aws analytics

Creating a Modern Analytics Architecture

Today, data is indispensable

In today's world, data and analytics are indispensable to organizations. Those who successfully generate business value from their data will outperform their peers. An Aberdeen survey saw organizations who implemented a modern data lake analytics platform outperform similar companies by 9% in organic revenue growth. These leaders were able to perform new types of analytics like machine learning over data sources such as log files, data from clickstreams, social media, and internet-connected devices stored in the data lake.

Modern data lake analytics help organizations to:

- Identify and act upon opportunities more quickly
- Grow business faster
- Attract and retain more customers
- Boost productivity
- Proactively maintain devices
- Make better-informed decisions

What are the barriers to getting to the data you need?

We hear from organizations all the time that they are looking to extract more value from their data but struggle to capture, store, and analyze all the data generated by today's modern and digital businesses. Data is growing exponentially, coming from new sources, increasingly diverse, and needs to be securely accessed and analyzed by any number of applications and people in shorter and shorter periods of time. The size, complexity, and varied sources of the data mean the same technology and approaches that worked in the past don't work anymore.

As the amount of data accumulates, customers store it in different silos, making it difficult to perform analytics. To make it easier, customers want all of their data in a single repository, i.e., a data lake. Organizations need to store data securely at any scale and at low cost, using the standards-based data formats of their choice. They want the flexibility to analyze the data in a variety of ways, using a broad set of analytic engines to ensure their needs will be met for their present and future analytics use cases. They also need to go beyond insights, from operational reporting on historical data to being able to perform real-time analytics and machine learning in order to accurately predict future outcomes.

The challenge with existing data infrastructures

Almost all organizations have built data warehouses for reporting and analytics purposes. They use data from a variety of sources, including their own transaction-processing systems and other databases. Many have also built Hadoop frameworks for analyzing what is commonly called “big data” or data that does not fit well in highly structured data warehouses. Building and running a data warehouse and a big data framework have been complicated and expensive.

Traditional data warehouse challenges

Traditional data warehousing systems create a range of issues and demands:

- Cost millions of dollars in upfront software and hardware expenses
- Take months in planning and procurement
- Difficult to set up
- Need time for implementation and deployment processes
- Require that you define your data models and ingest data
- Hire a team of data warehouse administrators
- Keep your queries running fast and protect against data loss
- Only highly normalized data needed for mission-critical analytics
- A lot of data (dark data) in many siloed data stores
- Dark data never makes it into a data warehouse for analysis
- Difficult to scale

When data volumes grow or you want to make analytics and reports available to more users, you have to choose between accepting slow query performance or investing time and effort on an expensive upgrade process. In fact, some IT teams discourage augmenting data or adding queries to protect existing service-level agreements. To mitigate this, organizations often set up multiple data marts. These contain copies of a subset of the data in the data warehouse. Specialized and long-running queries then don't impact the performance and SLAs of mission-critical business operations and decision-making. This complicates the data and analytics infrastructure and further locks organizations to the chosen vendors for their data warehouse and data marts.

A different analytics engine needed for new varieties of data

Traditional warehouses are also not effective in handling the variety of semi-structured and unstructured data coming from clickstreams, logs, social media, IoT sensors, and other modern data sources. These data types do not fit well within the normalized, structured data model. A different analytics engine is needed—one that can process these new data types like a Hadoop framework for big data. This greatly complicates the data strategy, where data is stored, who can access it, and which analytics engine works best on which data.

Batch data loading delays

Batch data loading poses a challenge. Extract, transform, and load (ETL) jobs typically run nightly when the analytics load on the data warehouse is minimal. Nightly ETL jobs don't work well for a global organization that needs 24/7 access to data. Nightly ETL jobs also delay time-to-insight for those users who need immediate answers. In today's fast-paced world, waiting until tomorrow may be too late.

Limited support for modern analytics

Traditional data warehouses either don't support sophisticated machine learning or predictive workloads or only support them in a limited fashion. Therefore, they're unable to support modern use cases such as real-time or predictive analytics and applications that need advanced machine learning.

Securing data requires workaround solutions

Security and data privacy are also an issue. Industries such as health care and financial services that work with highly sensitive data require the data warehouse to be compliant with ISO, HIPAA, FedRAMP, and more. General Data Protection Rules (GDPR) further add to the burden on IT to ensure that sensitive customer data is encrypted in all states—at rest and in motion. Some of these regulations also require organizations to react quickly to retrieve and update or delete a record at short notice. Traditional data warehouses often require organizations to implement expensive workaround solutions, which often leave the sensitive data out of the analysts' reach.

