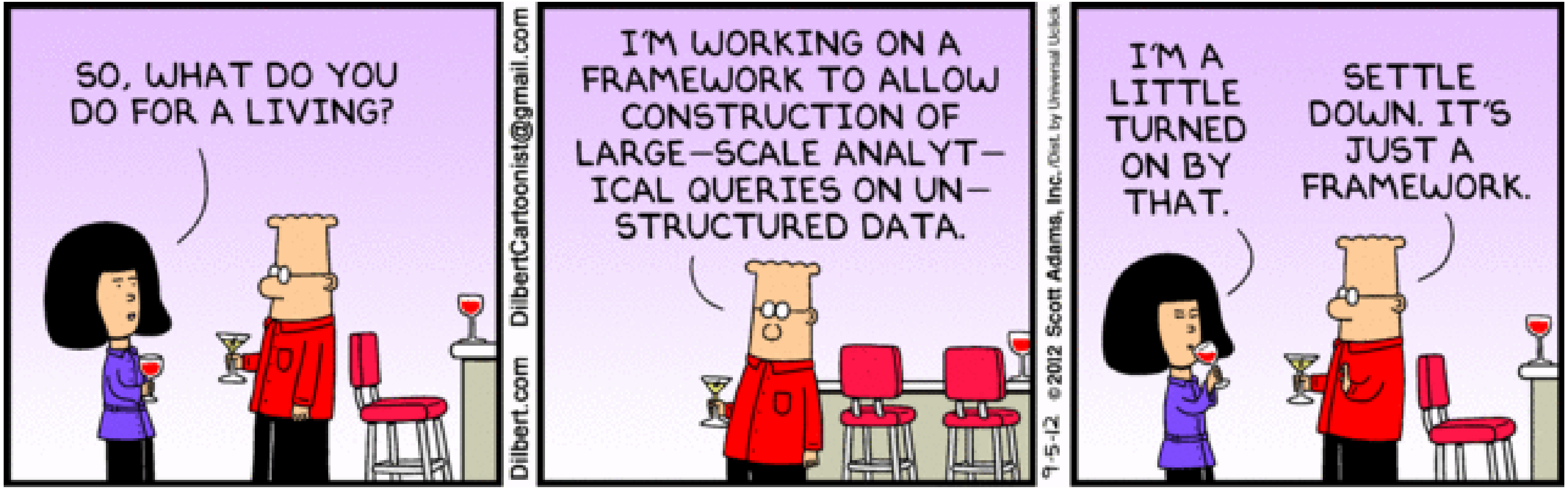
LYNCHPIN

# Big Data: Does Size Matter?

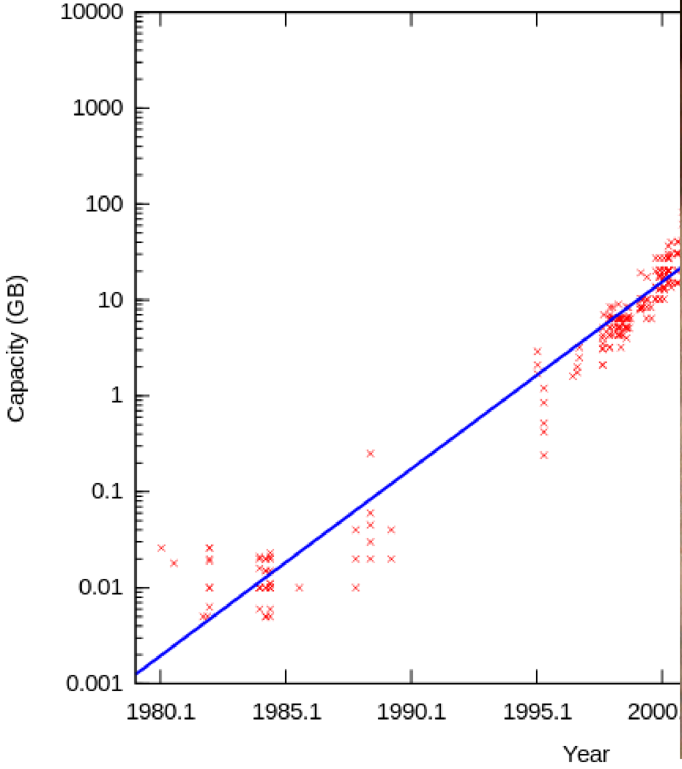
Andrew Hood
Managing Director

LYNCHPIN


SO, WHAT DO YOU DO FOR A LIVING?
I'M WORKING ON A FRAMEWORK TO ALLOW CONSTRUCTION OF LARGE-SCALE ANALYTICAL QUERIES ON UNSTRUCTURED DATA.
I'M A LITTLE TURNED ON BY THAT.
SETTLE DOWN. IT'S JUST A FRAMEWORK.
Dilbert.com DilbertCartoonist@gmail.com
9-5-12 ©2012 Scott Adams, Inc./Dist. by Universal Uclick

