Gartner

Building an Analytics and AI Architecture Using Amazon Web Services

Published 30 November 2022 - ID G00778080 - 50 min read

By Analyst(s): Zain Khan, Mayank Talwar

Initiatives: Analytics and Artificial Intelligence for Technical Professionals; Data Management Solutions for Technical Professionals; Evolve Technology and Process Capabilities to Support D&A

Amazon Web Services offers a wide array of analytics and AI offerings, but are they right for your use case? This document helps data and analytics technical professionals to understand the different analytics and AI tools from AWS in order to build an end-to-end analytics and AI architecture.

Overview

Key Findings

- AWS has a broad range of services catered to analytics, ML and AI domains. The myriad of options can also cause confusion, and it can become challenging to select the right tool for each use case.
- AWS is developing self-service analytics capabilities through serverless, federation and no-code features to empower BI analysts and citizen data scientists to perform data integration, exploration and discovery without needing advanced data engineering knowledge.
- QuickSight is evolving into a mature BI tool and now features embedding and augmented analytics capabilities.
- SageMaker continues to expand and build new features into its platform and now offers dedicated platforms for data scientists and citizen data scientists.
- AWS analytics offerings remain confined to work with AWS data sources, thereby restricting multicloud or hybrid use cases.

Recommendations

Data and analytics technical professionals looking to build or modernize their analytics architecture on AWS should:

- Ease sophisticated data ingestion and curation tasks by using the federation, low-code, no-code and serverless features of Athena, Glue DataBrew and SageMaker Data Wrangler.
- Utilize ML-infusion in Athena, QuickSight and Redshift to augment advanced analytics in business intelligence workflows and ease ML development and adoption.
- Maximize the productivity of data science teams by aligning ML tools with the appropriate skill set. Citizen data scientists can use SageMaker Canvas, and data scientists can use SageMaker Studio. SageMaker Studio Lab provides a free development environment targeted at students and researchers.

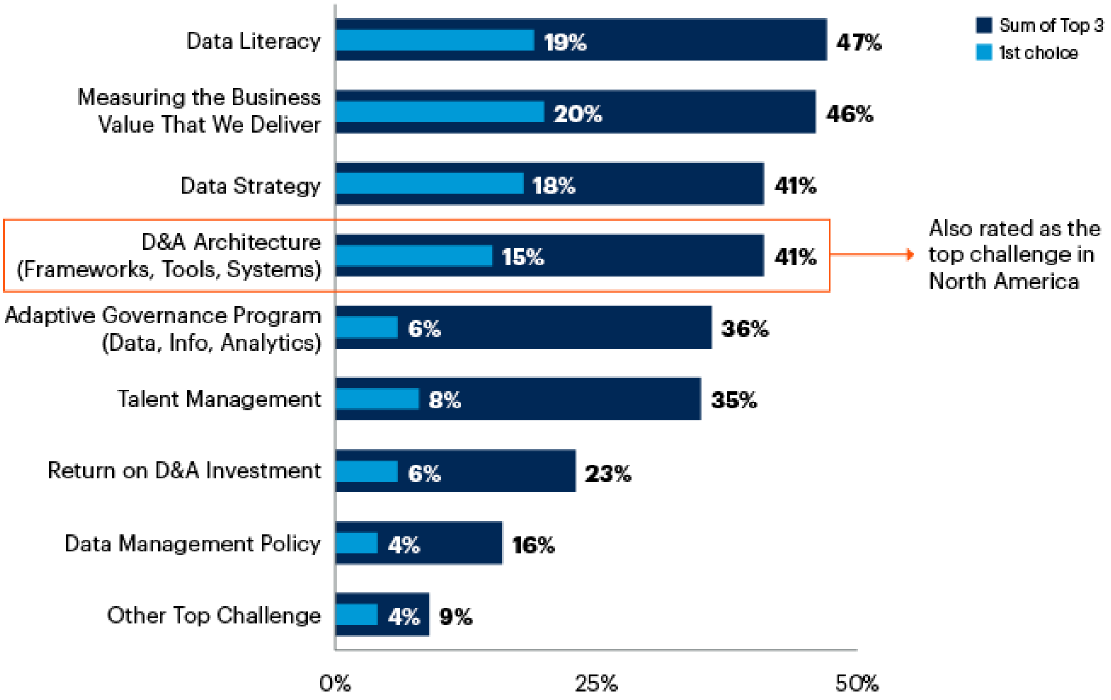
Analysis

Introduction

Data and analytics (D&A) technical professionals and leaders identify data and analytics architecture as one of their major challenges. For data and analytics leaders in Gartner's Chief Data Officer Pre-Survey 2022 (Figure 1), data and analytics architecture tied with data strategy as the third greatest challenge on a global level. D&A architecture was the top challenge in North America.

Download the Figures for This Material

Figure 1: Top 3 Data and Analytics Challenges for Analytics Leaders



Top 3 Data and Analytics Challenges for Analytics Leaders
Ranking 1-3

n = 157, CDOs and D&A leaders

Q: Below are some of the challenges data and analytics leaders face. Which are you most challenged by? Rank 1-3
Source: Gartner 2021 Chief Data Officer Pre-Survey
778080_C

The situation can be exacerbated by the wide array of analytics and AI products and overlapping features offered for different use cases. As a result, it can become challenging to select the correct cloud provider and then align the right tools to the organization's use cases.

Amazon Web Services offers cloud services covering cloud infrastructure, platform and software services. Included within its portfolio are tools covering the data management, analytics and AI space as well.

What Does This Research Cover?

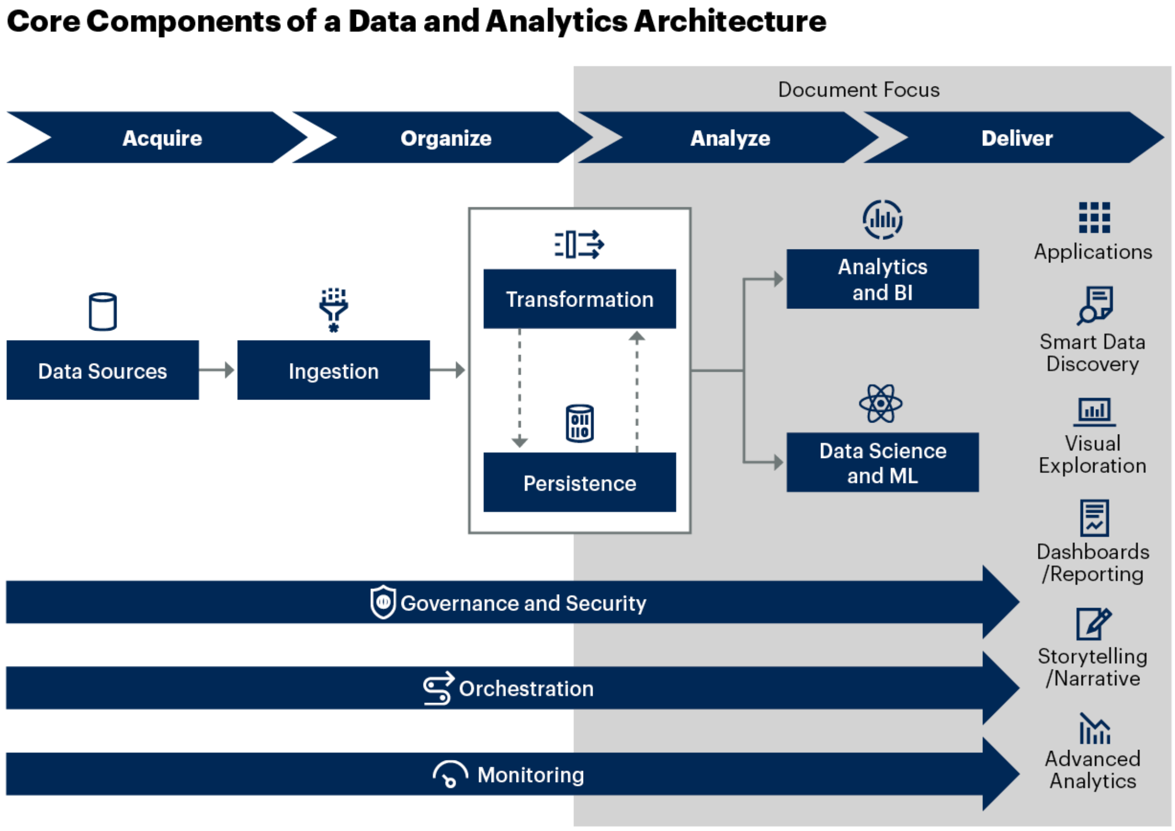
This research offers brief explanations and comparisons between different services for business intelligence, machine learning and AI on the AWS platform as well as guidance on when to use each service and which persona fits them. Even though the prime focus is on analytics and AI offerings, there is also a brief discussion on the transformation and ingestion services as they now offer self-serviced capabilities that can be used by data scientists and citizen data scientists.

This research does not aim to offer a comparison between AWS services and those of other cloud providers or third-party tools. For a comparison of AWS with Microsoft's and Google Cloud Platform's ML offerings, read Solution Comparison for Cloud Data Science and Machine Learning Platforms.

For more details on the tools used within the data management domain, please consult the partner document Building a Data Management Architecture on Amazon Web Services.

Gartner offers a data, analytics and architecture blueprint for analytics use cases. It can be seen in Figure 2. It divides the architecture between four domains (acquire, organize, and analyze and deliver) and six subdomains (data sources, ingestion, transformation, persistence, analytics and BI, and data science and ML). Governance, security, orchestration and monitoring are employed throughout the process and cover the entire architecture.

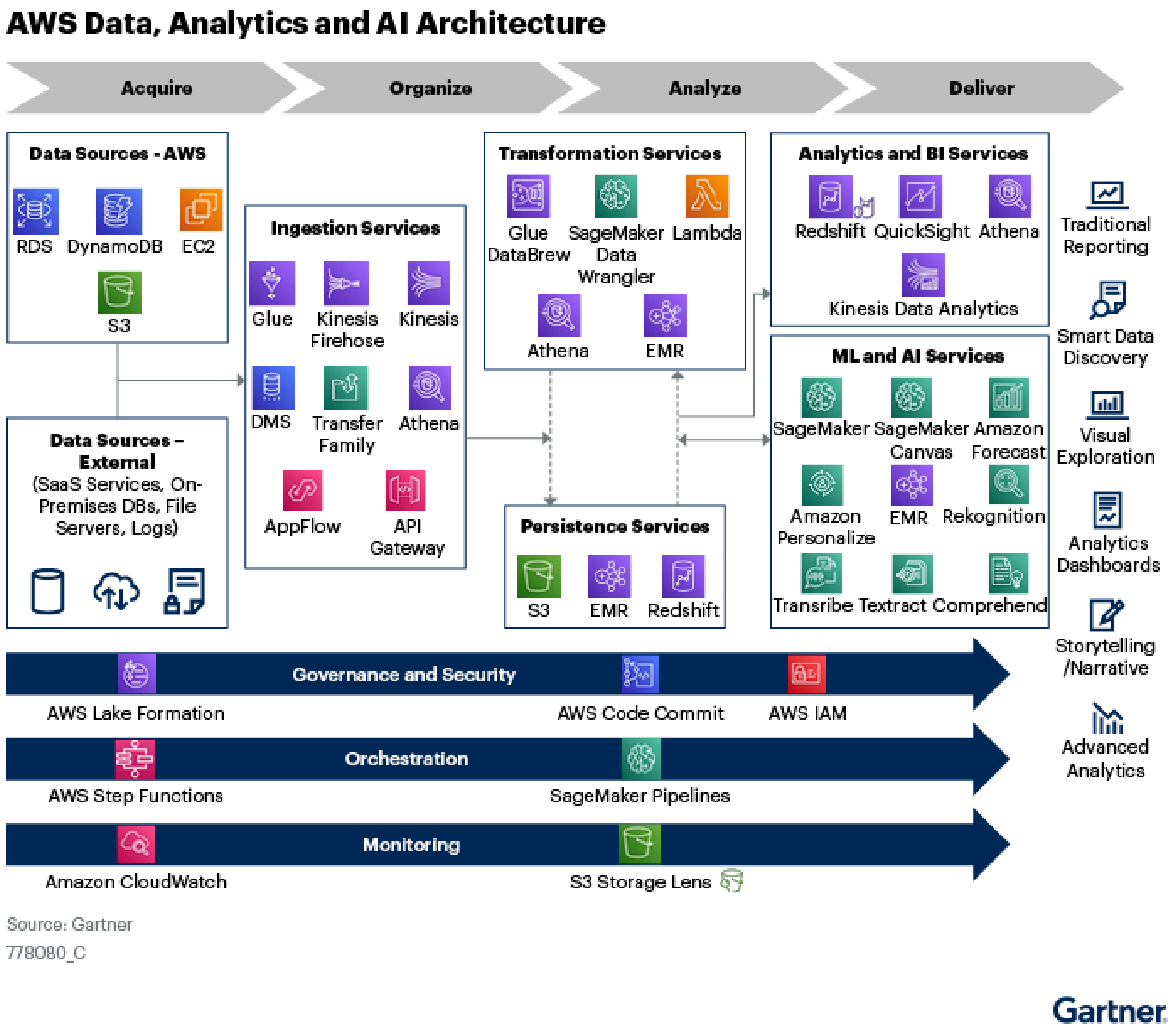
Figure 2: Core Components of a Data, Analytics and AI Architecture



Source: Gartner
763681_C

Figure 3 shows the AWS analytics and AI architecture based on Gartner's blueprint. It should be noted that the products shown are not exhaustive. The intent is to show the mainstream analytics offerings.

