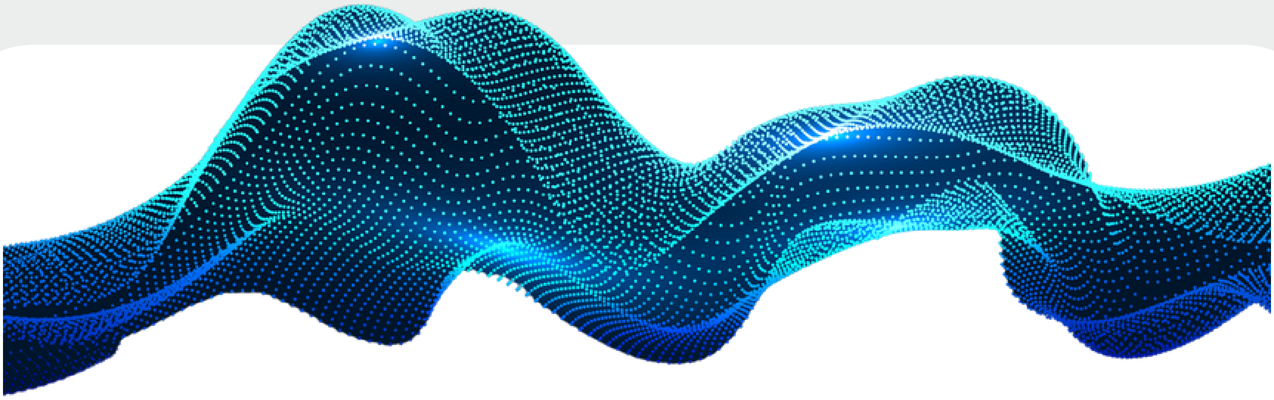


OPERATIONAL DATA LAKES: A COMPREHENSIVE GUIDE

WHITE PAPER

This white paper aims to provide a comprehensive overview of operational data lakes, covering the key considerations and best practices for building and maintaining an effective data lake.

www.Apica

OVERVIEW

Data lakes have become an increasingly popular choice for storing and analyzing large volumes of structured and unstructured data. They provide a centralized repository that allows organizations to store all their data at any scale, including data from different sources, types, and formats. This makes them an ideal platform for big data, analytics, and machine learning applications.

Table of Contents

Introduction

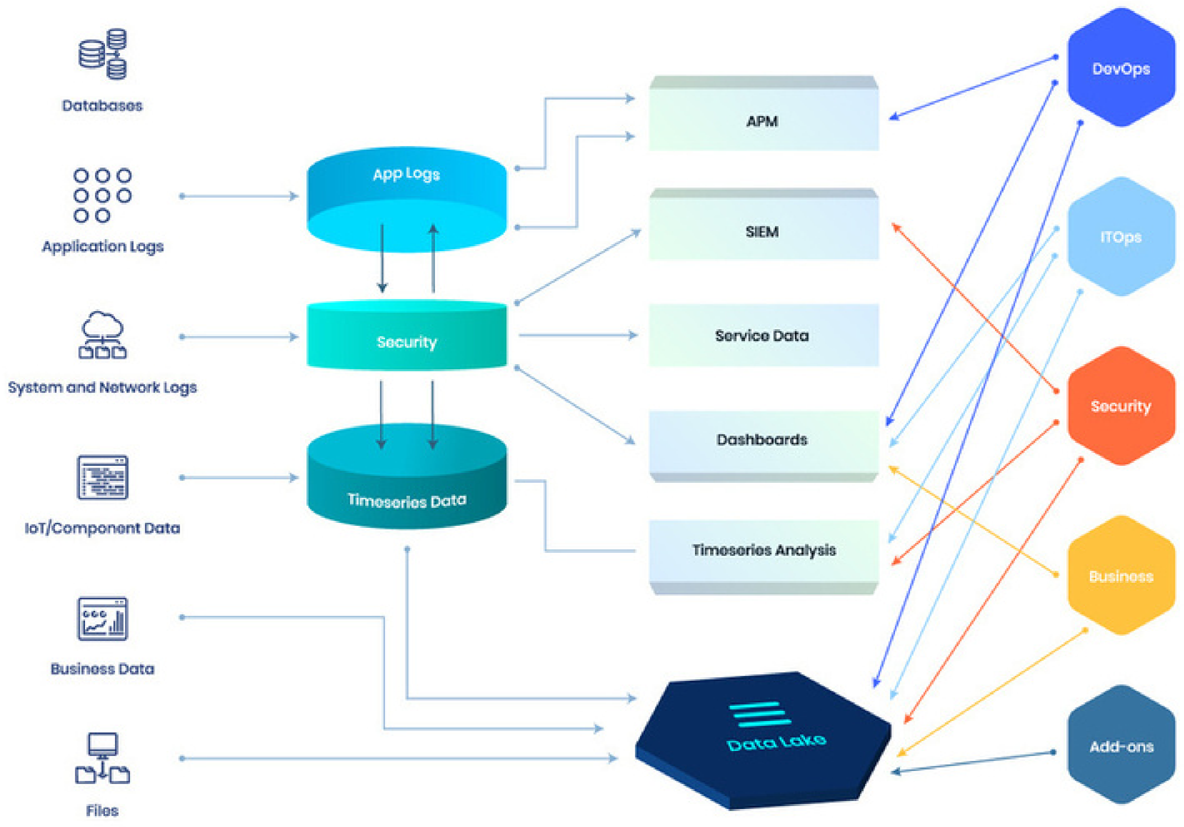
Data lakes have become an increasingly popular choice for storing and analyzing large volumes of structured and unstructured data. They provide a centralized repository that allows organizations to store all their data at any scale, including data from different sources, types, and formats.

