BIG DATA: A COMMERCIAL INTRODUCTION

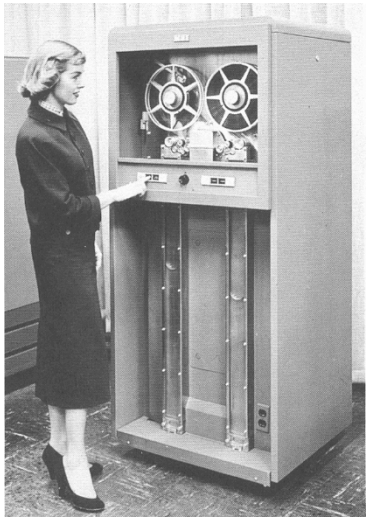
Phil Claridge, Virtual CTO, Mandrel Systems

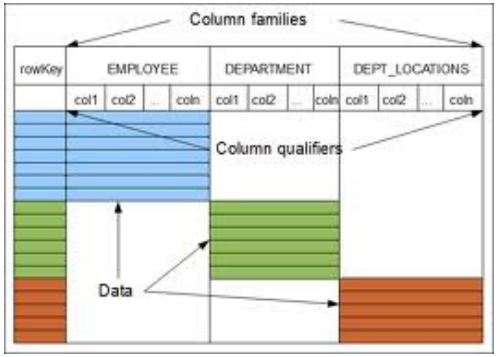
Version 1.5

phil@mandrel.com
www.philclaridge.com

What is Big Data ?

- What is big data
  - Much more data than can be processed on a single machine or database.
- Not a new problem
  - Mainframe computers had big data problems 40 years ago.
- What is different now
  - The volumes and velocity of generated data.
  - The emerging technologies.
  - The commercial opportunities.
  - The legal concerns.





Big Data – Now & Future

Events & People

- Bank transactions
- Shop purchases.
- Itemized phone bills
- Tweets
- Web page hits.
- Scientific – genetics / CERN.