Figure 3: AWS Data, Analytics and AI Architecture



Source: Gartner
778080_C

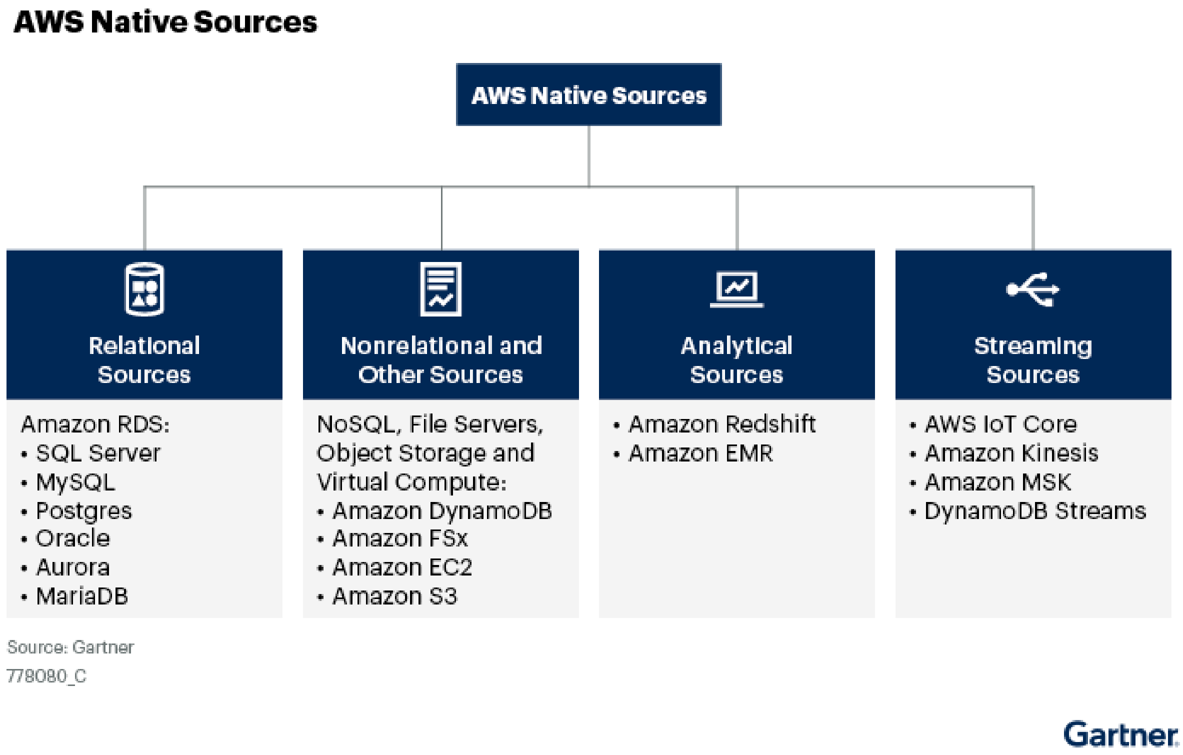
Acquire

The acquisition part of a data, analytics and AI architecture involves determining the data sources and the types of data they contain (structured, semistructured or unstructured) as well as determining the qualities of data (batch or streaming). These data sources can be external to AWS (third-party services such as SaaS ERP/CRM systems or databases) or they can be native (AWS services). This research covers only native AWS services and sources. Once the data sources have been identified, appropriate ingestion services are employed to bring the data within the AWS analytics ecosystem.

Data Sources

AWS offers a wide array of data sources. Figure 4 gives an overview of the most common sources for data and analytics. For more details, read Building a Data Management Architecture on Amazon Web Services.

Figure 4: AWS-Native Sources



Source: Gartner
778080_C

Ingestion Services

AWS offers a wide variety of services catered to different data sources and data properties. AWS Glue is a managed serverless Apache Spark offering that can connect to Amazon RDS, Amazon Redshift, Amazon S3 and Amazon DynamoDB. It now offers a studio environment, called AWS Glue Studio, that grants visual authorship, thereby allowing data scientists and BI analysts to discover and ingest data as well.

For streaming sources, such as social media and IoT sources, Kinesis Data Streams and Kinesis Video Streams can ingest real-time data and video streams, respectively. Kinesis Data Streams can then be connected with Amazon Kinesis Data Firehose for persisting data in S3 or Redshift. For ingesting files (such as CSV and Excel) from file servers, use AWS Transfer Family services, which offer SFTP, FTPS and FTP for secure file transfer into S3. Database Migration Service (DMS) is used to ingest data from databases hosted within the AWS environment and addresses both homogeneous and heterogeneous database migrations. It also caters to change data capture (CDC) and can be paired with AWS Schema Conversion Tool (SCT) for schema inference and migration.

There can be instances when data scientists need direct access to ERP or CRM systems like Salesforce. AWS offers Amazon AppFlow, which is a managed integration service that is used for transferring data between SaaS applications, like Salesforce, SAP, Zendesk, Slack and ServiceNow, and AWS services like Amazon S3 and Redshift. AppFlow now also integrates with AWS Glue DataBrew, thereby offering a visual no-code data integration and processing pipeline catered specifically for data scientists and BI analysts. [1]

More details on ingestion methods and services can be found in Building a Data Management Architecture on Amazon Web Services.

Organize

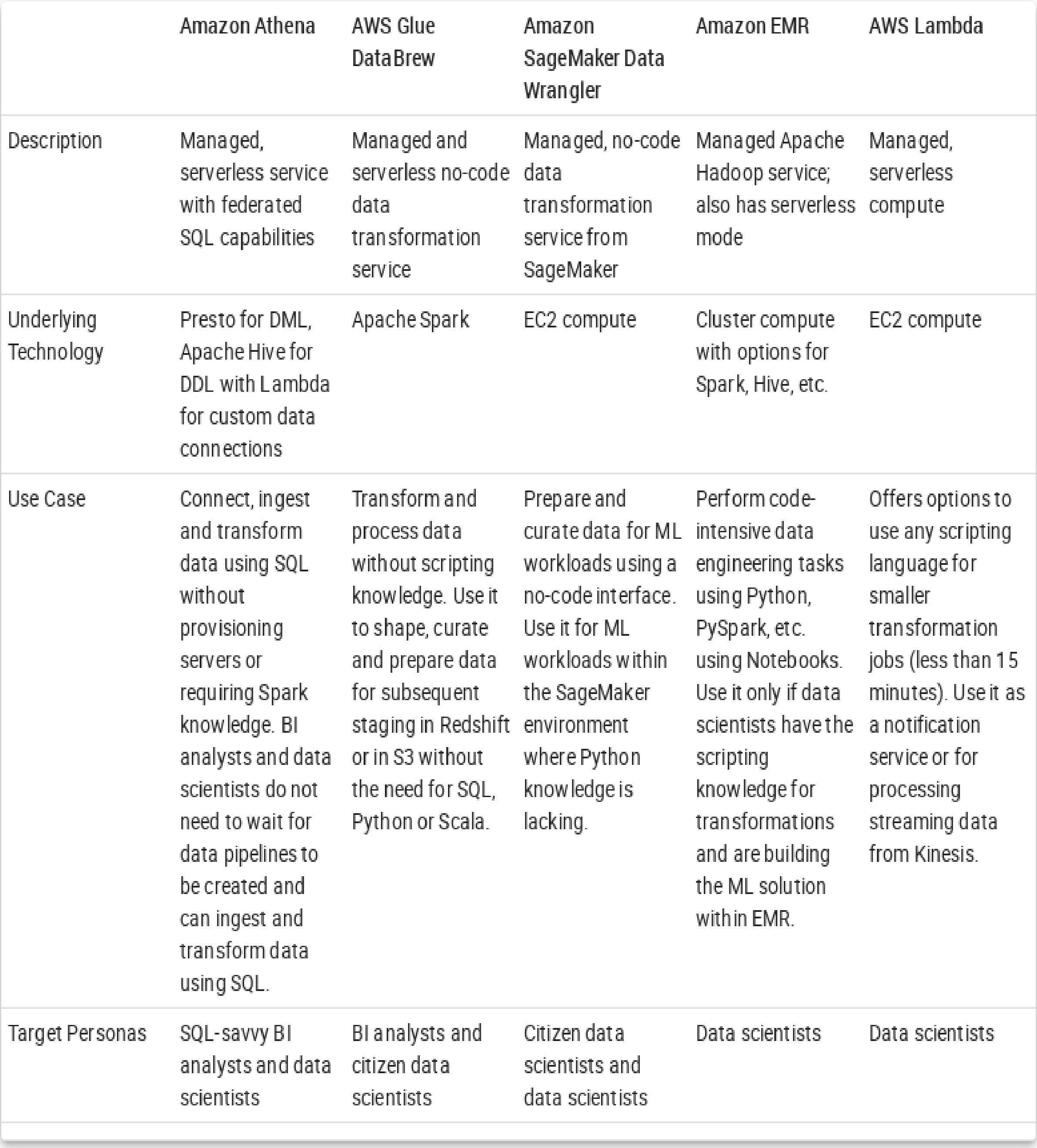
Once data sources have been identified and ingestion tools and processes determined, data needs to be either stored first (extract, load and transform [ELT]) or transformed prior to storing (extract, transform and load [ETL]). Traditionally, ETL/ELT development and data integration tasks have been restricted to data engineers. However, AWS is augmenting its services with serverless, low-code/no-code and federation features that ease data engineering tasks for citizen data scientists and BI analysts. This section explores data transformation capabilities as they relate to analytics personas. For a data engineering perspective, consult the companion document Building a Data Management Architecture on Amazon Web Services.

Transformation Services

Transformation services are responsible for cleaning, deduplicating, integrating and standardizing data prior to analytics consumption. AWS offers a range of services with different compute engines, use cases and target personas. Table 1 shows the comparison of these services as they relate to technical professionals in the analytics and advanced analytics space.

Table 1: How Should Data and Analytics Technical Professionals Decide Between the Different Transformation Services on AWS?

(Enlarged table in Appendix)



	Amazon Athena	AWS Glue DataBrew	Amazon SageMaker Data Wrangler	Amazon EMR	AWS Lambda
Description	Managed, serverless service with federated SQL capabilities	Managed and serverless no-code data transformation service	Managed, no-code data transformation service from SageMaker	Managed Apache Hadoop service; also has serverless mode	Managed, serverless compute
Underlying Technology	Presto for DML, Apache Hive for DDL with Lambda for custom data connections	Apache Spark	EC2 compute	Cluster compute with options for Spark, Hive, etc.	EC2 compute
Use Case	Connect, ingest and transform data using SQL without provisioning servers or requiring Spark knowledge. BI analysts and data scientists do not need to wait for data pipelines to be created and can ingest and transform data using SQL.	Transform and process data without scripting knowledge. Use it to shape, curate and prepare data for subsequent staging in Redshift or in S3 without the need for SQL, Python or Scala.	Prepare and curate data for ML workloads using a no-code interface. Use it for ML workloads within the SageMaker environment where Python knowledge is lacking.	Perform code-intensive data engineering tasks using Python, PySpark, etc. using Notebooks. Use it only if data scientists have the scripting knowledge for transformations and are building the ML solution within EMR.	Offers options to use any scripting language for smaller transformation jobs (less than 15 minutes). Use it as a notification service or for processing streaming data from Kinesis.
Target Personas	SQL-savvy BI analysts and data scientists	BI analysts and citizen data scientists	Citizen data scientists and data scientists	Data scientists	Data scientists

Source: AWS blogs and Gartner research notes

It should be noted that AWS Glue DataBrew now supports writing prepared data into Tableau Hyper format.

Persistence Services

AWS offers a wide variety of options of persisting data for data engineering and analytics use cases. For complete details on their features and capabilities, consult Building a Data Management Architecture on Amazon Web Services.

