DATA LAKES:
BELOW THE SURFACE

SVA | Consulting
Strategic Technology Solutions

Measurable Results.™

SVAC_EB02_0219

Published by:

SVA Consulting, LLC
1221 John Q. Hammons Drive, Madison, WI 53717
18650 W. Corporate Drive, Suite 200, Brookfield, WI 53045
SVAconsulting.com | info@SVAconsulting.com

Data Lakes: Below the Surface

The information contained herein is confidential and proprietary to SVA Consulting, LLC. It is not to be disclosed, in whole or in part, without the expressed written consent of SVA Consulting, LLC. It shall not be duplicated or used, in whole or in part, for any purpose other than to evaluate the suggestions of SVA Consulting, LLC.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The explosion of data volume, variety, and velocity in recent years has prompted many organizations to re-evaluate traditional data warehousing approaches. Relational data warehouses require a lot of upfront work to profile and analyze data, define rigid schemas, and build out processes to cleanse and conform data into those schemas. In addition, business users are demanding better and faster access to a broader range of data than ever before. The time and effort involved with incorporating new data streams into a data warehouse is often at odds with the expectations of the business.

This disconnect between business and technology has spurred the development of a new data storage and organization paradigm: the **Data Lake**.

If you think of a data mart as a bottle of spring water and a data warehouse as a 24-pack of those bottles, then think of a Data Lake as being the stream where those spring water bottles were derived from - raw, ready to be sampled, tested, or processed into an end product.

At its core, a Data Lake is a storage repository that contains raw data from various sources, stored in its original, unaltered form. Files are organized into consistent, self-descriptive directory structures and are tagged with metadata that define data lineage, security levels, and other bits of information.

Storing data into a Data Lake gives businesses the ability to take full ownership of their data sooner, and to future-proof themselves from changes to any particular database technology, reporting tool, or processing application in their data pipeline.

Cloud-based object storage technologies are traditionally used for Data Lakes, which allows for a complete decoupling of data storage from any compute resources. Once stored, data can be processed further downstream in an event-driven or batch manner to satisfy various use-cases.

This "data ownership-first" methodology is the next evolution of self-service analytics for business users, enabling users to get the data they need, at the level of "rawness" they need, without having to wait on the IT organization to deliver new or updated solutions:

- **Data scientists, analysts, and developers can tap directly into the raw data layer of the Data Lake to perform experiments, create proof-of-concept projects, and link up new processing pipelines to data storage events.**
- **Sales and marketing analysts can analyze a "speed layer" of data streaming in from social media, website clickstream, retail point-of-sale, and other sources in near real-time.**
- **Business users can utilize familiar data reporting and analysis tools to query downstream data warehouses that are hydrated from the Data Lake.**

Data processing traditionally involves three different phases. Data is **extracted** from a source, **transformed,** aggregated, or altered into a desired shape, and then **loaded** into a destination. This "Extract-Transform-Load" (ETL) pattern has historically been the backbone of business intelligence systems.

Data Lakes take a slightly different approach, using an "Extract-Load-Transform" (ELT) pattern. The key difference between ELT and ETL patterns is that ELT strives to land the data first, then transforms it as a secondary step.

DATA LAKE USE CASES

- **Playground for data scientists and data analysts**
  - Create proofs-of-concept against source data without the need to formalize schema or process
    - For example, a supply chain analyst could begin mining for insights across three different sources of logistics data on their own, without having to wait for data modelers and ETL specialists to construct a data pipeline
- **Repository for fast-moving streaming or Internet-of-Things (IoT) data**
  - Speed layer optimized for fast writes
  - Data can be captured from factory machinery, for example, and analyzed for patterns and anomalies in real-time
- **Starting ground for data mining, machine learning, and artificial intelligence work**
  - Access to raw source data as well as curated processed data provides a wealth of sources for unearthing the hidden power of your data
- **TTStrong foundation for governance and audit**
  - Metadata stores and data catalogs enhance data discoverability and traceability
    - Policies can define data retention, deletion, and access
- **Data source for hydrating a data warehouse**
  - Provide a more consistent, stable target for data warehouse ETL
- **Fast path to data ownership**
  - Integrate, organize, and go

