aws | ExitCertified®

Supercharge Your Data Analytics with AWS

The cloud is big business, and it's getting bigger. As of 2022, over **60% of all corporate data** is being stored in the cloud.[1] By 2025, experts predict this will increase to **100 Zettabytes, or one trillion gigabytes, comprising half of all data accumulated**.[2] The insights buried in data are invaluable for helping companies make better business decisions, engage with customers and influence the development of products and services.

Although enterprises have been using cloud data storage for years, many haven't made full use of the data analysis opportunities offered by the cloud. Cloud vendors began developing their own managed services to make the development and management of cloud analytics easier by handling low-level administrative duties like provisioning and managing physical and virtual machines. As the leading cloud vendor, Amazon Web Services (AWS) provides a range of advanced tools that help companies study their data, helping them to make quicker business decisions, save costs, and store larger volumes of data.

WHAT'S INSIDE

COMPREHENDING CLOUD ANALYTICS

Data analytics is the process of analyzing millions of bits of data to identify patterns, anticipate future developments, and produce actionable business intelligence (BI). For example, a company could determine that people purchase more of one product in certain months, helping them to price products for slow seasons and determine the types of products they should sell in the future. Historically, analytic engineers conducted analyses with data housed on-premises using big data frameworks such as **Apache Hadoop** and **Apache Spark**. The process was slow and scalability was limited, unless organizations were prepared to invest significantly in additional hardware that would increase the data storage space and analytics capabilities of their on-premises storage.

These analytics frameworks, which revolutionized big data processing, relied on the fact that parallelism—using multiple computers rather than one to attack a job—was the key to processing large jobs. An early computer programmer named Grace Hopper preached that efficiency lies in numbers. She said in the pioneer days, rather than trying to grow a bigger ox to pull a heavy load, settlers added more oxen to pull it. She thought IT should follow that lead and said, “We shouldn’t be trying for bigger computers but for more systems of computers.”

However, in traditional on-premises data centers, it was expensive to purchase enough machines to get large analytics workloads processed in the time that business needs dictated. The scalability and economics of the cloud make those timelines possible. For example, if a job can be sufficiently parallelized and it takes 10 machines 10 hours to complete, then it might take 100 machines only 1 hour. In traditional data centers, the upfront expense to purchase a cluster of 100 machines may be prohibitive, but in the cloud, whether you run one machine for 100 hours, 10 machines for 10 hours, or 100 machines for one hour, all three solutions cost the same – 100 machine-hours.

Companies have discovered they could be more agile with **cloud analytics** as they can quickly add and remove the machines that analyze data, which reduces costs, and can perform better than is possible with on-premises infrastructure. But all cloud platforms are not alike, so you need to be sure to use a cloud analytics platform that is right for you. Since cloud storage platforms are not all created equal, it's important to know what to look for in a provider.

Picking a Cloud Storage Platform

Currently, AWS dominates the public cloud service arena, with about one-third of the market share.[3] AWS offers low-cost, easily accessible storage, along with a slate of helpful tools.

There are many reasons to choose AWS for your storage vendor:

Robust disaster recovery:

AWS has well-defined disaster recovery solutions that ensure your data is recoverable regardless of the magnitude of a disaster.

Powerful flexibility:

With AWS, companies can choose the operating system, programming languages, web application platform, database, and services they prefer, without being locked into specific solutions.

Decentralized backups:

The AWS cloud structure enables data backups across multiple regions using snapshots taken at specific points in time to ensure data safety regardless of location.

Superb scalability:

The massive infrastructure at AWS enables the analytics platform to supply computing resources regardless of the quantity of data storage required. Storage and processing can be independently scaled up or down as required while continuing to deliver high performance.

Affordable pricing:

The pay-for-use pricing model allows customers to purchase only the resources they need at any given time. Instead of paying for the amount of storage provisioned, companies pay only for the resources they actually use. This approach helps enterprises manage the cost of their data and increases profitability.

