Putting “big data” to use

HP Discover London 2015

Analytics, technology and information risk

Colin Butcher CEng FBCS CITP
Technical director, XDelta Limited
www.xdelta.co.uk

xdelta

Personal background

- Systems architect specialising in mission critical systems
- Engineering background (printing, power generation)
- Wide range of experience (aerospace, healthcare, finance, transport, power and energy)
- Strong interest in mentoring and teaching

xdelta

XDelta – what we do

- Lead mission-critical systems projects:
  - Strategic planning
  - Technical leadership
  - Project direction
- Minimise risk of disruption to business:
  - Design for change while in continuous operation
  - Prepare in advance for ease of transition
- Ensure long term success through skills transfer

xdelta

Agenda

- Why is “big data” so interesting ?
- Technology – performance, availability, security
- Information risk
- Discussion

xdelta

Part 1 - The information game

- It’s nothing new!
- Making “better” decisions
- Getting information in time to make use of it before someone else does so
- Knowing that it’s good information, not bad or misleading

xdelta

“Data! Data! Data!” he cried impatiently. “I cannot make bricks without clay.”

- Sherlock Holmes – the “analytics processor”
- Gathering data from many sources
- Observation, knowledge, minute attention to detail
- Filtering out the irrelevant, making sense of conflicting data
- Arriving at “the truth”

xdelta

How can we define “big data” ?

Big data is an all-encompassing term for any collection of **data sets** so **large** and **complex** that it becomes difficult to process them using traditional **data processing applications**.

xdelta

What does this imply ?

- Lots of data
- Lots of data in lots of places
- Lots of different kinds of data
- More data than anybody has ever seen or imagined
- Unwieldy - almost impossible to manage

xdelta

Big data is an ill defined subject

- The ~~five~~ six “V”’s:
  - Volume – lots and lots of it
  - Variety – many data types and sources
  - Variability – wildly different qualities from different sources
  - Velocity – most of it transient and changing over time
  - Veracity – some of it could be deliberately misleading
  - Visibility – we may not have access to all of it

xdelta

Analytics

- Explaining what happened
- Predicting what is likely to happen
- Operating on
  - Data at rest – stored somewhere, not necessarily here
  - Data in motion – streaming, catch it as it flies past
  - Multiple disparate data sources
- Looking through historical data - what is hidden there ?

xdelta

Applying technology to solving problems

- Integrated data sources and huge effort:
  - Intelligence and cryptography
  - Engineering: stress analysis, fluid dynamics, “big physics”
  - Healthcare: genome sequencing
  - Weather prediction
  - Finance industry
  - Marketing

xdelta

Decreasing cost of entry

- Confluence of many things at “affordable” cost:
  - Communications
  - Processing
  - Storage
  - Integration of many data sources
- Pervasive and easy access to information
- Available to all, not just nation states and large corporations

xdelta

A world of non-stop rapid change

- Costs of entry are lower
- Possibilities are greater
- Risks are bigger
- Timescales are shorter
- Data sources are more diverse
- Results are wider spread

xdelta

The “Wild West”

- A global “mass market”
- No constraints, no standards
- Legislation failing to keep up with pace of change
- Blurring of national boundaries and currencies
- Who’s held responsible when things go wrong ?

xdelta

What's next ?

- Many uses we hadn't previously imagined
- Shift in the way people interact and behave
- New ideas emerging without being constrained by old ways of doing things
- Ever faster spread of information and disinformation

xdelta

Part 2 – Advances in technology

- Storage – greater capacity, faster
- Communication – faster, new protocols
- Processing – increased parallelism

- Merging of functionality – storage and memory

- Locality – where is everything ?
- Mobility – access from wherever we are
- Security – controlling access and detecting problems

