A LITTLE BEE BOOK

What is a
Data Lake?

This book belongs to:

____

____

____

A LITTLE BEE BOOK

What is a **Data Lake?**

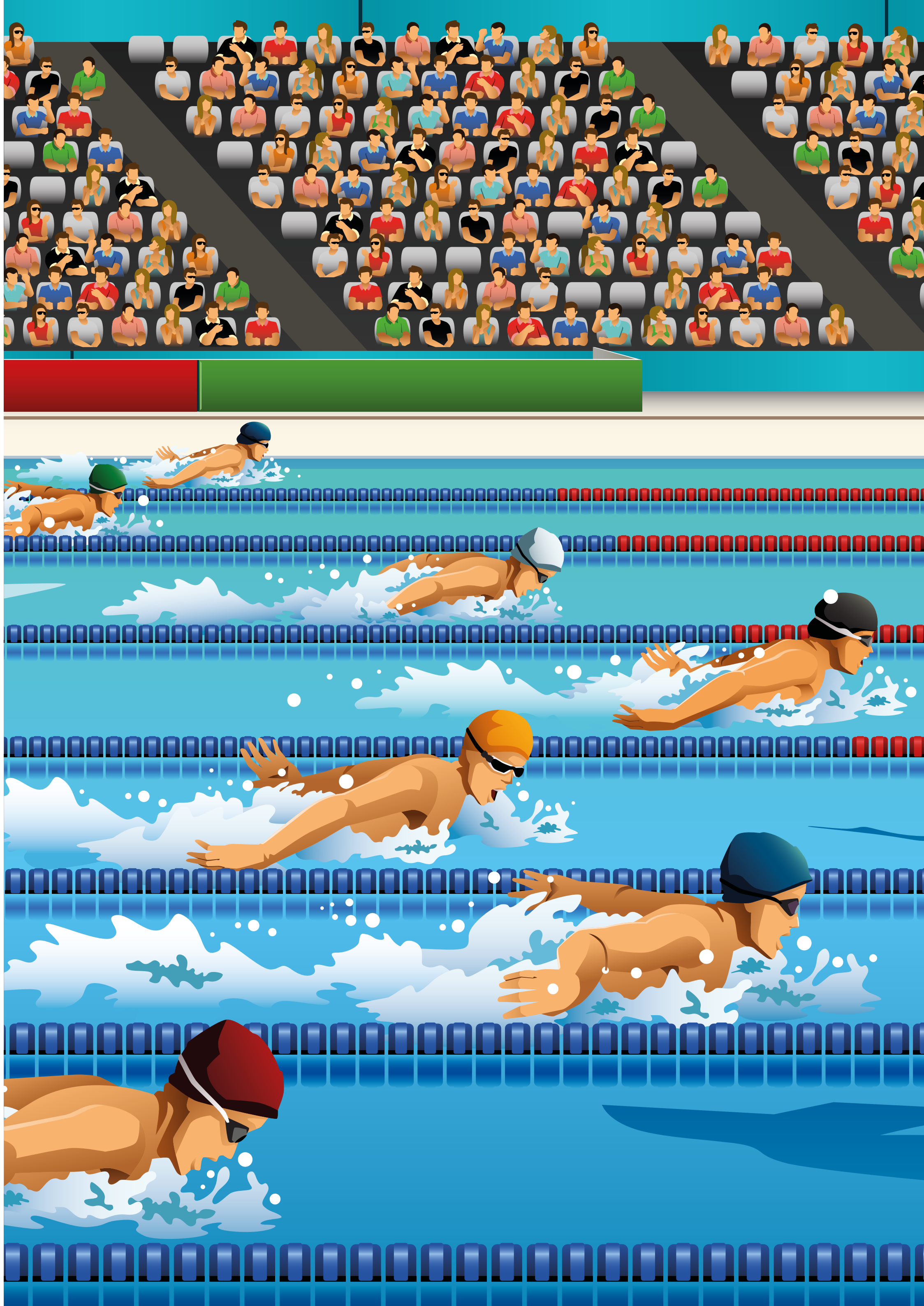
Adapted from an Enterprise Value Point Primer
by **Philip Thomas**

BACK ◀ ▶ NEXT

For more copies of this book, or to read others in the series, visit: **littlebeelibrary.com**

Organisations that wish to make widespread use of data and analytics need to be able to store and manage data in a way that provides agility for new and innovative analysis, while being robust and scalable over the long term.

A Data Lake is a shared data environment which is designed to meet this need by providing for long-term storage and management of data of all types, and provision of data for analysis.



