Essential Guide to Data Lakes

DESIGNING DATA LAKES TO OPTIMIZE ANALYTICS

MATILLION

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Businesses today generate large amounts of data, and desire to utilize it in order to gain business insights that will help drive competitive advantage. Data collected typically originates from diverse source systems and then needs to be transformed to better understand your core business, customers, and market dynamics. Due to the ever-increasing volumes of data, businesses need to continually find new and more efficient ways to consolidate and transform various data sources in order to make better, more informed decisions and in turn, gain competitive advantage.

How can businesses consolidate all their data? Centralizing all data within a data warehouse can prove effective for a number of use cases. While this approach may work for some businesses, others may require advanced scalability, accessibility, and a need to better control costs. A data lake, which allows all data types in any volumes to be stored and made available without the need to transform it before being ready for analysis, can address these unique requirements by providing a cost-effective resource for scaling, storing and accessing large volumes of diverse data types.

INTRODUCTION

The concept of a data lake has been around for a while, however, the term 'data lake' was first introduced by James Dixon, CTO at Pentaho in 2010. He outlined the shortcomings of the 'Data Mart' when handling businesses' real-life data needs. He claimed 80-90% of businesses needed to handle unstructured data as well as growing data volumes, and that such requirements were straining the Data Mart model. The solution to these challenges, Dixon suggested, was a **data lake**.

> **"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 (2010)**

Since then, the industry has experienced significant growth in the number of data lake offerings and implementations that seek to help organizations unlock the value of their unstructured and semi-structured data, requiring ever more advanced and sophisticated analytics techniques. Businesses with growing analytics functions also need to ensure data access, and governance is easier and more efficient. This eBook will show how a modern data lake architecture can be designed to meet these objectives.

Key

When reading the eBook, look out for colored boxes for additional resources, how to's, best practices, and hints and tips.

🟦 Further details and resources

■ 'How to' information and examples

🟧 Best practices and optimizations

🟩 Hints and tips from Matillion

CHAPTER 1

WHAT IS A DATA LAKE?

CHAPTER 1: WHAT IS A DATA LAKE

1.1 Data Lake Definition

A data lake allows massive amounts of data to be stored in its native format. The structure and processing rules do not need to be defined until the data is needed.

1.2 What is a data lake and why does it matter?

According to recent research, the average business is seeing the volume of its data grow at a rate that exceeds 50% per year. Additionally, these businesses are managing an average of 33 unique data sources used for analysis.[1]

As data volumes, varieties, and velocities increase, the ability to securely store, process and govern that data becomes more difficult. A data lake architecture sets out principles and characteristics enabling organizations to meet these challenges by providing a centralized repository that allows the storage of business data no matter the volume, variety, or velocity at which it is generated. This single repository can then service many different types of analytical workloads from visualizations and dashboards, to machine learning and beyond.

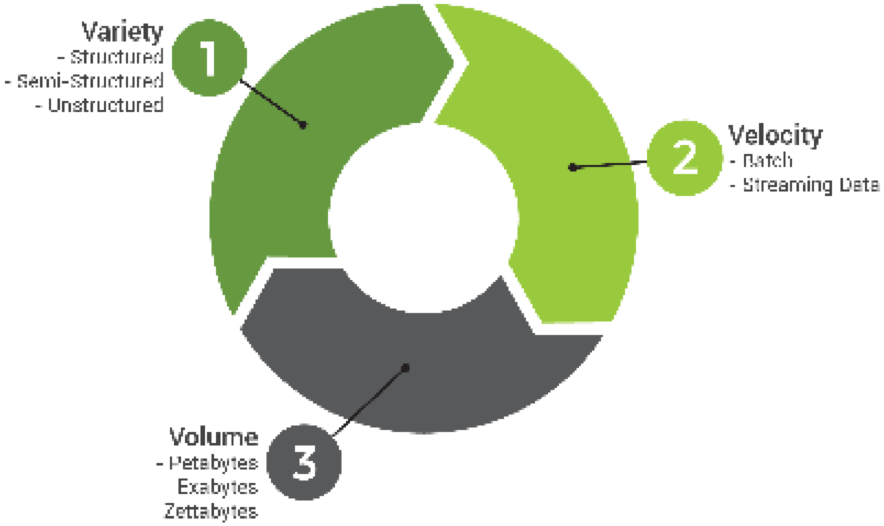


Figure 1: The Three V's of Big Data

[1] https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf

1.3 Characteristics of a Data Lake

Consolidation

The centralization of siloed data is one of the principles of a data lake architecture. This centralization brings a number of benefit's, including being easier to govern and manage, as well as making it easier to innovate non-disruptively around heterogeneous data sets.

Collect and Store All Data at Any Scale

Data lakes allow the collection and storage of data at any scale. In terms of the 3 V's of Big Data, cloud object storage services provide virtually unlimited space at very low costs. Your business' data can be collected in real-time using streaming technology at high velocity, if so required. Furthermore, these platforms perform the difficult task of unlocking value in unstructured data, such as the automatic transcription of audio recordings, thus accommodating a wide variety of modern data sources.