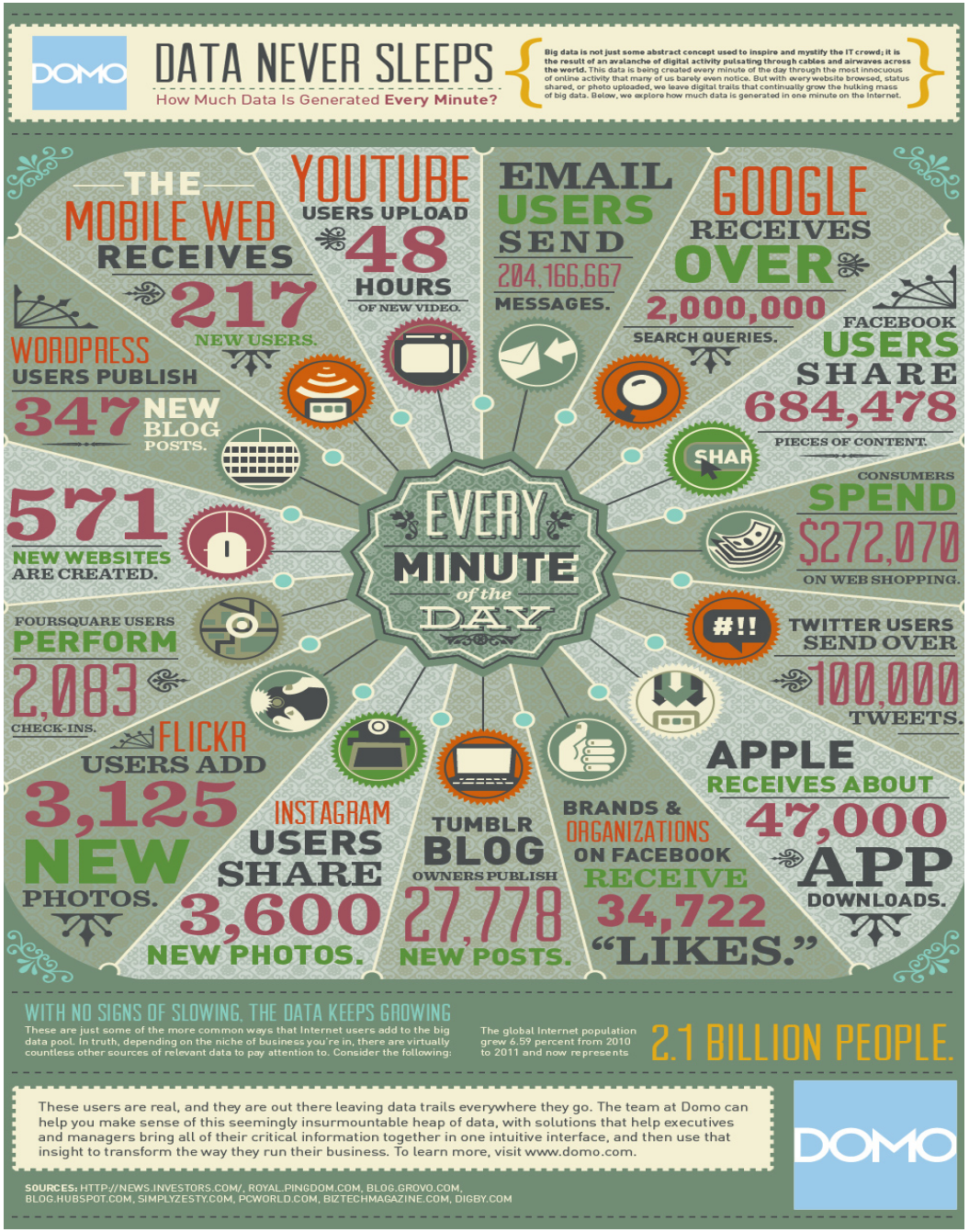
Machines To Machine

- Automatic trading.
- Surveillance.
- Smart energy
- Home sensors (elderly/ medical).
- CCTV processing.
- ANPR / Road tolls.
- ‘The Internet Of Things’

Big Data & The Internet

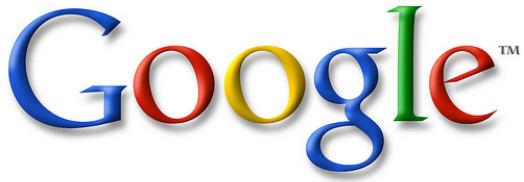
- Per minute:
  - 2,000,000 searches
  - 100,000 tweets
  - 34,722 new likes
  - 48 hours of video

http://www.visualnews.com/2012/06/19/how-much-data-created-every-minute/

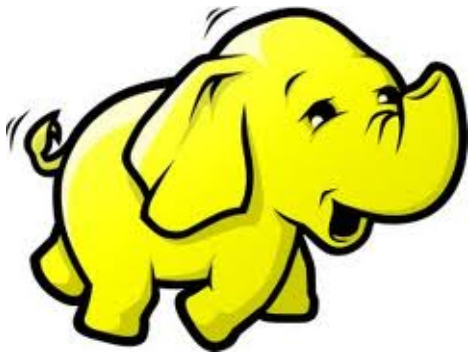


Google "Kickstarts" Big Data

- Google wanted to build search index and end user profiles.
  - Developed their own technology published a paper on *'Big Table'.*
  - Infrastructure called Colossus.



- Yahoo and others needed to compete.
  - Developed Open Source Java version of Google's Big Table.
  - This is a processing engine called *Hadoop* running across multiple machines running a storage engine called *Hbase.*
  - Most database and business intelligence technology providers rush to integrate Hadoop.



- But there is more to big data than Google derived Big Data development.
  - See Forbes big data landscape reproduced next …
    - http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence

ORACLE | Hyperion
SAP Business Objects RJMetrics
Microsoft | Business Intelligence
IBM COGNOS birst
MicroStrategy
Autonomy
bime
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS TIBCO KARMASPHERE
panopticon
Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx visual.ly AYATA

Data As A Service

factual
kaggle
knoema
GNIP DATASIFT Windows Azure Marketplace INRIX LexisNexis LOQATE

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
ParAccel
EMC² GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

