Introduction to AWS Analytics

AWS offers a powerful suite of analytics services designed to handle diverse data needs, from real-time streaming to big data processing. The AWS Analytics whitepapers explore key services like Amazon Athena for querying, Amazon EMR for big data processing, and Amazon Kinesis for real-time data streaming. Other highlights include Amazon OpenSearch Service for search and analytics, Amazon Redshift for data warehousing, and AWS Glue for ETL processes.

Topics

- Amazon Athena
- Amazon CloudSearch
- Amazon DataZone
- Amazon EMR
- Amazon FinSpace
- Amazon Kinesis
- Amazon Kinesis Data Firehose
- Amazon Managed Service for Apache Flink
- Amazon Kinesis Data Streams
- Amazon Kinesis Video Streams
- Amazon OpenSearch Service
- Amazon OpenSearch Serverless
- Amazon Redshift
- Amazon Redshift Serverless
- Amazon QuickSight
- AWS Clean Rooms
- AWS Data Exchange
- AWS Data Pipeline
- AWS Entity Resolution
- AWS Glue
- AWS Lake Formation
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

Athena is easy to use. Simply point to your data in Amazon S3, define the schema, and start querying using standard SQL. Most results are delivered within seconds. With Athena, there's no need for complex extract, transform, and load (ETL) jobs to prepare your data for analysis. This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets.

Athena is out-of-the-box integrated with AWS Glue Data Catalog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions and maintain schema versioning.

Amazon CloudSearch

Amazon CloudSearch is a managed service in the AWS Cloud that makes it simple and cost-effective to set up, manage, and scale a search solution for your website or application. Amazon CloudSearch supports 34 languages and popular search features such as highlighting, autocomplete, and geospatial search.

Amazon DataZone

Amazon DataZone is a data management service that you can use to publish data and make it available to the business data catalog through your personalized web application. You can access your data more securely regardless of where it is stored—on AWS, on premises, or in SaaS applications such as Salesforce. Amazon DataZone simplifies your experience across AWS services such as Amazon Redshift, Amazon Athena, AWS Glue, AWS Lake Formation, and Amazon QuickSight.

Amazon EMR

Amazon EMR is the industry-leading cloud big data platform for processing vast amounts of data using open source tools such as Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi, and Presto. Amazon EMR makes it easy to set up, operate, and scale your big data environments by automating time-consuming tasks such as provisioning capacity and tuning clusters. With Amazon EMR, you can run petabyte-scale analysis at less than half of the cost of

traditional on-premises solutions and over 3x faster than standard Apache Spark. You can run workloads on Amazon EC2 instances, on Amazon Elastic Kubernetes Service (Amazon EKS) clusters, or on-premises using Amazon EMR on AWS Outposts.

Amazon FinSpace

Amazon FinSpace is a data management and analytics service purpose-built for the financial services industry (FSI). FinSpace reduces the time you spend finding and preparing petabytes of financial data to be ready for analysis from months to minutes.

Financial services organizations analyze data from internal data stores such as portfolio, actuarial, and risk management systems as well as petabytes of data from third-party data feeds, such as historical securities prices from stock exchanges. It can take months to find the right data, get permissions to access the data in a compliant way, and prepare it for analysis.

FinSpace removes the heavy lifting of building and maintaining a data management system for financial analytics. With FinSpace, you collect data and catalog it by relevant business concepts such as asset class, risk classification, or geographic region. FinSpace makes it easy to discover and share data across your organization in accordance with your compliance requirements. You define your data access policies in one place and FinSpace enforces them while keeping audit logs to allow for compliance and activity reporting. FinSpace also includes a library of 100+ functions, such as time bars and Bollinger bands, for you to prepare data for analysis.

Amazon Kinesis

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information. Amazon Kinesis offers key capabilities to cost-effectively process streaming data at any scale, along with the flexibility to choose the tools that best suit the requirements of your application. With Amazon Kinesis, you can ingest real- time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning (ML), analytics, and other applications.

