Dr. Barry Devlin is a founder of the data warehousing industry, defining its first architecture in 1985. A foremost authority on business intelligence (BI), big data and beyond, he is respected worldwide as a visionary and thought-leader in the evolving industry. Barry has authored two ground-breaking books: the classic "Data Warehouse--from Architecture to Implementation" and "Business unIntelligence--Insight and Innovation Beyond Analytics and Big Data" (http://bit.ly/BunI_Book) in 2013.

Barry has over 30 years of experience in the IT industry, previously with IBM, as a consultant, manager and distinguished engineer. As founder and principal of 9sight in 2008, Barry provides strategic consulting and thought-leadership to buyers and vendors of BI and Big Data solutions. He is an associate editor of TDWI's Journal of Business Intelligence, and a regular keynote speaker, teacher and writer on all aspects of information creation and use.

Barry operates worldwide from Cape Town, South Africa.
AGENDA

1. The Data Lake from the Shore
   - Historical perspective
   - Monsters - mythical and real

2. Diving into the Data Lake

3. Life in the Data Lake

WHAT IS A DATA LAKE OR RESERVOIR?

A Data Lake is characterized by:
1. Collect everything ...contains all data, both raw sources over extended periods of time as well as any processed data.
2. Dive in anywhere ...enables users across multiple business units to refine, explore and enrich data on their terms.
3. Flexible access ...enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines.

Big data reservoirs...large holding areas where all kinds of data can be stored and analyzed.

Madhu Neer, Oracle
Evolution of a Meme – A Salutary Tale (Chapter 1)

**Keep it simple, short, visual**

- Words have meanings
- Metaphors make images

"If you think of a data mart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples."

James Dixon, CTO, Pentaho (Blog, 2010)

Evolution of a Meme – A Salutary Tale (Chapter 2)

**Answer questions thoughtfully**

- A consistent architecture?

"A single data lake houses data from one source. You can have multiple lakes... if you store data from many systems and join across them, you have a Water Garden, not a Data Lake."

[Booz Allen Hamilton]...putting data from multiple data sources into what they call a ‘Data Lake’ is a minor change to the original definition... In practice these conceptual differences won’t make much, if any, impact when it comes to the implementation."

James Dixon, CTO, Pentaho (Blog, Sept. 2014)
**Data Lake – More Definitions Pose More Questions**

- Should we really store everything?
- Is all data of equal value?
- Is quality and consistency no longer needed?
- Build it and they will come?
- What problem are we trying to solve?

**A data lake is a large object-based storage repository that holds a vast amount of raw data in its native format until it is needed.**

Cory Jansen, Technopedia.com

**A data lake is a massive, easily accessible, centralized repository of large volumes of structured and unstructured data.**

Margaret Rosset, WhatIfs.com

---

**Other Comments on the Data Lake: All Water and Little Substance?**

- Gartner report, G00264950, 23 July 2014
  - Nick Heudecker, Andrew White
  - [https://www.gartner.com/doc/2805917](https://www.gartner.com/doc/2805917)
  - “The main risk of using data lakes is the absence of metadata and an underlying mechanism to maintain it... the lack of which can turn a data lake into a ‘data swamp’”

- Michael Stonebraker: “Why the ’Data Lake’ is Really a ’Data Swamp’”, 22 December 2014
  - “The moral of this story is ‘don’t underestimate the difficulty of data curation.’ If you do, you will revisit... the experience of enterprises in the 1990’s concerning data warehouses.”
Evolution of a meme – a salutary tale (chapter 3)
Try for world domination

“A Data Lake can also be used to solve the problems of history and trending for workflow applications and state machines. What if these applications write their initial state into the Data Lake and then also write the change of every attribute in there as well? ...let’s log all the application events coming from the user interface tier as well.