DATA LAKE ANTI-PATTERNS

- **Not a replacement for a Data Warehouse or other structured reporting tools**
  - Data that is heavily relational, such as financial reporting, should still flow into a proper data mart or data warehouse downstream
- **Not a replacement for an online transactional processing (OLTP) system**
  - Data Lake storage platforms can offer assurances about data integrity, but they're designed to deal with files and not individual row-level transactions
- **Not ideal for system-to-system integrations out of the box**
  - A Data Lake provides the foundation for a data integration hub, but additional controls and logic need to be put in place to take advantage
- **Lack of oversight, proper cataloging, or metadata tagging can turn your Data Lake into a "Data Swamp"**
  - The "store data first and ask questions later" concept is powerful, but you still need to ensure that what you're storing has business value and is consumable

Data Lake Layers

Data Lakes can be separated into one or many “layers”:

1. Speed Layer

- Optional landing zone for real-time data feeds, or data that requires simple validation checks
- Separates “new” data from “raw” data
- This layer may contain a queue or streaming-data technology (like Amazon Kinesis) that batches or preprocesses real-time data before it’s persisted in the raw layer

2. Raw Layer

- The core layer of a Data Lake
- Data in its original, raw, and most granular form
- Immutable - designed to be append-only
- Tagged with metadata and organized by category, source, date/time, etc.
- Data landing in this layer can trigger downstream actions in an event-based manner

3. Processed Raw Layer

- Data may be cleansed, conformed, and/or aggregated for consumption
- Data formats normalized into one (i.e. CSV, JSON)
- May also consist of binary compressed formats (ORC/Parquet/Avro)

4. Curated Layer

- Data Warehouse or Data Mart
- Data is cleansed and conformed into a defined schema
- This layer should be able to be recreated on-demand from the raw layer

Data Catalog / Metadata

Keeping your Data Lake tidy and organized is critical to its success. It's just as important to define and store information **about** your data as it is to store the data itself.

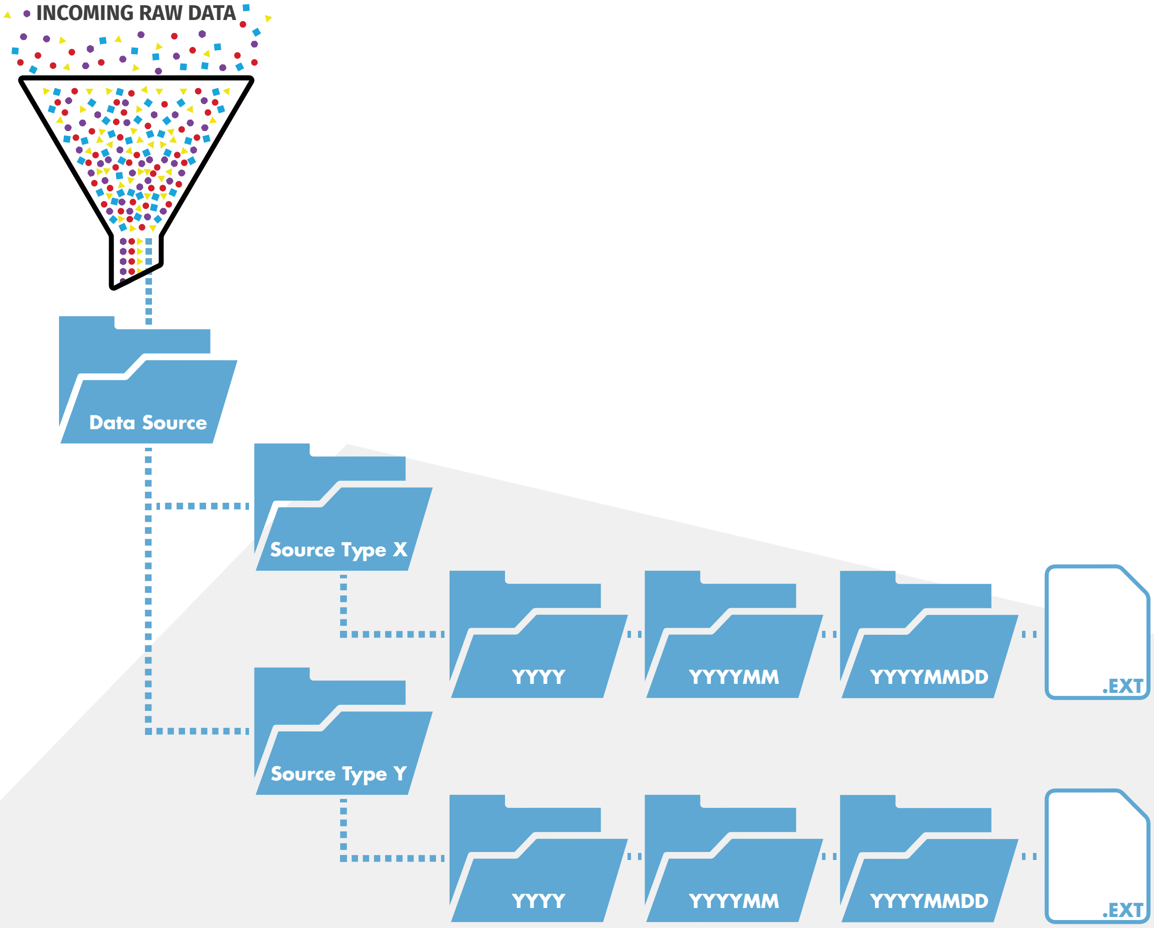
- **Metadata about individual files in the Data Lake are kept in object-level tags or a data catalog.**
- **Files cannot go into the Data Lake without being tagged with some kind of metadata.**
- **The catalog contains metadata about objects in the lake such as data lineage, schema profile, data-type profile, statistics, and search keywords.**
- **Schema changes to the files/streams can be tracked over time in a catalog.**
- **In a self-service scenario, the catalog should be searchable, allowing users to discover new data sets and find the ones they need.**