However, simply having a data lake is not enough to derive value from it. To truly realize the benefits of a data lake, organizations must operationalize it. This involves building processes and systems to ensure that the data lake is properly governed, secured, and integrated with other systems and tools in the organization.

This white paper aims to provide a comprehensive overview of operational data lakes, covering the key considerations and best practices for building and maintaining an effective data lake.

What is a Data Lake?

A data lake is simply a centralized storage location for all of your structured and unstructured information at any scale. Without initially organizing your data, you can use a variety of analytics to improve your decision-making, from dashboards and visualizations to big-data processing, real-time analytics, and machine learning.



Data Lake Architecture

Data lakes are a great solution for organizations that have complex data management needs or need to quickly access large volumes of information. Data lakes are highly scalable and cost-effective, empowering organizations to easily query and analyze vast amounts of data in real time. Additionally, data lakes can be used to develop insights, uncover hidden trends and correlations, and enable more informed decisions.

Data lake architectures typically include two main components: a data storage layer and a processing layer.

The storage layer is responsible for storing the raw data in its native format. Meanwhile, the processing layer enables organizations to process and analyze the data. It usually consists of a set of tools and technologies to enable search, discovery, analytics, and machine learning capabilities.

With data lakes, organizations can quickly and accurately access large amounts of data without having to transfer it into another format or build out complex infrastructure. They also do not need to devote huge amounts of time and resources to managing the data.

Instead, they can use the data to gain insights into their business operations and trends, helping them make better decisions. Additionally, data lakes are cost-effective as they often require fewer resources and less time to maintain than traditional methods of storing and processing data.

What is an Operational Data Lake?

An operational data lake is a data lake that is designed, built, and maintained to support the ongoing needs of an organization. It is more than just a repository for storing data; it is a system that is integrated with other systems and tools in the organization, and it is governed and secured to ensure that the data is accessible, trustworthy, and compliant with regulations.

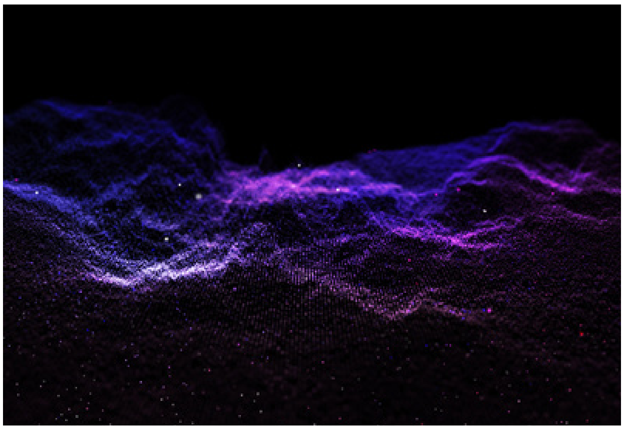
An operational data lake is a centralized repository that stores both structured and unstructured data in its raw format until it is needed. The data lake is used to store large amounts of data, which can then be easily retrieved and analyzed to provide valuable insights.

Operational data lakes are often used in big data scenarios where organizations need to collect, process, and store large amounts of data from multiple sources in real time.

The data stored in an operational data lake can be analyzed and processed quickly, making it ideal for use in real-time decision-making processes, such as fraud detection, customer behavior analysis, and real-time supply chain optimization.