We have the ‘Union of the State’... [and] the ability to rewind the Union of the State to any point in time... [an] Enterprise Time Machine”

James Dixon, CTO, Pentaho (Blog, Jun. 2015)

So, what is a Data Lake (or Reservoir)?

Marchitecture vs Architecture

- Defined at a very high level
- Often driven (mainly) by the Hadoop industry
- Characterized by simple mantras
  - Silos bad; reservoirs good
  - Collect everything; dive in anywhere
- Data Lake often seen as unarchitected → Data Swamp

- Reservoirs are (usually) more formally architected Lakes
  - (Depending on who you ask!)
- Data Reservoir = data delivery and management ecosystem
  - Based on multiple technologies
  - SQL, NoSQL and Hadoop stores
  - Functional components to manage, populate and use data
  - Context and metadata
  - Substantial architectural thought

Terms increasingly used interchangeably; Lakes becoming more architected
**Agenda**

1. The Data Lake from the Shore

2. Diving into the Data Lake
   - Architectural overview
   - Principal components

3. Life in the Data Lake

---

**The drive to the Data Lake looks attractive**

- Data volumes are growing exponentially!
  - We need to use much cheaper storage
  - And why not migrate existing storage too

- Data velocity is growing exponentially!
  - We need cheap, parallel processing
  - At scale

- Data variety is increasing dramatically!
  - We need the agility of non-relational systems
  - Fast and big

- Doesn’t this Hadoop approach offer all that... and for free? 😊

- But, do we really have a viable architecture and design?
**Sample Data Lake Architecture** (From Capgemini / Pivotal)

www.capgemini.com/blog/capping-it-off/2014/08/you-have-to-manage-your-data-lake-the-fallacy-of-technology-being-magic

**Data Reservoir/Lake Conceptual Architecture (IBM)**
DATA RESERVOIR/LAKE LOGICAL ARCHITECTURE

IBM DIAGRAM SOURCES

- IBM Redguide “Governing and Managing Big Data for Analytics and Decision Makers”, REDP-5120
  (http://www.redbooks.ibm.com/redpieces/abstracts/redp5120.html)
- IBM Redbook “Designing and Operating a Data Reservoir”
  (http://www.redbooks.ibm.com/abstracts/sg248274.html?Open)
CONCERNS WITH DATA LAKE ARCHITECTURES

- Light on detail
  - Aspirational in the extreme
  - Overall or in certain areas

- Product-centric
  - Based on what the vendor knows and/or wants to sell

- "Boxologies"
  - A list of low-level functions arranged as boxes

- Is this really architecture?
- We need to work from first principles

IDEAL ARCHITECTURE: 3 CONCEPTUAL “THINKING SPACES”

- Conceptual architecture: basis for conversation between business and IT
  - Integrated, Distributed, Emergent, Adaptive, Latent

- Design foundation for the biz-tech ecosystem
- Initial translation of business needs into architectural options and constraints

- Can also be read as a story: People process information
The conceptual information space

- Contains all information used by the business
- Three information axes based on key information characteristics

The tri-domain information model

- Process-mediated data
  - "Traditional" operational & informational data
  - Via data entry and cleansing processes
- Machine-generated data
  - Output of machines and sensors
  - The Internet of Things
- Human-sourced information
  - Subjectively interpreted record of personal experiences
  - From Tweets to Videos

[In the context of these domains, "data" signifies well-structured and/or modeled and "information" is more loosely structured and human-centric.]
INTRODUCING INFORMATION PILLARS

- Single architecture for all types of data/information
  - Consistent but separate data management and governance
- Data can flow as fast as needed
  - No unnecessary storage or transformations
  - (Contrast layered architecture)
  - Process-mediated data may retain some layering
- Pillars vs. silos
- Context-setting information (metadata) spans all pillars
- Three principal pillars
  - But more if needed – at physical level

- Mix/match technology to optimize for different data characteristics
  - Relational, NoSQL, Hadoop, etc.