Amazon Kinesis enables you to process and analyze data as it arrives and respond instantly instead of having to wait until all your data is collected before the processing can begin. Amazon Kinesis currently offers four services: Kinesis Data Firehose, Managed Service for Apache Flink, Kinesis Data Streams, and Kinesis Video Streams.

Amazon Kinesis Data Firehose

Amazon Kinesis Data Firehose is the easiest way to reliably load streaming data into data stores and analytics tools. It can capture, transform, and load streaming data into Amazon S3, Amazon Redshift, Amazon OpenSearch Service, and Splunk, enabling near real-time analytics with existing business intelligence tools and dashboards you're already using today. It is a fully managed service that automatically scales to match the throughput of your data and requires no ongoing administration. It can also batch, compress, transform, and encrypt the data before loading it, minimizing the amount of storage used at the destination and increasing security.

You can easily create a Kinesis Data Firehose delivery stream from the AWS Management Console, configure it with a few clicks, and start sending data to the stream from hundreds of thousands of data sources to be loaded continuously to AWS—all in just a few minutes. You can also configure your delivery stream to automatically convert the incoming data to columnar formats such as Apache Parquet and Apache ORC, before the data is delivered to Amazon S3, for cost-effective storage and analytics.

Amazon Managed Service for Apache Flink

Amazon Managed Service for Apache Flink is the easiest way to analyze streaming data, gain actionable insights, and respond to your business and customer needs in real time. Amazon Managed Service for Apache Flink reduces the complexity of building, managing, and integrating streaming applications with other AWS services. SQL users can easily query streaming data or build entire streaming applications using templates and an interactive SQL editor. Java developers can quickly build sophisticated streaming applications using open source Java libraries and AWS integrations to transform and analyze data in real-time.

Amazon Managed Service for Apache Flink takes care of everything required to run your queries continuously and scales automatically to match the volume and throughput rate of your incoming data.

Amazon Kinesis Data Streams

Amazon Kinesis Data Streams is a massively scalable and durable real-time data streaming service. Kinesis Data Streams can continuously capture gigabytes of data per second from hundreds of thousands of sources such as website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events. The data collected is

available in milliseconds to enable real-time analytics use cases such as real-time dashboards, real-time anomaly detection, dynamic pricing, and more.

Amazon Kinesis Video Streams

Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, ML, playback, and other processing. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest streaming video data from millions of devices. It also durably stores, encrypts, and indexes video data in your streams, and allows you to access your data through easy-to-use APIs. Kinesis Video Streams enables you to playback video for live and on-demand viewing, and quickly build applications that take advantage of computer vision and video analytics through integration with Amazon Rekognition Video, and libraries for ML frameworks such as Apache MxNet, TensorFlow, and OpenCV.

Amazon OpenSearch Service

Amazon OpenSearch Service (OpenSearch Service) makes it easy to deploy, secure, operate, and scale OpenSearch to search, analyze, and visualize data in real-time. With Amazon OpenSearch Service, you get easy-to-use APIs and real-time analytics capabilities to power use-cases such as log analytics, full-text search, application monitoring, and clickstream analytics, with enterprise-grade availability, scalability, and security. The service offers integrations with open-source tools such as OpenSearch Dashboards and Logstash for data ingestion and visualization.

It also integrates seamlessly with other AWS services such as Amazon Virtual Private Cloud (Amazon VPC), AWS Key Management Service (AWS KMS), Amazon Kinesis Data Firehose, AWS Lambda, AWS Identity and Access Management (IAM), Amazon Cognito, and Amazon CloudWatch, so that you can go from raw data to actionable insights quickly.

Amazon OpenSearch Serverless

Amazon OpenSearch Serverless is a serverless option in Amazon OpenSearch Service. As a developer, you can use OpenSearch Serverless to run petabyte-scale workloads without configuring, managing, and scaling OpenSearch clusters. You get the same interactive millisecond response times as OpenSearch Service with the simplicity of a serverless environment.