Locate, Curate, and Secure Data

Having a centralized data lake makes it easier to keep track of what data you have, who has access to it, what type of data you are storing, and what it's being used for. The cost of non-compliance with various regulations, such as GDPR and the Consumer Privacy Act, both in terms of fines and reputational damage is so great that organizations need a way to meet these requirements without stifling innovation.

Increased Agility

Having a centrally curated data lake allows your business to innovate in new ways of processing data. You can introduce new use cases without having to re-engineer your architecture. For example, you may be dealing with batch data. However, extending the architecture to include real-time data streaming shouldn't impact the use cases you already have in place. Similarly, you may be using a data warehouse as a computational resource. Adding Spark workloads to incorporate machine learning use cases can use the same underlying data, made possible by separating storage and compute resources, as well as allowing flexible access to different applications.

CHAPTER 2

HOW DATA LAKES ADD VALUE

CHAPTER 2: HOW DATA LAKES ADD VALUE

2.1 Getting Value From a Data Lake

A key value proposition of data lakes is the ability to store data of unknown value, importance or utility for almost negligible cost, data that would otherwise be discarded due to the unjustified costs associated with storage and security. This is because as analytics capabilities within an organization mature, the potential uses cases for such data can be revealed. This historical data can be used to train machine learning models and answer questions in the future.

However, arguably the most value an organization can get out of a data lake is to use it as an engine for innovation. By making data access simpler, faster and more efficient for users and facilitating experimentation with different processing technologies businesses can discover game changing insights that fuel competitive advantage. Research from independent analyst firm, Aberdeen, shows that businesses who implement data lakes are able to translate the associated analytical benefits into significant ROI. Participating companies reported a ***9% increase in organic revenue growth*** and ***4% increase in operating profit*** compared to similar businesses who did not.

Data Lake Use Cases

Augmented data warehouse

For data that is not queried frequently, or is expensive to store in a data warehouse, federated queries make the different storage types transparent to the end user.

Support for IoT data

Data lakes are excellent choices for storing the high volume high frequency data from streams. Easily adapted to a lambda architecture they can support near real-time analysis.

Advanced analytics

Quicker access to untransformed data is useful for data scientists, particularly when feature engineering for machine learning models.

Regulatory compliance

A centralized repository of an organization's data makes it easier to apply role-based security, catalog data sets, and track lineage. It allows simpler processes for subject access and removal requests.

CHAPTER 3

OVERCOMING DATA LAKE CHALLENGES

CHAPTER 3: OVERCOMING DATA LAKE CHALLENGES

3.1 Data Lake Challenges

Although a data lake is a great solution to manage data in a modern data-driven environment, it is not without its significant challenges.

Looking again at how we define a data lake: ***allows for the ingestion of large amounts of raw structured, semi-structured, and unstructured data that can be stored en masse and called upon for analysis as and when needed.*** We can see this definition carries inherent risks and can lead to the dreaded **Data Swamp** which many organizations have fallen prey to.

PwC quotes Sean Martin, CTO of Cambridge Semantics as saying "We see customers creating big data graveyards, dumping everything [into the data lake] and hoping to do something with it down the road. But then they just lose track of what's there. The main challenge is not creating a data lake but taking advantage of the opportunities it presents." [2]

This observation is reinforced by independent analyst firm Gartner, which states "the data lake will end up being a collection of disconnected data pools or information silos all in one place...Without descriptive metadata and a mechanism to maintain it, the data lake risks turning into a data swamp."[3]

3.2 Mitigating Data Lake Challenges

To mitigate these risks, a governance layer needs to be built into the architecture to answer the following questions and formulate the four pillars of data governance.

4 Pillars of Data Governance

- What data do you have and where it is stored? (**Data Catalog**)
- Where has data come from and what has happened to it? (**Data Lineage**)
- Is data accurate and fit for purpose? (**Data Quality**)
- Is data protected from unauthorized access? (**Data Security**)

[2] https://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf

[3] https://www.gartner.com/en/newsroom/press-releases/2014-07-28-gartner-says-beware-of-the-data-lake-fallacy

This governance layer is a combination of process and tooling which generally increases the total cost of ownership (TCO) of a solution but makes a return on investment (ROI) more likely.

In order to manage the implementation of policies and processes, organizations should appoint a stakeholder who is responsible for data lake governance. These roles may include Data Steward, Enterprise Data Architect, or Data Protection Officer depending on the size and structure of an organization. Having an accountable person who is also a subject matter expert not only ensures that these tasks do not fall by the wayside as a project rolls on but any solutions proposed are informed by the relevant technical knowledge.

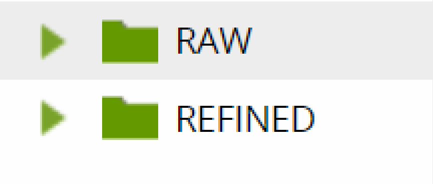
One of the first non-trivial tasks in designing your data lake architecture is to determine the storage taxonomy, we have put together an example of how you might design your data lake storage below.

