Data Lifecycle and Analytics in the AWS Cloud

A Reference Guide for Enabling Data-Driven Decision-Making

Contents

Public Sector Case Studies

Purpose of this guide

Data is an organization's most valuable asset and the volume and variety of data that organizations amass continues to grow. The demand for simpler data analytics, cheaper data storage, advanced predictive tools like artificial intelligence (AI), and data visualization is necessary for better data-driven decisions.

The *Data Lifecycle and Analytics in the AWS Cloud* guide helps organizations of all sizes better understand the data lifecycle so they can optimize or establish an advanced data analytics practice in their organization. The document guides readers through five stages of the data lifecycle, including **data ingestion, data staging, data cleansing, data analysis (including AI/machine learning (ML) inference and deep learning tools) and visualization, data archiving, and overall data security.**

This guide is written primarily for IT professionals, as well as for chief data officers, data scientists, data analysts, and data engineers. Technical professionals of diverse training and experience can learn how to extract more value from their data by taking advantage of the Amazon Web Services (AWS) Cloud to support data-driven decision-making. Business leaders and managers will also benefit from the guide's overview sections and customer case studies.

One could spend countless hours searching online to understand the many services available to turn data into insights. This reference guide aggregates important definitions, workflows, and the relevant AWS services for each stage of the workflow, to simplify the learning process. Here are some data lifecycle and management questions this guide will help you answer:

- How do you collect and analyze high-velocity data across a variety of data types – structured, unstructured, and semistructured?
- How do you scale up IT resources to run thousands of concurrent queries against your data – and then scale back down automatically to lower costs?
- How do you analyze your data across platforms, so users can view, search, and run queries on multiple data repositories?
- How do you cost-effectively store petabytes of data and share them on-demand with users around the world?
- How do you get your data to answer questions about past scenarios and patterns, while predicting future events?

1 Introduction

Data within organizations is measured in petabytes, and it grows exponentially each year.

IT teams are under pressure to quickly orchestrate data storage, analytics, and visualization projects that get the most from their organizations' number one asset: their data. They're also tasked with ensuring customer privacy and meeting security and compliance mandates. These challenges are cost-effectively addressed with cloud-based IT resources, as an alternative to fixed, conventional IT infrastructure (e.g. owned data centers and computing hardware managed by internal IT departments). By modernizing their approach to data lifecycle management, and leveraging the latest cloud-native analytics tools, organizations reduce costs and gain operational efficiencies, while enabling data-driven decision-making.

What is Big Data?

A dataset too large or complex for traditional data processing mechanisms is called **"big data".** Big data also encompasses a set of data management challenges resulting from an increase in volume, velocity, and variety of data. These challenges cannot be solved with conventional data storage, database, networking, compute, or analytics solutions. Big data includes structured, semi-structured, and unstructured data. **"Small data,"** on the other hand, refers to structured data that is manageable within existing databases. Whether your data is big or small, the lifecycle stages are universal. It's the data management and IT tools that will differ in terms of scale and costs.

To support advanced analytics for big and small data projects, cloud services support a variety of use cases: **descriptive analytics** that address what happened and why (e.g. traditional queries, scorecards, and dashboards); **predictive analytics** that measure the probability of a given event in the future (e.g. early alert systems, fraud detection, preventive maintenance applications, and forecasting); and **prescriptive analytics** that answer, "What should I do if 'x' happens?" (e.g. recommendation engines). With AWS, it's also technically and economically feasible to collect, store, and share larger datasets and analyze them to reveal actionable insights.

AWS offers a complete cloud platform designed for big data across data lakes or big data stores, data warehousing, distributed analytics, real-time streaming, machine learning, and business intelligence services. These cloud-based IT infrastructure building blocks – along with AWS Cloud capabilities that meet the strictest security requirements – can help address a wide range of analytics challenges.

What is the Data Lifecycle?

As data is generated, it moves from its raw form to a processed version, to outputs that end users need to make better decisions. All data goes through this data lifecycle. Organizations can use AWS Cloud services in each stage of the data lifecycle to quickly and cost-effectively prepare, process, and present data to derive more value from it. The five data lifecycle stages include: **data ingestion, data staging, data cleansing, data analytics and visualization, and data archiving.**

1. **The first stage is data ingestion.** Data ingestion is the movement of data from an external source to another location for analysis. Data can move from local or physical disks where value is locked (e.g. in an IT data center), to the cloud's virtual disks. There, it can be closer to end users and where machine learning and analytics tools can be applied. During data ingestion, high value data sources are identified, validated, and imported while data files are stored and backed up in the AWS Cloud. Data in the cloud is durable, resilient, secure, cost-effectively stored, and most importantly, accessible to a broad set of users. Common data sources include transaction files, large systems (e.g. CRM, ERP), user-generated data (e.g. clickstream data, log files), sensor data (e.g. from Internet-of-Things or mobile devices), and databases. **AWS services available in this stage include** Amazon Kinesis, AWS Direct Connect, AWS Snowball/ Snowball Edge/Snowmobile, AWS DataSync, AWS Database Migration Service, and AWS Storage Gateway.

2. **The second stage is data staging.** Data staging involves performing housekeeping tasks prior to making data available to users. Organizations house data in multiple systems or locations, including data warehouses, spreadsheets, databases, and text files. Cloud-based tools make it easy to stage data or create a data lake in one location (e.g. Amazon S3), while avoiding disparate storage mechanisms. **AWS services available in this stage include** Amazon S3, Amazon Aurora, Amazon RDS, Amazon DynamoDB.

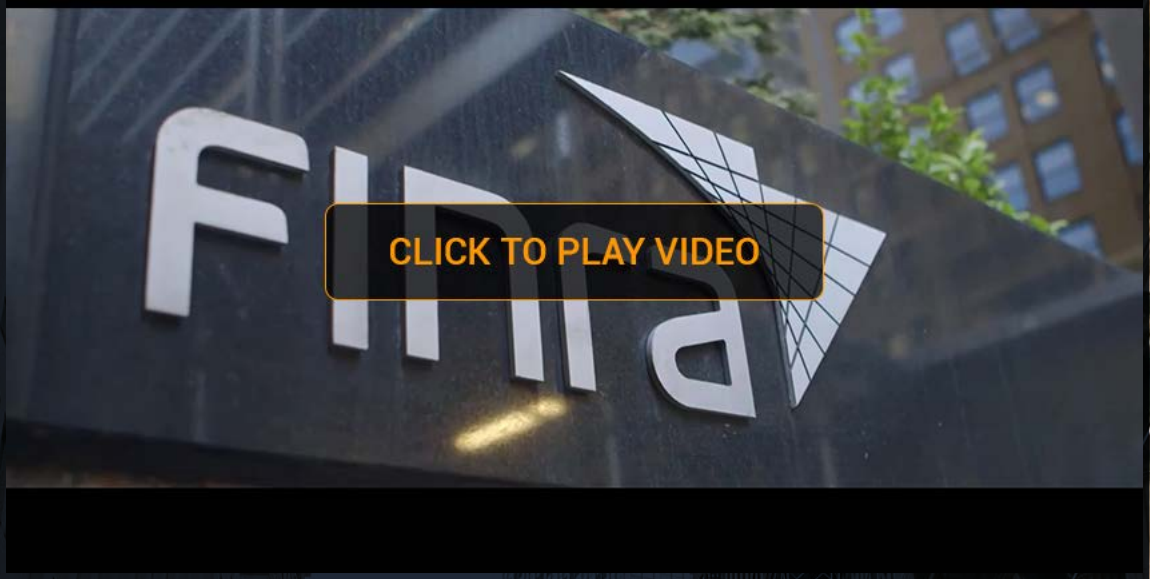
3. **The third stage is data cleansing.** Before data is analyzed, data cleansing detects, corrects, and removes inaccurate data or corrupted records or files. It also identifies opportunities to append or modify dirty data to improve the accuracy of analytical outputs. In some cases, data cleansing involves translating files, turning speech files to text, digitizing audio and image files for processing, or adding metadata tags for easier search and classification. Ultimately, data cleansing transforms data so it's optimized for code (e.g. Extract, Transform, Load (ETL)). **AWS services available in this stage include** AWS Glue (ETL), AWS Glue Data Catalog, Amazon EMR, and Amazon SageMaker Ground Truth.

