DATA MANAGEMENT • BIOSTATISTICS • PROGRAMMING • MEDICAL WRITING • PHARMACOVIGILANCE

Quanticate

A Passion For Excellence

Global Solutions from the World's Largest Data-Focused CRO

The myth of the Big Data silver bullet

Nick Burch, CTO

Confidential, Copyright © Quanticate

Quanticate
A Passion For Excellence

Setting the Scene

Confidential, Copyright © Quanticate

About me

- Nick Burch, CTO at Quanticate
- Quanticate is a Data-Focused CRO
- Quanticate helps with Clinical Trials, esp. data
- We're not a Big Data vendor!
- But we're increasingly using Big Data systems as we do our “normal” work on Clinical Trials
- Frequent speaker at Big Data events

Confidential, Copyright © Quanticate

What is Big Data anyway?

- Kinda the whole point of the conference!
- Definitions vary, but...
- Loosely defined as more than can be processed on a handful of machines with traditional methods
- Typically comes with a lower cost for storage
- Typically comes with high scalability
- Typically comes with trade-offs up-front

Confidential, Copyright © Quanticate

Why now?

- VC funding in Big Data is at an all-time high
- Many of these new Big Data companies have large valuations, and aggressive growth plans
- Moving from pure-technology plays into business focused and non-tech suitable offerings
- Market is maturing with clear winners showing
- Support and consultancy more widely available

Confidential, Copyright © Quanticate

Why here?

- When I first started going to Big Data events, it was all about how to do the basics!
- Used to be all “techie to techie” talks
- Initially driven by overwhelming business needs

- Now it’s easier to get started, including for non-technical people, eg Data Scientists

Confidential, Copyright © Quanticate

But not...

- I can't tell you the “ideal” Big Data solution, as it is different for everyone
- There is no silver bullet... Despite vendor claims!
- Exactly how to pick a solution – it's a process, not an equation
- Much about Quanticate – we're users of this, not developers

Confidential, Copyright © Quanticate

Key things

- What sorts of Big Data things are available
- What kinds of questions you need to ask yourself
- What kinds of questions you need to ask of potential vendors
- Some other things to consider

- (Why conference sponsors tend to hate me...)

Confidential, Copyright © Quanticate

Quanticate
A Passion For Excellence

Key kinds of Big Data solutions

Confidential, Copyright © Quanticate

Some broad classes

- Low level – distributed block storage, distributed locks, consensus algorithms, leader election etc
- Job scheduling, tracking and execution – things like Apache: Hadoop, Messos, Spark, Kafka
- Data tracking, data workflow, data lifecycle, metadata management - “data plumbing”
- Security, identity, auditability
- Operations information (status, availability etc)

Confidential, Copyright © Quanticate

More interesting broad classes

- Column Stores
- Document Stores
- Object Databases
- Graph Databases
- Key-Value Stores
- Big Data Warehouse Systems
- Distributed Computation Systems

Confidential, Copyright © Quanticate

Or looking another way

- Transactional or Eventually Consistent
- Partition Tolerant / High Availability
- High Write performance
- High Read performance
- Streaming processing
- Batch processing
- High scalability

Confidential, Copyright © Quanticate

Variety

- Wide range of solutions
- Tailored to different problem domains
- Solving those well
- But not always so generally
- Widely used in “big name firms”
- But is what's right for Amazon right for you?
- Is what's right for Facebook solving your issues?

Confidential, Copyright © Quanticate

Quanticate
A Passion For Excellence

Requirements

Confidential, Copyright © Quanticate

It all used to be so simple...

- For a time, data storage with computers was hard, and everything was custom
- Then we moved towards relational databases, queried and populated using SQL
- DBAs helped us organise our data
- Requirements were just about cost, scalability, speed and support, everything was SQL

Confidential, Copyright © Quanticate

NoSQL and Big Data

- NoSQL movement is 7 or 8 years old now
- Label covers a number of Big Data systems which are non-relational, don't use SQL for query, but allow large volumes and/or high speed and/or distributed
- Not all Big Data systems are NoSQL, eg Hadoop, Spark, Mesos
- Not all NoSQL systems are Big Data

Confidential, Copyright © Quanticate

The return of SQL...

- SQL is the language of choice for working with relational database systems
- Originally, SQL = Relational
- But SQL is actually a general data query system, designed to be used by non programmers
- Many of the NoSQL features are supported by SQL (though not relational databases)
- SQL proving popular for querying NoSQL today!