DESIGNING DATA LAKE STORAGE EXAMPLE

Most data lake implementations use **cloud object storage** as the underlying storage technology. It is recommended for the following reasons:

- **Durable** - You can typically expect to see eleven 9s availability
- **Scalable** - Object storage is particularly well suited to storing vast amounts of unstructured data, and storage capacity is virtually unlimited.
- **Affordable** - You can store data for approximately $0.01 per GB/ Month
- **Secure** - Granular access down to the object level
- **Integrated** - Most processing technologies support object storage as both a source and a **data sink.**

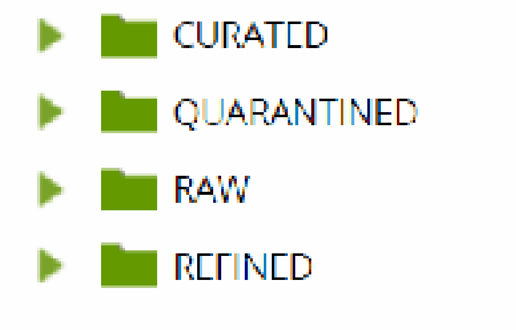
A common misconception and potential mistake is that data lakes are one, giant, centralized bucket where everything lands.   In this example, we assume our architecture has 2 tiers of storage, RAW and REFINED. The initial storage structure will be as shown below:



▶ RAW
▶ REFINED

Both user access and pricing tiers can be defined individually for each.

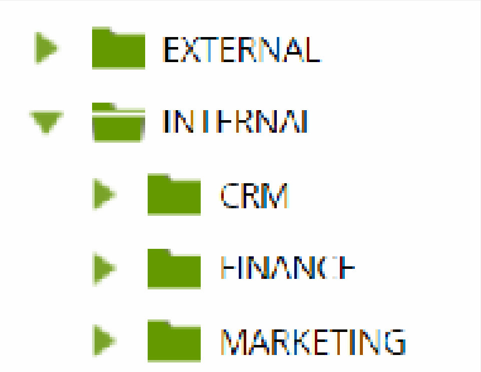
As discussed, this is only the minimum recommended configuration and many implementations will have more tiers, shown in the bucket.



▶ CURATED
▶ QUARANTINED
▶ RAW
▶ REFINED

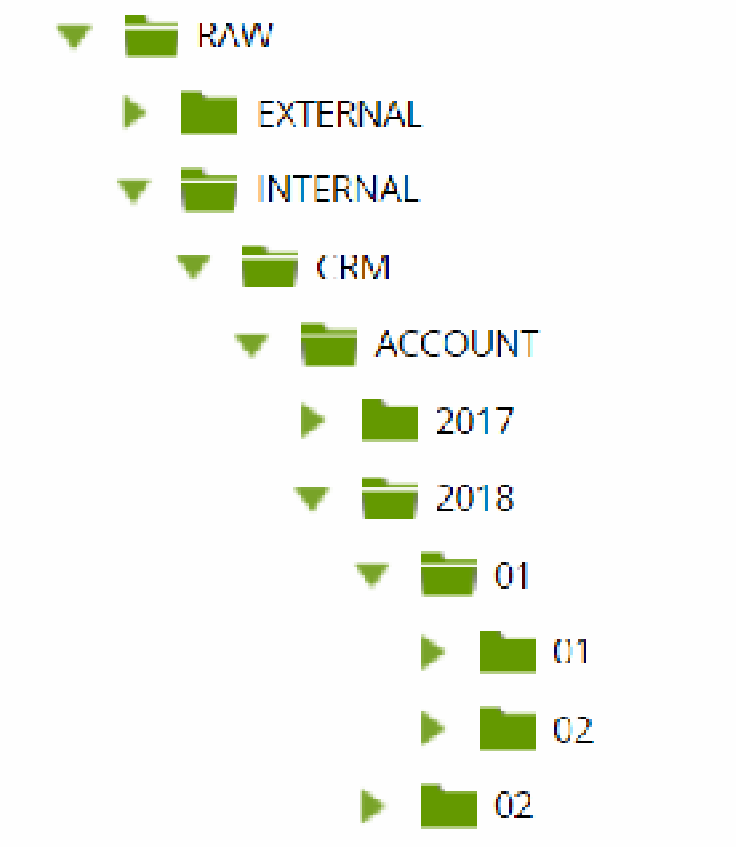
In this example, the QUARANTINED bucket has limited access and can only be seen by individuals responsible for ensuring that sensitive data does not proliferate into the rest of the data lake.   The other extra bucket we have is CURATED data, this is typically the output of some data processing such as an ETL job.