4. **The fourth stage is data analytics and visualization.** The real value of data can be extracted in this stage. Decision-makers use analytics and visualization tools to predict customer needs, improve operations, transform broken processes, and innovate to compete. The ability for mission owners and executives to rely on data reduces error-prone and costly guesswork. **AWS services available in this stage include** Amazon Athena, Amazon Redshift, Amazon QuickSight, Amazon SageMaker, Amazon Comprehend, Amazon Comprehend Medical, and AWS DeepLens.

5. **The fifth stage is data archiving.** The AWS Cloud facilitates data archiving, enabling IT departments to invest more time in other stages of the data lifecycle. These storage solutions have achieved numerous compliance standards, security certifications, and provide built-in encryption, enabling compliance from day one. **AWS services in this stage include** Amazon S3 Glacier, Amazon S3 Glacier Deep Archive, and AWS Storage Gateway.

Case in Point:

FINRA



The Financial Industry Regulatory Authority (FINRA) is a Wall Street regulator that monitors daily stock trading to protect investors. It runs market surveillance on billions of equity trades – with their highest daily volume reaching 135 billion stock transactions or events in October 2018. FINRA has over 20 petabytes of data on which to run analytics, so that insider trading or SEC violations can be traced or tracked. To handle the volume of trades and depth of analysis, FINRA decided to establish its infrastructure in the AWS Cloud.

FINRA runs its mission critical financial applications on AWS. With a data lake in S3 and use of Amazon Redshift – alongside Hive in Amazon EMR and Presto – data analysts can run queries on tips, alerts, and sweeps using Amazon ECS. With 90% of its data in the AWS Cloud, FINRA has generated a 2X return on investment from its expenditures on the AWS Cloud.

FIND OUT MORE →

9

2 Common data management challenges to address in the data lifecycle

Many organizations are sitting on valuable data and performing little-to-no analysis on it. While some organizations recognize and capitalize on this value, others are hindered by concerns of draining resources by building complex analytics projects.

Organizations of all sizes face challenges as they seek to derive meaning and value from their data. At each stage of the data lifecycle, the "five Vs" of data management can help dictate which tools are required to address a particular problem. This includes the **volume, velocity, variety, veracity, and value** of your data. Challenges associated with the five Vs include a high volume of data, data in multiple systems and formats, increasingly diverse users with differing analytics needs, the requirement to support emerging predictive and real-time analytics, semi-structured or unstructured data, and a lack of in-house data science expertise.

Data, data, everywhere – a growing volume. Organizations are amassing and storing ever-increasing amounts of data, yet only a fraction of that data enters the analytics stage. Data is often housed in multiple systems or locations, including data warehouses, log files, and databases. As this **volume** of data grows, classifying it becomes critical. Is it qualitative or quantitative? What are the storage costs and can this storage mechanism scale? How much of this data is being used, and how much is being ignored? What are the common sources of this data? These questions help determine the cost-benefit analysis of employing various options for handling high volumes of data.

Velocity of data. In addition, data is being generated at a greater velocity. Data velocity impacts highly meaningful applications, including public safety and emergency alerts, cybersecurity or physical security breaches, customer service applications, IoT sensor data that triggers immediate responses, and early indicators that drive interventions. Is the data being generated in real-time, in batches at frequent intervals, or as a result of events? As you amass data at faster rates, it calls for a modern IT approach – one that simplifies analysis, reduces storage costs, and untethers data from conventional data centers for easier analysis. These challenges continue to outstrip conventional, in-house IT infrastructure resources.

Questionable data veracity or dirty data. Data **veracity** refers to the integrity, reliability, accuracy, and trustworthiness of data. Data that is unbiased, de-duplicated, consistent, and stable is ideal. It helps ensure that analytics outputs based on that data are also accurate. In the instance of a survey, is there a way to validate that users accurately entered their zip codes, and for the survey owner to remove or correct bad zip codes?

A broad variety of data. The **variety** of data in an organization spans structured, unstructured, or semi-structured data. These formats often dictate how data is stored, processed, and analyzed. For example, structured data is often housed in relational databases with defined rows and columns (e.g. social security numbers, birth dates, zip codes). Semi-structured data does not follow the formal arrangement of databases or data tables, but has tags to indicate a hierarchy of records and fields within the data. NoSQL databases and JSON documents are considered semi-structured documents and may be stored as flat files and object stores.

Lastly, unstructured data may include audio files, images, videos, and other file types (e.g. meeting notes). Unstructured data requires the additional step of adding metadata tags for easier search and retrieval. Without a conscious effort to include the unstructured data in the analysis pipeline, an organization risks missing out on relevant insights. This unstructured data, when left untapped, is labeled 'dark data.'

The variable value of data. Not all data is of equal **value**. Assigning and ranking the value of data is required for prioritization, before embarking on a project. It's important to consider what outcomes the data is driving, who uses it, whether it's essential for mission-critical work; and how often multiple users will need it for security, compliance, or analysis.

Diversity of data users, stakeholders, and stewards. A multitude of organizational stakeholders can benefit from data to do their jobs well. To accomplish this, each may rely on a slightly different combination of data visualization tools and analytics packages. This begets the need to break down silos within a data management practice so that teams can share data, collaborate, and drive greater insights. Placing a business intelligence (BI) tool on top of a dataset to run reports is a common use case, however, data scientists and developers may conduct analyses against a variety of data sources including the use of APIs.

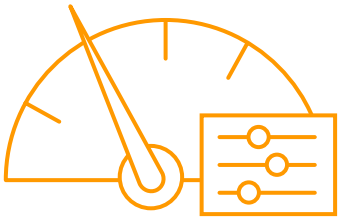
Evolving from retrospective analytics to real-time and predictive analytics. Data analysis isn't purely retrospective. Data can also be used to develop and deploy models for real-time predictions. More than before, organizations want to extract predictions from their data and use machine learning to detect patterns previously unseen in their historical data.

Lack of in-house data science expertise. Data scientists typically extract insights from data by applying engineering tools to a variety of data sources. They also prepare and process data, know "what" information is most valuable for organizational decision-making, and work with data engineers to answer the "why." Still, many organizations lack the in-house data science expertise to support necessary large-scale data analytics projects.