Handling big data

- Too big for a single computer system to process
- Too big for a single person to comprehend
- Too big to know if it's been compromised
- Too big to backup and restore
- Need standards to make things easier
- Need new ways of working

xdelta

Parallelism and scalability

- Can our workload break down into parallel streams of execution ?
  - Algorithm design, e.g. Map / Reduce
- How to automate the breaking down of the workload, the distribution of work elements and the combining of results ?
  - Distributed processing and data, e.g. Hadoop framework
- Non-linear scaling:
  - Overheads eat up performance and capacity as you add more resources

xdelta

Exploiting technology

- Abstraction layers
- Operating systems
- Virtualisation
- Cloud services
- Languages / compilers (e.g. Java, C++)
- Frameworks (collections of routines and templates)

xdelta

Availability and performance

- How do we build a highly available system from a collection of inherently unreliable components ?

- A system that doesn't meet its performance requirements is a system that's not working properly. Performance related failures are often transient and exceedingly difficult to fully understand and resolve.

- Systems have to have sufficient capacity and performance to deal with the workload in an acceptable period of time under normal, failure and recovery conditions.

Availability and security

- Denial of service - a system under attack is not available
- How do we protect against inbound attacks ?
- How do we protect against data exfiltration ?
- How do we detect and alert problems ?

xdelta

Big computing

- Is the technical infrastructure big enough to cope ?
  - Network bandwidth and latency
  - Processing capacity
  - Storage capacity and performance

- Is the overall system capable of scaling up ?
  - How is it managed and monitored ?
  - How well is the software designed ?
  - Does it match your kind of workload ?

xdelta

Management and monitoring

- Management:
  - Managing changes to infrastructure
  - Monitoring of infrastructure
  - Systems administration
  - User administration

- Security:
  - Protection of information in transit
  - Controlling access to systems and information
  - Knowing if there's been a breach

xdelta

Part 3 – Information risk

- Assessment
- Security
- Privacy
- Governance
- Responsibility

xdelta

The world according to “Tigger”

- Opportunities abound:
  - Data-driven Science
  - Data-driven Government
  - Data-driven Medicine
  - Data-driven Finance
  - Data-driven Marketing & Sales
  - Data-driven Operations
  - Disruption of Business Models

xdelta

The world according to “Eyeore”

- It’s all going to go horribly wrong…
- Security
  - **Confidentiality**: Who is looking at your data ? Why ?
  - **Integrity**: Who is fiddling with it ? Why ?
  - **Availability**: What if you lose it ? Or lose access to it ?
  - **Privacy**: Whose data is it anyway ?
- A breach can seriously damage:
  - your organisation
  - your career