Confidential, Copyright © Quanticate

Requirements: To Consider

- Data loading – how much, how often?
- Volume – how much now, how much growth?
- Querying – batch? real-time? small subsets? large swathes? simple fetches? Aggregates?
- Availability and Complexity – does downtime matter, and how much work to keep going?
- Reproducibility, Data Integrity – lab data may differ from DNA sequences or crystal structures

Confidential, Copyright © Quanticate

Requirements: To Consider

- Heterogeneity – How similar is all your data?
- Consistency – Even when the same type, how consistent is it between data sets?
- Structured vs Un-structured data?
- Changes – How do you anticipate needs changing over time?
- IT Support – How will it fit with what you already have, what IT will support, how hard for them?

Confidential, Copyright © Quanticate

Quanticate
A Passion For Excellence

Validation

Confidential, Copyright © Quanticate

Validating your solution

- Big Data solutions are tested before release
- But tested != validated...
- No FDA certified solutions for Big Data
- Validation is Domain Specific, just because it works for one (eg Clinical Trials), doesn't mean it's fine for all others (eg Banking, Sales)
- Big Data systems are too large to test by hand
- Documentation, process, automation

Confidential, Copyright © Quanticate

Industries are special!

- But not always in a good way....
- Requirements for validation are well known within one industry, can vary greatly between
- Many words have different meanings between Industries, eg CSV in Pharma != CSV elsewhere
- Make sure you understand your regulations
- And doubly make sure your suppliers do to!

Confidential, Copyright © Quanticate

Quanticate
A Passion For Excellence

Pharma and Big Data

Confidential, Copyright © Quanticate

Very mixed use

- Drug Discovery was using Big Data, back before it was a named thing!
- Drug Development only just starting to use

- Requirements wise, could pretty much be two different industries, even within the same firms!

Confidential, Copyright © Quanticate

Discovery & Big Data

- “Folding @ Home” - launched in 2000, distributed system, protein folding simulation
- Big Data widely used for simulations and filtering
- “We know X works, but why?”
- “Could Y / change-in-Y be the cause of Z?”
- “Which of these might fit / interact with this cell / hormone / virus” etc

Confidential, Copyright © Quanticate

Development & Big Data

- “We know X has an effect, but overall does it effectively and safely work on real people?”
- Most Pharma development companies think they have Big Data problems, most don’t....
- Trials with thousands of patients really don’t generate that much data!

Confidential, Copyright © Quanticate

Development & Big Data

- Wearables and Continuous Monitoring do generate lots of data
- Large Population studies do, especially if looking for small effects
- Virtual Trials & Outcomes Research, using existing datasets can
- Public Phamacovigalence eg Twitter mining
- Key issues – Validation, Permissions, Privacy

Confidential, Copyright © Quanticate

Quanticate
A Passion For Excellence

Questions for Vendors

Confidential, Copyright © Quanticate

Do they have answers for...

- How will this solve my problems today?
- And what about the ones we foresee tomorrow?
- How can we validate this?
- And do your really understand CSV?
- How will this work with my structured data?
- And my increasing unstructured data too?
- How can my IT team deploy this?
- If you vanish, who can help me then?

Confidential, Copyright © Quanticate

Quanticate
A Passion For Excellence

Other Resources

Confidential, Copyright © Quanticate

For those who like papers...

- http://static.googleusercontent.com/media/research.google.com/en/us/archive/bigtable-osdi06.pdf
- http://static.googleusercontent.com/media/research.google.com/en/us/archive/spanner-osdi2012.pdf
- http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf
- http://www.allthingsdistributed.com/files/amazon-dynamo-sosp2007.pdf
- https://www.cs.cornell.edu/projects/ladis2009/papers/lakshman-ladis2009.pdf
- http://www.vldb.org/pvldb/2/vldb09-938.pdf
- https://accumulo.apache.org/papers/accumulo-benchmarking-2.1.pdf

Confidential. Copyright © Quanticate

Conferences

- ApacheCon US – Miami, FL – 16-18 May
- Berlin Buzzwords – Berlin, DE – 11-13 June
- OSCON – Austin, TX – 8-11 May
- Strata + Hadoop World – various

- Ask Vendors where they're speaking
- Ask at lunch where else people are going!

Confidential, Copyright © Quanticate

DATA MANAGEMENT • BIOSTATISTICS • PROGRAMMING • MEDICAL WRITING • PHARMACOVIGILANCE

Quanticate

A Passion For Excellence

Global Solutions from the World's Largest Data-Focused CRO

Thanks!

Confidential, Copyright © Quanticate