Amazon S3

Amazon S3 is object storage and is considered the de facto storage and persistence layer for machine learning workloads. Most often, data is stored iteratively in S3 with different layers depending on the refinement of data (e.g., bronze layer for raw data, silver layer for integrated and curated data and gold layer for final transformed data). Data scientists should work with data engineers to ensure data has been partitioned, cleaned, transformed and stored in an optimal format (e.g., Apache Parquet) before accessing. For more details on file formats, read Working With Semistructured and Unstructured Datasets.

When Should Data and Analytics Technical Professionals Use Amazon S3?

Because S3 can store different types of data and offers limitless storage capabilities, it is most often chosen to store input data for ML development in SageMaker. Use S3 to store both input and output datasets and metrics within S3 in the appropriate buckets. S3 can also be used to store structured data (e.g., CSV) for Athena querying and visualization in QuickSight. Compared with Redshift or EMR, S3 offers the most support for storing any type of data: structured, semistructured or unstructured. It can also house images and video clips for further analysis by Amazon Comprehend or Amazon Rekognition in AI application and development use cases. S3's close integration with AWS Lake Formation enables strong governance, security and lakehouse features. S3 can also be used to store ML model training data, artifacts, feature stores and predictions.

Amazon Redshift

Redshift is AWS' proprietary data warehouse offering and is built for petabyte-scale business intelligence workloads involving complex SQL analytics. It stores data within its nodes to offer maximum performance and scalability and now offers a serverless option to quickly provision clusters without specifying node types. It offers the ability to store data in S3 through the usage of RA3 nodes.

When Should Data and Analytics Technical Professionals Use Amazon Redshift?

Use it to house structured and semistructured data for business intelligence workloads and data warehouse use cases. Redshift can be accessed by SageMaker and QuickSight as well. Most often, data will be stored in its final reformed format within Redshift to be exposed in QuickSight for interactive BI exploration. Note that Redshift cannot work with unstructured data, so if your use case is around storing log, audio, image, PDF or video files, then use S3. Redshift also offers the ability to query data in S3 through Redshift Spectrum as external tables and reduces the need to physically move data. To independently scale compute from storage, use RA3 nodes because they give the ability to store data in S3.

It should be noted that streaming data ingestion is also now supported, but is in public preview. [2]

Amazon EMR

Amazon EMR stores data within its cluster nodes (via HDFS) or in Amazon S3 (via EMRFS) and is suited for petabyte-scale data processing needs using the Hadoop stack (Pig, Sqoop, Hive, Presto, HBase, etc.). It needs highly trained big data professionals to conduct daily activities, but makes node selection and scaling easy. It also offers the serverless option, which makes it easier for analytics professionals to quickly set up nodes and clusters.

When Should Data and Analytics Technical Professionals Use Amazon EMR?

Use it for storage when data science workloads are being performed within the EMR environment and services from the Hadoop stack are required. However, EMR also offers options to decouple storage from compute through EMRFS, where data can be persisted on S3 while compute can run on the EMR nodes. Just like Redshift is for data warehousing and business intelligence, EMR can be used to house data for data science workloads. However, EMR can store all sorts of files — structured, semistructured and unstructured — and now also offers Hudi integration. This means it can also be used as a lakehouse implementation without needing Lake Formation over S3.

It should be noted that these services are mostly used in concert to deliver a complete analytics solution. Data can be staged in S3 initially and then moved to Redshift or EMR for the appropriate analytics or data science use case.

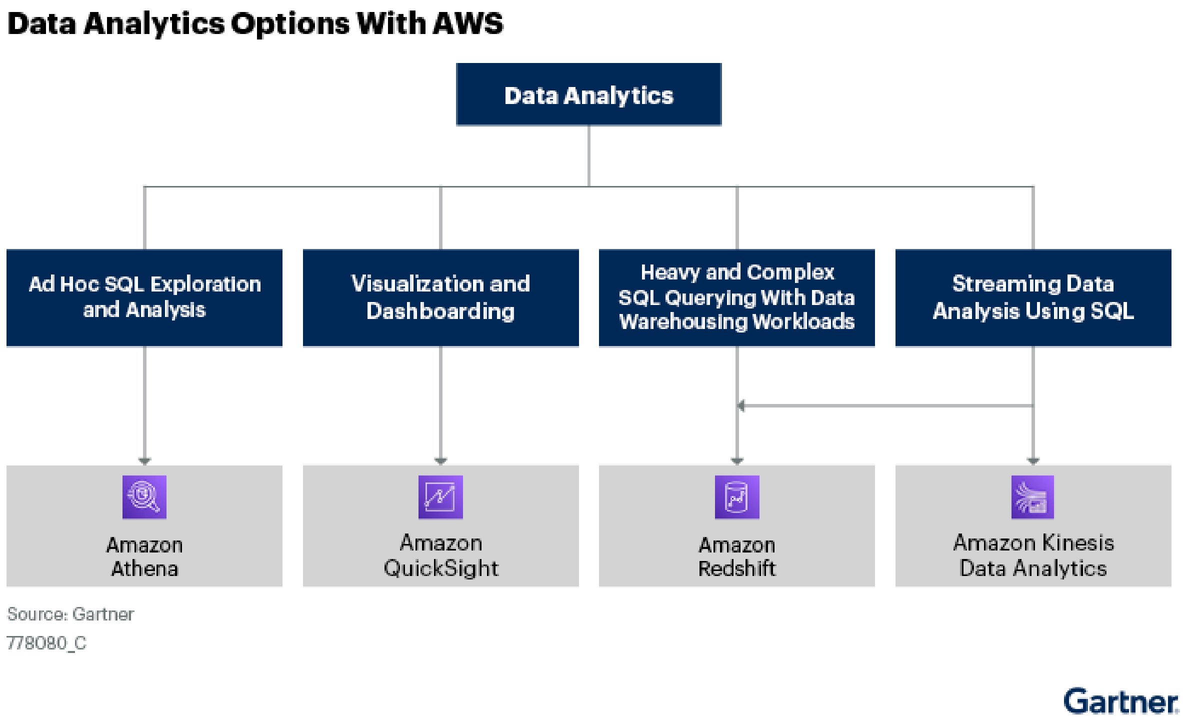
Analyze and Deliver

Once data has been ingested, loaded and transformed in the appropriate manner, it is analyzed through Redshift, Athena and EMR and is ultimately delivered using visualizations in QuickSight or as ML solutions in SageMaker. For a complete scoring guide on AWS analytical data stores, please consult Solution Scorecard for AWS Cloud Analytical Data Stores.

Analytics and BI Services

AWS offers a wide variety of options to analyze data for different scenarios. Figure 5 aligns the different analytics needs with the most common AWS services used for each.

Figure 5: Data Analytics Options With AWS


Data Analytics Options With AWS
Data Analytics
Ad Hoc SQL Exploration and Analysis
Visualization and Dashboarding
Heavy and Complex SQL Querying With Data Warehousing Workloads
Streaming Data Analysis Using SQL
Amazon Athena
Amazon QuickSight
Amazon Redshift
Amazon Kinesis Data Analytics
Gartner

Source: Gartner
778080_C

For programmatic analysis, AWS offers Redshift, Athena and Kinesis Data Analytics. For visual analytics, AWS offers QuickSight. Table 2 shows the use cases for these services as well as the personas required for each.

Amazon QuickSight

QuickSight is the primary business intelligence and data visualization offering from AWS and offers a custom in-memory engine called SPICE for responsive query performance. It is serverless, scales automatically, and offers a pay-per-use model and row-level security. It can connect to structured and semistructured data housed in S3, Athena, SageMaker, Redshift and on-premises databases through Java Database Connectivity/Open Database Connectivity (JDBC/ODBC). It now also offers connections to Snowflake and Teradata. However, it does not offer support for Parquet or open table formats like Delta Lake. Apart from object stores and databases, QuickSight also offers direct connection support for SaaS platforms such as JIRA and ServiceNow.

QuickSight now has competitive features such as:

- **ML insights:** Users can perform anomaly detection and forecasting and build automatic narratives within their dashboards without requiring in-depth statistics or coding.
- **QuickSight Q:** Q is the NLP service used to build dashboards and reports in response to natural language business questions. It provides autocompletions and spell-checks, as well as suggesting acronyms and synonyms.
- **SageMaker integration:** QuickSight integrates with SageMaker using the “Augment with SageMaker” option to run inferences on SageMaker models.
- **SPICE:** SPICE stands for Super-fast, Parallel, In-memory Calculation Engine and is used as an alternative to a direct query. It processes analytical queries faster, and data stored in SPICE can be reused without incurring further costs.
- **Dashboard sharing and embedding:** QuickSight offers the ability to share read-only snapshots of dashboard analysis with other users in the same account using email links. It should be noted that the underlying data is not shared — only the preserved state of the dashboard at that particular time. Users can view and filter the dashboard, but changes are lost when the session is closed. APIs can be used to share the dashboard with anyone on the internet. Dashboards and the Q search bar can also be embedded into websites and apps using QuickSight API.
- **Emailing reports:** Dashboards can also be shared in the form of reports through emails on a scheduled basis. However, this requires the users to have the same QuickSight subscription; have the dashboard already shared; and exist as readers, authors or admins.

- **Printing and PDF:** Dashboards can be printed for offline viewing and can also be converted to PDF documents. PDFs can also be attached to dashboard email reports.

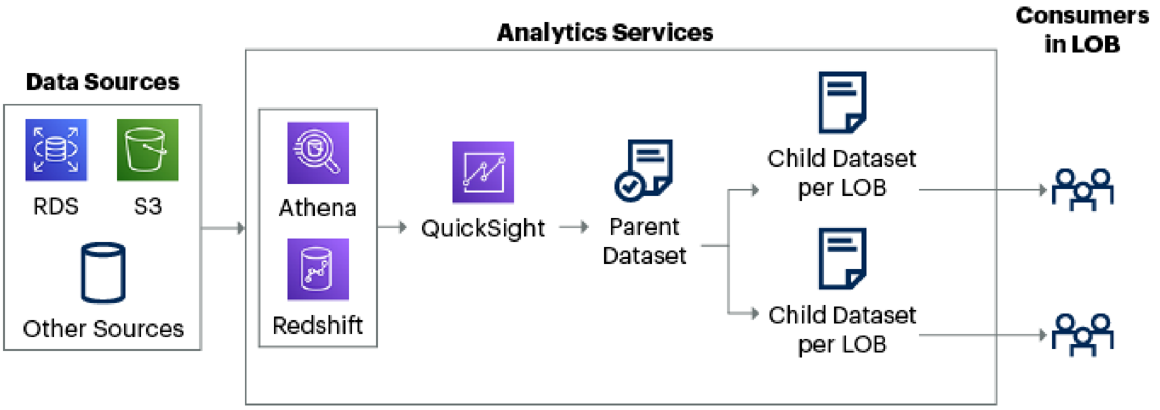
QuickSight offers two licensing types — Standard and Enterprise — with extra costs for using Q. However, Q can be tried for free for a limited time from within a standard dashboard. ML-powered insights and embedding is only available in Enterprise Edition. It is possible to upgrade from Standard to Enterprise, but a downgrade is not possible.

QuickSight also divides the user base into admin, author and reader. An author can create and share dashboards, while a reader consumes them. An admin organizes users and groups and is responsible for purchasing SPICE capacity.

One of the major use cases for any BI tool has been in centralizing metrics. QuickSight enables this through its datasets. A dataset, as the name suggests, is a collection of data extracted from a source. After an initial dataset has been created, it can be used as a parent dataset to create further datasets. The parent dataset can be the master metrics repository, with each child dataset catered to a line of business. Figure 6 shows a visual flow of this design.

Figure 6: QuickSight Datasets Can Be Used as Data Marts

QuickSight Datasets Can Be Used as Data Marts



Source: Gartner
778080_C

For more details on QuickSight, review the following resources from AWS:

- Supported Data Sources for QuickSight
- Gaining Insights with Machine Learning in Amazon QuickSight
- QuickSight Pricing
- Importing Data into SPICE
- Sharing Amazon QuickSight Dashboards

Amazon Athena

Amazon Athena is a managed, serverless, pay-per-query service used to analyze data using SQL on Amazon S3 with support for joins, window functions and arrays. Because it is serverless, there is no need to provision servers, and it scales up and performs parallelism to speed up query performance. It uses Presto for data definition language (DDL) and Hive for data manipulation language (DML) statements and supports different file formats including CSV, JSON, Avro and Parquet. Athena integrates with AWS Glue Data Catalog and can launch crawlers to populate the metadata repository as well as populate tables and partitions.