The complexity of big data systems

Big data platforms have experienced similar issues and are compounded by the volume and variety of data, complex algorithms needed to perform analytics, and a shortage of skilled workers. In addition, analyzing data across data warehouse and big data systems is complex and time-consuming.

Optimal data storage

As the amount of data accumulates, organizations have stored it in different silos, making it difficult to perform analytics. To make it easier, organizations want all of their data in a single repository, i.e., a data lake. They need to store data securely at any scale and at low cost, using the standards-based data format of their choice. And they want the flexibility to analyze the data in a variety of ways, using a broad set of analytic engines, ensuring their needs will be met for their existing and future analytics use cases.

What is a data lake?

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as is, without having to first transform or structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions. Data needed for business decision-making can now be processed, cleansed, and loaded from the data lake into the data warehouse.

The need for both a data warehouse and a data lake

Depending on the requirements, a typical organization will require both a data warehouse and a data lake as they serve different needs and use cases. As organizations with data warehouses see the benefits of data lakes, they are evolving their warehouse to include them in order to enable diverse query capabilities, data science use cases, and advanced capabilities for discovering new information models. Gartner named this evolution the "Data Management Solution for Analytics" or "DMSA."

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Nonrelational and relational from IoT devices, websites, mobile apps, social media, and corporate applications
Schema	Designed prior to the data warehouse implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher-cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e., raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI, and visualizations	Machine learning, predictive analytics, data discovery, and profiling

Create an analytics pipeline

Before data can be analyzed, it needs to be collected, processed and stored. You can think of this as an analytics pipeline that extracts data from source systems, processes the data, and then loads it into data stores where it can be analyzed. Analytics pipelines are designed to handle large volumes of incoming data from heterogeneous sources such as databases, applications, and devices.

1. Collect data
2. Process data
3. Store data
4. Analyze and visualize data
5. Predict future outcomes

For an illustration, see Figure 1, following.

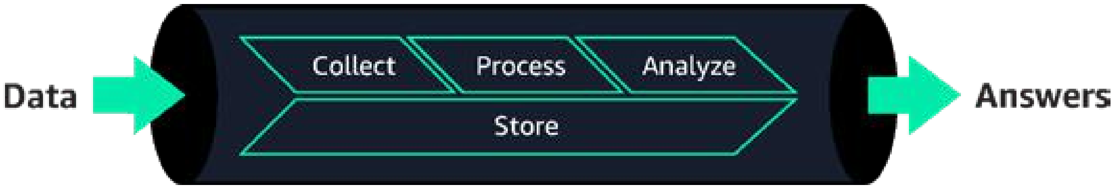

Figure 1: Analytics Pipeline

Collect data

Consider the different types of data—transactional data, log data, streaming data, and Internet of Things (IoT) data. Each type may be stored in data stores best suited for the data and its use. Some data stores are optimized for transactional or relational data and others for nonrelational or unstructured data. Your strategy should be to use a purpose-built database that best fits the data and the applications that produce or consume the data.

- **Transactional Data:** Data such as e-commerce purchase transactions and financial transactions, is typically stored in relational database management systems (RDBMS) or NoSQL database systems. The choice of database solution depends on the use case and application characteristics. An RDBMS solution is suitable for recording transactions and when transactions may need to update multiple table rows. A NoSQL database is suitable when the data is not well structured to fit into a defined schema or when the schema changes very often.

- **Log Data:** Reliably capturing system-generated logs will help you troubleshoot issues, conduct audits, and perform analytics using the information stored in the logs. A data lake is a popular storage solution for log data that is used for analytics.

- **Streaming Data:** Web applications, mobile devices, and many software applications and services can generate staggering amounts of streaming data—sometimes terabytes per hour—that need to be collected, stored, and processed continuously. This data is quite varied, often described as semi-structured or unstructured data.

- **IoT Data:** Devices and sensors around the world send messages continuously. Organizations see a growing need today to capture this data and derive intelligence from it.

Process data

The collection process gathers or extracts data from data sources, transforms the data, and stores the data in a separate destination such as another database, a data lake, or an analytics service like a data warehouse where it can be processed or analyzed.

