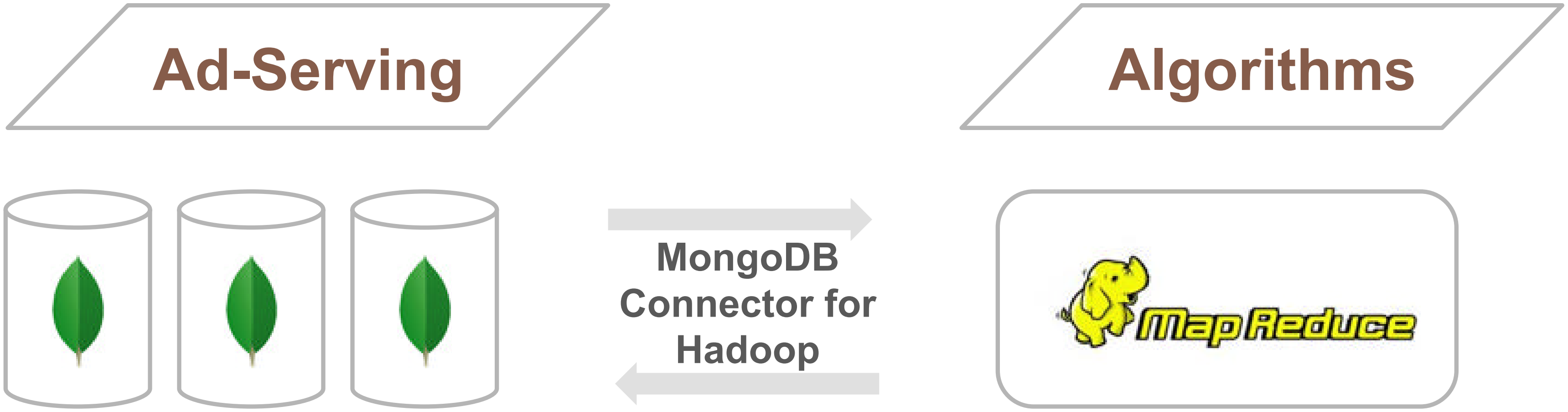
Churn Analysis

- Insurance policies
- Demographic data
- Customer web data
- Call center data
- Real-time churn detection

- Customer action analysis
- Churn prediction algorithms

Machine Learning

criteo



Ad-Serving

Algorithms

MongoDB Connector for Hadoop

- Catalogs and products
- User profiles
- Clicks
- Views
- Transactions

- User segmentation
- Recommendation engine
- Prediction engine

MongoDB + Hadoop Connector

- Makes MongoDB a Hadoop-enabled file system
- Read and write to live data, in-place
- Copy data between Hadoop and MongoDB
- Full support for data processing
  - Hive
  - MapReduce
  - Pig
  - Streaming
  - EMR

MongoDB Connector for Hadoop

mongoDB
hadoop

mongoDB

@mjasay