Some of the features of Athena include:

- **Federated querying:** Athena can now access, through Lambda, non-S3 sources such as DynamoDB, Redshift, RDS and on-premises databases. This can allow data in S3 to be jointly analyzed with other sources without requiring complex data integration pipelines.
- **SageMaker integration:** Athena SQL can be used to invoke SageMaker ML models without requiring Python knowledge. However, ML models cannot be created or trained through Athena. The model runs inference based on values the SQL passes, and it returns the inference results.
- **Lakehouse SQL:** Athena supports open table formats like Delta Lake, Hudi (on EMR) and Iceberg, thereby providing a SQL engine with ACID compliance over a data lake. It also integrates with Lake Formation, thereby allowing time-travel, row- and column-level security, and predicate pushdown.
- **SQL-based ETL:** Athena SQL offers Create Table AS (CTAS) and INSERT INTO statements to transform and convert data to different file formats, such as Parquet. This eliminates the need for data engineers and complex Spark knowledge.

It should be noted that Athena limits the runtime of DML queries to 30 minutes and DDL queries to 600 minutes, and queries will be terminated if they exceed the time limit. Also, in Lake Formation, Athena cannot be used to perform DML operations.

Best practices for Athena include compression, reading smaller amounts of data (use filters), limiting column selection (avoid SELECT *) and avoiding memory-intensive operations such as window functions (use PARTITION BY with windows functions). For more details on Athena, check out the following links:

- Performance Tuning in Athena
- Using CTAS and INSERT INTO for ETL and Data Analytics
- Using Governed Tables in Lake Formation
- Using ML With Athena

Amazon Redshift

Amazon Redshift is the primary AWS offering for MPP-style data warehousing needs for petabyte-scale business intelligence and machine learning. Redshift was, initially, designed to give data and analytics professionals all the configuration options (node size and types, clustering and resizing, concurrency scaling), which meant it was targeted more toward data engineers and data architects. More details on these features can be found in the partner document Building a Data Management Architecture on AWS. Recently, Redshift setup has been simplified with the introduction of serverless and federated features, thereby enabling BI analysts and data scientists to set up and use it as well. A comparison of these options is found in Table 2.

Table 2: Comparison Between Redshift and Redshift Serverless

	Redshift	Redshift Serverless
Underlying Compute	Nodes and clusters with options to select node types (e.g., RA3)	Workspaces and Redshift Processing Units (RPUs) — no nodes or cluster config required
Features	Cluster resizing, pause and resume, node type selection, manual concurrency scaling, port selection, manual maintenance, billed for non-idle clusters as well	No cluster resizing (only RPU capacity can be adjusted), pay-per-run, no maintenance, options to share data with provisioned cluster, static port 5439
Use Case	Fine-grained setup, management and administration of data warehouse	Low maintenance, self-serve setup with quick access to Redshift
Persona and Skills Requirement	Data engineers and data architects with strong SQL and data warehouse architecture and compute management skills	Citizen data engineers, BI analysts and data scientists who have good SQL skills but lack data warehousing administration and setup knowledge

Source: AWS blogs and Gartner research notes

AWS continues to add new features to Redshift, and some of them are as follows:

- **Serverless:** As mentioned above, Redshift provisioning is now quite straightforward and does not require setting up, configuring and managing clusters while retaining the standard Redshift features (apart from Spectrum and AQUA). It features automatic scaling, pay-per-use (instead of paying for non-idle clusters) and makes it easy for analytics professionals to set up different environments for development, testing and production use cases.

- **Data sharing:** Redshift now supports data sharing with other Redshift clusters and AWS accounts. This feature can be used to query data from other clusters without moving the data. Data can be shared with external consumers and across business groups for collaboration

- **Redshift Spectrum:** Data in the data lake, residing on S3 (in different formats), can be queried as external tables through Spectrum. It supports use cases for data scientists when they do not have direct access to the data lake but need access through SQL tools. All BI tools support this feature as well because Redshift is ODBC/JDBC compliant.

- **Query Editor V2:** V2 is a web-based tool for data exploration with visualization capabilities. It offers the capabilities to create graphs and charts as well as share data and collaborate with your team by using saved queries. V2 also allows querying data in S3 from the data lake. It is primarily geared toward BI analysts and data scientists who like to visualize their query results

- **Federation and cross-querying:** Federation allows Redshift to query non-S3 and Redshift clusters and includes sources such as RDS. Cross-querying allows Redshift to query across databases in a cluster. Both these features can be used to combine data from S3, RDS and other Redshift databases for analytics needs and reporting.

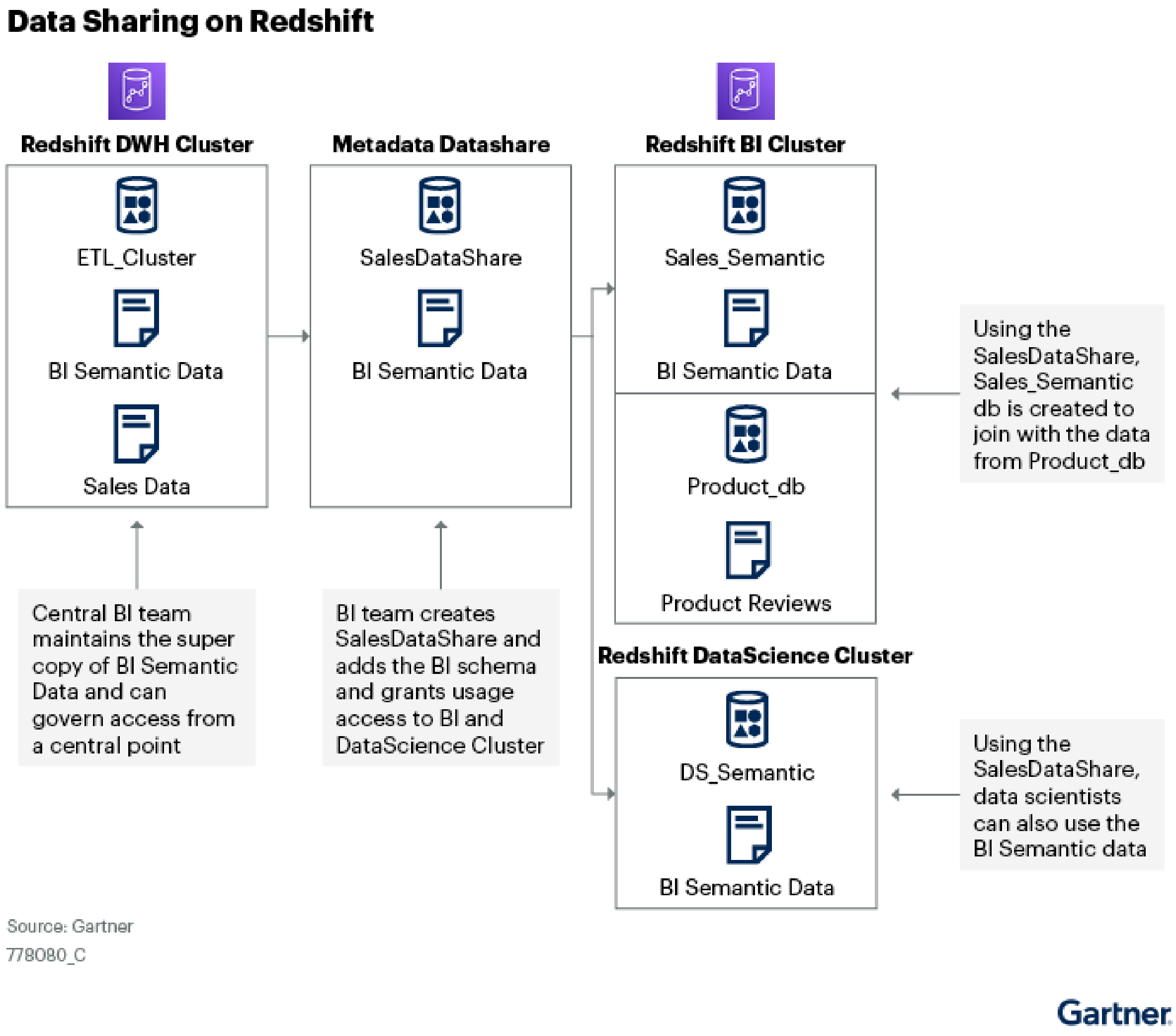
- **Redshift ML:** Redshift works with SageMaker Autopilot to train, develop and generate predictions for ML use cases. Redshift ML supports both supervised and unsupervised ML algorithms. However, its support for algorithms is limited to regression, binary classification and multiclass classification (for supervised learning) and K-means clustering (for unsupervised learning). This can serve as a useful starting point for conducting POCs and helping BI analysts develop quick ML models without using Python or exporting data to SageMaker.

- **AQUA:** For SQL queries that contain LIKE or SIMILAR TO predicates, Redshift automatically triggers Advanced Query Accelerator (AQUA). However, queries without predicates are not supported, and AQUA only runs on ra3.xlplus, ra3.4xlarge and ra3.16xlarge node types.

- **Streaming analytics (in preview):** Redshift has support for streaming data ingestion in preview. This will allow analysis from Kinesis Data Streams without the need to persist data in S3 first. BI analysts should use this feature to analyze streaming data in the event they are using Redshift as the primary analytics solution. [3]

One of the use cases for data warehouses is to build Semantic layers for workload isolation, downstream consumption by BI or other analytics tools and data security. This can be made possible through data sharing and cross-querying features in Redshift. Figure 7 shows an overview of such a setup. For more details on Semantic layers, please read Demystifying Semantic Layers for Self-Service Analytics.

Figure 7: Data Sharing on Redshift


Data Sharing on Redshift
Redshift DWH Cluster
ETL_Cluster
BI Semantic Data
Sales Data
Metadata Datashare
SalesDataShare
BI Semantic Data
Redshift BI Cluster
Sales_Semantic
BI Semantic Data
Product_db
Product Reviews
Redshift DataScience Cluster
DS_Semantic
BI Semantic Data
Central BI team maintains the super copy of BI Semantic Data and can govern access from a central point
BI team creates SalesDataShare and adds the BI schema and grants usage access to BI and DataScience Cluster
Using the SalesDataShare, Sales_Semantic db is created to join with the data from Product_db
Using the SalesDataShare, data scientists can also use the BI Semantic data
Gartner

Source: Gartner
778080_C

Adopted from AWS

For details on how Redshift compares with other cloud data warehouses, consult Solution Comparison for Cloud Data Warehouse Platforms.

Redshift offers connectivity to a host of BI tools for data visualization. For Power BI optimizations with Redshift, consult Inter- and Multicloud Analytics: Optimize Amazon Redshift, Google BigQuery and Snowflake for Power BI.

Amazon Kinesis Data Analytics

Kinesis Data Analytics (KDA) enables SQL-analytics on streaming data. It offers both Apache Flink and SQL runtimes and is a pay-per-use service that scales automatically. It is most often used in conjunction with Kinesis Data Streams and Kinesis Data Firehose. It should be noted that AWS now promotes the Kinesis Data Analytics Studio (Apache Flink runtime) over Kinesis Data Analytics SQL because it provides support for Python and Scala in addition to SQL. For more details on KDA, see Building a Data Management Architecture on AWS.