BACK ◀ ▶ NEXT

There are some key differences between a Data Warehouse and a Data Lake:

1. Data Lakes Retain All Data

Data Warehouses are designed after extensive analysis of data sources, understanding of business processes and profiling of data. The result is a highly structured data model with clear agreement about what data to include and to not include. In contrast, the Data Lake retains ALL data.

2. Data Lakes Support All Data Types

Data Warehouses consist of data extracted from transactional systems. Non-traditional data sources such as web server logs, sensor data, social network activity, text and images are largely ignored. A Data Lake embraces these non-traditional data types.

$x^2 + \frac{b}{a}x = -\frac{c}{a}$
$x^2 + \frac{b}{a}x + \left(\frac{b}{2a}\right)^2 = -\frac{c}{a} + \left(\frac{b}{2a}\right)^2$

3. Data Lakes Adapt Easily to Changes

One of the chief complaints about Data Warehouses is how long it takes to change them. Considerable time is spent during development getting the warehouse's structure right. In the Data Lake all data is stored in its raw form.

A Data Lake provides many benefits, it:

- Allows rapid landing and storage of data
- Is a managed and governed environment, to ensure data can be found, is understood, and unnecessary duplication is avoided
- Allows easy access to data for analysis
- Enables exploratory analysis to provide new insight or "fail fast"
- Provides a highly scalable and affordable infrastructure capable of analysis at scale
- Provides a queryable archive for low-touch data at an attractive cost.

A Data Lake includes a variety of data stores including Big Data technologies such as Hadoop, Spark, Graph databases and Data Warehouses.

Hadoop typically plays a key role in the Data Lake, but a pure Hadoop Data Lake is insufficient to satisfy the entire spectrum of needs likely to be faced by an organisation.

Governance capabilities are key to an effective Data Lake. An unmanaged, ungoverned Data Lake quickly turns into a Data Swamp, with widespread duplication of data, difficulty finding data and lack of trust in data due to inability to understand its lineage and quality.

A Data Lake might be the solution for your company if you are:

- Wanting to extend your existing Data Warehouse with new types of data
- Investing in or exploring Big Data and/or Hadoop within your organisation
- Looking to simplify self-service access to data for analysts, data scientists, business users or developers
- Thinking about scaling your Hadoop or Big Data environment into production usage
- Facing challenges associated with meeting security, management and regulatory requirements around your data landscape.

However, many clients embark on building a Data Lake without fully appreciating the governance and management challenges. IBM has a strong heritage in data integration and governance, and a mature product set to satisfy these needs.

The Information Governance Catalog addresses needs for governance of information within a Data Lake, including ability to search for data, to understand business and technical definitions of data using the Business Glossary and to understand data lineage with the Metadata Workbench.

This can be implemented within an existing or emerging Data Lake based on IBM or third-party technology.

IBM Analytics have a strong portfolio of products to implement and support a Data Lake.

InfoSphere Information Server is an integrated suite of capabilities for data quality management and data integration, of which the Information Governance Catalog is a part. It includes capabilities for data integration on Hadoop.

DataWorks family provide SaaS-based integration and governance capabilities.

InfoSphere MDM, and complementary entity matching technologies, for management of master data and analytical reference data of all types.

Optim and **Guardium** for information lifecycle management, security and privacy.

Aspera for file transfer where a cloud or hybrid solution is deployed.

BigInsights provides native Hadoop with IBM enterprise integration, ease-of-use and integrated analytical tools for use within a Data Lake, including as a service.

DB2 (including BLU acceleration, IDAA, PureData for Analytics/Operational Analytics) for storage and management of relational data within the Lake.

dashDB, Cloudant and **Bluemix** database as a service offering for cloud-based data management.

In addition, IBM Global Business Services and IBM Software Services offer:

- Quick-start services to rapidly deploy infrastructure and software
- Proof-of-concept delivery
- “Big Data Stampede” delivery, to rapidly deliver pilot implementations and prove the value of a newly established capability
- Strategy and Analytics Consulting services to assist clients in defining their business and technical strategy & implementation roadmap.

IBM Systems and Technology Group offer specific infrastructure solutions designed and tailored for “Big Data” and Analytics workloads.

Talk to IBM to find out more about:

- A solution architecture for the entire Data Lake ecosystem
- The ability to offer a broad, integrated platform of capabilities
- An enterprise-class Hadoop, based on open standards (Apache, Open Data Platform), with enhancements for enterprise scalability and resilience, ease of use and analytics
- A mature and comprehensive technology for data acquisition, integration, quality management, governance and security – with wide support for data stores and technologies including Hadoop, and with capabilities underpinned by a common repository for seamless sharing of metadata.