The QUARANTINE and RAW buckets we can consider as ingestion buckets and they should be organized by source system as shown in the example below:



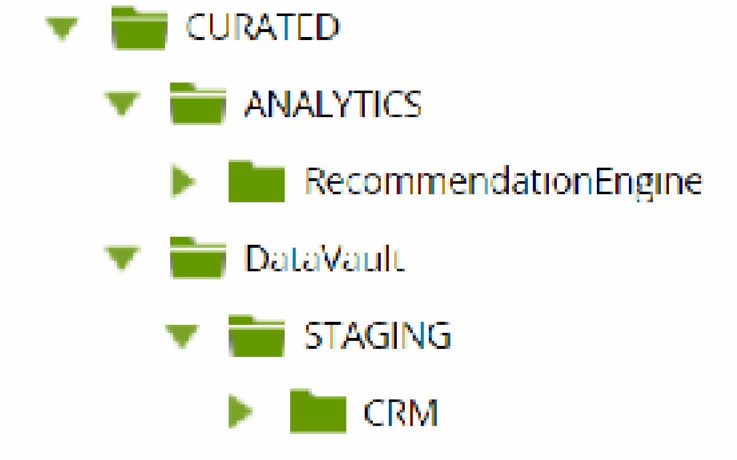
▶ EXTERNAL
▼ INTERNAL
  ▶ CRM
  ▶ FINANCE
  ▶ MARKETING

The data can be further partitioned by entity, year, month and day as shown.



```
RAW
  EXTERNAL
  INTERNAL
    CRM
      ACCOUNT
        2017
        2018
          01
            01
            02
          02
```

If the QUARANTINE and RAW buckets are considered ingestion buckets, then the REFINED and CURATED buckets are considered consumption buckets and as such they should follow a structure that aligns more to the target data model or subject area than the source system.



```
CURATED
  ANALYTICS
    RecommendationEngine
  DataVault
    STAGING
      CRM
```

CHAPTER 4

DATA LAKE VS. DATA WAREHOUSES

CHAPTER 4: DATA LAKE VS. DATA WAREHOUSES

In this chapter, we will look at the differences and similarities between a data lake and both traditional on-premises and cloud data warehouses.

4.1 Comparison to a Traditional Data Warehouse

A data lake is not a direct replacement for a datawarehouse, they are supplemental technologies that serve different use cases with some overlap. In fact, most organizations that have a data lake will also have a data warehouse. The following section will compare the properties of a data lake to a traditional BI Architecture (data warehouse and separate ETL Server).

Data in Data Lakes is Stored in its Native Format

Data can be loaded faster and accessed quicker since it does not need to go through an initial transformation process. For traditional relational databases, data would need to be processed and manipulated before being stored. This requires the development of a transformation process, as well as the testing and execution of that process. By loading data directly into the data lake in its source format, you can skip the transformation step for now.

Access Data Flexibility in a Data Lakes

Data scientists, engineers, and analysts can access data much quicker than would be possible in a traditional BI architecture. This increases agility and provides greater opportunities for data exploration and proof of concept activities, as well as self-service business intelligence. Although a characteristic of a data lake is accessibility, this doesn’t mean that everyone has unbridled access; privacy and security can and should be considered when granting permissions and access.

Data Lakes Provide Schema-on-Read Access

Traditional data warehouses employ Schema-on-Write. This requires an upfront data modeling exercise to define the schema for the data. All data requirements, from all data users, need to be known upfront to ensure the models and schemas produce usable data for all parties. As new requirements are unearthed, those models may need to be redefined.

Schema-on-Read, conversely, allows the schema to be developed and tailored on a case-by-case basis. The schema is developed and projected on the data sets required for a particular use case. This means that the data required is processed as needed. Once the schema has been developed it can be kept for future use or discarded when no longer required.

Data Lakes Provide Decoupled Storage and Compute

When you separate storage from compute you are fully able to optimize your costs by tailoring your storage requirements to the access frequency. For example, most cloud storage services provide tiered storage. This allows your business to archive raw data on less expensive tiers while allowing faster access to transformed, analytics-ready data. Being able to run experiments and exploratory analysis with new technologies is much easier thanks to such data preparation. Traditional data warehouses and ETL servers have tightly coupled storage and compute, meaning if we need to increase storage capacity, we also need to increase compute and visa-versa.

Data Lake	Traditional On-Premises Data Warehouse
• Data stored in native format	• Data requires transformation
• Can store unlimited data forever	• Expensive to store large volumes
• Schema-on-read	• Schema-on-write
• Decoupled storage & compute	• Tightly coupled storage & compute

4.2 Comparison to Modern Cloud Data Warehouse

ESG research shows roughly 35-45% of organizations are actively considering cloud for functions like Hadoop, Spark, databases, data warehouses, and analytics applications[4], and this is a trend that is increasing due to the benefits of cloud computing such as massive economies of scale, reliability and redundancy, security best practices and easy to use managed services. Cloud data warehouses combine these benefits with traditional data warehouse functionality to deliver increased performance and capacity and reducing the administrative burden of maintenance.

Matillion simplifies data transformation for cloud data warehouses, with increased speed of development and execution. Offered as a pay-as-you-go service, Matillion delivers increased savings to companies of all sizes. **www.matillion.com**

[4] https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/ESG-White-Paper-AWS-Apr-2017+(FINAL).pdf

The table below compares an aggregate feature set of the major cloud data warehouses.