Enhanced security:

AWS servers and data centers support multiple layers of operational security. Security can be customized for governance and auditing policies and can meet geography-specific regulations. Additionally, AWS regularly performs infrastructure assessments to check for vulnerabilities.

Widespread availability:

The AWS Cloud currently covers 87 zones in 27 geographical regions worldwide and has plans for 21 more zones and 7 more regions in the future, so access is assured regardless of company location.[4]

Although many companies have taken steps toward migrating their data to the cloud, they are not yet making use of the benefits of cloud analytics to drive timely, strategic decision-making. For organizations currently considering or undertaking a transition to the cloud, AWS offers a full suite of **analytics tools** and services to supplement its AWS analytics stack.

DISCOVERING THE AWS ANALYTICS STACK

The **AWS analytics stack** contains a range of solutions for organizations of all sizes. From the Extract, Transform, and Load (ETL) phase to the processing, warehousing, operational analytics and visual data preparation stage, the stack provides purpose-built services at a low cost. Additionally, the AWS analytics stack integrates with BI solutions such as **Databricks**, **Cloudera**, **Tableau**, **IBM Cognos**, **Microsoft Power BI**, as well as open source tools such as Git and Kubernetes.

AWS also offers an array of additional **cloud-based analytics** solutions that integrate well with the rest of the AWS ecosystem. These solutions include compute, storage, analysis, networking, mobile development, IoT, enterprise applications, and security—which can be customized for governance and auditing policies and can meet geography-specific regulations.

Companies can use third-party tools, as well as native AWS tools, to enable additional functions, such as logging, alerts and incident management. AWS tools can either replace your company's existing third-party tools or work alongside them to deliver your desired outcomes. The AWS stack can also co-exist with other applications your team uses, such as Apache Hadoop, Databricks and Cloudera.

There are many AWS solutions that can help supercharge your cloud data analytics:

AMAZON ATHENA	Offers an interactive query service to analyze data using standard SQL to analyze large-scale data sets.
AMAZON ELASTIC COMPUTE CLOUD (EC2)	Provides secure, resizable compute capacity in the cloud.
AWS DATASYNC	Automates and accelerates migration of data from on-premises data centers to AWS.
AWS DATABASE MIGRATION SERVICE (DMS)	Migrates databases to AWS while remaining fully operational.
AWS LAKE FORMATION	Allows rapid set up of a secure data lake, eliminating manual management and monitoring.
AMAZON EMR	Delivers a big data platform to support petabyte-scale analysis using open source tools like Apache Hadoop and Apache Spark.
AWS GLUE	Provides serverless data integration capabilities to prepare data for analytics, machine learning (ML) and application development.

AMAZON KINESIS	Offers a managed service that scales elastically for real-time processing of streaming data.
AWS LAMBDA	Enables the running of code at scale without a server for virtually any application or backend service.
AMAZON REDSHIFT	Delivers a scalable cloud data warehouse to run BI tools using standard SQL.
AMAZON QUICKSIGHT	Provides interactive BI dashboards publishable on any device that can also be embedded into applications.
AMAZON SAGEMAKER	Supplies a fully managed environment for building, training and deploying ML models at scale.
AMAZON S3	Offers a range of data storage classes with scalability, security, and performance for virtually any use case.

In addition to this suite of services, the AWS analytics stack offers specific solutions to meet the requirements of various industry verticals. For example, **Amazon HealthLake**[5] is a HIPAA-eligible data storage service available to healthcare providers and life sciences companies. This service enables providers to perform individual patient or population health data analysis at scale.

MOVE TO CLOUD ANALYTICS WITH EASE

The underlying promise of big data is to enable data-driven decision-making. But as traditional on-premises data warehouses became incapable of accommodating the large quantities of data that needed to be stored, it became a challenge to capture, clean, and analyze all the information. That challenge is one of the main reasons organizations of all sizes have been migrating their data warehouses to the cloud over the past ten years. Many organizations in the cloud have moved their data to data lakes, centralized repositories, which unlike data warehouses, store vast quantities of both structured and unstructured data. Both data warehouses and data lakes have been gathering data for years, and companies now want to put that data to use. A Lake House, a new type of data architecture, combines a data warehouse and a data lake, making it easier for you to use your data.