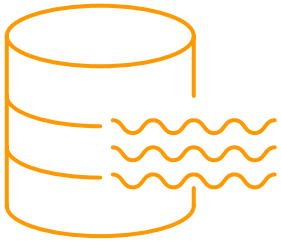
3 The data lifecycle in detail

Let's take a deeper look at the five stages of the data lifecycle involved in the preparation, processing, and presentation of data for decision-making*:

Data Sources
ERP
Transactions
Web logs
Clickstream
Data bases
Connected devices

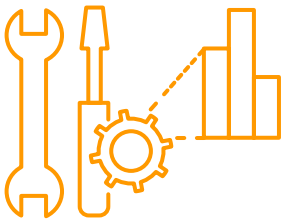


Data Ingestion
Amazon Kinesis
AWS Direct Connect
AWS Snowball
AWS Snowball Edge
AWS Data Sync
Amazon S3 Transfer Acceleration
AWS Database Migration Service



Data Staging
Amazon S3
AWS Lake Formation
Amazon Aurora
Amazon RDS
Amazon Dynamo DB
Amazon Elasticache
Amazon Neptune
Amazon QLDB
Amazon Timestream

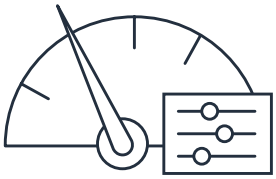
Data Cleansing
AWS Glue
AWS Lambda
AWS EMR



Data Analytics & Visalization
Amazon Athena
Amazon Redshift
Amazon Redshift Spectrum
Amazon Elasticsearch
Amazon QuickSight
AWS Lambda

Data Archiving
Amazon S3 Standard-IA
S3 One Zone-IA
Amazon S3 Glacier
Amazon S3 Glacier Deep Archive
Amazon S3 Intelligent Tiering
(Object Lifecycle Management)

*Note: Each data lifecycle stage represents a conceptual boundary and may differ across organizations. In addition, some AWS services can address more than one lifecycle stage. Finally, although the lifecycle appears linear, many data engineering and analytics processes are iterative.



Stage 1 – Data Ingestion

Data ingestion entails the movement of data from an external source, into another location for further analysis. Generally, the destination for data is some form of storage or a database (we discuss storage further in the Data Staging section of this guide). For example, ingestion can involve moving data from an on-premises data center or physical disks, to virtual disks in the cloud, accessed via an internet connection. Data ingestion also involves identifying the correct data sources, validating and importing data files from those sources, and sending the data to the desired destination. Data sources can include transactions, enterprise-scale systems such as Enterprise Resource Planning (ERP) systems, clickstream data, log files, device or sensor data, or disparate databases.

Key questions to consider: What is the volume and velocity of my data? For instance, ingesting website clickstream data, ERP data, or sensor data in an IoT scenario would warrant AWS Kinesis Data Streams However, ingesting a database would be best accomplished through the AWS Database Migration Service (DMS). What is the source and format of the data? For satellite data involving micro batch processing, S3 Transfer Acceleration is the more suitable ingestion solution. Still, the selection of the AWS ingestion service will depend on the source, volume, velocity, and format of the data at hand.

A real-time streaming data service: Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so that 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 clickstream data, and IoT telemetry data for machine learning, 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.

Three services within Kinesis help ingest data:

A video streaming service: Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, ML analysis, 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 play back video for live and on-demand viewing, and quickly build applications that take advantage of computer vision and video analytics through integration with Amazon Recognition Video, and libraries for ML frameworks, such as Apache MxNet, TensorFlow, and OpenCV.

Case in Point:

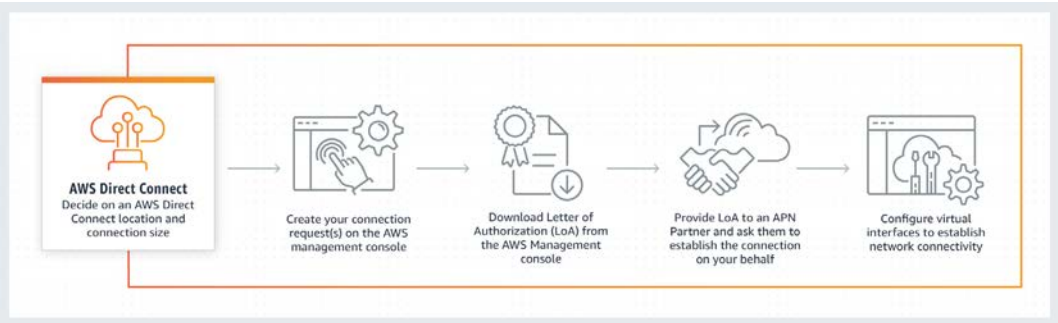
Brain Power

Automatic Analysis of Body Language Using Serverless and AI on AWS
Brain Power presents how they, together with AWS Professional Services, built a system to analyze body language to gauge attention, engagement, and enjoyment to help analyze clinical trial videos of children with autism and/or ADHD. The system uses technologies such as webcams and mobile devices to stream video directly to Amazon Kinesis Video Streams and later to Amazon Rekognition to detect face positions. Raw data is ingested into Amazon Kinesis Data Streams and consumed by AWS Lambda functions to analyze and mathematically compute attention and body motion metrics.

A real time data streaming service: Amazon Kinesis Data Streams (KDS) is a massively scalable and durable real-time data streaming service. KDS 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.

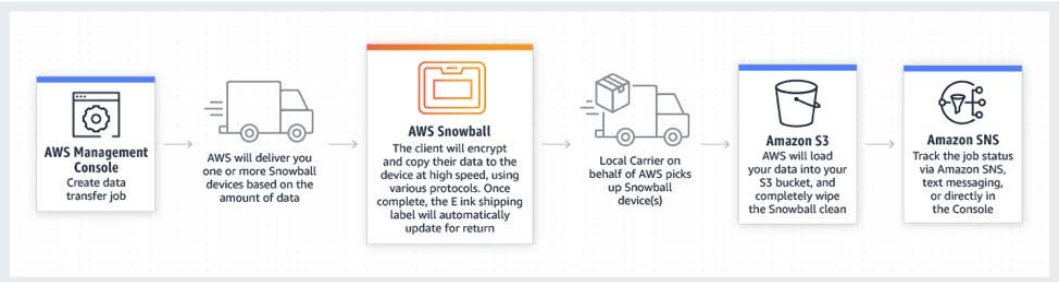
A capture-transform-load of streamed data: Amazon Kinesis Data Firehose is the easiest way to reliably load streaming data into data stores and analytics tools. It can capture, transform, and load streaming data into Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk, enabling near 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.

A dedicated network between AWS and on-premises IT: AWS Direct Connect is a cloud service solution that makes it easy to establish a dedicated network connection from your premises to AWS – bypassing your Internet service provider and removing network congestion. Transferring large data sets over the Internet can be time-consuming and expensive. Using AWS Direct Connect, you can establish private connectivity between AWS and your data center, office, or colocation environment, which, in many cases, can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.



A storage device for data transport: AWS Snowball is a data transport hardware device designed to securely transfer large amounts of data into and out of the AWS Cloud. AWS Snowball addresses common challenges with large-scale data transfers, including high network costs, long transfer times, and security concerns. Customers use Snowball to migrate analytics data, genomics data, video libraries, image repositories, and backups; as well as to archive part of data center shutdowns, for tape replacement, or for application migration projects. Transferring data with Snowball is fast, secure, and can cost as little as one-fifth of the price of transferring data via high-speed Internet.