FEEDING AND CARE OF INFORMATION PILLARS

Key process components

- **Instantiation** gathers measures, events and messages
  - From “real”, physical world
- Also responsible for transactions
  - Operational systems theoretically included in instantiation

- **Assimilation** integrates/reconciles data when necessary
  - Context-setting information (CSI) has key role

- Collaboration between instantiation and assimilation via CSI
Information pillars – high-level product overview

- Process-mediated data
  - Stability, Consistency & Reliability
  - RDBMS remain core technology
  - “New” storage and processing
    - Columnar (& compressed) to hybrid
    - Solid-state disk and in-memory
    - Massively parallel processing

- Machine-generated data
  - Speed, Size and Flexible Structure
  - NoSQL and Streaming / CEP
  - Hadoop 2.x
  - Relational can also address

- Human-sourced information
  - Soft, Large and Ill-defined
  - Hadoop, Hadoop & more Hadoop
    - Hadoop 2.x: more real-time
  - Plus… traditional ECM tools

Agenda

1. The Data Lake from the Shore

2. Diving into the Data Lake

3. Life in the Data Lake
   - Uptake in the market
   - Emerging tools
Data lake overall adoption, 2014-16

Data lake functional components, 2016

- Top 3: typical traditional operational / informational components
- How does top item fit with high percentage of replacement strategies in previous figures?
Data Lake obstacles, 2014-16

- Data freshness problems have come to fore in 2016
- Data management / linkage problems diminished
- Executive backing has improved

Data lake driving evolution in many tool areas

- Data integration / preparation / etc.
  - Driven by data scientists' wrangling
  - Moving from batch loads to streaming

- Data quality / governance
  - Dubious reliability of external data
  - Volume and velocity pose problems

- Data / information modeling
  - From needs-driven to information-driven

- Databases and data stores
  - Convergence of relational and NoSQL

- BI / analytic tooling
  - Statistics become key
  - Artificial intelligence approaches emerging

New vendors / open source to fore for now
- Difficult to predict winners / losers
- Traditional vendors are ramping up
- Data lake seen as complementary to / including data warehouse
Logical positioning of data lake and warehouse

- Serve different purposes
  - Functional – run / manage the business
  - Illustrative – predict / influence the future

- Both required
  - Optimized for different strengths
  - Warehouse = accuracy and consistency
  - Lake = timeliness and rawness

- Links between environments
  - Better than copying everything into one (or both)

<table>
<thead>
<tr>
<th>Functional</th>
<th>Illustrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accurate, consistent data</td>
<td>Timely, raw data</td>
</tr>
<tr>
<td>Discarded if outdated</td>
<td>Stored forever</td>
</tr>
<tr>
<td>Legally binding, traceable process</td>
<td>Creative, free-flowing process</td>
</tr>
</tbody>
</table>

Data Lake (or reservoir): Five key questions to ask before starting

1. What business problem are you trying to solve?
2. Should you really store everything?
3. Is all data of equal value?
4. How will quality and consistency be assured?
5. Is this "build it and they will come"?
CONCLUSIONS

- Data lake concept is maturing
  - Old marketing-driven view is outdated
  - Original issues still need to be addressed

- Data lake architecture
  - Beware of vendor-driven architectures
  - Business unIntelligence architecture also fully addresses data lakes

- Uptake growing, products emerging
  - Significant market interest
  - Vendors addressing shortcomings

- Key question
  - Make sure you know why you’re diving in!
ADDITIONAL RESOURCES

All white papers and events: www.9sight.com

Articles and Blogs at:

- Eckerson Group: http://eckerson.com/practices/information-architecture
- TDWI Upside: https://upside.tdw.org/articles/list/contributor-barry-devlin.aspx
- Cutter Consortium: https://www.cutter.com/experts/barry-devlin

Follow me on Twitter: @BarryDevlin

LinkedIn profile: https://www.linkedin.com/in/barrydevlin