Lake House Architecture on AWS

Organizations are starting to use a Lake House approach to storing data, providing the benefits of both a data lake and a data warehouse. Let's take a deeper look at the differences among the various ways to store data.

A data warehouse contains only structured data—data that fits into the rows and columns of a relational database. The primary goal of a data warehouse is to provide operational reporting across lines of business and product analytics to help companies make more informed decisions. Data warehousing requires an ETL process to edit, correct, and structure data before uploading it into the warehouse. Data analysts must create a predefined data format and fixed schema—a blueprint of how the data might relate to other tables or models—before they can run their queries. Creating and using this predefined format can take some time, preventing data scientists, analysts, and other lines of business from getting the data they need in a timely fashion. A data warehouse traditionally had no ability to perform ML, and typically, the data stored in a warehouse was append-only, making existing data immutable so it could be difficult to update or delete it.

The other type of storage, a data lake, is a centralized store that allows you to hold all your structured and unstructured data at any scale, no matter how much data you have. You can store your data as-is, without having to first structure the data, and run different types of analytics. Organizations use the data lake to combine the data that analysts store for individual projects and the transactional data that lines of business store, each in their own "siloed" data stores.

Without a data lake, teams have to communicate with the owner of the siloed data and request access. Sometimes the data can easily be shared, but other times it might contain confidential data that must first be obfuscated or redacted. This places a great burden on the owners of the siloed data, especially if there are a number of requests. With a centralized data lake strategy, data is pulled from the silos just once, and then data governance is applied, creating different levels of data that is appropriate for various audiences.

Organizations often require both a data lake and a data warehouse as they serve different needs. A data warehouse is a database optimized to quickly analyze relational data coming from transactional systems and line-of-business applications. A data lake stores relational data from those applications as well as non-relational data from mobile apps, IoT devices, and social media. While it may not be as quick to query as a data warehouse, a data lake can store all your data without massaging it or needing to know the type of information you may be seeking in the future. Years from now, as data scientists see they need data that was stored in the data lake years ago, they can obtain that data and use ML to uncover new insights.

The Lake House architecture makes data universally accessible through a single access point, regardless of whether the underlying data is stored in the data lake, a data warehouse, or in some other proprietary data store. A Lake House supports BI, SQL analytics, real-time data applications, data science and **ML**. A Lake House can quickly ingest large amounts of data from a variety of sources, such as line-of-business applications, ERP applications, point-of-sale systems, and CRM applications. It can hold data in its raw form, as well as in standard file formats like Apache Parquet and Apache Avro, or in purpose-built databases. The data in a Lake House is accessible to authorized users and can easily be moved into various data stores. Using various tools and real-time dashboards that deliver a comprehensive view of their data, analytical engineers and other IT professionals, as well as individuals outside of IT, can easily run data analytics to determine how well their business has been performing and what types of decisions they should make to improve future business outcomes.

OPTIMIZING THE BENEFITS OF AWS ANALYTICS

The AWS portfolio of managed services helps you build, secure, and scale end-to-end big-data applications. The platform's infrastructure supports real-time streaming and batch data processing for workloads that scale with demand. The AWS analytics stack also supports the entire spectrum of analytics from ingesting and cleaning data to generating BI outputs.

Managed Services

While you are free to run almost any kind of workload on virtual machines (VMs) in the AWS Cloud, there are a number of managed services that will help you avoid a lot of mundane administration tasks. For example, if you need to run a traditional Apache Hadoop cluster in the cloud, you could provision hundreds of VMs (Amazon EC2 instances) and install your Hadoop distribution of choice. But then you would be burdened with hundreds of VMs with Linux operating systems that need to be managed and patched. You would also have to periodically patch your Hadoop software. You would even have to prepare for the eventual failure of VMs and develop a plan to react to such a situation. Instead of doing all that, you could simply use the managed service that AWS provides for running Hadoop workloads - Amazon EMR. It takes care of the provisioning, high availability and basic management of the cluster of VMs. Or you could go one step further and use a fully serverless solution like AWS Glue to run your ETL processing, so you don't have to think about VMs at all.