If you are running MapReduce jobs on premises and storing data in the Hadoop Distributed File System (HDFS), you can now copy that data directly from HDFS to an AWS Snowball without using an intermediary staging file. Because HDFS is often used for big data workloads, this can greatly simplify the process of importing large amounts of data to AWS for further processing.



A storage device for data transport with compute: AWS Snowball Edge is a data migration and edge computing device that comes in two options. Snowball Edge Storage Optimized provides 100 TB of capacity and 24 vCPUs and is well suited for local storage and large scale data transfer. Snowball Edge Compute Optimized provides 52 vCPUs and an optional GPU for use cases such as advanced machine learning and full motion video analysis in disconnected environments. Customers can use these two options for data collection, machine learning and processing, and storage in environments with intermittent connectivity (such as manufacturing, industrial, and transportation), or in extremely remote locations (such as military or maritime operations) before shipping it back to AWS. These devices may also be rack mounted and clustered together to build larger, temporary installations.

When to use Snowball Edge? Besides storage, Snowball Edge includes compute capabilities. Snowball Edge supports specific Amazon EC2 instance types as well as AWS Lambda functions, so customers may develop and test in AWS then deploy applications on devices in remote locations to collect, pre-process, and return the data. Common use cases include data migration, data transport, image collation, IoT sensor stream capture, and machine learning.

A large scale storage device for data transport: AWS Snowmobile is a petabyte-scale data transfer service used to move extremely large amounts of data to AWS. You can transfer up to 100PB per Snowmobile, a 45-foot long ruggedized shipping container, pulled by a semi-trailer truck. Snowmobile makes it easy to move massive volumes of data to the cloud, including video libraries, image repositories, or even a complete data center migration. Transferring data with Snowmobile is secure, fast, and cost effective.

A long distance file transfer service: Amazon S3 Transfer Acceleration enables fast, easy, and secure transfers of files over long distances (100 or more miles) between your client and your Amazon S3 bucket. We will discuss Amazon S3 in greater detail in the next chapter on staging. During ingestion, Amazon S3 Transfer Acceleration leverages Amazon CloudFront's globally distributed AWS edge locations. An edge location is where end users can access services located in the AWS Cloud in closer proximity with reduced latency. AWS has 155 edge locations around the world, and is used in conjunction with content delivery network, Amazon CloudFront. As data arrives at an AWS edge location, data is routed to an Amazon S3 bucket over an optimized network path.

Case in Point:

DigitalGlobe

Using AWS Snowmobile, DigitalGlobe is able to deliver petabytes of data in weeks instead of months, while saving on costs and enabling the organization to deliver data to its customers in the shortest possible amount of time. DigitalGlobe is one of the world's leading providers of high-resolution Earth imagery, data, and analysis. It uses AWS Snowmobile to move up to 70 petabytes of archive data to the cloud, allowing it to move away from large file transfer protocols and delivery workflows.

FIND OUT MORE →

A data center gateway service: AWS Storage Gateway is a hybrid-storage service that enables your on-premises applications to seamlessly use AWS Cloud storage. You can use the service for backup and archiving, disaster recovery, cloud data processing, storage tiering, and migration. The service helps you reduce and simplify your data center and branch, or remote office storage infrastructure.

Your applications connect to the service through a virtual machine or hardware gateway appliance using standard storage protocols, such as NFS, SMB and iSCSI. The gateway connects to AWS storage services, such as Amazon S3, Amazon S3 Glacier, Amazon S3 Glacier Deep Archive, Amazon EBS, and AWS Backup, providing storage for files, volumes, snapshots, and virtual tapes in AWS.

A database migration service: AWS Database Migration Service helps migrate databases to AWS quickly and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to and from the most widely used commercial and open-source databases. As of early 2019, more than 120,000 databases had been migrated using AWS Database Migration Service.

AWS Database Migration Service supports homogenous migrations, such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle or Microsoft SQL Server to Amazon Aurora. With AWS Database Migration Service, you can continuously replicate your data with high availability and consolidate databases into a petabyte-scale data warehouse by streaming data to Amazon Redshift and Amazon S3.

Moving data across databases: 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.

Case in Point:

US Department of Veterans Affairs

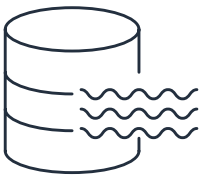
The US Department of Veterans Affairs (VA) processes hundreds of thousands of veterans appeals each year.

"Our appeals processing system, VACOLS, includes 20 million records stored in an Oracle 11g database. The system is more than 20 years old and is in the process of being modernized. During this time, we need to ensure that the data is securely replicated into the cloud for safekeeping. We're using AWS DMS to replicate the database into an RDS Oracle database in AWS GovCloud (US), in a Multi-AZ deployment. This setup ensures that VACOLS data is preserved, secured, and highly available in the cloud, which is a serious win for the VA and for our veterans, who rely on us for the safeguarding of their information."

– Alan Ning, Site Reliability Engineer – U.S. Digital Service, Veterans Affairs

A data transfer service between on-premises and AWS Cloud: AWS DataSync is a data transfer service that makes it easy for you to automate moving data between on-premises storage and Amazon S3 or Amazon Elastic File System (Amazon EFS). DataSync automatically handles many of the tasks related to data transfers that can slow down migrations or burden your IT operations, including running your own instances, handling encryption, managing scripts, network optimization, and data integrity validation. You can use DataSync to transfer data at speeds up to 10 times faster than open-source tools. DataSync uses an on-premises software agent to connect to your existing storage or file systems using the Network File System (NFS) protocol, so you don't have write scripts or modify your applications to work with AWS APIs. You can use DataSync to copy data over AWS Direct Connect or internet links to AWS. The service enables one-time data migrations, recurring data processing workflows, and automated replication for data protection and recovery. Getting started with DataSync is easy: Deploy the DataSync agent on premises, connect it to a file system or storage array, select Amazon EFS or S3 as your AWS storage, and start moving data. You pay only for the data you copy.

You can also transform and move AWS Cloud data into your data store using AWS Glue. AWS Glue is covered in more depth in Stage 3.



Stage 2 – Data Staging

Data staging provides the opportunity to perform data housekeeping and cleansing prior to making the data available for analysis. One of the most common challenges is that data is housed in multiple systems or locations, including data warehouses, spreadsheets, databases, and text files. Not only is the variety expanding; its volume, in many cases, is growing exponentially. Add to that, the complexity of mandatory data security and governance, user access, and the data demands of the analytics.

Key questions to consider: Which use cases is the organization looking to address with the data? With databases, an organization should not have to set up a data warehouse where an online transaction processing (OLTP) database would suffice. What is the size of the dataset? S3 remains the most robust destination for staging datasets. For very small sets of data, buffering in Amazon Kinesis Firehose could be optimal. Other relevant questions include, what is the structure of the data, and what are the cost considerations?

Object Storage as your Staging Environment

Amazon Simple Storage Service (Amazon S3) offers a robust destination for staging mechanism. Amazon S3 is object storage built to store and retrieve any amount of data from anywhere on the Internet. It's a simple storage service that offers an extremely durable, highly available, and infinitely scalable data storage infrastructure at very low costs.