The vector engine for Amazon OpenSearch Serverless, now in preview, adds a simple, scalable, and high-performing vector storage and search capability to help developers build ML-augmented search experiences and generative AI applications without having to manage vector database

infrastructure. Use cases for vector search collections include image search, document search, music retrieval, product recommendation, video search, location-based search, fraud detection, and anomaly detection.

Amazon Redshift

Amazon Redshift is the most widely used cloud data warehouse. It makes it fast, simple and cost-effective to analyze all your data using standard SQL and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against terabytes to petabytes of structured and semi-structured data, using sophisticated query optimization, columnar storage on high-performance storage, and massively parallel query completion. Most results come back in seconds. You can start small for just $0.25 per hour with no commitments and scale out to petabytes of data for $1,000 per terabyte per year, less than a tenth the cost of traditional on-premises solutions.

Amazon Redshift Serverless

Amazon Redshift Serverless makes it easier to run and scale analytics without having to manage your data warehouse infrastructure. Developers, data scientists, and analysts can work across databases, data warehouses, and data lakes to build reporting and dashboarding applications, perform near real-time analytics, share and collaborate on data, and build and train machine learning (ML) models. Go from large amounts of data to insights in seconds.

Amazon Redshift Serverless automatically provisions and intelligently scales data warehouse capacity to deliver fast performance for even the most demanding and unpredictable workloads, and you pay only for what you use. Just load data and start querying right away in Amazon Redshift Query Editor or in your favorite business intelligence (BI) tool and continue to enjoy the best price performance and familiar SQL features in an easy-to-use, zero administration environment.

Amazon QuickSight

Amazon QuickSight is a fast, cloud-powered business intelligence (BI) service that makes it easy for you to deliver insights to everyone in your organization. QuickSight lets you create and publish interactive dashboards that can be accessed from browsers or mobile devices. You can embed dashboards into your applications, providing your customers with powerful self-service analytics. Amazon QuickSight easily scales to tens of thousands of users without any software to install, servers to deploy, or infrastructure to manage.

AWS Clean Rooms

AWS Clean Rooms helps companies and their partners more easily and securely analyze and collaborate on their collective datasets–without sharing or copying one another's underlying data. With AWS Clean Rooms, customers can create a secure data clean room in minutes, and collaborate with any other company on the AWS Cloud to generate unique insights about advertising campaigns, investment decisions, and research and development.

AWS Data Exchange

AWS Data Exchange makes it easy to find, subscribe to, and use third-party data in the cloud. Qualified data providers include category-leading brands such as Reuters, who curate data from over 2.2 million unique news stories per year in multiple languages; Change Healthcare, who process and anonymize more than 14 billion healthcare transactions and $1 trillion in claims annually; Dun & Bradstreet, who maintain a database of more than 330 million global business records; and Foursquare, whose location data is derived from 220 million unique consumers and includes more than 60 million global commercial venues.

Once subscribed to a data product, you can use the AWS Data Exchange API to load data directly into Amazon S3 and then analyze it with a wide variety of AWS analytics and ML services. For example, property insurers can subscribe to data to analyze historical weather patterns to calibrate insurance coverage requirements in different geographies; restaurants can subscribe to population and location data to identify optimal regions for expansion; academic researchers can conduct studies on climate change by subscribing to data on carbon dioxide emissions; and healthcare professionals can subscribe to aggregated data from historical clinical trials to accelerate their research activities.

For data providers, AWS Data Exchange makes it easy to reach the millions of AWS customers migrating to the cloud by removing the need to build and maintain infrastructure for data storage, delivery, billing, and entitling.

AWS Data Pipeline

AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premises data sources, at specified intervals. With AWS Data Pipeline, you can regularly access your data where it’s stored, transform and process it at scale, and efficiently transfer the results to AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon EMR.

AWS Data Pipeline helps you easily create complex data processing workloads that are fault tolerant, repeatable, and highly available. You don't have to worry about ensuring resource availability, managing inter-task dependencies, retrying transient failures or timeouts in individual tasks, or creating a failure notification system. AWS Data Pipeline also allows you to move and process data that was previously locked up in on-premises data silos.