Storage & Ingestion

The maturation of cloud-based object storage and associated tooling is the primary driver that makes Data Lakes efficient and affordable for businesses of all sizes.

- **Data Lakes utilize cloud-based object storage for a variety of reasons:**
  - Low storage costs
  - Near-infinite scale
  - Geo-redundancy
  - Security and encryption features
  - Event triggering capabilities
  - Wide support by other technologies and tooling
- **Data can be ingested in various ways:**
  - Direct file drops
  - Pulled in via external streaming APIs
  - Pulled in via scheduled or triggered batch processes
  - Pushed in the lake via a Data Lake API
- **Any type of data can be loaded into a Data Lake:**
  - Structured or non-structured
  - Human-readable or binary formats
- **Data profiling and automation can be used to create new data ingestion pipelines upon receipt of new types of data without human intervention.**
- **Data retention and archival policies can be used to move archived data to cheaper storage tiers after they've been processed, or in-line with usage patterns.**

- Files are organized into a defined, consistent structure. Example:



- Data is first stored in “Data Source” folders to indicate which system or process it originated from
- One or more “Source Type” folders are utilized to separate data by type within a given source
  - For example, an ERP data source may provide general ledger data as well as purchase order data
  - All files within a particular “Source Type” folder should be able to be processed by the same code

- **In the Processed Raw layer of the Data Lake, compressed binary formats may be utilized to improve query performance and limit data transfer costs:**
  - Optimized Row Columnar (ORC) - optimized for heavy reads, column-based
  - Parquet - optimized for heavy reads, column-based
  - Avro - optimized for heavy writes, row-based
  - These files are machine readable, can be split across multiple disks, and have a self-describing schema
  - Using these compressed formats with schema-on-read query services, such as Amazon Athena, can be more performant and cost-effective than querying raw CSV or JSON data

Binary Data Formats

PUBLICATION / DISTRIBUTION

A Data Lake isn't of much use without a plan to make its contents available to the right people, processes, and tools in your enterprise. There are a number of ways to facilitate this:

- **Data Lake data can be broadcasted, published, or otherwise distributed from any of the "layers" previously listed, using a number of methods:**
  - Programmatically via a Data Lake API
  - Published onto a notification topic or enterprise message bus
  - Directly accessed via trusted tools
- **When Data Lake data is created, read, updated, or deleted (CRUD), various events fire behind the scenes. This robust event triggering is the real power behind the Data Lake concept - one CRUD operation on a file can, for example:**
  - Trigger an automated process to aggregate raw data into a summary form and save it elsewhere
  - Send a real-time alert to subscribing systems that data has changed
  - Update a dashboard that is reading from a streaming queue
  - Kick off an ETL process for a Data Warehouse system

SECURITY / ENCRYPTION

A Data Lake is an incredibly powerful tool for your enterprise, but it can also be an attractive attack vector for malicious parties. Data Lake storage providers generally offer robust security and encryption tooling, but it's still up to you and your team to ensure that best practices have been followed to secure your data assets.

- **Most cloud-based object data stores support encryption out of the box, and follow a "least privilege" approach by default.**
- **Data should only be read from the Data Lake over an encrypted connection.**
- **Securing types of files for particular audiences or groups goes hand-in-hand with the hierarchical storage structure outlined previously. Security can be provided at the "bucket" level, at the subfolder levels within the bucket, and even to the actual files themselves.**
- **Additionally, metadata tags can be used to define whether data is public, internally-facing, PHI/PII, financially sensitive, etc., and configuration and code can be set up to take advantage of those tags when trying to access data.**

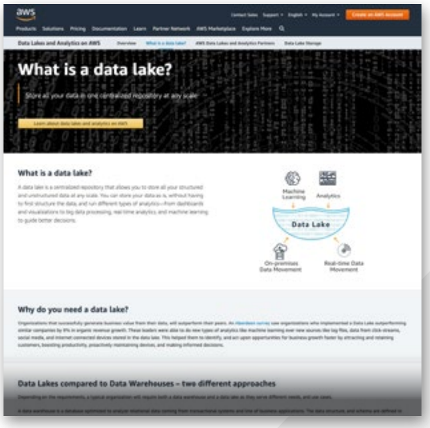
VENDOR SOLUTIONS AND TOOLS

The cornerstone of modern data management is the notion of “the right tool for the right job.” Leading cloud platforms offer an array of interconnected managed services that help you build out your Data Lake ecosystem. While this does not preclude the need for technical resources to implement your Data Lake, it does mean that those resources will be able to assemble your system from industry standard pieces vs. building everything from scratch.

- **Data Storage**
  - Amazon Web Services (AWS)
    - AWS Simple Storage Service (S3)
      - Versatile all-purpose object storage service
      - Geo-redundancy options available
      - Storage tier offerings available to balance performance vs. cost
  - Microsoft Azure
    - Azure Data Lake Storage
      - Built on top of Azure Blob Storage (as of Gen2)
      - Optimized for analytics and schema-on-read tools
      - Allows for hierarchical file organization and granular security
  - Google Cloud Platform (GCP)
    - Cloud Storage
      - Hot, cold, and archival options for storage
      - General feature parity with AWS and Azure, and integrates well with other GCP offerings
- **Data Processing**
  - Amazon Web Services
    - AWS Data Pipeline
    - AWS Glue (using Spark)
    - AWS Lambda
  - Microsoft Azure
    - Azure Data Factory
    - Azure Functions
    - Azure Databricks
  - Google Cloud Platform
    - Cloud Dataproc
    - Cloud Dataflow
    - Cloud Functions
    - Cloud Composer
- **Reporting / EDW / Querying**
  - Amazon Web Services
    - AWS Athena
    - AWS Redshift / Redshift Spectrum
    - AWS QuickSight
  - Microsoft Azure
    - Azure Synapse Analytics
    - Power BI
  - Google Cloud Platform
    - BigQuery
    - Cloud Datalab
    - Cloud Data Studio
- **Catalogs and Metadata**
  - Amazon Web Services
    - AWS Glue
  - Microsoft Azure
    - Azure Data Catalog
  - Google Cloud Platform
    - Data Catalog

ADDITIONAL READING

What is a Data Lake?