Amazon S3 provides a simple web service interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web. Using this web service, you can easily build applications that make use of Internet storage. Since Amazon S3 is highly scalable and you only pay for what you use, you can start small and grow your application as you wish, with no compromise on performance or reliability.

The tool is also designed to be highly flexible: store any type and amount of data that you want; read the same piece of data a million times or only for emergency disaster recovery; build a simple FTP application, or a sophisticated web application such as the Amazon.com retail web site. Amazon S3 frees developers to focus on innovation instead of figuring out how to store their data.

Amazon S3 enables any developer to leverage Amazon's own benefits of massive scale with no up-front investment or performance compromises. Developers can innovate with the confidence that no matter the size or scope of their businesses, it will be inexpensive and easy to make their data quickly accessible, always available, and secure. And by adding user-defined metadata to Amazon S3 objects, a user can gain assurance and a deeper understanding of their data. Learn more about Amazon S3 user-defined metadata.

We regularly see the following metadata included with Amazon S3 objects:

Technical	Operational	Business
Format	Source of data	Meaning and context
Structure	Frequency (daily/weekly/monthly/ad-hoc)	Data classification
Checksum	Size (Full extract, historical, incremental)	PII/PHI
Row count	Currency (Date information)	Data set
Date stamp	Confidence/accuracy	Data element
File type	Lineage	Key data element (Primary keys/foreign keys)
File size	Data Owner	Data domain/subject area

Creating a Data Lake

A data lake is an increasingly popular way to store and analyze data that addresses the challenges of dealing with massive volumes of heterogeneous data that is queried by multiple users within the organization. A data lake lets you store data as-is; there is no need to convert it to a predefined schema, allowing you to store all of your data – structured or unstructured – in one centralized repository.

Many organizations are realizing the benefits of using Amazon S3 as their data lake. For example, Amazon S3 is a highly durable, cost-effective object start that supports open formats of data while decoupling storage from compute, and it works with all AWS analytic services. Although Amazon S3 provides the foundation of a data lake, you can add other services to tailor the data lake to your business needs. (Check here for additional resources on data lakes.)

A data lake template: 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 of 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.

Typical steps of building a data lake

1) Setup storage 2) Move data 3) Cleanse, Prep, and Catalog Data 4) Configure and enforce security 5) Make data available for analytics

Data Security and Classification

Amazon S3 object tags provide data security and classification, and when used in combination with Amazon Identity and Access Management (IAM), can granularly control access to objects in Amazon S3. Object tags are user created key/value pairs that you can add to Amazon S3 buckets or objects. Just like jpeg image EXIF data, object tags are not added to the content of the S3 object. For example, an Amazon S3 object can have the following key/value pairs:

Key	⟷	Value
Classification	⟷	Sensitive
Use Case	⟷	Analytics
Project	⟷	Wireless Upgrade
PII	⟷	True

Case in Point:

Healthdirect Australia

Using AWS to Connect People with Healthcare
Healthdirect Australia walks viewers through a system that supports every health service, provider, and practitioner in Australia. The architecture is split into two sides—write-intensive and read-intensive—and leverages multiple AWS services including Amazon API Gateway, AWS Lambda, Amazon DynamoDB, Amazon Kinesis, Amazon S3, Amazon EMR, Amazon Elasticsearch Service, and Amazon Athena.

AWS Cloud Databases

For more structured data, AWS offers a variety of purpose-built cloud databases that deliver cost-savings, scale, and high performance.

A MySQL and PostgreSQL relational database: Amazon Aurora is a MySQL and PostgreSQL compatible relational database engine that combines the speed and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases. Amazon Aurora features a distributed, fault-tolerant, self-healing storage system that auto-scales up to 64TB per database instance. It delivers high performance and availability with up to 15 low-latency read replicas, point-in-time recovery, and continuous backup to Amazon S3, and replication across three Availability Zones (AZs).

A managed MySQL database: Amazon Aurora Serverless is an on-demand, auto-scaling configuration for Amazon Aurora (MySQL-compatible edition), where the database will automatically start up, shut down, and scale capacity up or down based on your application's needs. It enables you to run your database in the cloud without managing any database instances. It's a simple, cost-effective option for infrequent, intermittent, or unpredictable workloads.

Manually managing database capacity can take up valuable time and can lead to inefficient use of database resources. With Aurora Serverless, you simply create a database endpoint, optionally specify the desired database capacity range, and connect your applications. You pay on a per-second basis for the database capacity you use when the database is active, and migrate between standard and serverless configurations with a few clicks in the Amazon RDS Management Console.

A relational database: Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while automating time-consuming administration tasks such as hardware provisioning, database setup, patching and backups. It frees you to focus on your applications so you can give them the fast performance, high availability, security and compatibility they need.

Amazon RDS is available on several database instance types – optimized for memory, performance or I/O – and provides you with six familiar database engines to choose from, including Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database, and SQL Server. As discussed in the earlier chapter, you can use the AWS Database Migration Service to easily migrate or replicate your existing databases to Amazon RDS.

Amazon Relational Database Service (RDS) on VMware lets you deploy managed databases in on-premises VMware environments using the Amazon RDS technology enjoyed by hundreds of thousands of AWS customers. Amazon RDS provides cost efficient and resizable capacity while automating time-consuming administration tasks including hardware provisioning, database setup, patching, and backups, freeing you to focus on your applications. RDS on VMware brings these same benefits to your on premises deployments, making it easy to set up, operate, and scale databases in VMware vSphere private data centers, or to migrate them to AWS.

A managed key-value database: Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale. It's a fully managed, multiregion, multimaster database with built-in security, backup and restore, and in-memory caching for internet-scale applications. DynamoDB can handle more than 10 trillion requests per day and support peaks of more than 20 million requests per second.

More than 100,000 AWS customers have chosen DynamoDB as their key-value and document database for mobile, web, gaming, ad tech, IoT, and other applications that need low-latency data access at any scale. Create a new table for your application and DynamoDB handles the rest. Many of the world's fastest growing businesses such as Lyft, Airbnb, and Redfin as well as enterprises such as Samsung, Toyota, and Capital One depend on the scale and performance of DynamoDB to support their mission-critical workloads.

DocumentDB differs from DynamoDB in that it supports large documents, and is MongoDB compatible.

An in-memory retrieval web service: Amazon ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud. The service improves the performance of web applications by allowing you to retrieve information from fast, managed, in-memory caches, instead of relying entirely on slower disk-based databases.