AWS Entity Resolution

AWS Entity Resolution is a service that helps you match and link related records stored across multiple applications, channels, and data stores without building a custom solution. Using flexible, configurable ML and rule-based techniques, AWS Entity Resolution can remove duplicate records, create customer profiles by connecting different customer interactions, and personalize experiences across advertising and marketing campaigns, loyalty programs, and e-commerce. For example, you can create a unified view of customer interactions by linking recent events, such as ad clicks, cart abandonment, and purchases, into a unique match ID.

AWS Glue

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics. You can create and run an ETL job with a few clicks in the AWS Management Console. You simply point AWS Glue to your data stored in AWS, and AWS Glue discovers your data and stores the associated metadata (such as table definition and schema) in the AWS Glue Data Catalog. Once cataloged, your data is immediately searchable, queryable, and available for ETL.

AWS Glue Data Integration Engines provide access to data using Apache Spark, PySpark, and Python. With the addition of AWS Glue for Ray, you can further scale your workloads using Ray, an open-source unified compute framework.

AWS Glue Data Quality can measure and monitor the data quality of Amazon S3 based data lakes, data warehouses, and other data repositories. It automatically computes statistics, recommends quality rules, and can monitor and alert you when it detects missing, stale, or bad data. You can access it in the AWS Glue Data Catalog and in AWS Glue Data Catalog ETL jobs.

AWS Lake Formation

AWS Lake Formation is a service that makes it easy to set up a secure data lake in days. A data lake is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. A data lake enables you to break down data silos and combine different types of analytics to gain insights and guide better business decisions.

However, setting up and managing data lakes today involves a lot of manual, complicated, and time-consuming tasks. This work includes loading data from diverse sources, monitoring those data flows, setting up partitions, turning on encryption and managing keys, defining transformation jobs and monitoring their operation, re-organizing data into a columnar format, configuring access control settings, deduplicating redundant data, matching linked records, granting access to data sets, and auditing access over time.

Creating a data lake with Lake Formation is as simple as defining where your data resides and what data access and security policies you want to apply. Lake Formation then collects and catalogs data from databases and object storage, moves the data into your new Amazon S3 data lake, cleans and classifies data using ML algorithms, and secures access to your sensitive data. Your users can then access a centralized catalog of data which describes available data sets and their appropriate usage. Your users then leverage these data sets with their choice of analytics and ML services, such as Amazon EMR for Apache Spark, Amazon Redshift, Amazon Athena, SageMaker, and Amazon QuickSight.

Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Amazon Managed Streaming for Apache Kafka (Amazon MSK) is a fully managed service that makes it easy for you to build and run applications that use Apache Kafka to process streaming data. Apache Kafka is an open-source platform for building real-time streaming data pipelines and applications. With Amazon MSK, you can use Apache Kafka APIs to populate data lakes, stream changes to and from databases, and power ML and analytics applications.

Apache Kafka clusters are challenging to setup, scale, and manage in production. When you run Apache Kafka on your own, you need to provision servers, configure Apache Kafka manually, replace servers when they fail, orchestrate server patches and upgrades, architect the cluster for high availability, ensure data is durably stored and secured, setup monitoring and alarms, and carefully plan scaling events to support load changes. Amazon MSK makes it easy for you to build and run production applications on Apache Kafka without needing Apache Kafka infrastructure management expertise. That means you spend less time managing infrastructure and more time building applications.

With a few clicks in the Amazon MSK console you can create highly available Apache Kafka clusters with settings and configuration based on Apache Kafka's deployment best practices. Amazon MSK automatically provisions and runs your Apache Kafka clusters. Amazon MSK continuously monitors cluster health and automatically replaces unhealthy nodes with no downtime to your application. In addition, Amazon MSK secures your Apache Kafka cluster by encrypting data at rest.

Credit to: AWS Documentation