# Size is (Historically) Relative

LYNCHPIN



# Defining “Big Data”

| | |
|---|---|
| Volume | • Terabytes/Petabytes |
| Velocity | • Streaming Fast |
| Variety | • Unstructured |

# Defining “Big Data”

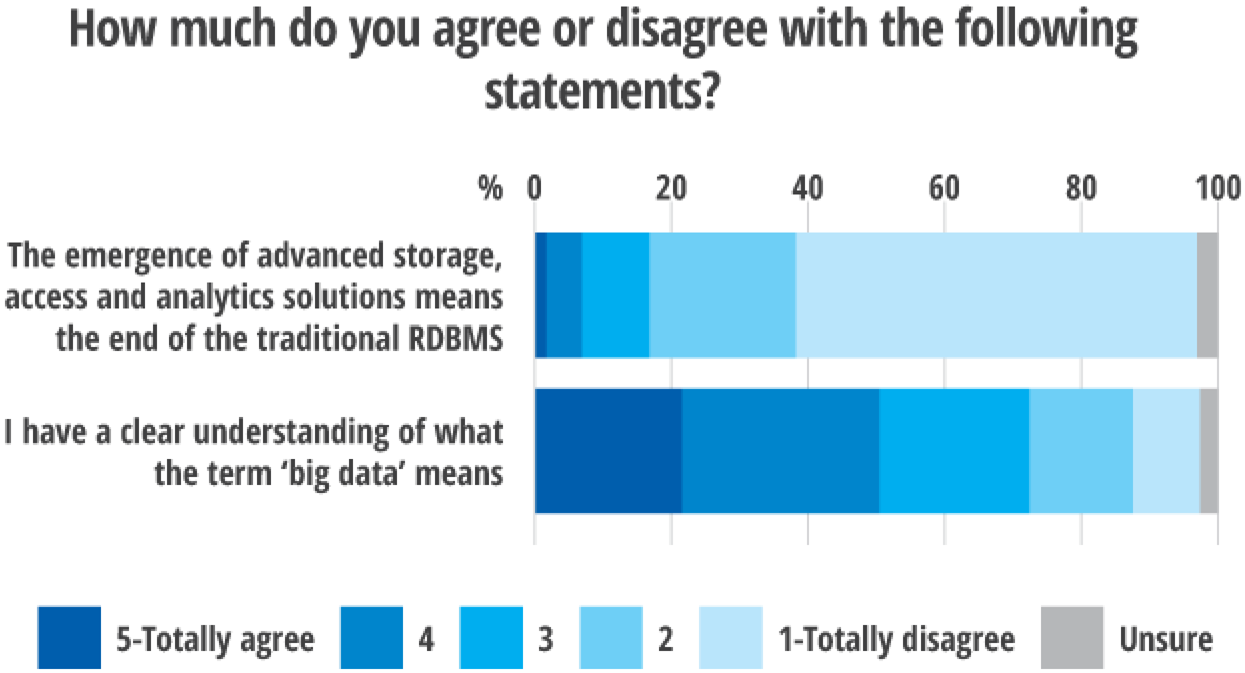
LYNCHPIN

- “Big data is any data that: *doesn’t* fit well into tables and that generally responds poorly to manipulation by SQL.”