AWS Databases At-a-Glance: Comparing and contrasting AWS services

	Amazon ElastiCache	Amazon DynamoDB + DAX	Amazon Aurora	Amazon RDS	Amazon Elasticsearch Service	Amazon Neptune	Amazon S3 + Amazon Glacier
Use Cases	In memory caching	K/V lookups, document store	OLTP, transactional	OLTP, transactional	Log analysis, reverse indexing	Graph	File store
Performance	Ultra high request rate, ultra low latency	Ultra high request rate, ultra low to low latency	Very high request rate, low latency	High request rate, low latency	Medium request rate, low latency	Medium request rate, low latency	High throughput
Shape	K/V	K/V and document	Relational	Relational	Documents	Node/ edges	Files
Size	GB	TB, PB (no limits)	Low TB	GB, mid TB	GB, TB	GB, mid TB	GB, TB, PB, EB (no limits)
Cost/GB	$$	¢¢-$$	¢¢	¢¢	¢¢	¢¢	¢-¢4/10
Availability	2 AZ	3 AZ	3 AZ	2 AZ	1-2 AZ	3 AZ	3 AZ
VPC support	Inside VPC	VPC endpoint	Inside VPC	Inside VPC	Outside or inside VPC	Inside VPC	VPC endpoint
	HOT DATA				WARM DATA		COLD DATA

A graph database: Amazon Neptune is a fast, reliable, fully-managed graph database service that makes it easy to build and run applications that work with highly connected datasets. The core of Amazon Neptune is a purpose-built, high-performance graph database engine optimized for storing billions of relationships and querying the graph with milliseconds latency. Amazon Neptune supports popular graph models Property Graph and W3C's RDF, and their respective query languages Apache TinkerPop Gremlin and SPARQL, allowing you to easily build queries that efficiently navigate highly connected datasets. Neptune powers graph use cases such as recommendation engines, fraud detection, knowledge graphs, drug discovery, and network security.

A managed ledger database: Amazon Quantum Ledger Based Database (QLDB) is a fully managed ledger database that provides a transparent, immutable, and cryptographically verifiable transaction log owned by a central trusted authority. Amazon QLDB tracks each and every application data change and maintains a complete and verifiable history of changes over time.

A managed time series database: Amazon Timestream is a fast, scalable, fully managed time series database service for IoT and operational applications that makes it easy to store and analyze trillions of events per day at 1/10th the cost of relational databases. Driven by the rise of IoT devices, IT systems, and smart industrial machines, time-series data – data that measures how things change over time – is one of the fastest growing data types. Time-series data has specific characteristics such as typically arriving in time order form, data is append-only, and queries are always over a time interval. While relational databases can store this data, they are inefficient at processing this data as they lack optimizations such as storing and retrieving data by time intervals.

Stage 3 – Data Cleansing

Before data is analyzed, the cleansing stage helps appropriately prepare data. During this stage, the ETL function transforms the data to a format that is optimized for code. For example, a field may contain data/time in a format that cannot be processed by the analytics software programs. Or there may be a name field where the first and last names need to be separated for processing. Other data cleansing processes include merging of data sources, aligning formats, converting strings to numerical data, or summarizing of data.

Key questions to consider: Which tool to use for the ETL function? If integrating with a third-party tool outside of the AWS Cloud, how is identity and access management addressed? Also, knowing which user you are cleansing the data for will determine how much cleansing is needed. The cleansing must be far more thorough for a BI analyst, than for a data scientist – who might value under-prepared data for its nuanced flavor.

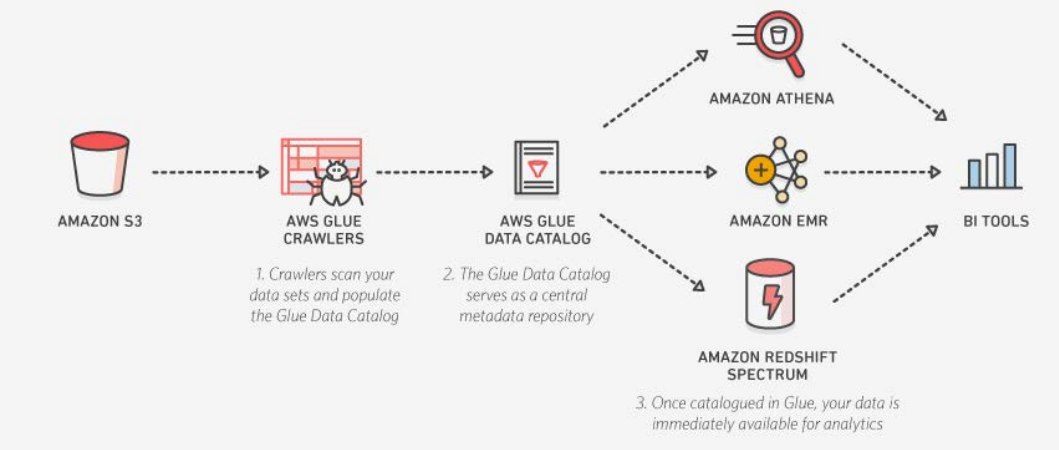
Extract, Transform and Load (ETL) in the Cloud

AWS Glue supports Spark ETL; whereas AWS EMR supports the entire Hadoop ecosystem, including Presto and Spark. Amazon Athena serves as an AWS service for self-managed Presto.

A fully managed ETL service: AWS Glue is a fully managed 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 on AWS, and AWS Glue discovers your data and stores the associated metadata (e.g. table definition and schema) in the AWS Glue Data Catalog. Once cataloged, your data is immediately searchable, queryable, and available for ETL. You can use AWS Glue to build a data warehouse to organize, cleanse, validate, and format data.

Simply select a data source and data target, and AWS Glue will generate ETL code in Scala or Python to extract data from the source, transform the data to match the target schema, and load it into the target. You can edit, debug, and test this code via the Console, in your favorite IDE, or notebook.

You can also run serverless queries on your data lakes, using AWS Glue. AWS Glue can catalog your Amazon S3 data, making it available for querying with Amazon Athena and Amazon Redshift Spectrum. With crawlers, your metadata stays in sync with the underlying data. Amazon Athena and Redshift Spectrum can directly query your Amazon S3 data lake using the AWS Glue Data Catalog. With AWS Glue, you can create a unified catalog of data for use with variety of AWS services.

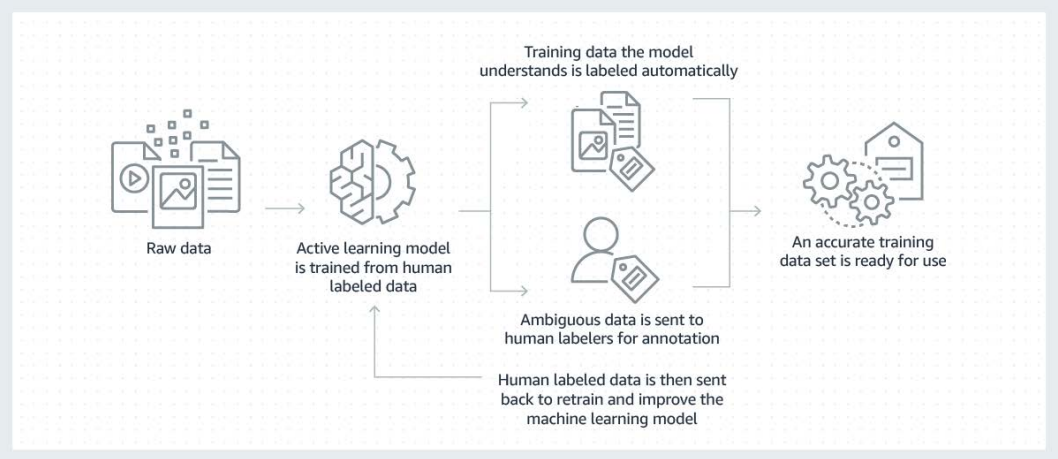


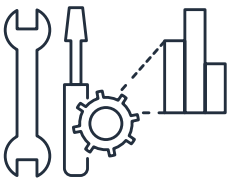
The AWS Glue Data Catalog contains references to data that is used as sources and targets of your extract, transform, and load (ETL) jobs in AWS Glue. To create your data warehouse, you must catalog this data. The AWS Glue Data Catalog is an index to the location, schema, and runtime metrics of your data. You use the information in the Data Catalog to create and monitor your ETL jobs. Typically, you run a crawler to take inventory of the data in your data stores, but there are other ways to add metadata tables into your Data Catalog. You can also use a crawler to populate the AWS Glue Data Catalog with tables.