	Data Lake	Cloud Data Warehouse	On-Premises Databases
Unstructured data (schema-less data)	Yes	No	No
Semi-Structured data (Self-describing schema)	Yes	Yes	No
Structured Data (Relational)	Yes	Better	Better
Independently scale storage and compute	Yes	Yes	No
Schema-on-Read	Yes	Yes (semi-structured)	No
Schema-on-Write	No	Yes	Yes

Figure 2: Comparing Data Lakes, Cloud Data Warehouses and On-Premises Databases

According to research conducted by Aberdeen, 26% of all decisions to invest in a data lake are to offload data warehouse workloads. One compelling feature of cloud data warehouses is the ability to query data residing in low-cost storage using the compute resources of the massively parallel processing (MPP) data warehouse (Query offloading). This, coupled with the ability to natively query semi-structured data (schema-on-read), is seeing more data lakehouse architectures where some of the data is stored in a data warehouse and other datasets are stored in the data lake.

“Data Lakehouse” Benefits

- Increases interoperability of data
- Schema-on-read of semi-structured data
- Join data lake files with data warehouse tables
- Increased query concurrency

Employ Matillion to transform data, using the power of your cloud data warehouse (CDW), by combining tables in your data lake and your CDW. Write the output of this transformation to either target based on your needs.

CHAPTER 5

BUILDING A DATA LAKE

CHAPTER 5: BUILDING A DATA LAKE

5.1 Object Storage or HDFS

Current real-world data lakes fall into one of 2 categories, Hadoop Distributed File System (HDFS) based either on-premises or cloud and Object Storage based (Cloud "buckets"). This eBook focuses on object storage-based data lake implementations such as Amazon S3, Azure Blob and Google Cloud Storage. This is not to say that HDFS based data lakes are without merit, so let's examine some pros and cons of object-based storage compared with HDFS.

The Cons

Performance. Object storage is generally lower-performance than HDFS for a number of reasons.

1. Object storage tends to have to higher I/O variance meaning it's inadequate for transactional data, such as databases that require a more consistent I/O.
2. Objects are immutable, which means no support for append or truncate operations. If you need to do an incremental update to some data you would need to overwrite the object, which can be slower in some circumstances.
3. Object storage will almost always have higher round trip request latency than HDFS as it is not physically attached to the compute resource.

The Pros

1. HDFS relies on local storage meaning to increase capacity we either add more nodes or add bigger drives. Cloud object storage, however, is unlimited and elastically scales without any configuration.
2. Cloud object storage is generally more durable than HDFS, offering up to 99.999999999% durability.
3. When a cluster is turned off data stored on HDFS is no longer available to query, with cloud object storage data persists when associated compute resources are turned off.
4. Because of the separation of data from the cluster, it is accessible to other processing engines such as data warehouses.
5. The object storage solutions provided by the cloud platforms mentioned are all HDFS compatible, so your existing Hadoop or Spark cluster can access the data as if it were HDFS.

Object-based storage is suited to a larger number of use cases and provides a more flexible data lake.

5.2 Data Lake Essential Elements

Storage

The reference architecture diagram **[Figure 4 - section 5.3, p.28]** shows two storage tiers. Although this is a minimum recommendation, it's not uncommon to add extra tiers or zones such as a quarantine zone, used in highly regulated industries where the data is particularly sensitive and needs to be manually checked by data stewards before moving to a zone with more user access. Another approach is to subdivide the second tier into multiple zones based on the use case. For example, there may be a data warehouse staging zone, a trusted master data zone, a machine learning training data zone, and so on. Each zone may have its own granular levels of access and file formats.

File Format	Properties	Use Cases
Orc	Columnar, schema stored in footer.	Read heavy analytical workloads, e.g. Hive Tables.
Parquet	Columnar, schema stored in footer.	Read heavy analytical workloads, e.g. Spark processing.
Avro	Row-major, schema and data separate.	Write heavy workloads, e.g. Apache Kafka.
CSV	Human readable, fixed schema.	Small volumes, consumer is an analyst.
JSON	Human readable, flexible schema.	Small volumes, consumer is an application.

Figure 3: Data storage properties and use cases

Compute

To move data from between storage tiers requires some compute resource to perform the transformation. The three most common ways to process data from one tier to another are:

1. SQL - There are various tools available that allow you to use an SQL syntax to transform data. This can either be serverless or use a data warehouse as the compute resource.

2. Spark - Execution is done on a cluster of virtual machines.

3. Serverless functions - Functions are run on cloud platform provisioned resources.

> Matillion allows you to transform your data using the power of your cloud data warehouse. By utilising the massively parallel processing engines to run SQL transformations on data in cloud object storage, giving you optimal storage and compute power that is cost efficient.

Similarly, once data is stored in the refined tier, multiple different processing technologies can use this data as input, run a job to perform the transformation, or train machine learning models. Then, once processing is complete, compute resources can be turned off or scaled back. This approach makes implementing new use-cases much easier, as you have a significant amount of the infrastructure already built to support them.

Governance

The governance layer comprises different tools and capabilities which allow an organization to comply with regulations, secure data, and manage access.

Data Quality

Data quality can be defined as data that is "fit for its intended uses in operations, decision making, and planning".

Data masking and de-identification. One of the goals of a governed data lake should be to limit the proliferation of personally identifiable information (PII), using a tool, vendor, or framework that stops PII data from entering the data lake. This could be achieved with machine learning or implementing some business rules to identify, tag, and remove the data.