Batch and real-time data

There are two types of processing workflows: batch and real time.

Batch data loading has been, and still is, pervasive. Nightly batch jobs extract data from one system, transform it into a ready-to-consume format for analytics, and load it into a destination. This introduces delays before data is available to those who need it.

Real-time processing performs inline data transformations in-memory while the data is still in transit before it is stored. These streaming technologies allow data to be ingested at a massive scale, in real time so you can perform analytics almost instantaneously.

- **Extract Transform Load (ETL):** ETL is the process of pulling or extracting data from multiple sources, transforming the data to fit a defined target schema (schema-on-write), and loading the data into a destination data store. ETL is normally a continuous, ongoing process with a well-defined workflow that occurs at specific times, such as nightly. Setting up and running ETL jobs can be a tedious task, and some ETL jobs may take hours to complete.

- **Extract Load Transform (ELT):** ELT is a variant of ETL where the extracted data is loaded into the target system before any transformations are made. The schema is defined when the data is read or used (schema-on-read). ELT typically works well when your target system is powerful enough to handle transformations and when you want to explore the data in ways not consistent with a predefined schema.

- **Real-Time Processing:** Real-time data streaming services enable you to collect, process, analyze, and deliver continuous streaming data at scale to your real-time applications and analytics solutions. The key here is that the data is available for analysis immediately, without waiting for a nightly batch ETL job to complete. Developers can easily build real-time applications and leverage the secure, highly available, durable, and scalable fully managed services. You can process streaming data sequentially and incrementally on a record-by-record basis or over sliding-time windows and use the processed data for a wide variety of analytics, including correlations, aggregations, filtering, and sampling.

Store data

You can store your data in either a data lake or an analytics tool like a data warehouse.

A data lake is a centralized repository for all data, including structured and unstructured. In a data lake, the schema is not defined, enabling additional types of analytics like big data analytics, full text search, real-time analytics, and machine learning. More and more, organizations are using data lakes as a central repository for all data so it can be used by downstream applications and analytics tools.

A data warehouse utilizes a predefined schema optimized for analytics, and the data is highly curated and serves as a single source of the truth from multiple data sources.

- **Data Lake:** Data lakes can handle the scale, agility, and flexibility required to combine different types of data and analytics approaches to gain deeper insights in ways that traditional data silos and data warehouses cannot. They give organizations the flexibility to use the widest array of analytics and machine learning services, with easy access to all relevant data, without compromising on security or governance.

- **Data Warehouse:** A data warehouse is a central repository of information coming from one or more data sources—or your data lake—where the data is transformed, cleansed, and deduplicated to fit into a predefined data model. A data warehouse is specially designed for data analytics, which involves reading large amounts of data to understand relationships and trends across the data. A database is used to capture and store data, such as recording details of a transaction. Using data warehouses, you can run fast analytics on large volumes of data and unearth patterns hidden in your data by leveraging BI tools. Data scientists query a data warehouse to perform offline analytics and spot trends. Users across the organization consume the data using ad-hoc SQL queries, periodic reports, and dashboards to make critical business decisions.

- **Data Mart:** A data mart is a simple form of a data warehouse focused on a specific functional area or subject matter and contains copies of a subset of data in the data warehouse. For example, you can have specific data marts for each division in your organization or segment data marts based on regions. You can build data marts from a large data warehouse, operational stores, or a hybrid of the two. Data marts are simple to design, build, and administer. However, because data marts are focused on specific functional areas, querying across functional areas can become complex because of the data distribution.

Analyze data

Unlock the real value of data

A modern analytics pipeline can utilize a variety of tools to unlock value hidden in the data. One size does not fit all. Any analytics tool should be able to access and process any data from the same source—your data lake.

Access data warehouse and the data lake easily

Data can be copied from your data lake into your data warehouse to fit a structured and normalized data model that takes advantage of a high-performance query engine. At the same time, some use cases require analysis of unstructured data in context with the normalized data in the data warehouse. Here, extending data warehouse queries to include data residing in both the data warehouse and the data lake, without the delay of data transformation and movement, is essential to timely insights.

Other big data analytics tools should be able to access the same data in the data lake. This allows everyone across the organization from business users to data scientists and everyone in between to have confidence in both the data and their analytics results.