A managed Hadoop and Spark framework: Amazon EMR provides a managed Hadoop environment that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances (virtual compute instances). You can also run other popular distributed frameworks such as Apache Spark, HBase, Presto, and Flink in Amazon EMR, and interact with data in other AWS data stores such as Amazon S3 and Amazon DynamoDB. EMR Notebooks, a managed environment based on the popular Jupyter Notebook, provide a development and collaboration environment for ad hoc querying and exploratory analysis for data scientists, developers and analysts. Amazon EMR securely and reliably handles a broad set of big data use cases, including log analysis, web indexing, data transformations (ETL), machine learning, financial analysis, scientific simulation, and bioinformatics.

You can launch an EMR cluster in minutes without needing to worry about node provisioning, cluster setup, Hadoop configuration, or cluster tuning. EMR takes care of these tasks so you can focus on analysis. The main processing frameworks available for Amazon EMR are Hadoop MapReduce and Spark.

Training datasets with labeling: Amazon SageMaker Ground Truth helps you build highly accurate training datasets for machine learning quickly. SageMaker Ground Truth offers easy access to public and private human labelers and provides them with built-in workflows and interfaces for common labeling tasks. Additionally, SageMaker Ground Truth can lower your labeling costs by up to 70% using automatic labeling, which works by training Ground Truth from data labeled by humans so that the service learns to label data independently.





Stage 4 – Data Analytics and Visualization

Data analytics and visualization is the stage of the lifecycle where the data preparation pays off in the form of actionable results for the organization. Where, on the one hand, analytics involves generating results from the data; visualization is about exploring data and communicating results from analysis to decision-makers.

Key questions to consider: Who are the consumers of the data at this stage within the organization? Are the results and presentations created in the stage aligned with the desired use cases? Would the decision-makers consuming the output find the insights compelling, understandable, and actionable?

Let's review some of the AWS services for data analytics and visualization:

An interactive query service: 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 also 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 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 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. You can also use Glue's fully-managed ETL capabilities to transform data or convert it into columnar formats to optimize cost and improve performance.

Case in Point:

Ivy Tech

Ivy Tech Community College of Indiana, the largest community college in the United States, wanted to better understand student engagement after enrollment. Large volumes of data produced from social media, online quizzes, adaptive learning tools, learning management systems, and online discussions, among other sources, offer an opportunity to personalize education for the individual learner.

With close to two million student records at 23 campuses, Ivy Tech started proactive remediation with 16,000 students. Ivy Tech achieved the largest decrease in dropout rates, as well as a reduction in low test scores, indicating that more students were passing their classes. Educators at Ivy Tech are now able to predict with 83% accuracy which students are likely to fail those classes – allowing them to anticipate remediation needs early.

FIND OUT MORE →

A scalable data warehouse: Amazon Redshift is a fast, scalable data warehouse that makes it simple and cost-effective to analyze all your data across your data warehouse and data lake. Redshift delivers ten times faster performance than other data warehouses by using machine learning, massively parallel query execution, and columnar storage on high-performance disk. You can setup and deploy a new data warehouse in minutes, and run queries across petabytes of data in your Redshift data warehouse, and exabytes of data in your data lake built on Amazon S3. You can start small for just $0.25 per hour and scale to $250 per terabyte per year, less than one-tenth the cost of other solutions.

Data Lakes compared to Data Warehouses

Depending on the requirements, a typical organization will require both a data warehouse and a data lake as they serve different needs and use cases. See table below for comparison.

As organizations with data warehouses see the benefits of data lakes, they are evolving their warehouses to include data lakes to enable diverse query capabilities, data science use cases, and advanced capabilities for discovering new information models. Gartner names this evolution the "Data Management Solution for Analytics," or DMSA."

Data Warehouse

A data warehouse is a database optimized to analyze relational data coming from transactional systems and line of business applications. The data structure, and schema are defined in advance to optimize for fast SQL queries, where the results are typically used for operational reporting and analysis. Data is cleaned, enriched, and transformed so it can act as the "single source of truth" that users can trust.

Data Lake

A data lake is different, because it stores relational data from line of business applications, and non-relational data from mobile apps, IoT devices, and social media. The structure of the data or schema is not defined when data is captured. This means you can store all of your data without careful design or the need to know what questions you might need answers for in the future. Different types of analytics on your data like SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights.

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

With Redshift Spectrum, Amazon Redshift customers can easily query their data in Amazon S3 (our object storage service). Like Amazon EMR, you get the benefits of open data formats and inexpensive storage, and you can scale out to thousands of nodes to pull data, filter, project, aggregate, group, and sort. Like Amazon Athena, Redshift Spectrum is serverless and there's nothing to provision or manage. You just pay for the resources you consume for the duration of your Redshift Spectrum query. Like Amazon Redshift itself, you get the benefits of a sophisticated query optimizer, fast access to data on local disks, and standard SQL. And like nothing else, Redshift Spectrum can execute highly sophisticated queries against an exabyte of data or more – in just minutes.

Case in Point:

University of Maryland University College (UMUC)

University of Maryland University College (UMUC) is an open-access institution. Almost 85% of the student population at UMUC takes courses online and many are working adults. UMUC had many legacy applications that needed updating, so when the time came to look for replacements and upgrades, they chose AWS for their IT infrastructure. UMUC uses a broad array of AWS services for analytics, including Amazon Relational Database (RDS), Amazon Redshift, and support for Oracle on AWS.

Since analytics is central to the university, UMUC runs predictive models for improving student outcomes by identifying at-risk students. Once identified, these students are directed to the pathways best suited to help them succeed. Administrators use dashboards for student performance, course details, and enrollment numbers, to help determine which courses need to be re-designed.

"The price-value proposition of [Amazon] Redshift is incredible," according to David Catalano VP of Analytics. "From an analytics perspective, [Amazon] Redshift is very disruptive." Besides cost savings, Amazon Redshift has also offered performance improvements through analysis. Compared to legacy applications, the university found the ETL process to be twice as efficient. With queries, it has witnessed five-to-ten times the efficiency gains using Amazon Redshift as compared to their last set of applications.

FIND OUT MORE →

A managed compute service: AWS Lambda is an AWS compute service that lets you run code without provisioning or managing servers. You pay only for the compute time you consume – there is no charge when your code is not running. You can use AWS Lambda to perform data validation, filtering, sorting, or other transformations for every data change in a DynamoDB table (or any other database) and load the transformed data to another data store. With Lambda, you can run code for virtually any type of application or backend service – all with zero administration.