Couchbase
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

Technologies

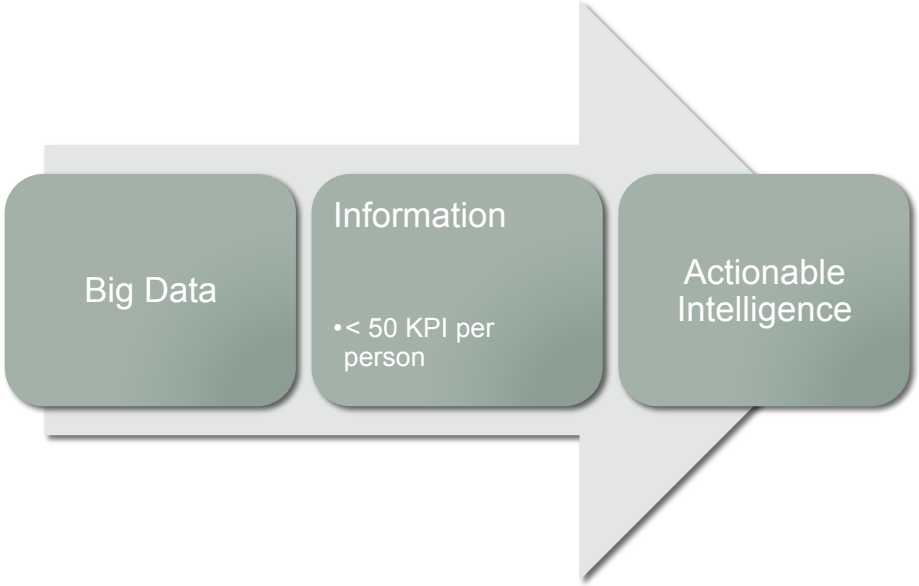
hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

dave@vcdave.com
blogs.forbes.com/davefeinleib

Why Keep Big Data

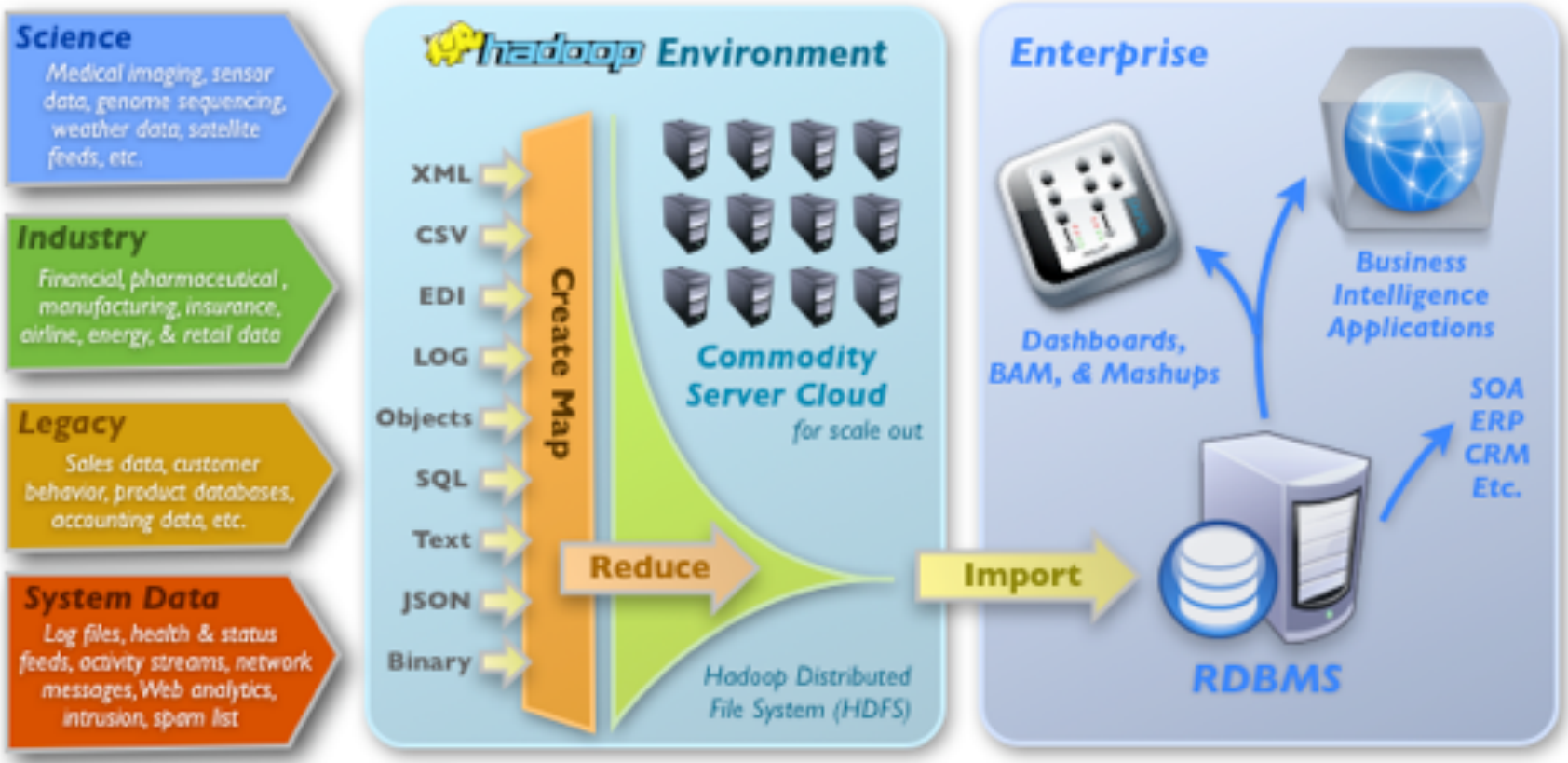
- You don't know the questions you want to ask
  - Data you choose to keep. So save everything just in case. E.g.
    - Every item you buy at Tesco.
    - Every Google search.
- You have huge volumes of data you must keep.
  - Medical records.
  - Share trades.
- Top four benefits (from Ventana)
  - Allow us to retain and analyze more data (74%)
  - Increase the speed of analysis (70%)
  - Produce more accurate results (61%)
  - Reduce or eliminate manual processes (59%)