Master data management. One of the barriers to unlocking value in an organization's data is removing the redundancies, conflicts, and errors in the data. Having a centralized repository for an organization's data dovetails nicely with an MDM solution and creates a single source of truth from heterogeneous sources.

Data Security

There are several challenges when it comes to securing data in a data lake, including controlling data access and preventing data breaches. The place to start is by reviewing your shared responsibility model with your cloud platform and SaaS vendors. Generally speaking, the cloud service providers are responsible for the cloud and you are responsible for data in the cloud, but it is your responsibility to check and understand your security needs and set up.

When building your data lake, you need to make sure the following aspects of data security are covered:

1. **Role based access control at the appropriate level of granularity**
   - Access to buckets, folders, and objects within the data lake
   - Permissions to execute, modify, and read jobs
   - Access to administration consoles, features, and utilities
2. **Network isolation**
   - Firewall Rules
   - Network ACLs (Access Control Lists)
   - Security Groups
3. **End-to-end encryption**
   - Valid SSL Certificates for encryption in transit
   - Enabled encryption at rest

Data Catalog

A data catalog allows you to automatically crawl and compile both metadata and index data sets within a data lake, allowing them to be searchable. This metadata can be used for audit purposes or to dynamically drive data transformations.

Another feature that is available in some data catalog tools is the ability to tag data as containing PII, either manually or using a machine learning algorithm, to automatically profile and identify sensitive data.

Data Lineage

“Data lineage includes the data's origins, what happens to it and where it moves over time.”[5] Being able to have the full audit trail of where a specific attribute came from, how it was transformed, and then used in analytics is a key requirement to many of the regulatory demands organizations now face. In addition, the data provided by the lineage can help engineers debug issues that arise when executing workloads.

In summary, using a combination of governance tools allows you to be compliant and mitigate security risks. Consider the example of a request from a customer to identify thier data in your possession and remove any data you have of theirs. The tasks involved might be similar to the following:

De-identifying Data using Data Governance Tools

1. Use a data catalog to list all of my data sets and search for those data sets that might contain PII.

2. Use role-based security to access the data.

3. View the lineage of the attributes to find out where that record originated and how it was transformed to provide an overview of all workloads that use that data.

4. Use a data masking function to either obscure the data or remove it completely.

5. Use an MDM tool to make sure the delete is pushed back to the originating system.

[5] https://www.techopedia.com/definition/28040/data-lineage

Orchestration

Orchestration is the glue that holds everything together. Once you have built data loading jobs, you can build workflows that reformat data, copy data into your data warehouse, and transform that data into your data model. You will want to automate all of this and add some logging and error handling capabilities. This is where you will need an orchestration capability. The major cloud platforms all offer services such as Amazon Data Pipeline, Azure Data Factory and Google Cloud Composer.

In addition to providing powerful SQL transformation, Matillion has a suite of orchestration components allowing you to trigger processes in other technologies through remote command line or API execution, all as part of a holistic scheduled workflow.

Consolidate ALL your data once and for all!

Purpose-built data transformation for cloud data warehouses

- ✓ Simplify data transformation across your cloud data warehouse and data lake
- ✓ Increase the speed of development and execution
- ✓ Pay-as-you-go service to deliver increased savings
- ✓ Best practice features to make the most of your cloud data infrastructure

GET A DEMO

Get a demo to see first-hand how Matillion can help you consolidate data in your data warehouse and data lake

MATILLION

5.3 Reference Architecture Diagram

Putting all of the essential elements together we can produce a reference architecture as shown in Figure 4 below:

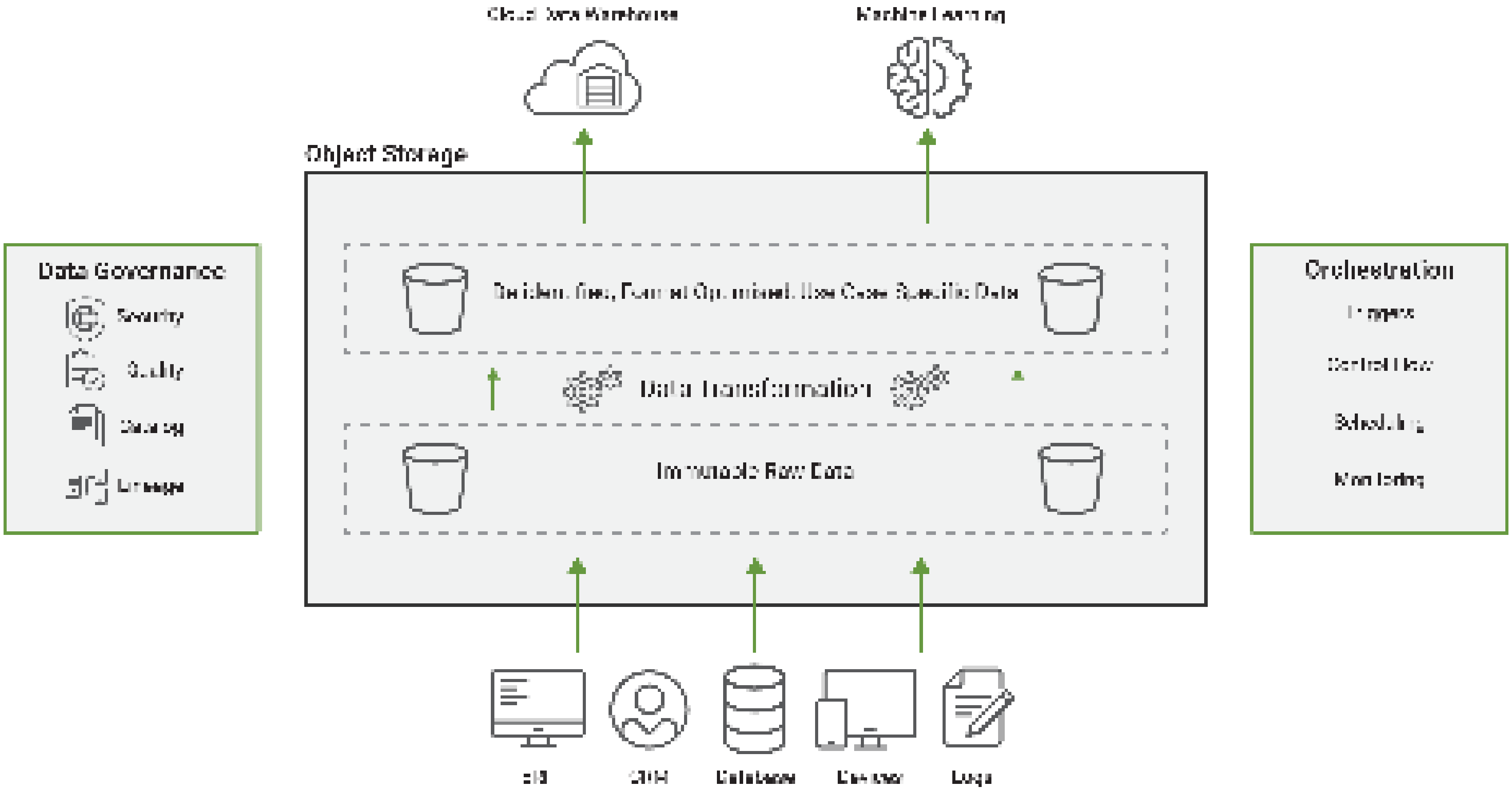


Figure 4: Data Lake Reference Architecture Diagram

CHAPTER 6

DATA LAKE BEST PRACTICES

CHATPER 6: DATA LAKES BEST PRACTICES

6.1 Data Loading

Use an immutable raw data store. This allows you to perform any new exploratory analysis without having to go back and extract data from the source system. It also allows you to rebuild the state of your business at a point in time, which can be useful for testing predictive models.

It is a good practice to use a quarantine zone for removing PII and other sensitive data before landing into the raw data zone. Data masking techniques can be used to de-identify data, but this can pose a problem for certain use cases. For example, when you have your own data and third-party data that you need to match, for example credit bureau data. This can still be accomplished by hashing the PII data and then dropping those identifiable attributes such as first name, last name, date of birth etc. When you receive your third-party data, you can generate a hash from those same fields and match that same customer.

Data masking is not without its risks. Data exposed by your business can be cross referenced with other data your organization might not even hold and be used to de-anonymize and identify individuals. Taking this into consideration, it is often the best idea to hash and then completely exclude certain attributes altogether.

6.2 Data Processing

Decoupling storage and compute resources is one of the key design goals of a cloud-based data lake, and enables you to optimize costs, innovate quickly, and experiment with different processing technology and vendors. This decoupling allows you to truly pick the best tool for the job. The ability to pay only for the compute resource you need allows start-up businesses to compete with the biggest enterprises without the huge capital expenditures. Licensing costs can also be reduced in this manner, removing a barrier to competition by utilizing serverless and managed service compute.

Common tasks that are performed in the data lake that require compute resource are:

- Preparing data before allowing user access, for example de-identifying datasets.
- Optimizing file formats for analytics e.g. CSV to PARQUET

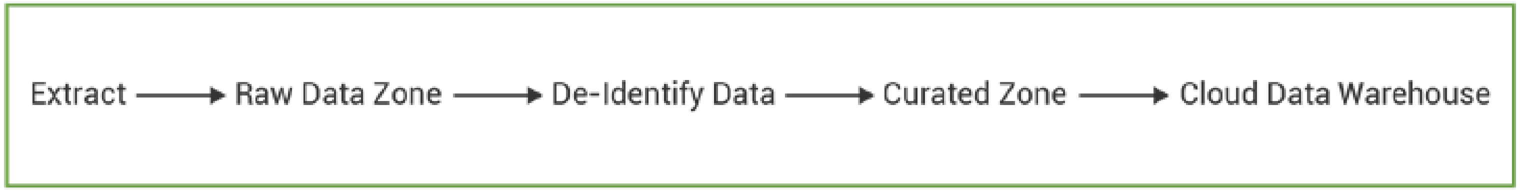


Figure 5: Typical data pipeline

6.3 Governance and Security

Proper governance is a key enabler of deriving value from a data lake. Follow these best practices to improve governance and security.