*Mark Whitehorn*

*Chair of Analytics at the University of Dundee*

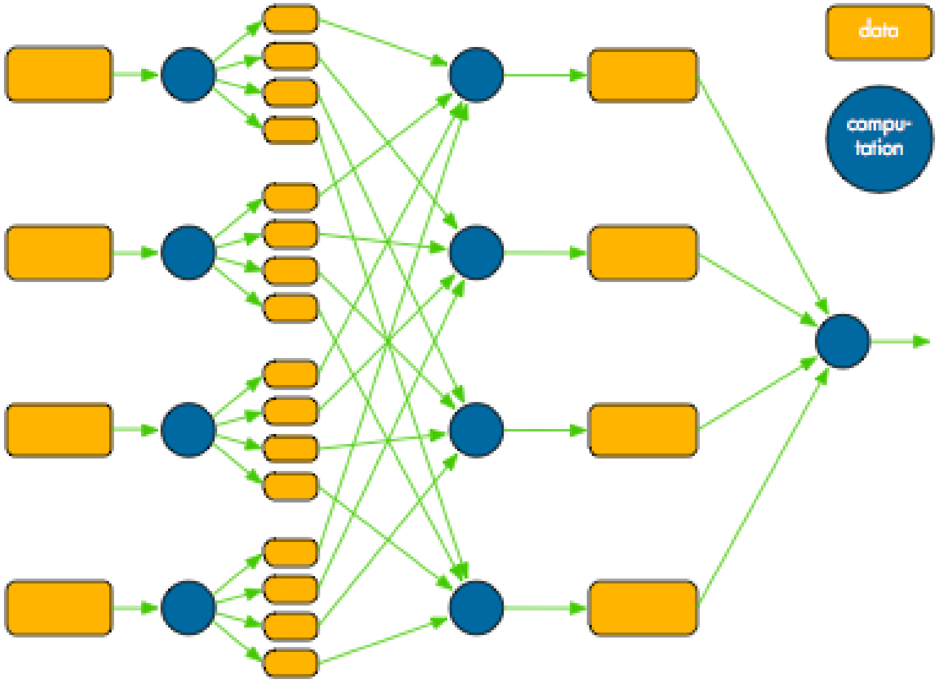
# View of the IT Professional



How much do you agree or disagree with the following statements?

The Register/FreeformDynamics Aug/Sep 2012
[http://www.theregister.co.uk/2012/10/08/big_data_revolution/](http://www.theregister.co.uk/2012/10/08/big_data_revolution/)

# Big Data Innovations



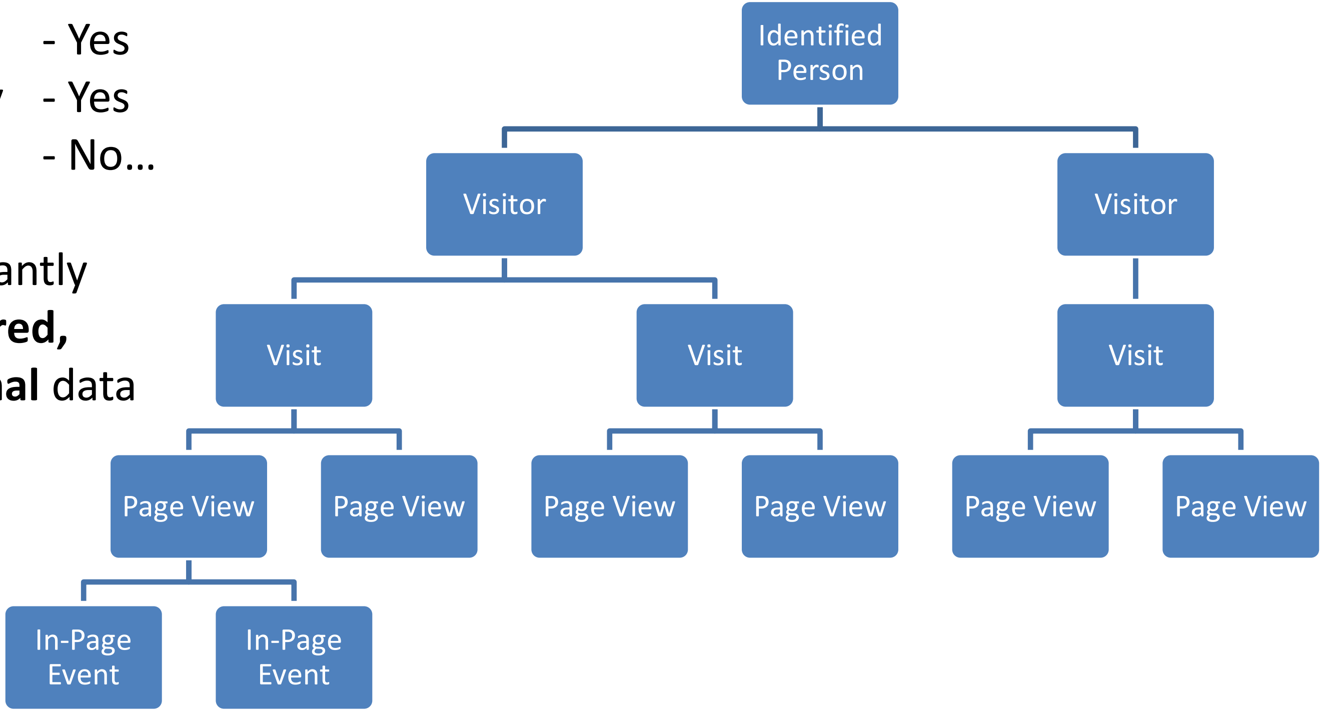
- **MapReduce**
  - Developed by Google
  - Ideal for distributed computation
  - Works very well for building search engines…

