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

The myth of the Big Data silver bullet

Nick Burch, CTO
Setting the Scene
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
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
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
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
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
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...)

www.quanticate.com
Key kinds of Big Data solutions
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)
More interesting broad classes

- Column Stores
- Document Stores
- Object Databases
- Graph Databases
- Key-Value Stores
- Big Data Warehouse Systems
- Distributed Computation Systems
Or looking another way

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

www.quanticate.com
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!
Requirements: To Consider

- Data loading – how much, how often?
- Volume – how much now, how much growth?
- 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
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?
Validation
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

www.quanticate.com
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!
Pharma and Big Data
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!
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
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!
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 Pharmacovigilence eg Twitter mining
• Key issues – Validation, Permissions, Privacy
Questions for Vendors
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?
Other Resources
For those who like papers...


www.quenticate.com

Our Values: Excellence • Customer Focus • Team Work • Passion • Integrity

Our Services: Biostatistics • Clinical Programming • Clinical Data Management • Medical Writing • Pharmacovigilance
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!
Global Solutions from the World’s Largest Data-Focused CRO

Thanks!