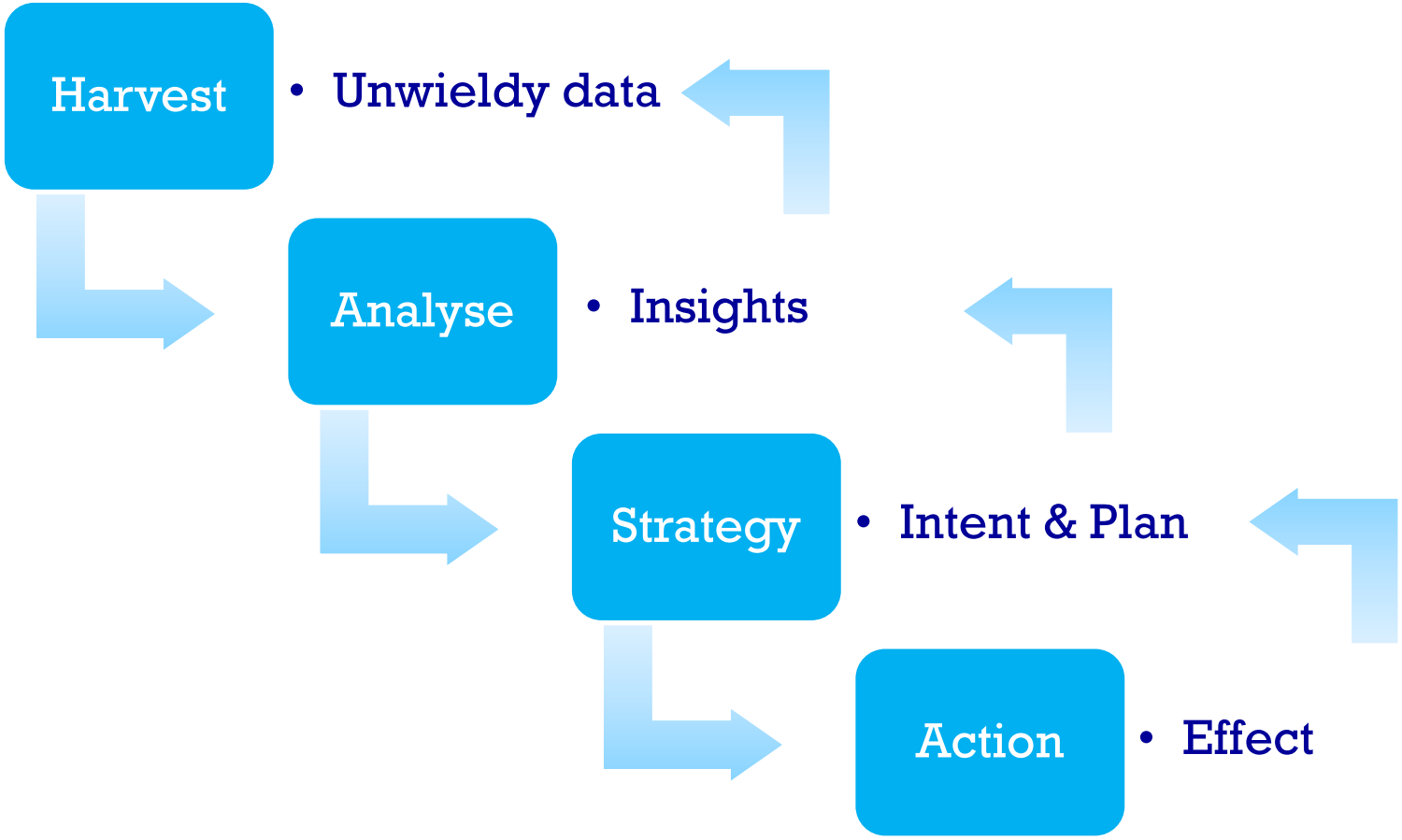
xdelta

Governance

- Understand the risks
- Manage and mitigate the risks
- Accept the responsibilities
- Invest appropriately in people and technology

xdelta

Understand where the risks are



xdelta

Ownership

- Who owns the risk ?
- Who cares ?
- Who decides ?
  - Chief Digital Officer (CDO) ?
  - Chief Marketing Officer (CMO) ?
  - Chief Operations Officer(COO) ?
  - Chief Information Officer(CIO) ?
  - Chief Technical Officer (CTO) ?
  - Senior Information Risk Owner (SIRO) ?

xdelta

Who takes responsibility ?

- There is no right answer
- Make a conscious decision, don’t default it
- Become comfortable with uncertainty
- Manage the inevitable tensions

xdelta

Summary

- Ill defined subject with great potential
- Scale can vary wildly
- Huge opportunities
- Huge risks
- Major investment over a long time
- Need outstanding leaders

xdelta

Further information

- CESG (the Information Security arm of GCHQ)
  - *Information risk management guidance:*
    https://www.gov.uk/government/collections/risk-management-guidance

- ICO (Information Commissioners Office)
  - *Publication and discussion on Big Data and Security:*
    https://ico.org.uk/for-organisations/guide-to-data-protection/big-data/

xdelta

Putting “big data” to use

HP Discover London 2015

Thank you for your participation

Colin Butcher CEng FBCS CITP
Technical director, XDelta Limited
www.xdelta.co.uk

xdelta