- Appoint an individual who is accountable for data lake governance
- Define a roles & responsibility matrix
- Leverage a centralized data catalog
- Perform regular audits

6.4 Cost Optimization

Object storage services often have the ability to optimise costs based on the temperature of data. You can get much cheaper storage costs for infrequently accessed data whilst enjoying increased performance for your hottest data sets. Moving data between buckets for the purpose of cost optimisation should be incorporated into your data pipelines and workflows.

An example of a hot data set would be one that has a low latency requirement with high request rates. From this you can infer other properties such as a higher cost storage. An example of a cool data set would be high volume data with low request rates, from this you can infer other properties such as low cost storage and higher latency. See [Figure6].

Where possible use automated serverless functions to handle the archiving of data to lower tier storage.

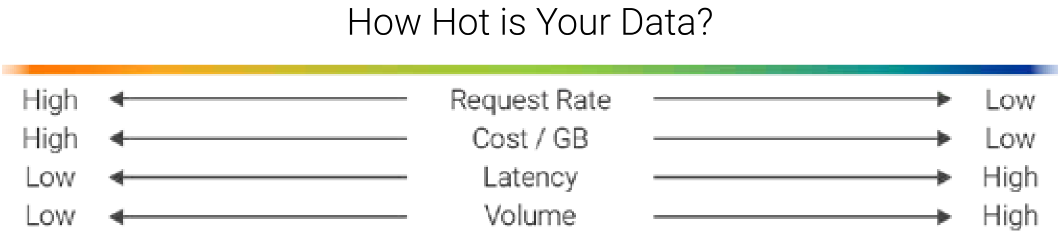


Figure 6: Data Temperature

CHAPTER 7

KEY TAKEAWAYS

CHAPTER 7: KEY TAKEAWAYS

Data lake initiatives often come about because a particular use case is not feasible in the current infrastructure or because of an objective to modernize data access and processes. Whatever the reason for your data lake project, keep the following takeaways at the forefront of your decision making.

Implement Reports that Deliver Value

McKinsey & Co. found that "Too often, the enthusiastic inclination is to apply analytics tools and methods like wallpaper—as something that hopefully will benefit every corner of the organization to which it is applied. But such imprecision leads only to large-scale waste, slower results (if any), and less confidence, from shareholders and employees alike, that analytics initiatives can add value."[6] It is important that organizations examine the feasibility and value of each use case and concentrate on delivering those that will add the most value. The data lake is a flexible architecture that easily allows the addition of varied use cases over time.

Avoid Data Swamps

Gartner states that through 2018 80% of data lakes will not include effective metadata management capabilities, making them inefficient.[7] The benefits of a data lake architecture have been presented throughout this eBook, however, the key to leveraging those benefits and ROI is the ability to be able to fully utilize the data lake. This comes from proper governance.

Leverage the Benefits of the Cloud

Cloud economics, allowing you to pay by the query or by the second, have brought down the cost of using SQL-based compute resources. This combined with the low cost of cloud object storage, enables data transformation on a scale not feasible with traditional on-premises architectures. Furthermore, independent and elastic scaling allows you to respond in an agile manner to the different demands of different types of workloads.

Include the Essential Elements

In an attempt to realize a quicker time to value, it can be tempting to concentrate on the "core" features, loading, storing and data processing. Building a data lake without orchestration management will mean relying on something akin to cron daemons to organize and schedule your data pipelines. This lacks the ability to have robust error handling, dependency checking, event driven logic, logging and monitoring capabilities. Similarly, neglecting the governance layer can leave you exposed to data security risks, but can also decrease trust from business users due to poor data quality, which potentially leads to project failure.

[6] https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/ten-red-flags-signaling-your-analytics-program-will-fail

[7] https://www.gartner.com/imagesrv/media-products/pdf/LogTrust/LogTrust-1-3F7HE3J.pdf

CONCLUSION

CONCLUSION

Data lakes allow all types of data in any volumes to be stored and made available without the need to transform it before being ready for analysis. These advantages overcome some challenges you may be experiencing with just a data warehouse. Achieve advanced scalability, accessibility, and get a better control on your infrastructure costs by extending your cloud data capabilities with a data lake.

About Matillion

Matillion provides industry-leading data transformation products for cloud data warehouses. Delivering a true end-to-end data transformation (not just data prep or movement from one location to another), Matillion provides an instant-on experience to get you up and running in just a few clicks, a pay-as-you-go billing model to cut out lengthy procurement processes, and an intuitive user interface to minimize technical pain and speed up time to results. Matillion is available globally for Amazon Redshift, Snowflake, and Google BigQuery on leading cloud infrastructures.

Find out more at **www.matillion.com**.

Simplicity.
Speed.
Scalability.
Savings.

Purpose-built data transformation for cloud data warehouses

- ✓ Our intuitive UI and approach to data transformation makes complex taks simple
- ✓ We deliver the fastest time to value, from launch to development to production
- ✓ Built to take advantage of the power and features of Amazon Redshift, Snowflake, and Google BigQuery
- ✓ Pay-as-you-go with no long term commitments

GET A DEMO

Get a demo to see first-hand the power of Matillion data transformation

MATILLION

V1.1 (001)