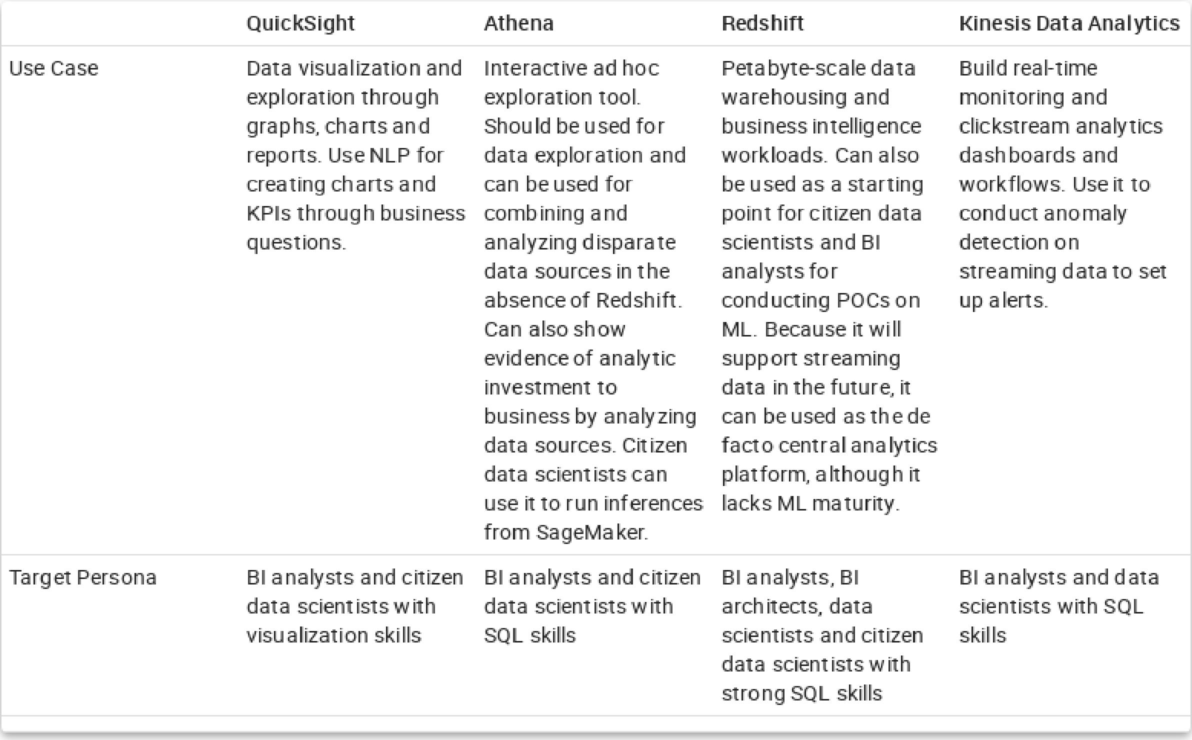
Use KDA for the following analytics use cases:

- **Streaming metrics generation:** KDA can be used to aggregate streaming and clickstream data to understand how data is trending over time and create time-series analysis. These can then be shown in QuickSight after they have been staged in S3 through Amazon Kinesis Data Firehose.
- **Data visualizations on streaming data**: Use Kinesis Data Analytics Studio to perform visualizations on incoming data and see real-time trends.
- **ML on streaming data:** Amazon KDS supports some ML use cases on anomaly detection (and explanation) and hot spot analysis through RANDOM_CUT_FOREST, RANDOM_CUT_FOREST_WITH_EXPLANATION and HOT SPOTS functions.
- **Log analytics with Amazon OpenSearch:** KDA can be combined with Data Streams and Firehose to deliver near-real-time data to OpenSearch for data visualization and search analytics.

For more details on the comparison between streaming services on the major cloud platforms, read Streaming Analytics in the Cloud: A Comparative Analysis of Amazon, Microsoft and Google (also see Table 3).

Table 3: Comparison of Use Cases and Personas/Skills Required for Each Analytics Services

(Enlarged table in Appendix)



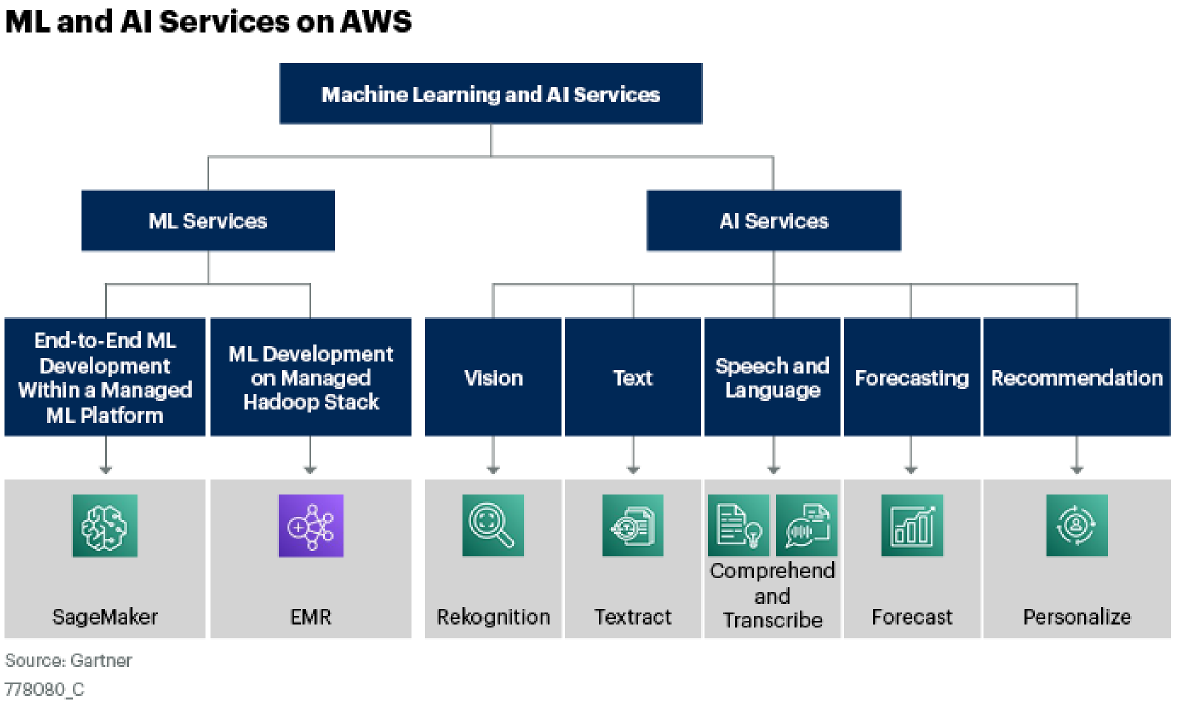
	QuickSight	Athena	Redshift	Kinesis Data Analytics
Use Case	Data visualization and exploration through graphs, charts and reports. Use NLP for creating charts and KPIs through business questions.	Interactive ad hoc exploration tool. Should be used for data exploration and can be used for combining and analyzing disparate data sources in the absence of Redshift. Can also show evidence of analytic investment to business by analyzing data sources. Citizen data scientists can use it to run inferences from SageMaker.	Petabyte-scale data warehousing and business intelligence workloads. Can also be used as a starting point for citizen data scientists and BI analysts for conducting POCs on ML. Because it will support streaming data in the future, it can be used as the de facto central analytics platform, although it lacks ML maturity.	Build real-time monitoring and clickstream analytics dashboards and workflows. Use it to conduct anomaly detection on streaming data to set up alerts.
Target Persona	BI analysts and citizen data scientists with visualization skills	BI analysts and citizen data scientists with SQL skills	BI analysts, BI architects, data scientists and citizen data scientists with strong SQL skills	BI analysts and data scientists with SQL skills

Source: AWS notes and blog series

Machine Learning and AI Services

AWS offers a wide variety of ML and AI development services, both as prepackaged solutions and as development solutions for creating customized AI solutions. Figure 8 shows the division of these services.

Figure 8: ML and AI Services on AWS


ML and AI Services on AWS
Machine Learning and AI Services
ML Services
AI Services
End-to-End ML Development Within a Managed ML Platform
ML Development on Managed Hadoop Stack
Vision
Text
Speech and Language
Forecasting
Recommendation
SageMaker
EMR
Rekognition
Textract
Comprehend and Transcribe
Forecast
Personalize

Source: Gartner
778080_C

Gartner

Machine Learning Services

Amazon EMR

Amazon EMR is a managed Hadoop service from AWS and it offers the entire product range of the Hadoop ecosystem. It includes services like Hive, Spark, Presto and Sqoop. Data and analytics technical professionals employ EMR primarily for data science workloads, and it was the de facto DSML platform in AWS prior to SageMaker.

Some of the useful features and use cases for EMR are as follows:

- **Serverless:** EMR, traditionally, has been considered as a difficult service to manage due to node and cluster administration. Becoming serverless makes it easy for data scientists to quickly set up EMR and begin ML development.
- **Iceberg and Hudi lakehouse:** Hudi and Iceberg enable lakehouse capabilities for Spark within EMR. Data scientists can now perform rollbacks and analyze streaming data without worrying about CDC issues. Note that Iceberg is only available from EMR 6.7 onwards, whereas Hudi comes preinstalled on 5.28 onwards.

- **ML development options:** EMR offers a wide variety of choices when it comes to developing ML solutions and includes EMR Studio, EMR Notebooks, JupyterHub and Zeppelin notebooks. Out of these, EMR Studio is the most versatile option because it offers management features along with support for Python, PySpark, R, Spark R, Scala and Spark SQL. It also features workspace collaboration where code can be written and run simultaneously with other team members. EMR Studio comes preinstalled when working with EMR Serverless.

- **Configuration options:** If not going serverless, then EMR offers a lot of options for compute and workflow optimizations. This includes node size and instance types, clustering management and also the ability to use EMRFS. EMRFS decouples storage from compute and persists data on S3. However, this requires highly trained data engineers, and data scientists are not expected to perform these tasks.

- **Sagemaker integration:** Previously, data scientists had to develop ML models on Spark within EMR and then export the data to SageMaker for inference. However, EMR and SageMaker are now closely integrated. EMR clusters can be directly connected, managed and visualized from within SageMaker, thereby providing Spark compute to SageMaker.

- **Central analytics:** EMR offers data engineering and ML capabilities within a single tool. However, it lacks in business intelligence and data warehousing capabilities. Moreover, it requires highly trained data engineers to continuously manage and maintain.

EMR can be used for the following reasons:

- For fine-grained management and cost optimizations on the underlying ML infrastructure. EMR can be customized at the node level and provides options for reserving instances to reduce costs. Even though a serverless option is offered, fine-grained customizations can provide more optimization for ML development.

- Where data science workflows can be customized and are not tool or platform dependent.

- When citizen data science is not needed, but instead, technical professionals have the required knowledge on Hadoop and Spark and can code using Python or Scala, for instance.

- Where workloads were, initially, based on the on-premises Hadoop stack and there is strong interest to continue development on a managed cloud offering.

Gartner recommends carefully evaluating SageMaker's integration with EMR to take advantage of both platforms for optimum workload balancing and scaling. [4]

Amazon SageMaker

SageMaker is the de facto managed service for developing ML solutions. It integrates with S3, Redshift and RDS and with external data sources such as Snowflake and Databricks. It can consume batch and streaming data and offers support for structured, semistructured and unstructured data types. It should be noted that, in most cases, data from non-S3 sources will be staged in S3 prior to ingestion in SageMaker.

SageMaker has several options catered to different personas and development styles. Figure 9 shows the range of SageMaker options now available.

Figure 9: SageMaker Options and Use Cases

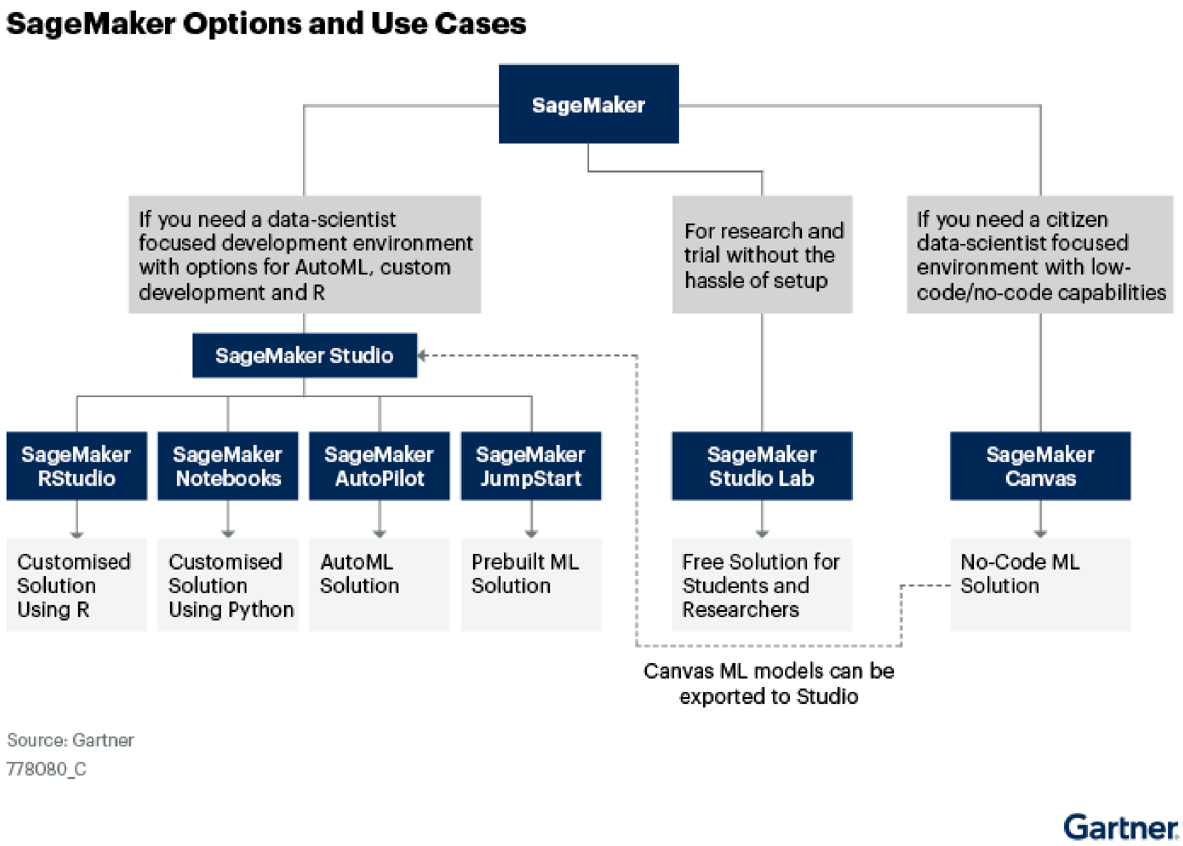

SageMaker Options and Use Cases
SageMaker
If you need a data-scientist focused development environment with options for AutoML, custom development and R
For research and trial without the hassle of setup
If you need a citizen data-scientist focused environment with low-code/no-code capabilities
SageMaker Studio
SageMaker RStudio
SageMaker Notebooks
SageMaker AutoPilot
SageMaker JumpStart
SageMaker Studio Lab
SageMaker Canvas
Customised Solution Using R
Customised Solution Using Python
AutoML Solution
Prebuilt ML Solution
Free Solution for Students and Researchers
No-Code ML Solution
Canvas ML models can be exported to Studio
Source: Gartner
778080_C
Gartner

Figure 9 should be taken as an initial overview, and it should be noted that SageMaker Studio caters to a broad base within the data and analytics technical professional space. Apart from SageMaker Studio and SageMaker Canvas, AWS also offers:

- **SageMaker Studio Lab** which offers a free, separate SageMaker environment without requiring an AWS account and is primarily catered for students and researchers. Users can simply sign up using their email address. It is based on JupyterLab and provides T3.XL as the CPU (12 hour limit) and G4D.XL as the GPU (4 hour limit) with 15 GB of space. It also integrates with GitHub and has terminal access. ML models can also be exported to SageMaker Studio.