Big Data – Useless Unless Processed



- Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
- Convert 350 billion annual meter readings to better predict power consumption
- Scrutinize 5 million trade events created each day to identify potential fraud
- Analyze 500 million daily call detail records in real-time to predict customer churn faster.

Using Hadoop in the Enterprise



From http://www.ebizq.net/blogs/enterprise

Two Processing Approaches

"Filter The Pond"

- Pour all of your data into an archive.
- Formulate questions.
- Gather results.
- Answers may take hours or days – not 'real time'.
- But you can go back on history ands ask new questions.

"Filter The Pipe"

- Sift though data as it arrives.
  - E.g. personal or vehicle location.
- Generally only good for questions you have anticipated.
- Real time notifications possible.
- Products called "complex event processing"

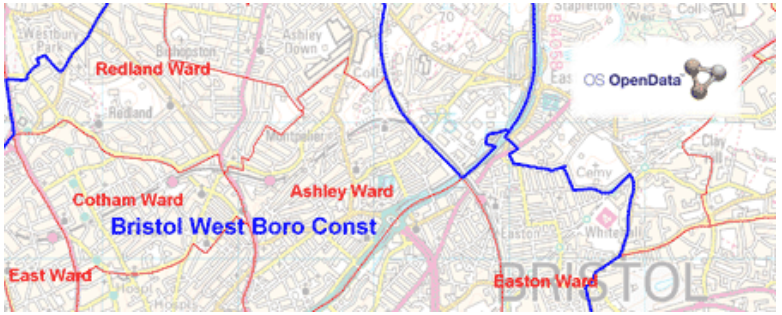
- Lack of real time support for 'Big data' is major technical challenge.

Big Data - Governance

- Separate what you must and would like to keep.
- Critical to evaluate risks and benefits of big data
  - Many companies are choosing to delete some big data
    - E-mail deletion policies.
- You have to protect what choose to you keep.
  - Customers may have rights to see what you keep.
  - Government may legislate access.
  - Disasters can happen … many disasters involve water (slow to fix)

Free Big Data Sets

- Free Big Data sets are being released.
- For example:
  - Ordinance Survey's "Open Data"
    - All UK boundaries
    - Lat/long of all UK postcodes
  - Transport for London
    - Traffic, underground, bus arrival, cycle docking.
    - http://www.tfl.gov.uk/businessandpartners/syndication/default.aspx
  - HESA
    - Student information.
    - See information publised vua www.bestcourse4me.com
- See also information indexed from
  - UK government data – 8,700+ datasets
    - http://data.gov.uk
  - Guardian Datablog
    - http://www.guardian.co.uk/news/datablog
- Pressure on governments to release more data.
- Be aware of emerging rights for individual to access data.
  - E.g. energy companies may have to release customer's data in electronic form for usage consumption comparison.