An operational data lake can be implemented using various technologies, such as Apache Hadoop, Apache Spark, and Amazon S3. These technologies provide a scalable and flexible infrastructure for storing and processing big data. Additionally, organizations can use various data management and analysis tools, such as Apache Hive, Apache Pig, and Apache Storm, to extract insights from the data stored in the data lake.

In summary, an operational data lake is a centralized repository that stores large amounts of raw data from multiple sources, which can then be easily retrieved and analyzed to provide valuable insights. It is designed for real-time big data processing and is implemented using various big data technologies.



Why do you need an Operational Data Lake?

There are several reasons why an organization might need an operational data lake, but before exploring those, let’s have a look at the core pain points that are associated with Log Analytics:

- The issue of latency comes first. The data must travel through further hops before it reaches the final systems.
- Second, because there isn't a single source of truth, data transformation may occur in several places at once, making it extremely difficult to comprehend what has happened to the data.
- Additionally, there are simply too many hurdles to get through in order to transition to machine learning on top of this login information.
- Due to the user's dependence on so many distinct interfaces, they wind up with an uncounted number of tabs while trying to export data from one application to another.

Now that we know the main issues that the customers are facing, operational data lakes arise as a well-rounded solution. The need for an operational data lake for an organization arises for the following reasons:

To store and manage large volumes of structured and unstructured data:
Data lakes provide a centralized repository for storing data of all types and sizes, including data from multiple sources and formats. This can help organizations better manage their data assets and make them more accessible to authorized users.

To support data analytics and machine learning:
An operational data lake can provide a rich and diverse data set for analytics and machine learning, enabling organizations to gain insights and make data-driven decisions.

To facilitate data integration and collaboration:
The operational data lake can act as a focal point for combining data from many sources and systems, dismantling data silos, and promoting better communication and decision-making.

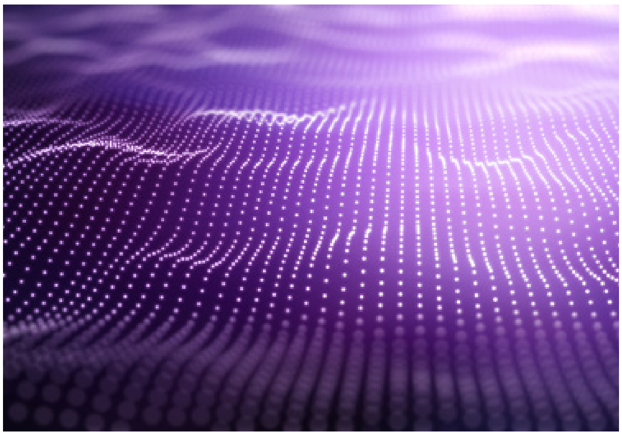
To enforce data governance and security:
With the aid of an Operational data lake, organizations establish and enforce policies and procedures for data management, including data quality, data lineage, and data security. This can help ensure that the data is trustworthy and compliant with relevant regulations and standards.

To support a wide range of use cases:
Operational Data Lakes ought to support a wide range of use cases, including data integration, data analytics, data governance, data security, and data management. It should be able to handle the diverse data types, structures, and formats that an organization generates and consumes.

Key considerations for Building an Operational Data Lake

There are several key considerations to keep in mind when implementing an operational data lake, such as:

- **Data Ingestion:** An operational data lake should be able to handle a wide range of data sources and formats, including structured, semi-structured, and unstructured data. It should also be able to handle large volumes of data and support real-time data ingestion.
- **Data Transformation:** An operational data lake should provide robust data transformation capabilities, including the ability to clean, enrich, and transform data as needed.
- **Data Storage:** An operational data lake should provide scalable and cost-effective storage for data of all types and sizes. It should also support multiple storage options, including object storage, file storage, and relational databases.
- **Data Access:** An operational data lake should provide easy and secure access to data for authorized users, including data scientists, analysts, and business users. It should also support multiple access methods, including SQL, APIs, and visualization tools.
- **Data Governance:** A data lake that is in use should include comprehensive data governance capabilities, including high-quality data, lineage, and security.



It also needs to comply with any required standards or regulations.

Indexing is both the Key and the Killer

Indexing is extremely resource-hungry but it's imperative to sustain an operational data lake. It is a Data Lake feature that adds performance to your data lake, allowing end users to search for information and retrieve the results in milliseconds. Indexing becomes necessary when the scale is the order of TBs or hundreds of TBs per day.

Most Big Data platforms, including Apica, require indexing to operationalize the data lake. The more resources are involved in indexing, the more the expense will be. All the major Big Data Platforms like Splunk, HADOOP, and BigQuery focus intensely on indexing and thus their customers suffer from hefty charges later down the line.