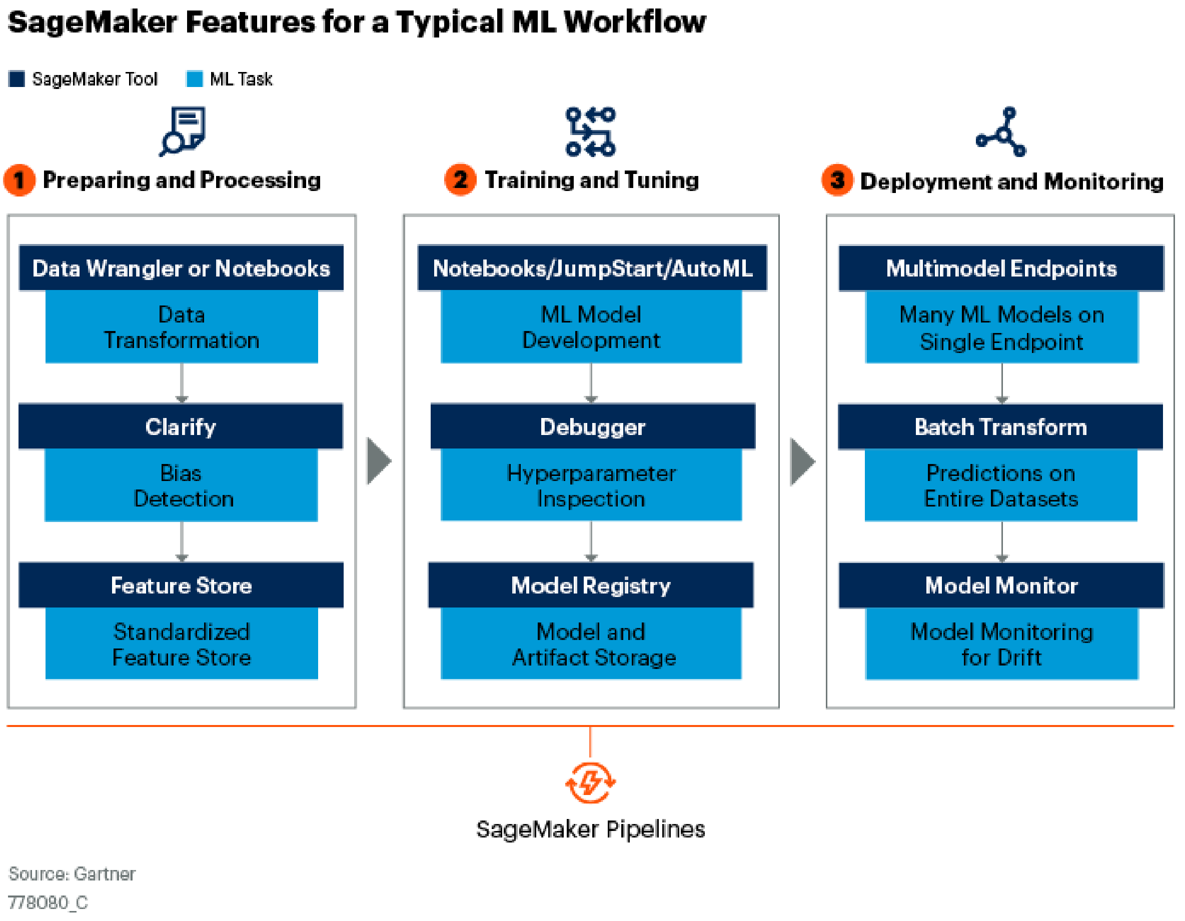
- **SageMaker JumpStart** is an option from within the SageMaker Studio that provides pretrained, open-source ML models and solution templates for common use cases. It can be used to quickly get started with executable example notebooks.

- **SageMaker Autopilot** is an AutoML offering and is available as an option from within SageMaker Studio. It automates the ML development pipeline, including data preparation, building, training and tuning the ML models from tabular datasets.

SageMaker Studio

SageMaker Studio offers a wide variety of services catered to the different phases of the ML development life cycle. Figure 10 shows some of these services.

Figure 10: SageMaker Features for a Typical ML Workflow



SageMaker Features for a Typical ML Workflow
■ SageMaker Tool
■ ML Task
1 Preparing and Processing
2 Training and Tuning
3 Deployment and Monitoring
Data Wrangler or Notebooks
Data Transformation
Clarify
Bias Detection
Feature Store
Standardized Feature Store
Notebooks/JumpStart/AutoML
ML Model Development
Debugger
Hyperparameter Inspection
Model Registry
Model and Artifact Storage
Multimodel Endpoints
Many ML Models on Single Endpoint
Batch Transform
Predictions on Entire Datasets
Model Monitor
Model Monitoring for Drift
SageMaker Pipelines
Gartner

Source: Gartner
778080_C

Source Adopted from AWS

These features and some of their use cases can be understood if the workflow is divided into three sections, as shown in Figure 10:

- **Preparing and processing:** SageMaker Data Wrangler offers more than 300 transformations in an entirely low-code/no-code visual interface and is suited for both citizen data scientists and data scientists who want to quickly clean data. For more customized transformations, preparation can be done in Jupyter Notebook with options to use Spark as the compute choice. Clarify is a model testing and explainability tool, based on SHAP, that can help in bias detection. Use Feature Store to store features once data has been transformed and tested for bias. For more details on feature stores, read Feature Stores for Machine Learning (Part 1): The Promise of Feature Stores.

- **Training and tuning:** Studio notebooks offer options for custom ML development with open-source frameworks and are suited for technically inclined data scientists. AutoML and JumpStart provide quickstart options to fastrack ML development. Debugger is used to test hyperparameters and automatically detects errors in the training data. Use Model Registry for storing models and their metadata. Once approved by Model Owners, a model can move into production
- **Deployment and monitoring:** Models can be deployed in various ways and multimodel endpoints should be used to deploy a large number of ML models on a single endpoint. Batch Transform should be used for predictions on entire datasets. AWS also offers Amazon API Gateway, which is a managed service for creating, publishing and monitoring APIs. For machine learning use cases with Amazon SageMaker, API Gateway can be used to expose a SageMaker inference endpoint as a REST API. This makes it possible for the REST API to be integrated directly with a SageMaker endpoint and would avoid the use of any intermediary compute resources like AWS Lambda. Once deployed, models have to be monitored, and Model Monitor automatically detects drifts and gives alerts. Amazon SageMaker Pipelines should be used to automate and orchestrate the pipeline.

It should be noted that AWS also offers CloudFormation as an Infrastructure as Code (IaC) offering that can be used to deploy SageMaker endpoints using code. This is a good practice within the MLOps domain because it creates reusable scripts for quickly deploying ML solutions.

For more information on SageMaker, please consult the following resources:

- Prepare ML Data with Amazon SageMaker Data Wrangler
- Create, Store, and Share Features With Amazon SageMaker Feature Store
- Deploy Models for Inference

SageMaker Canvas

SageMaker Canvas is an entirely visual/no-code AutoML offering that has been designed for citizen data science. It provides an interface with drag-and-drop functionality with options to share the model with data scientists for integration with SageMaker notebooks.

SageMaker Canvas's features are as follows:

- **Data types:** Currently supports categorical, numerical, text and datetime data types. This restricts ML development to the tabular and time-series domains, while image, text, video and image data and computer vision cases are not supported.
- **Data files and sources:** Only supports CSV files and does not have support for semistructured or open table formats like Parquet, JSON or Delta. It can connect to three external sources: Amazon Redshift, Snowflake and Amazon S3.
- **Data preparation:** Multiple datasets can be combined and joined without needing SQL knowledge. However, the maximum dataset file size limit is 50 GB. Apart from standard data preparation tasks like column renaming, Canvas also runs statistics on the datasets and also highlights key features within the dataset.
- **Model building:** Offers two build options: Standard and Quick Build. Use Standard for more extensive model training, while Quick Build should be used for timeframes of less than 15 minutes. Standard also gives more visibility into the ML training and provides information on different models, metric scores and training jobs. A “best model” can be selected from the range of models, and the model’s underlying hyperparameters can also be seen (such as the best algorithm, compute instance types and sizes)
- **Predictions:** Currently, Canvas only supports predictions using either batch inference or single data point.

Technical professionals should carefully assess the use case and skills they possess and then determine which option suits them best:

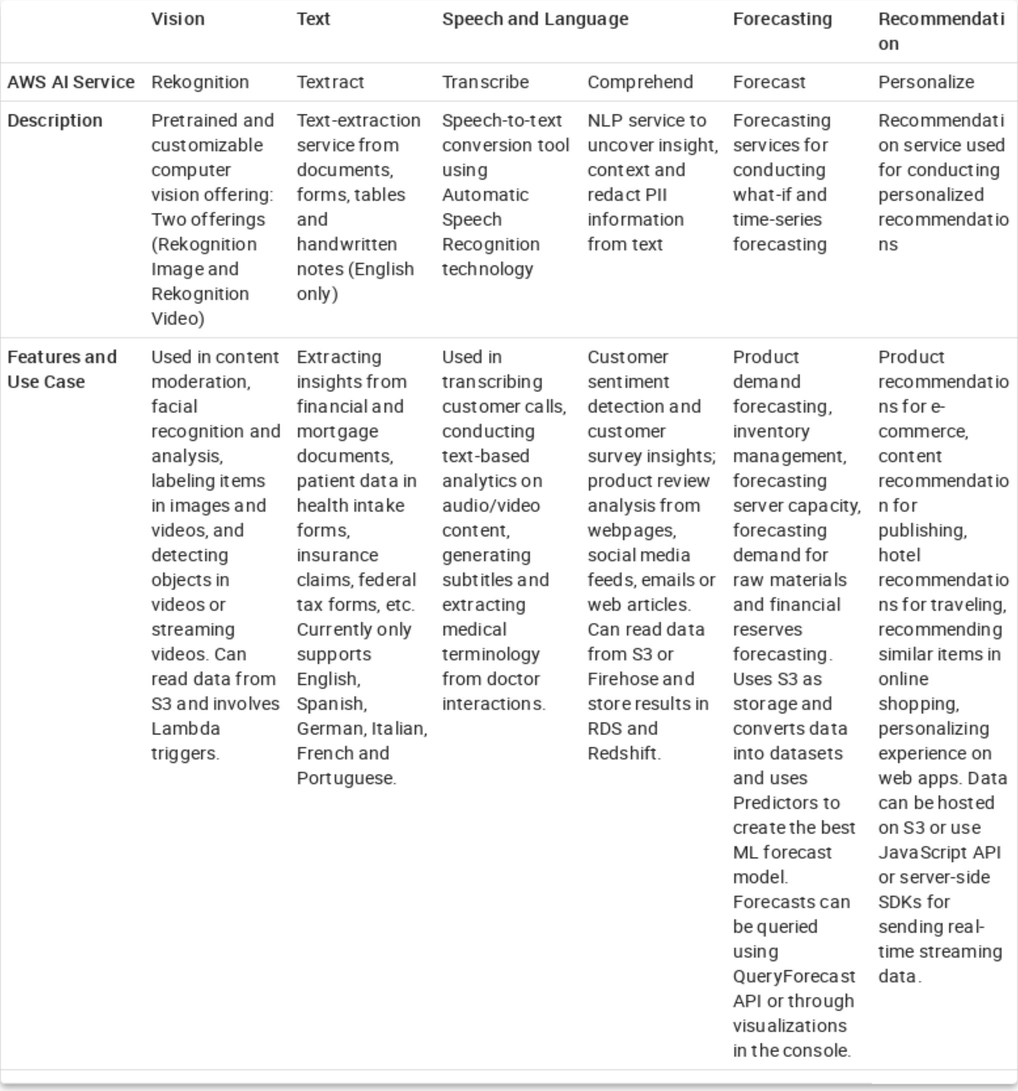
- It should be noted that the different options identified above may not, necessarily, be used in isolation. For instance, analytics professionals interested in exploring SageMaker as a tool but without access to AWS can simply use their email to sign up for SageMaker Studio Lab. ML models developed within Studio Lab can also later be reused within the greater AWS SageMaker environment.
- Similarly, Canvas can be used to create quick prototypes or even as an initial test-and-play setup by citizen data scientists and BI analysts. ML models created within Canvas can be exported as Python scripts to data scientists, who can then integrate them or build a more customized solution from a boilerplate code. Take note that not all ML models can be exported.