# Is Web Analytics “Big Data”?

LYNCHPIN

| | |
|---|---|
| Volume | - Yes |
| Velocity | - Yes |
| Variety | - No… |

It’s blatantly **structured, relational** data

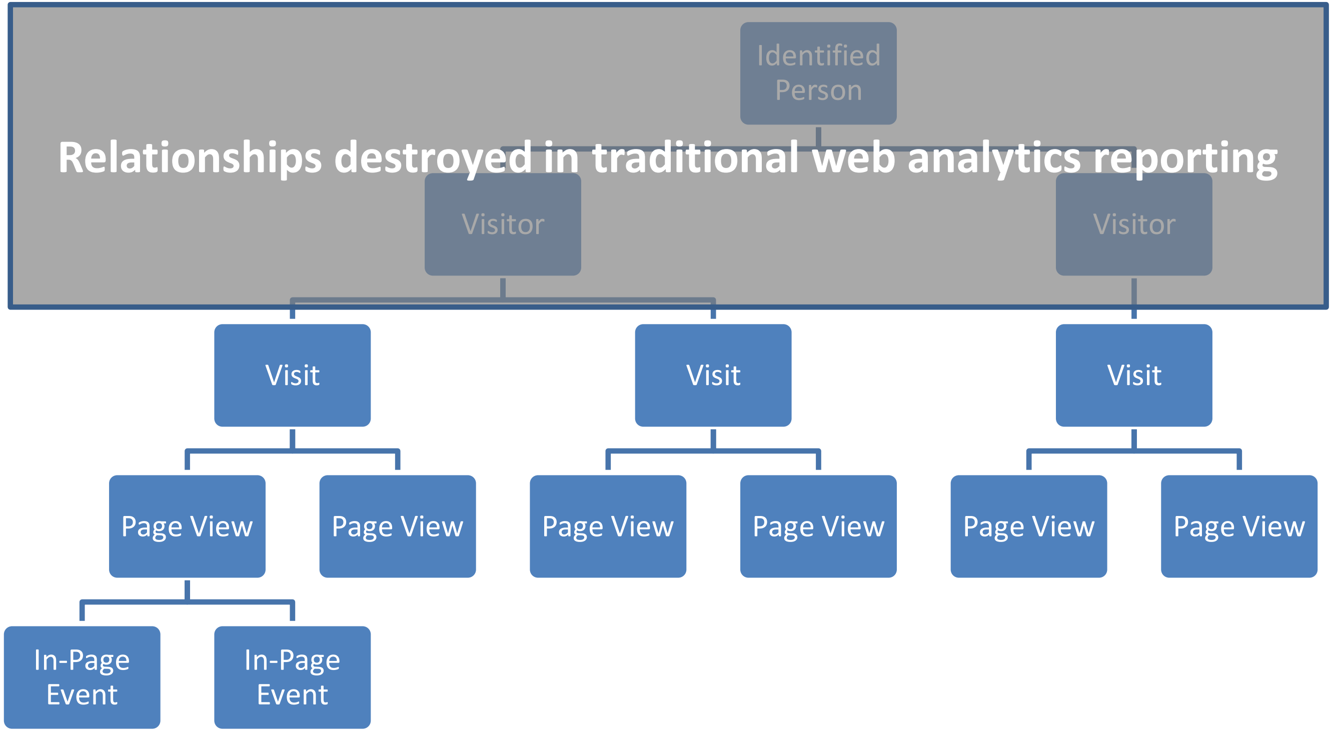


# So if “Big Data” isn’t the answer…

…how do we get more value out of web data?