Server Security

A suite of server-side security tools integrates with solutions such as server information and event management (SIEM) and incident management programs to complement native cloud security services. AWS ensures its data operations meet relevant regulatory standards, including PCI, HIPAA and specific national requirements.

Business Intelligence

The ultimate goal of big data is to help your organization make informed decisions through actionable data that will help it meet business objectives. With cloud analytics, business intelligence (BI) can be democratized throughout an organization. Business users can obtain the data they need without requiring deep programming knowledge.

Operational Flexibility

While the AWS stack includes a complete range of tools, users don't have to adopt every single component of the stack. They can continue using their existing BI tools like Tableau or IBM Cognos as they work seamlessly with AWS. However, if users prefer to use the AWS BI tool, Amazon QuickSight, they can use it. Users can use certain AWS tools and integrate their preferred external tools with the AWS stack for maximum operational flexibility.

Machine Learning

Machine Learning is a form of applied analytics that teaches your systems to make better decisions based on past events. AWS has a number of ML services that can be used to enhance your applications with predictive capabilities, helping any developer or data scientist to discover patterns in data and in build models to generate future predictions. While ML has long been under the purview of a few specialized experts, AWS provides various levels of ML services, so both developers with no ML experience and the most expert data scientists can uncover insights into your data. For example, developers can use a mature ML service like Amazon Rekognition to index a large number of images and then search for images by pre-defined categories. No data science experience is necessary because the developer is using an ML model that is built and trained by Amazon. For experienced data scientists, Amazon SageMaker gives them all the tools they need to build, train and deploy ML models at scale.

THE FINAL ANALYSIS

Storing enterprise data with AWS offers your company the opportunity to harness sophisticated cloud technology to augment your data analytics. AWS gives you access to cloud-based data analytics services that are fast and affordable because you pay only for the resources you use, not the resources that you plan to use and provision upfront. And because the cloud is durable, it's the best place to store and process large volumes of data. The AWS analytics stack can help you manage big data at whatever level you need to. Discover how it has helped some of the largest enterprises in the world.

GET THE TRAINING YOU NEED TO HARNESS AWS ANALYTICS

As the cloud becomes ever more prominent in enterprise computing, IT professionals and organizations must prepare to take advantage of the analytics capabilities it provides. Some data management skills are transferable from on-premises environments to the cloud, while other skills may need to be learned. Training offers the opportunity to upskill, cross-skill and re-skill your IT professionals to support analytics in various cloud architectures.

ExitCertified's **CloudCentrix** suite of training courses helps companies get the most value from their cloud journey through commercial and open source technologies. The training suite includes AWS-specific analytics tools as well as other widely available solutions to help organizations gain insight from their data.

CloudCentrix training can help your organization unlock the information buried in data stores of all types and can deliver insights via business intelligence applications. ExitCertified offers analytics training through all leading cloud and technology vendors, including **Authorized AWS Training**.

Find out more about our analytics training courses to gain insights from your data.

Contact Us

www.exitcertified.com

edu-customerexp@techdata.com

1 (800) 803-3948

Sources

1 **Statista**, "Share of corporate data stored in the cloud in organizations worldwide from 2015 to 2022", Mar 28, 2022. Source.

2 **Arcserve**, "The 2020 Data Attack Surface Report," 2020. Source.

3 **Srgresearch**, "Q2 Cloud Market Grows by 29% Despite Strong Currency Headwinds; Amazon Increases its Share," July 28, 2022. Source.

4 **AWS**, "AWS Global Infrastructure," 2022. Source.

5 **AWS**, "Amazon HealthLake," Source.