AI Services

AI services are predeveloped ML solutions targeting voice, language, text, audio and video domains, among others. Consumers of these applications do not need to know ML development or scripting. While Figure 8 shows the division of AI services, Table 4 shows the descriptions and use cases for these services.

Table 4: AI Services on AWS

(Enlarged table in Appendix)



	Vision	Text	Speech and Language		Forecasting	Recommendation
AWS AI Service	Rekognition	Textract	Transcribe	Comprehend	Forecast	Personalize
Description	Pretrained and customizable computer vision offering: Two offerings (Rekognition Image and Rekognition Video)	Text-extraction service from documents, forms, tables and handwritten notes (English only)	Speech-to-text conversion tool using Automatic Speech Recognition technology	NLP service to uncover insight, context and redact PII information from text	Forecasting services for conducting what-if and time-series forecasting	Recommendation service used for conducting personalized recommendations
Features and Use Case	Used in content moderation, facial recognition and analysis, labeling items in images and videos, and detecting objects in videos or streaming videos. Can read data from S3 and involves Lambda triggers.	Extracting insights from financial and mortgage documents, patient data in health intake forms, insurance claims, federal tax forms, etc. Currently only supports English, Spanish, German, Italian, French and Portuguese.	Used in transcribing customer calls, conducting text-based analytics on audio/video content, generating subtitles and extracting medical terminology from doctor interactions.	Customer sentiment detection and customer survey insights; product review analysis from webpages, social media feeds, emails or web articles. Can read data from S3 or Firehose and store results in RDS and Redshift.	Product demand forecasting, inventory management, forecasting server capacity, forecasting demand for raw materials and financial reserves forecasting. Uses S3 as storage and converts data into datasets and uses Predictors to create the best ML forecast model. Forecasts can be queried using QueryForecast API or through visualizations in the console.	Product recommendations for e-commerce, content recommendation for publishing, hotel recommendations for traveling, recommending similar items in online shopping, personalizing experience on web apps. Data can be hosted on S3 or use JavaScript API or server-side SDKs for sending real-time streaming data.

Source: Gartner

When should AI services be used instead of ML services like SageMaker and EMR?

A fine distinction should be drawn between SageMaker/EMR and AI services. Predictions, forecasting and recommendations can be developed in SageMaker/EMR. However, this would require data scientists who will build solutions from scratch. Using AI services like Amazon Forecast and Amazon Personalize removes the need to develop using data scientists and eases the usage of AI services for a wider audience like business users, BI analysts and citizen data scientists. AI services can also be used to extract data and store in the persistence layer. Examples can include Amazon Textract extracting text from unstructured documents like patient intake forms and using Amazon Comprehend for entity extraction, adding context and saving data in S3 for further analyses or loading into Redshift for structured data analytics. [5]

Monitoring and Orchestration

Monitoring and orchestration involves automating, scheduling and managing alerts on analytics and advanced analytics solutions and platforms. AWS offers monitoring solutions like Amazon CloudWatch and Amazon S3 Storage Lens, while orchestration services include AWS Step Functions.

Some of the tools such as SageMaker and QuickSight also offer orchestration and monitoring services (e.g., SageMaker Model Monitor and SageMaker Pipelines) for ML monitoring and orchestration. For details on XOps, including MLOps for AI, please read Demystifying XOps: DataOps, MLOps, ModelOps, AIOps and Platform Ops for AI.

Amazon CloudWatch

Amazon CloudWatch is used in monitoring application performance and for monitoring AWS resources that are running in AWS in real time. It is used to collect metrics, create customized dashboards and set alarms to notify users to take actions when specified metrics reach specified thresholds. CloudWatch integrates with and can access any tool within the AWS ecosystem.

CloudWatch can monitor SageMaker to create real-time metrics and store the statistics for 15 months and metric searches for two weeks. It can monitor the number of invocations on endpoints, time taken for model to respond and complete the inference, time taken to download model from S3, number of loaded models in containers in multimodel endpoints and other metrics.

QuickSight also sends metrics to CloudWatch so that admins can stay informed on the availability and performance of the QuickSight environment in real time. Metrics can be monitored for visuals, dashboards and dataset ingestions and can include metrics like dashboard view count and dashboard load time. It can also deduce load metrics as well, such as the number of failed ingestions, ingestion latency and the number of successful row ingestions.

Amazon S3 Storage Lens

As data required for analytics increases, so do the number of S3 buckets because they are used to house not only raw data but also processed, refined data as well as ML artifacts and feature stores. It can become difficult for analytics professionals and leaders to manage these buckets.

Use S3 Storage Lens to gain organization wide visibility into object storage, with point-in-time metrics, trend lines and recommendations to help discover anomalies, identify cost inefficiencies and better manage S3 buckets. Storage Lens includes an interactive dashboard, which can be found in the S3 console. Use it to perform filtering and drill-down into metrics on data stored in S3. Dashboards can be used to export the data to an S3 bucket for further analysis with QuickSight or Redshift.

Amazon QuickSight

QuickSight offers threshold alerts to stay informed about changes in data. Threshold alerts can be set in the form of Gauges and KPI visuals. Notifications can be sent in the form of emails if data limits are exceeded. Examples include a dashboard that contains a Gauge for monitoring the loss ratios in an insurance company. Usually insurance companies want to operate under certain percentages for loss ratios (e.g., 60%). In this case, an alert can be created when the loss ratio falls below the established threshold, and emails can be sent to senior management without the need for manual reporting.

Amazon SageMaker

SageMaker offers Model Monitor as a tool within its ecosystem for monitoring ML models without using code. It offers options for prediction outputs and captures metadata such as time stamp, model name and endpoints so that model predictions can be monitored based on metadata. It offers built-in analysis in the form of statistical rules for detecting model drifts and gauging model quality. Metrics emitted by Model Monitor can be visualized in SageMaker Studio, so model performance can be monitored visually. Monitoring jobs can also be scheduled to analyze predictions during certain time periods

Model Monitor integrates with Clarify to provide more visibility into potential model bias. Even if the original data was clean, it can change over time, which can result in model drifts and cause biases to develop. Integrate Clarify with CloudWatch to configure alerts if model bias begins to develop.

SageMaker Experiments stores the different ML experiments and tracks parameters, metrics, datasets and other artifacts related to the different ML training jobs. SageMaker Pipelines is the orchestration solution for ML development within SageMaker. Use it to create automated training workflows as part of a repeatable process in model orchestration. The entire model build workflow can be automated. This includes data preparation, feature engineering, model training, tuning and validation.

It should be noted that CloudWatch is primarily intended for monitoring at the infrastructure and application level. Analytics tools like SageMaker and QuickSight offer monitoring at the data level within their respective environments. Example use case differentiation can be with QuickSight sending an email to a senior business executive if temperature thresholds within a manufacturing facility exceed a certain limit. Meanwhile, CloudWatch can be used to send an alert to the business executive if the allotted usage threshold of QuickSight has been breached.

AWS Step Functions

Step Functions is an AWS orchestration offering for implementing and automating workflows with two options — Standard and Express. Express is suited for shorter workflows (fewer than five minutes), while Standard can be used for longer workflows like ETL pipelines.

Step Functions can be used with EventBridge (another service for running tasks on schedules) to execute an end-to-end ETL and ML workflow. It can consist of checking data in S3, performing ETL through Glue and storing the data in S3. It can then train and deploy an ML model on this data using Lambda functions that trigger SageMaker jobs and wait for completion. Finally, Lambda functions can be triggered to generate predictions that are saved to S3.

No single tool can provide all the functionalities of end-to-end monitoring and orchestration. Services offered by AWS are meant to be used in conjunction with each other to gain the maximum benefits. AWS also offers Infrastructure as Code (IaC) service called AWS CloudFormation which can spin up analytics environments in the form of code without requiring manual setup each time, thereby speeding up environment setup. Covering CloudFormation is beyond the scope of this document, but analytics professionals should explore options on automation using CloudFormation, in addition to the other tools mentioned.

Governance and Security

Governance and security lies at the heart of any analytics initiative and ensures the right people have the right access to the right data in a secure manner. For details on the best approaches for analytics governance, read Data and Analytics Governance Approaches for the Technical Professional. This paper briefly touches on these topics as they relate to the analytics space, but more details can be found in Building a Data Management Architecture on AWS.

AWS Lake Formation

Lake Formation is the default AWS offering for building secure and governed data lakes on S3. It integrates with Glue and enables collecting and cataloging data from databases and object storage into S3. Use it to enforce granular controls at the column, row and cell levels. Services like Athena and QuickSight can have BI analysts and citizen developer access restricted and controlled through Lake Formation.

Certain regulatory compliances and company policies may restrict access for data scientists. SageMaker can be integrated with Lake Formation to restrict access to ML systems, ensure data governance and data lineage and enforce regulatory compliance. Similarly, fine-grained controls can be installed to protect feature store data and grant access based on an individual’s role. [6]

Amazon Identity and Access Management (IAM)

Another method in governance is to segregate the user base with respect to different roles and assign permissions accordingly. This ensures that, even at the application level, not every persona has access to build or share everything.

There are options for securing work within SageMaker. AWS Identity and Access Management (IAM) can be used to create privileges and preventive controls for many aspects of the data science environments. These are called IAM policies and can be used to restrict access to SageMaker resources like SageMaker Studio notebooks. Create roles and define access rules accordingly, like a cloud administrator role with full admin access to the cloud for the cloud engineering and infrastructure team. This role will be responsible for creating data science team environments, onboarding users to SageMaker and creating further IAM roles such as data science admin or data scientist role to restrict access. The data scientist role will be limited in scope and can be restricted to launching SageMaker Studio, sharing Jupyter Notebook. This can prevent data scientists from misusing certain instant types such as heavy GPU instances.

QuickSight offers three different roles: admin, author and reader. Only an admin can create users, groups and purchase SPICE capacity, while an author can build dashboards. Readers can only consume the reports and dashboards. Row-level security can also be enabled on datasets to restrict further access to a reader. When embedding QuickSight into applications and granting access to users not part of the QuickSight subscription, tags can also be added along with row-level security to specify which data users can see depending on who they are.

It should be noted that AWS IAM can be used for services within the infrastructure, application development and data management domains as well, but an in depth explanation of these uses is beyond the scope of this paper.

Data Protection and Auditability

By default, S3 stores encrypted data at rest. Because most of the analytics offerings source most of their data from S3, it is important to protect data at the source. S3 has multiple ways of encrypting data. One of them is through the AWS Key Management Service (KMS). KMS can be used to automatically encrypt objects with a customer managed key (CMK) that is stored in KMS. This mechanism ensures that the data protection is under the control of the customer at all times. However, it should be noted that this can provide an extra overhead for the customer and requires skilled staff. This overhead involves maintaining the encryption keys, and it can be problematic to recover lost keys. Similar to S3, Amazon Elastic Container Registry (ECR) can be used to store customer-built Docker images with CMK enabled. Encryption keys can also be extended to the SageMaker environment by encrypting the volumes of all SageMaker EC2-based resources such as processing jobs, notebooks and endpoints. Functionalities can be built to allow the provisioning of SageMaker resources only if a KMS has been specified to encrypt the volumes.

It is essential to show traceability and auditability of the analytics solutions, especially in ML environments. One of the ways to enforce traceability is to use AWS CodeCommit, which is the equivalent of a repository used for commits and changes. Code should not be approved, nor should models be moved to production, unless and until code has been checked in. Senior or lead data scientists can monitor and ensure code is being committed properly. Apart from code versioning, data versioning is important as well, and that is where SageMaker Experiments comes into the picture. It should be used to automatically keep track of different versions of training data, experiments and user profile of the user launching jobs.

