What is a “Data Lake”, anyway?

Alec Gardner, GM Advanced Analytics, Teradata ANZ

Wednesday 10th June 2015
One of the Big Data labels that we risk over-loading to complete abstraction is the idea of a "Data Lake"...

"A large object-based repository that holds data in its native format"

"...store all data present and future and create a centralised data archive location."

"Sometimes called the bit bucket or the landing zone"

"As more and more applications are created that derive value from... new types of data... the Data Lake forms"

"All Water and Little Substance"
One of the Big Data labels that we risk over-loading to complete abstraction is the idea of a "Data Lake"...
Big Idea #1: “store all data” (whatever “all” means)

Big Idea #2: “un-washed, raw data” (NoETL / late-binding)

Big Idea #3: “resolve the nagging problem of accessibility and data integration”

Data accessibility and integration? Isn’t that what the Data Warehouse is for?

…and at least part of the discussion sounds familiar
Is the Data Lake a new architectural construct? Or are we building next-generation data silos?

Simple, single subject area Dimensional Data Marts – with all of the dimensions pre-joined to the fact table? One-per-workload / application?

Is this really the future of Enterprise Analytics? Or circa 1995 silo, departmental Decision Support Systems warmed-over and re-platformed?
Up-front data modelling and data integration are difficult, time-consuming and expensive; maybe we just shouldn’t bother with them any more?
Explicit, or implicit, there is always, always, always (at least one) schema;
“pay me now, or pay me later (and over and over)”
For the foreseeable future, we will need multiple Information Management strategies - and multiple Information Management technologies - integration becomes a critical concern.
The “new” Big Data

The “big 5” challenges of making the “transactions, to interactions” journey

#1: The requirement to manage multi-structured data and data whose structure changes continuously means that there is no single Information Management strategy that works equally well across the entire Big Data space.

#2: Understanding Interactions requires path / graph / time-series Analytics in addition to traditional “set-based” Analytics, so that there isn’t a single parallel processing framework or technology that works equally well across the entire Big Data space.

#3: The economic challenge of capturing, storing, managing and exploiting Big Data sets that may be large; getting larger quickly; noisy; of (as yet) unproven value; and infrequently accessed.

#4: There might be a needle in one of these haystacks - but if it takes 6-12 months and $1M just to go look, I’ll never know.

#5: Getting past “so what” to drive real business value (because old business process + expensive new technology = expensive, old business process)
The “Logical Data Warehouse” is the industry’s adaptation to “Big Data”

How will you deploy? How many / which platforms will you need? How will you integrate them? And which data need to be centralised and integrated?

The Enterprise Data Warehouse Era

1. Multi-structured data
2. Interaction / observation Analytics
3. Flat / falling IT budgets, exploding data volumes
4. Agile Exploration & Discovery
5. Operationalisation

“Give me integrated, high quality data.”

The Logical Data Warehouse (a.k.a.: Unified Data Architecture) Era

1. Multi-structured data
2. Interaction / observation Analytics
3. Flat / falling IT budgets, exploding data volumes
4. Agile Exploration & Discovery
5. Operationalisation

“Centralise and integrate the data that are widely re-used and shared, but integrate all of the analytics.”
The Data Lake as it should be: a centralised, consolidated store of raw data from multiple sources

Agile acquisition...

...of raw, multi-structured data...

...efficient non-relational computation...

...and cost-effective storage of large and noisy data-sets

Now that is new, interesting and potentially very, very useful...
Repeatable process and reproducible results are the basis of experimental science...

...successful, real-world laboratories are not ungoverned environments.
The Enterprise Information Value Chain

Where did this data come from and when was it created?
What did its creator believe that it represented?
Full copy or delta file?
Part of a collection? Etc., etc., etc.

Any chain is only as strong as its weakest link
STOP PRESS: Laws of Physics Unchanged by Big Data!

Left to its own devices, does nature tend to give us a single, beautiful lake?

Or a messy patchwork of lakes, plural?

The new information management strategies and technologies are not a cure for information entropy; a well-ordered Data Lake will not just spontaneously emerge from a collection of consolidated Hadoop applications.
Architecture and Technology are only enablers

Web / clickstream
Who navigates to the website, what do they do in each session – and then afterwards within other channels?

Voice/text
Who is complaining to the call center & about what?

E-mail / Graph
Which brokers are colluding to rig markets – and with whom?

Sentiment
What are customers saying about the company / products / services on social media sites?

Process / Path Analytics
What’s the optimal process for claims or collections activity?
All Data And The Big Bit Bucket Myth
Operationalising “Exploration & Discovery” Insight

Type 1
- Discovery Analytics identifies a “broken” process;
- Fix the business process – and you are done;
- E.g.: Large US Telco used Path Analytics to understand path-to-call-centre journey.

Type 2
- Discovery Analytics enables us to identify / extend an existing model that is already being scored in the IDW;
- E.g.: a leading US Retail Bank was able to enhance its existing churn models by understanding the “path to churn”.

Type 3
- New model is dependant on data that is not yet available in the IDW and/or on Analytics (e.g.: path, graph, time-series, etc.) that cannot easily be supported there;
- Operationalisation requires run-time integration of multiple, distributed components.
Summary and Conclusions