Sure you get instant retrieval but at what cost? What if there was a way or solution to achieve almost the same level of efficiency without breaking the bank? This is exactly where Apica's operation data lake comes in. But more on that later. Let's first have a look at the benefits of operational data lakes.

Benefits of an Operational Data Lake

There are several key benefits to operationalizing a data lake:

- **Data Integration:** An operational data lake can serve as a central hub for integrating data from different sources and systems. This can help organizations break down data silos and facilitate better collaboration and decision-making.
- **Data Analytics:** An operational data lake can provide a rich and diverse data set for analytics and machine learning, enabling organizations to gain insights and make data-driven decisions.
- **Data Governance:** An operational data lake can help organizations establish and enforce policies and procedures for data management, including data quality, data lineage, and data security.
- **Data Security:** An operational data lake can help organizations secure their data by providing robust access control and data protection measures.

Data Management: An operational data lake can provide a centralized system for managing data assets, including data ingestion, data transformation, data storage, and data archiving.

Why do enterprises need an on-prem/self-hosted solution vs SaaS for an operational data lake?

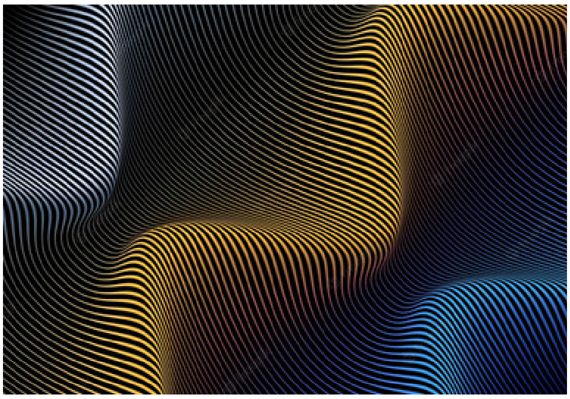
Both on-prem and SaaS environments have their fair share of pros and cons. On one hand, you get more control, data security, performance, and cost efficiency with on-prem. On the other hand, there are a couple of advantages to using a SaaS solution for an operational data lake as well, including lower upfront costs, easier maintenance, and access to the latest features and updates.

Most on-premise deployments are typically more expensive than cloud-based deployments, as they require you to purchase and maintain the hardware and infrastructure necessary to run the software, as well as provide ongoing maintenance and support. However, some organizations may prefer the control and security of an on-premise deployment, which can justify the additional cost.

Ultimately, the decision between an on-premises or SaaS solution will depend on the specific needs and resources of your enterprise.

There are a few reasons why an organization might prefer an on-premises or self-hosted solution for an operational data lake, rather than a Software-as-a-Service (SaaS) solution:

1. **Control:** An on-premises solution gives the enterprise complete control over the hardware and software infrastructure, as well as the data itself. This can be important for compliance or security reasons, or simply to have more control over the performance and reliability of the system.
2. **Customization:** An on-premises solution can be customized to the enterprise's specific needs, which may not be possible with a SaaS solution.
3. **Cost:** In some cases, an on-premises solution may be more cost-effective than a SaaS solution, especially for large enterprises with significant data volumes or complex data processing requirements.
4. **Integration:** An on-premises solution can be more easily integrated with other on-premises systems and data sources, which may be important for certain organizations.



The Key Players in the Operational Data Lake Domain

A developing organization can save information that it is certain will be important in analytics data lakes, like an Azure data lake, without the trouble, delay, or cost of cleaning and organizing it beforehand. Data lakes are also easier to scale than organized or structured storage.

However, Operational data lakes can be expensive, especially for larger organizations with a lot of data. It is important to carefully consider the cost-benefit of using one service over the other.

Now, there are multiple options available to enterprises when it comes to implementing an operational data lake, including- Splunk, Elastic, SnowFlake, and BigQuery.

There are a few potential drawbacks to using an operational data lake from the big platforms:

Complexity: Setting up and managing an operational data lake can be complex, especially for those who are not familiar with the technology. It may require specialized training and expertise to use effectively.

Cost: Implementing and maintaining an operational data lake can be expensive, especially for larger organizations with a lot of data.

It is important to carefully consider the cost-benefit of using an operational data lake versus other options.

As an enterprise, you may find services like Splunk and Elastic to be helpful initially. However, in the long run, the overhead cost simply becomes too hefty to handle. In typical security requirements, a general Splunk deployment can balloon overnight to ridiculous levels (up to 13TB per day). Thus the prices with the operational data lake will have skyrocketed before you'll know it.