Activity logs should also be maintained for the analytics applications. CloudWatch Logs and CloudTrail can receive logs from any part of the analytics and AI architecture.

Strengths

AWS continues to develop its services at great speed. Some of the strengths of its platform and services include:

- **Diversity of analytics options:** AWS offers a diverse range of options for analytics and advanced analytics use cases — SageMaker and EMR for machine learning, QuickSight for business intelligence, and Athena and Redshift for SQL-based analytics
- **Ease of use:** AWS has introduced self-serve capabilities through no-code, serverless and federation features. These services include Glue DataBrew and SageMaker Data Wrangler for visual data preparation and Redshift Serverless for setting up Redshift. This eases the usage for these tools for diverse personas within the analytics space such as citizen data scientists and BI analysts.
- **Mature ML platform:** SageMaker continues to offer features for different personas and use cases. Canvas can be used by citizen data scientists for self-serve ML, and SageMaker Studio offers tools for different parts of the ML development life cycle.
- **ML-infusion into analytics offerings:** ML is no longer an EMR or SageMaker domain. Redshift, Athena and QuickSight have ML capabilities built into them, and analytics professionals have various options for using ML within the tool of their choice.

Weaknesses

Although AWS offers lots of options for analytics, some of its weaknesses are as follows:

- **Abundance of tools and capability overlap:** It can be daunting for analytics professionals to select the right tool for the right use case because almost all the analytics tools have been designed to do a bit of everything. Glue DataBrew and SageMaker Data Wrangler both offer no-code data preparation, but each caters to different needs and is powered by different processing engines. It would be helpful to have a single tool that does data preparation for all analytics workloads with the options to select the desired engine. Athena and Redshift both support federation and can access non-S3 sources, but they require performance and cost considerations that bring more complexity in tool and use-case selection.

- **SageMaker still needs improvement:** SageMaker has different development environments for R and Python and does not offer coding options from within a single environment. Spinning up separate environments only for a programming language can be cumbersome. Endpoints tend to have a long startup time, which can complicate testing and debugging. SageMaker can incur costs if the endpoints and notebooks are not terminated — options to shut down resources require installing extensions. SageMaker notebooks do not support collaboration or paired programming, discussion threads, and integration with collaborative tools like Slack.

- **Relatively weaker BI tool:** QuickSight does not perform well on Magic Quadrant for Analytics and Business Intelligence Platforms and shows as a Niche Player. QuickSight would do better with an enhanced data preparation component or a tighter integration with Glue DataBrew. This can empower BI analysts to curate and transform data from within QuickSight without needing to use other tools. There is an extra cost of using +Q features in addition to paying extra for the Enterprise Edition, and customers cannot switch back from Enterprise to Standard Edition, which increases risk of lock-in. Moreover, QuickSight does not support streaming use cases and only offers batch support.

- **Lack of multicloud support:** AWS tools are designed to work within the AWS ecosystem and do not cater to data sources from other cloud vendors. Glue DataBrew, QuickSight, Athena and Redshift federation features do not allow connections to Azure or GCP data sources. SageMaker will host data in S3, even while connecting to external data sources like Snowflake. This forces data to be persisted in AWS storage.

Guidance

AWS is now augmenting its analytics and advanced analytics offerings with no-code, serverless and federation capabilities. This makes it easier for analytics professionals with low skill and scripting expertise to be able to use them and ramp up quickly. However, the overabundance of options can often prove challenging and adds complexity in decision making when organizations try to determine the best use fit for the tools.

In order to create an optimal end-to-end analytics and AI architecture, the data management domain cannot be ignored and should be set up to support analytics and advanced analytics use cases. Determine the correct ingestion and storage patterns for use by analytics offerings, and create partnerships between data engineers and data scientists. Data scientists can now partner with data engineers and are able to use tools like Glue DataBrew to create the transformed data of their choice. BI analysts can work with data engineers by using Athena to show the early promise of a Redshift endeavor prior to completion. They can use the federation and serverless options to quickly ingest and analyze data from disparate sources and visualize the results in QuickSight to show early POC value to senior management.

On the ML front, use SageMaker and its offerings toward multiple personas within the advanced analytics space. SageMaker Canvas offers a no-code visual interface, which makes it easy to get started with ML development. Use it to create POCs and show the promise of ML to the business or perform initial development and then extract scripts to data scientists for custom development. Citizen data scientists can also use Data Wrangler to provide visual data cleaning and transformation, thereby preserving data scientists for core ML model development and training. Model owners can observe their models' performance using Model Monitor and can be alerted of any mode drifts that can impact the business. ML engineers can use multiple deployment options, such as multimodal endpoints and use pipelines to automate the entire ML development workflow.

Lastly, keep an eye on the costs. AWS offers a plethora of tools to monitor costs and budget. It is easy to get carried away due to the plethora of options and the enhanced features offered, but analytics professionals should work closely with data management professionals to ensure that FinOps is an important element within the AI and analytics architecture.

Acronym Key and Glossary Terms

AWS	Amazon Web Services
SFTP	Secure File Transfer Protocol
FTPS	File Transfer Protocol Secure
STP	Spanning Tree Protocol
CDC	Change Data Capture
ELT	Extract, Load and Transform
ETL	Extract, Transform and Load
EMR	Elastic Map Reduce
DDL	Data Definition Language
DML	Data Manipulation Language
EMRFS	EMR File System
JDBC	Java Database Connectivity
ODBC	Open Database Connectivity
NLP	Natural Language Processing
MPP	Massively Parallel Processing
SHAP	SHapley Additive exPlanations
NLQ	Natural Language Query

Evidence

P-21016a Chief Data Officer Pre-Survey, 2022. 2021 Gartner Chief Data Officer Role Pre-Survey: This survey supplements findings from the annual Gartner Chief Data Officer Survey. The objective was to explore the role of the CDO, including top challenges and the gaps in meeting those challenges. The survey was conducted online from 1 July through 15 August 2021 among 157 respondents from across the world. A majority of participants were chief data and/or analytics officers (71%). The survey sample was gleaned from Gartner's curated annual CDO Survey list composed of clients and nonclients. Regions were primarily represented by North America (43%) and EMEA (35%). The survey was developed collaboratively by a team of Gartner data and analytics analysts and Gartner's Research Data, Analytics and Tools team. Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

[1] Extract, Prepare, and Analyze Salesforce.com Data Using Amazon AppFlow, AWS Glue DataBrew, and Amazon Athena, AWS.

[2] Redshift Streaming Ingestion (Preview), AWS.

[3] Real-Time Analytics With Amazon Redshift Streaming Ingestion, AWS.

[4] Create and Manage Amazon EMR Clusters From SageMaker Studio to Run Interactive Spark and ML Workloads, AWS.

[5] Amazon Textract Architecture Blog, AWS.

[6] Control and audit data exploration activities with Amazon SageMaker Studio and AWS Lake Formation, AWS.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Solution Path for Building Modern Analytics and BI Architectures

A Guidance Framework for Deploying Data and Analytics in the Cloud

Solution Comparison for Cloud Data Science and Machine Learning Platforms

Streaming Analytics in the Cloud: A Comparative Analysis of Amazon, Microsoft and Google

Table 1: How Should Data and Analytics Technical Professionals Decide Between the Different Transformation Services on AWS?

	Amazon Athena	AWS Glue DataBrew	Amazon SageMaker Data Wrangler	Amazon EMR	AWS Lambda
Description	Managed, serverless service with federated SQL capabilities	Managed and serverless no-code data transformation service	Managed, no-code data transformation service from SageMaker	Managed Apache Hadoop service; also has serverless mode	Managed, serverless compute
Underlying Technology	Presto for DML, Apache Hive for DDL with Lambda for custom data connections	Apache Spark	EC2 compute	Cluster compute with options for Spark, Hive, etc.	EC2 compute
Use Case	Connect, ingest and transform data using SQL without provisioning servers or requiring Spark knowledge. BI analysts and data scientists do not need to wait for data pipelines to be created and can ingest and transform data using SQL.	Transform and process data without scripting knowledge. Use it to shape, curate and prepare data for subsequent staging in Redshift or in S3 without the need for SQL, Python or Scala.	Prepare and curate data for ML workloads using a no-code interface. Use it for ML workloads within the SageMaker environment where Python knowledge is lacking.	Perform code-intensive data engineering tasks using Python, PySpark, etc. using Notebooks. Use it only if data scientists have the scripting knowledge for transformations and are building the ML solution within EMR.	Offers options to use any scripting language for smaller transformation jobs (less than 15 minutes). Use it as a notification service or for processing streaming data from Kinesis.
Target Personas	SQL-savvy BI analysts and data scientists	BI analysts and citizen data scientists	Citizen data scientists and data scientists	Data scientists	Data scientists

Source: AWS blogs and Gartner research notes

Table 2: Comparison Between Redshift and Redshift Serverless

	Redshift	Redshift Serverless
Underlying Compute	Nodes and clusters with options to select node types (e.g., RA3)	Workspaces and Redshift Processing Units (RPUs) — no nodes or cluster config required
Features	Cluster resizing, pause and resume, node type selection, manual concurrency scaling, port selection, manual maintenance, billed for non-idle clusters as well	No cluster resizing (only RPU capacity can be adjusted), pay-per-run, no maintenance, options to share data with provisioned cluster, static port 5439
Use Case	Fine-grained setup, management and administration of data warehouse	Low maintenance, self-serve setup with quick access to Redshift
Persona and Skills Requirement	Data engineers and data architects with strong SQL and data warehouse architecture and compute management skills	Citizen data engineers, BI analysts and data scientists who have good SQL skills but lack data warehousing administration and setup knowledge

Source: AWS blogs and Gartner research notes

Table 3: Comparison of Use Cases and Personas/Skills Required for Each Analytics Services

	QuickSight	Athena	Redshift	Kinesis Data Analytics
Use Case	Data visualization and exploration through graphs, charts and reports. Use NLP for creating charts and KPIs through business questions.	Interactive ad hoc exploration tool. Should be used for data exploration and can be used for combining and analyzing disparate data sources in the absence of Redshift. Can also show evidence of analytic investment to business by analyzing data sources. Citizen data scientists can use it to run inferences from SageMaker.	Petabyte-scale data warehousing and business intelligence workloads. Can also be used as a starting point for citizen data scientists and BI analysts for conducting POCs on ML. Because it will support streaming data in the future, it can be used as the de facto central analytics platform, although it lacks ML maturity.	Build real-time monitoring and clickstream analytics dashboards and workflows. Use it to conduct anomaly detection on streaming data to set up alerts.
Target Persona	BI analysts and citizen data scientists with visualization skills	BI analysts and citizen data scientists with SQL skills	BI analysts, BI architects, data scientists and citizen data scientists with strong SQL skills	BI analysts and data scientists with SQL skills

Source: AWS notes and blog series

Table 4: AI Services on AWS

	Vision	Text	Speech and Language		Forecasting	Recommendation
AWS AI Service	Rekognition	Textract	Transcribe	Comprehend	Forecast	Personalize
Description	Pretrained and customizable computer vision offering: Two offerings (Rekognition Image and Rekognition Video)	Text-extraction service from documents, forms, tables and handwritten notes (English only)	Speech-to-text conversion tool using Automatic Speech Recognition technology	NLP service to uncover insight, context and redact PII information from text	Forecasting services for conducting what-if and time-series forecasting	Recommendation service used for conducting personalized recommendations
Features and Use Case	Used in content moderation, facial recognition and analysis, labeling items in images and videos, and detecting objects in videos or streaming videos. Can read data from S3 and involves Lambda triggers.	Extracting insights from financial and mortgage documents, patient data in health intake forms, insurance claims, federal tax forms, etc. Currently only supports English, Spanish, German, Italian, French and Portuguese.	Used in transcribing customer calls, conducting text-based analytics on audio/video content, generating subtitles and extracting medical terminology from doctor interactions.	Customer sentiment detection and customer survey insights; product review analysis from webpages, social media feeds, emails or web articles. Can read data from S3 or Firehose and store results in RDS and Redshift.	Product demand forecasting, inventory management, forecasting server capacity, forecasting demand for raw materials and financial reserves forecasting. Uses S3 as storage and converts data into datasets and uses	Product recommendations for e-commerce, content recommendation for publishing, hotel recommendations for traveling, recommending similar items in online shopping, personalizing experience on web

Predictors to create the best ML forecast model. Forecasts can be queried using QueryForecast API or through visualizations in the console.	apps. Data can be hosted on S3 or use JavaScript API or server-side SDKs for sending real-time streaming data.

Source: Gartner