Big Data: Does Size Matter?

Andrew Hood
Managing Director
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.
Size is (Historically) Relative

Just one MicroSD card stores more than the rest combined...

25 years of storage
Defining “Big Data”

Volume
- Terabytes/Petabytes

Velocity
- Streaming Fast

Variety
- Unstructured
Defining “Big Data”

• “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 emergence of advanced storage, access and analytics solutions means the end of the traditional RDBMS

I have a clear understanding of what the term ‘big data’ means

The Register/FreeformDynamics Aug/Sep 2012
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”?

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

Relationships destroyed in traditional web analytics reporting
Narrow View of Attribution

PPC → Email → Sale!

Affiliate → Natural Search
Narrow View of Attribution

PPC
Affiliate
Email
Natural Search
Customer
Retention Campaign
Repeat Purchase
Lifetime Value

15 October 2012
© Lynchpin Analytics Limited, All Rights Reserved
Relational Databases

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)
Statistics

• “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

• 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

• 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