Expertise and Scalability: Sophisticated tools like Splunk and Elastic can be complex to set up and manage, especially for those who are not familiar with them. It may require specialized training and expertise to use effectively. Not to mention that Splunk may not be suitable for very large data sets, as it can be resource-intensive and may not scale well.

Security and compliance: Ensuring the security and compliance of an operational data lake can be challenging, as it may involve managing access to sensitive data and adhering to various regulations and standards.

Integration: Integrating an operational data lake with other systems and tools can be difficult, which can be a drawback if you need to integrate your data lake with other parts of your infrastructure. Some tools may not integrate well with other systems and tools, which can be a drawback if you need to integrate your data lake with other parts of your infrastructure.

Apica's Architecture

The architecture of a typical operational data lake basically consists of the following components. With Apica, every aspect can scale on demand be it ingesting, indexing, search, or parallelism:

Data sources: These are the various systems and sources that provide data to the operational data lake, such as transactional databases, log files, IoT devices, and so on. Apica's architecture is source agnostic in that regard. Therefore, no matter what the source is, you ingest and forward logs and metrics

Data ingestion: This is the process of collecting and transporting data from various data sources into the operational data lake. This may involve using tools such as batch processing, streaming, or a combination of both.

Data storage: The operational data lake stores data in a central repository, often using a distributed file system such as HDFS (Hadoop Distributed File System) or S3 (Amazon Simple Storage Service).

Apica has a built-in CloudFormation template that you can configure and use to provision a stack with the Apica S3 exporter Lambda function. The S3 bucket that stores server access logs sends an event when a log file is added. When an event is triggered, the Lambda function reads and processes the access log and sends it to your Apica instance.

Data processing: This is the process of transforming and manipulating the data in the operational data lake, using tools such as SQL, Python, or Spark. This may involve cleaning and organizing the data, as well as performing analytics or machine learning tasks.

Data access: The process of providing access to the data in the operational data lake for querying and analysis, involving tools such as SQL or BI (Business Intelligence) platforms to query the data. With Apica, you can collect stats, perform real-time processing on the data and generate searchable reports.

Overall, the architecture of Apica's operational data lake is designed to enable the collection, storage, processing, and access of large amounts of data in a flexible and scalable way.

Additionally, besides data governance advantages like security, compliance, and access control, Apica offers the following benefits with its operational data store:

On any cloud: With the active observability data fabric from Apica, you may monitor your whole application and cloud infrastructure through a single interface and gather logs, metrics, traces, events, and security logs. We ensure broad compatibility with platforms by building on open standards and protocols.

Low-Cost: Apica is eliminating the difficulties associated with putting in place and realizing a high ROI from full-stack visibility. Only Apica's data fabric provides control, insights, compliance, and security in one platform - which is unmatched in the industry.

Infinite Scaling: From the very beginning, you can construct your data pipelines with limitless storage that can serve as an unending store for throughput discrepancies. Apica's InstaStore allows for the quick replay of any data to a target upon request. With InstaStore, you can bid farewell to the data block and data loss issues forever.

Simplicity: When it comes to complexity, managing a data lake can easily become a chore. With Apica's straightforward architecture a philosophy that allows infinite scaling, you can easily handle your operational data lake.

Ownership: Set up and customize your dashboards, notifications, event rules, and data sources with all the flexibility you need.

Bottomline

An operational data lake is a centralized repository that allows an enterprise to store, process and analyze all its structured and unstructured data at any scale.

It provides a single source of truth for an organization's data and enables organizations to operationalize big data by making it available for use in real-time applications and workflows.

Enterprises need an operational data lake because it provides the following key benefits:

Data Integration: It enables organizations to integrate and centralize data from various sources into a single location, reducing the time and cost associated with managing multiple data silos.

Scalability: An operational data lake is designed to handle large volumes of data and can scale as data grows, making it ideal for enterprises with increasing data needs.

Flexibility: It allows organizations to store both structured and unstructured data, providing greater flexibility in the types of data that can be analyzed.

Cost-effective: An operational data lake eliminates the need for expensive proprietary data warehousing solutions, providing cost savings for organizations.

Real-time Processing: With the ability to handle real-time data streams, an operational data lake enables organizations to make real-time data-driven decisions.

Improved Insights: By centralizing and integrating data from various sources, organizations can gain deeper insights into their operations and customers, leading to better decision-making and improved business outcomes.

In conclusion, an operational data lake provides enterprises with a scalable, flexible, and cost-effective solution for managing, processing, and analyzing all their data in real-time. It enables organizations to turn big data into actionable insights and make data-driven decisions to improve their operations and drive business growth.