- **Interactive Analysis:** Interactive analysis typically uses standard SQL query tools to access and analyze data. End users want fast results and the ability to modify queries quickly and rerun them.

- **Data Warehousing:** Data warehousing provides the ability to run complex analytic queries against large volumes of data—petabytes—using high-performance, analytics-optimized query engine.

- **Data Lake Analytics:** A new breed of data warehouse is emerging that extends data warehouse queries to a data lake to process structured or unstructured data in the data warehouse and data lake and scale up to exabytes without moving data.

- **Big Data Analytics:** Big data processing uses the Hadoop and Spark frameworks to process vast amounts of data.

- **Operational Analytics:** Operational analytics focuses on improving existing operations and uses data such as application monitoring, logs, and clickstream data.

- **Business intelligence (BI):** BI software is an easy-to-use application that retrieves, analyzes, transforms, and reports data for business decision-making. BI tools generally read data that is stored in an analytics service like a data warehouse or big data analytics system. BI tools create reports, dashboards, and visualizations and enable users to dive deeper into specific data on an ad-hoc basis. The results give organizations the power to accelerate and improve decision-making, increase operational efficiency, identify new opportunities and revenue potentials, identify market trends, and report KPIs.

Apply machine learning

As organizations generate, store, and analyze increasing amounts of data, there is a desire to use these vast data sets in automated ways to drive business results. They are increasingly relying on machine learning to automate tasks, provide personalized services to end users and customers, and increase the efficiency of operations by analyzing their data. Machine learning often feels a lot harder than it should because the process to build and train models and deploy them into production is complicated and slow.

Machine learning process

First, you need to collect and prepare your training data to discover which elements of your data set are important. Then, you need to select which algorithm and framework to use. After deciding on your approach, you need to teach the model how to make predictions by training, which requires a lot of compute. Then, you need to tune the model so it delivers the best possible predictions, which is often a tedious and manual effort. After you've developed a fully trained model, you need to integrate the model with your application and deploy this application on infrastructure that will scale. All of this takes a lot of specialized expertise, access to large amounts of compute and storage, and a lot of time to experiment and optimize every part of the process.

Easily set up machine learning

Machine Learning Services: ML Services enable developers and data scientists to quickly and easily build, train, and deploy machine learning models at any scale. It removes the complexity that gets in the way of successfully implementing machine learning across use cases and industries—from running models for real-time fraud detection to virtually analyzing biological impacts of potential drugs to predicting stolen-base success in baseball.

- **Artificial Intelligence (AI) Services:** AI Services provide ready-made intelligence for your applications and workflows. They easily integrate with your applications to address common use cases such as personalized recommendations, modernizing your contact center, improving safety and security, and increasing customer engagement.

- **Machine Learning Frameworks:** ML Frameworks enable you to experiment with and customize machine learning algorithms. They provide machine learning practitioners and researchers with the infrastructure and tools to accelerate deep learning in the cloud, at any scale.

Data lakes—The way forward for future innovation

Data lakes are going mainstream

Data lakes in the cloud are becoming a mainstream strategy for many organizations, providing promises of greater flexibility in the way data is handled and made available to decision-makers. A data lake can store raw and processed data in any format to be transferred and transformed at a later date as applications and end users demand. The thinking behind the concept is that the analytics or questions to be applied against the data may have not yet been identified, and by holding the data in a readily accessible environment, it is open for future innovation.

Avoid data silos at enterprise level

However, as with any major enterprise data initiative, the concept has to be sold to the enterprise. Data lakes absorb data from a variety of sources and store it all in one place, with all the necessary requirements for integration and security. Data lakes are a response to the eternal problem of data silos, attempting to bypass these various, fragmented environments to finally maintain data all in one place. The data lake also reduces the requirement for immediately processing or integrating the wide variety of data formats that comprise big data.

To learn more, visit Data Lakes and Analytics on AWS.

ABOUT AWS

For 13 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform. AWS offers over 165 fully featured services for compute, storage, databases, networking, analytics, robotics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media, and application development, deployment, and management from 61 Availability Zones (AZs) within 20 geographic regions, spanning the U.S., Australia, Brazil, Canada, China, France, Germany, India, Ireland, Japan, Korea, Singapore, Sweden, and the U.K. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs.
To learn more about AWS, visit **https://aws.amazon.com**.