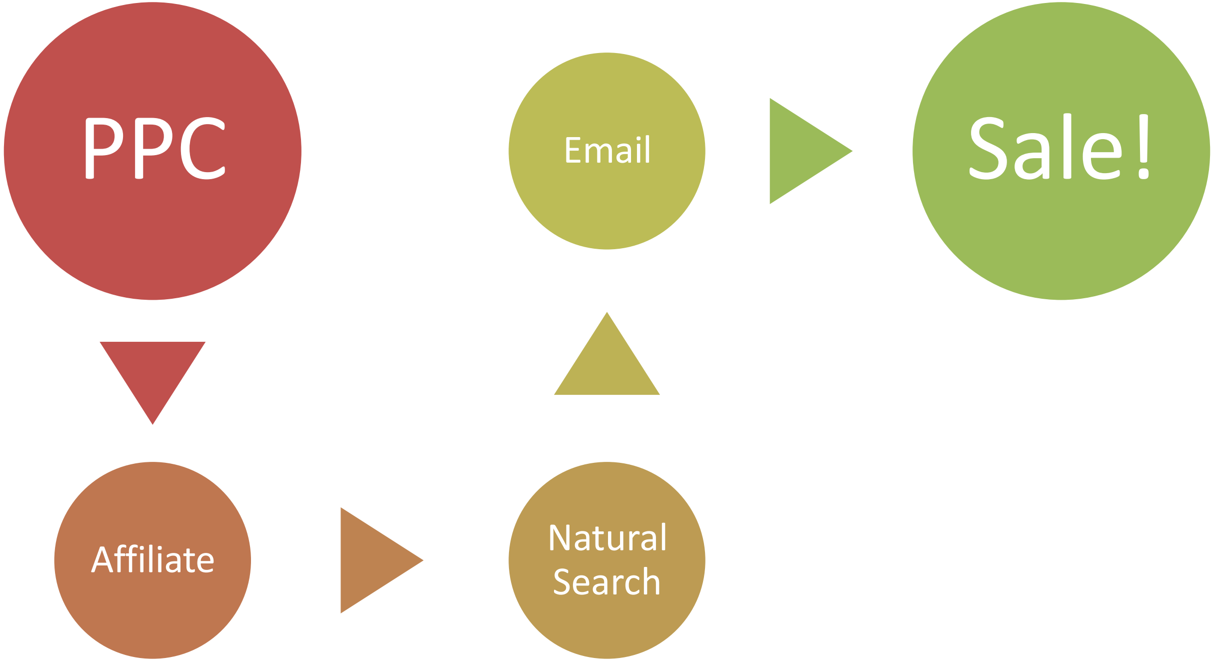
1. Move beyond session-based models/metrics
2. Extend our view of “attribution”
3. Use relational databases properly
4. Apply some good old statistics