Data Lakes and Analytics on AWS

What is a data lake?

Store all your data in one centralized repository at any scale

Learn about data lakes and analytics on AWS

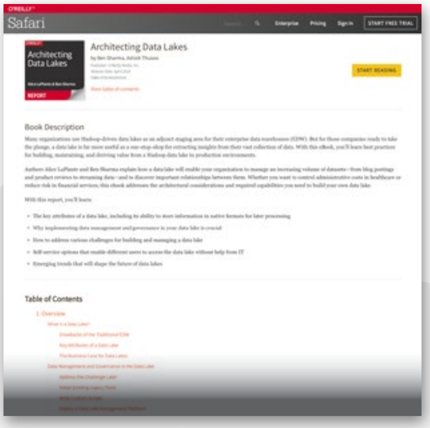
What is a data lake?

Why do you need a data lake?

Data Lakes compared to Data Warehouses – two different approaches

https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake

Architecting Data Lakes



Safari

Architecting Data Lakes

by Ben Sharma, Ashish Thusoo

START READING

Book Description

With this report, you'll learn:

- The key attributes of a data lake, including its ability to store information in native formats for later processing
- Why implementing data management and governance in your data lake is crucial
- How to address various challenges for building and managing a data lake
- Self-service options that enable different users to access the data lake without help from IT
- Emerging trends that will shape the future of data lakes

Table of Contents

1. Overview

https://www.oreilly.com/library/view/architecting-data-lakes/9781492042518

Zones in a Data Lake

SQL Chick

Zones in a Data Lake

December 30, 2017

As we are approaching the end of 2017, many people have resolutions or goals for the new year. How about a goal to get organized....in your data lake?

The most important aspect of organizing a data lake is optimal data retrieval.

It all starts with the zones of your data lake, as shown in the following diagram:

Hopefully the above diagram is a helpful starting place when planning a data lake structure.

Want to Know More?

You Might Also Like...

Popular Posts:

https://www.sqlchick.com/entries/2017/12/30/zones-in-a-data-lake

Data Lake

DataLake

Martin Fowler

Data Lake is a term that's appeared in this decade to describe an important component of the data analytics pipeline in the world of Big Data. The idea is to have a single store for all of the raw data that anyone in an organization might need to analyze. Commonly people use Hadoop to work on the data in the lake, but the concept is broader than just Hadoop.

Find similar articles at these tags

database big data

https://martinfowler.com/bliki/DataLake.html

Big Data File Formats Demystified

datanami

DATA SCIENCE • AI • ADVANCED ANALYTICS

May 16, 2018

Big Data File Formats Demystified

Alex Woodie

What's The Same

Sponsored Whitepapers

Sponsored Multimedia

Contributors

https://www.datanami.com/2018/05/16/big-data-file-formats-demystified

Click Above Images to Follow the Links

CONCLUSION

As data opportunities continue to explode across industries, only those businesses who have chosen to treat their data as an enterprise asset will benefit. Implementing a Data Lake is an important step in establishing data ownership for your business, and setting up your teams for success in an ever-changing landscape of data sources and tooling. The best part? Public cloud platforms have removed the barrier to entry for starting a Data Lake - they are low-cost and low-risk to create, and can scale over time to nearly any size.

Whether you're a startup business looking to establish a data footprint, or an established enterprise looking to build a data-driven culture within your business, a Data Lake is a strong foundation to help you achieve your goals.

Author

Jeffery H. Lewis
Data Solutions Development Manager
lewisj@svaconsulting.com

Contact Us

Business leaders hire us to help them tackle complex challenges associated with business growth and transformation. While our clients rely heavily on technology to remain viable, they don't care to become technology experts. Instead, they choose to partner with accomplished business professionals who know how to make technology work for them. Our approach is centered on positively impacting our clients' businesses, helping them achieve *Measurable Results.*™

If you are seeking an advisor and partner that you can trust, look to SVA Consulting. With over 90% of our business coming from referrals, our track record shows there is value in partnering with us.

☎ (800) 366-9091 | SVAconsulting.com