Big Data In You Pocket

- How does your business change when massive quantities of data can be held in your hand?

- For example ..
  - Registration of every car with tax and insurance status.
  - Every item on plant on the road or rail network.
  - Picture of everyone in the UK (future?)
  - Batch number and usage date of every package of drugs..

- What are the unexpected consequences?

The Future - Information Trading?

- Example
  - Supermarket trades purchase information with supplier.
  - Telco sells profiles top advertiser.
- Emerging business opportunity?
- Real world challenges
  - Lack of clear business models.
  - Information distribution control / privacy.
  - Merging data – no common identities.

It Can All Go Wrong

By Unexpected Consequences

- In 2006 AOL released detailed search data.
- It was anonymised.
  - But searches could be associated together.
- Users were identified.
- AOL's CTO was fired and lawsuits followed.

- http://en.wikipedia.org/wiki/Data_breach

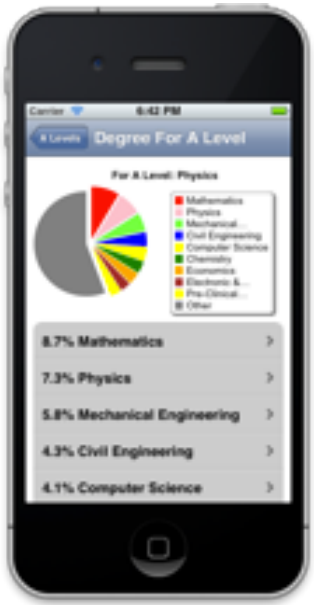
Hacking / Criminal Activity

- Password hacking.
  - LinkedIn.
  - Sony
  - eHarmony
- Verizon cataloged 855 data theft incidents in 2011

- http://www.bbc.co.uk/news/technology-17428618

Big Data - Personal Perspectives

- Best Course 4 Me
  - Take education and career information for every UK student.
  - Provide best advice for A levels, degrees based on detailed student and grade information.
  - For any parents with teenagers go to …
    - http://www.bestcourse4me.com
  - Phil helped architect this charitable solution and is now condensing the big data to a number of mobile applications for iPhone and Android.



- Telco (Geneva)
  - Processing every call, SMS, and data session for Tier 1 telcos - 100,000s events/second.
  - Phil was Chief Architect for the Geneva billing system used by BT, O2 and 120 other carriers worldwide.

Some Recent Big Data Quotes

- Oracle exec says "we're not competing with Amazon for Netflix, we're competing with Amazon for Boeing,”
- "Most software looks more like a whirlpool than a pipeline.”
- When someone says "Big Data," I always check to see if I still have my wallet.

- http://highscalability.com

Thank You

Please contact me if you want more insight on big data or developing for the cloud?

Phil Claridge, Virtual CTO
Mandrel Systems, www.mandrel.com

Also phil@mandrel.com and www.philclaridge.com and
http://www.linkedin.com/in/claridge

Summary Bio …

Currently Phil is having great fun working within Mandrel Systems a Cambridge, UK consultancy and bespoke/boutique software development organisation. Within Mandrel he provides his skills as part time 'Virtual CTO', 'Chief Architect' and consultant to a number of interesting pre-IPO companies.

Notably Phil was Chief Architect for Geneva (billing software licensed with over 120 carriers installed worldwide) and represented Geneva in the trade sale to Convergys for $692m. Originally graduating in electronics, Phil's broad experience spans hardware and embedded software design, LAN and WAN product design, technology outbound sales, M&A and IPR licencing and management, and more recently large-scale software product design serving millions of users. Phil therefore has a rare perspective in being able to influence real world products ranging over hardware device to high scale software solutions, embracing new technologies including multi-tenant cloud-based deployments and big data solutions.

Recommended Reading

- Ventana Research
  - http://www.ventanaresearch.com/uploadedFiles/Content/Landing_Pages/Ventana_Research_Big_Data_Benchmark_Research_Presentation.pdf
- Bringing Big Data To The Enterprise
  - http://www-01.ibm.com/software/data/bigdata/
- David Feinleib's Blog
  - http://blogs.forbes.com/davefeinleib/