# Curse of Session-Based Models



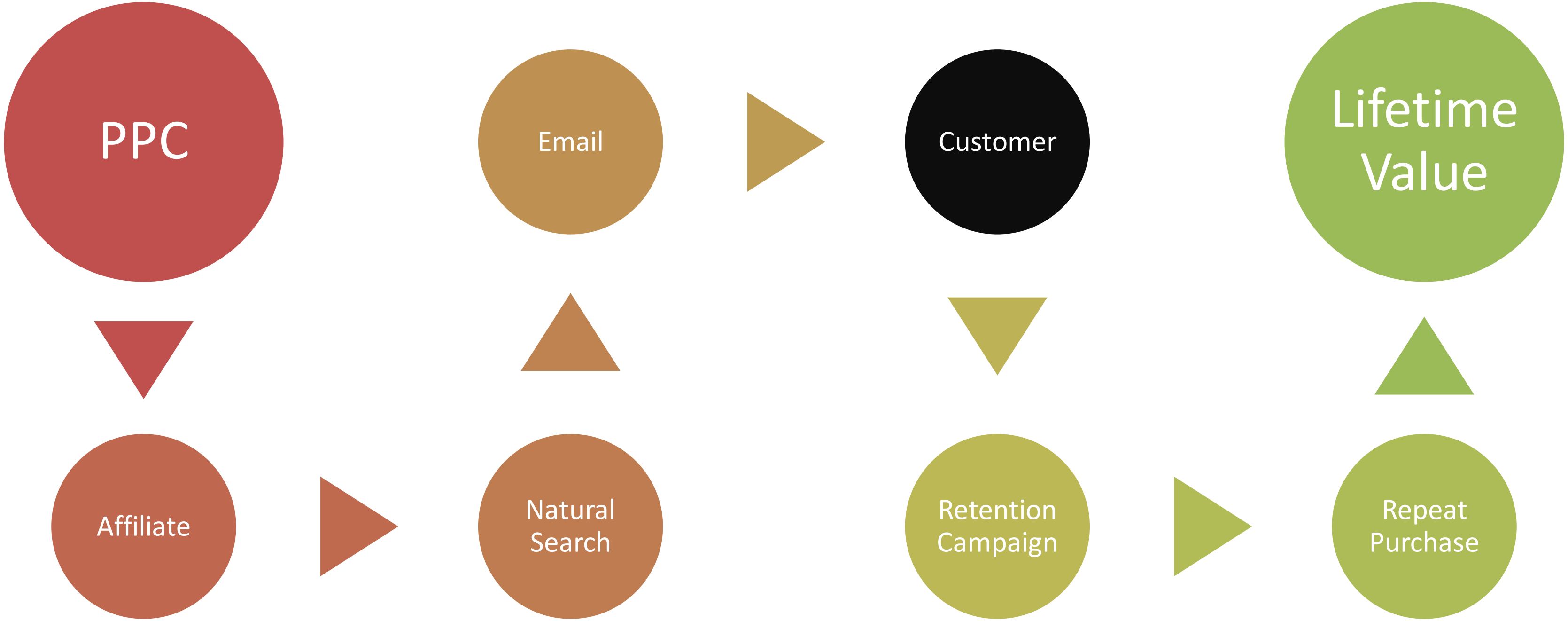
# Narrow View of Attribution

LYNCHPIN



# Narrow View of Attribution

LYNCHPIN



PPC

Affiliate

Natural Search

Email

Customer

Retention Campaign

Repeat Purchase

Lifetime Value

# Relational Databases

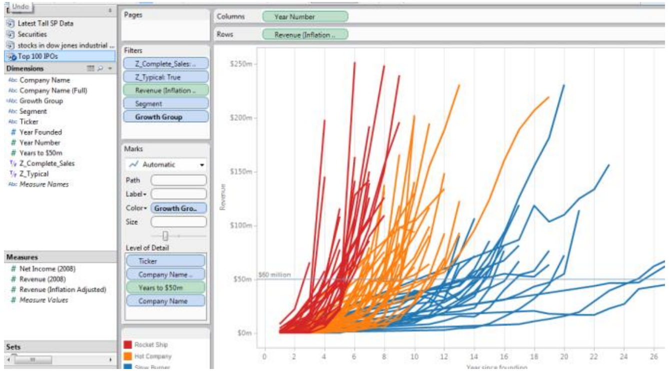
LYNCHPIN

## Free and Powerful

- E.g. PostgreSQL
  - 15 years old
  - Runs on Windows, Mac, UNIX
  - Feature competitive with Oracle
  - Cost: £0
  - In 2008, Yahoo! already had a 2 Petabyte data warehouse based on PostgreSQL processing 24 billion events per day

## Easy to Use

- Not everyone speaks SQL
- Whole host of data interrogation/visualisation tools out there (e.g. Tableau)



Columns: Year Number

Rows: Revenue (Inflation ..

Filters: Z_Complete_Sales: .. | Z_Typical: True | Revenue (Inflation .. | Segment | Growth Group

Marks: Automatic

Color: Growth Gro..

Level of Detail: Ticker | Company Name .. | Years to $50m | Company Name

# Statistics

LYNCHPIN

- “Big Data” stores do not have magical built-in analytical capabilities
  - (Exception: some standardised algorithms for things like fraud detection are emerging)
- Making sense of data big and small is going to need some established statistical techniques:
  - Propensity modelling
  - Association/correlation analysis
  - Identifying statistically significant changes/trends

# Convergence

LYNCHPIN

- Common complaint in digital is the struggle to recruit decent “web analysts”
- By contrast, there is an established industry of data analytics with skills in…
  - Relational databases
  - Statistical modelling
- If less of our web data was locked up in proprietary data models, those skills suddenly become exceptionally valuable

# Summary

LYNCHPIN

- Take a reality check on “big”
  - CPU and storage capabilities growing *much* faster than data points in clickstream
- Not everything is unstructured
  - In fact, most web data is highly structured and relational (the opposite of “big data”)
- Established systems and skills are going to be key to unlocking more value in the short-medium term
  - Relational databases and BI slice-and-dice tools
  - Statistical modelling techniques