A managed Elasticsearch service: Amazon Elasticsearch Service is a fully managed service that makes it easy for you to deploy, secure, and operate Elasticsearch at scale with zero down time. The service offers open-source Elasticsearch APIs, managed Kibana, and integrations with Logstash and other AWS Services, enabling you to securely ingest data from any source and search, analyze, and visualize it in real time. Amazon Elasticsearch Service lets you pay only for what you use – there are no upfront costs or usage requirements. With Amazon Elasticsearch Service, you get the ELK stack you need, without the operational overhead.

For predictive analytics use cases, AWS provides a broad set of machine learning services, and tools that run on your data lake on AWS. Our services come from the knowledge and capability we've built up at Amazon, where ML has powered Amazon.com's recommendation engines, supply chain, forecasting, fulfillment centers, and capacity planning.

A business intelligence (BI) tool: 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. Amazon 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. With the industry's first pay-per-session billing model, you only pay for what you use. This allows you to give all of your users access to the data they need without expensive per-seat licenses.

Case in Point:

UK Home Office

Building a Custom Cost Analytics Service for a Self-Managed Kubernetes Environment
In this video the UK Home Office, discusses a custom-built Cost Analytics service that internal customers use to consume reports around team utilization of their shared Kubernetes infrastructure. The session includes a architectural description of their ETL Pipeline, from collection of cost and utilization data to analysis in Amazon Quicksight.

Analytics with AI and ML

Amazon Machine Learning (Amazon ML) makes it easy for anyone to use predictive analytics and machine-learning technology. Amazon ML provides visualization tools and wizards to guide you through the process of creating machine learning (ML) models without having to learn complex ML algorithms and technology. After your models are ready, Amazon ML makes it easy to obtain predictions for your application using API operations, without having to implement custom prediction generation code or manage any infrastructure.

Amazon ML can create ML models based on data stored in Amazon S3, Amazon Redshift, or Amazon RDS. Built-in wizards guide you through the steps of interactively exploring your data, to training the ML model, to evaluating the model quality and adjusting outputs to align with business goals. After a model is ready, you can request predictions in either batches or using the low-latency real-time API.

Amazon ML is ideal for discovering patterns in your data and using these patterns to create ML models that can generate predictions on new, unseen data points. For example, you can:

- Enable applications to flag suspicious transactions – Build an ML model that predicts whether a new transaction is legitimate or fraudulent.
- Forecast product demand – Input historical order information to predict future order quantities.
- Personalize application content – Predict which items a user will be most interested in, and retrieve these predictions from your application in real-time.
- Predict user activity – Analyze user behavior to customize your website and provide a better user experience.
- Listen to social media – Ingest and analyze social media feeds that potentially impact business decisions.

Build, train, and deploy machine learning: Amazon SageMaker gives you the ability to build, train, and deploy machine learning models quickly. Amazon SageMaker is a fully-managed service that covers the entire machine learning workflow to label and prepare your data, choose an algorithm, train the algorithm, tune and optimize it for deployment, make predictions, and take action. Developers use Amazon SageMaker for building, training, and deploying ML since it comes with everything they need to connect to their training data, select, and optimize the best algorithm and framework, and deploy their model on auto-scaling clusters of Amazon EC2.

Infrastructure and tools for machine learning: AWS Deep Learning AMIs provide machine learning practitioners and researchers with the infrastructure and tools to accelerate deep learning in the cloud, at any scale. You can quickly launch Amazon EC2 instances pre-installed with popular deep learning frameworks such as Apache MXNet and Gluon, TensorFlow, Microsoft Cognitive Toolkit, Caffe, Caffe2, Theano, Torch, PyTorch, Chainer, and Keras to train sophisticated, custom AI models, experiment with new algorithms, or to learn new skills and techniques.

An additional set of nine AI services are shown in the diagram below:

Recommendations

Personalize experiences for your customers with the same recommendation technology used at Amazon.com.

AMAZON PERSONALIZE »

Forecasting

Build accurate forecasting models based on the same machine learning forecasting technology used by Amazon.com.

AMAZON FORECAST »

Image and Video Analysis

Add image and video analysis to your applications to catalog assets, automate media workflows, and extract meaning.

AMAZON REKOGNITION »

Advanced Text Analytics

Use natural language processing to extract insights and relationships from unstructured text.

AMAZON COMPREHEND »

Document Analysis

Automatically extract text and data from millions of documents in just hours, reducing manual efforts.

AMAZON TEXTRACT »

Voice

Turn text into lifelike speech to give voice to your applications.

AMAZON POLLY »

Conversational Agents

Easily build conversational agents to improve customer service and increase contact center efficiency.

AMAZON LEX »

Translation

Expand your reach through efficient and cost-effective translation to reach audiences in multiple languages.

AMAZON TRANSLATE »

Transcription

Easily add high-quality speech-to-text capabilities to your applications and workflows.

AMAZON TRANSCRIBE »

A Word about AWS Marketplace

AWS Marketplace is a curated digital catalog that enables qualified independent software vendors to sell software solutions to AWS customers. AWS Marketplace lets AWS customers find and use products and services offered by members of the AWS Partner Network. It offers a range of trials, annual subscriptions, and pay-as-you-go pricing options. Some offerings are available in software-as-a-service (SaaS) form and are billed based on consumption units specified by the seller. The SaaS model (described in New – SaaS subscriptions on AWS Marketplace) give sellers the flexibility to bill for actual usage: number of active hosts, number of requests, GB of log files processed, and so forth.

Stage 5 – Data Archiving

For the data archiving or re-use stage, AWS offers a flexible set of cloud storage services for archiving, making the process easy to manage, and allowing organizations to focus on the storage of data – rather than on managing tape systems and libraries. You can choose Amazon S3 Glacier or Amazon S3 Glacier Deep Archive for affordable, non-time sensitive cloud storage, or Amazon S3 for faster storage, depending on your needs. AWS Cloud storage solutions have achieved numerous compliance standards, security certifications, and provide built-in encryption. This offers the assurance that data stored in AWS meets all the requirements for your business.

Key questions to consider: Will the organization archive data for analytics or compliance? What are the cost and performance benefits of using cloud, versus traditional archiving? How might this augment or complement your current archiving practices?

Long term preservation of data: Amazon S3 Glacier is storage designed for low-cost storage services for data archival and long-term preservation. To keep costs low yet suitable for varying retrieval needs, the S3 API provides multiple options for access to archives, ranging from a few minutes to 48 hours. Some examples of archive use cases include digital media archives, financial and healthcare records, genomic sequence data, long-term backups, and data that must be retained for regulatory compliance.

A tape replacement service: S3 Glacier Deep Archive is a new Amazon S3 storage class that provides secure, durable object storage for long-term data retention and digital preservation. S3 Glacier Deep Archive offers the lowest price of storage in AWS, and reliably stores any amount of data. Amazon S3 Glacier stores data for as little as $0.004 per gigabyte per month and S3 Glacier Deep Archive is as little as $0.00099 per gigabyte per month (note: pricing varies by AWS region – quoted prices reflect the US-East region). It is the ideal storage class for customers who need to make archival, durable copies of data that rarely, if ever, need to be accessed.

With S3 Glacier Deep Archive, customers can eliminate the need for on-premises tape libraries, and no longer have to manage hardware refreshes and re-write data to new tapes as technologies evolve. All data stored in S3 Glacier Deep Archive can be retrieved within 12 hours. Amazon S3 Glacier Deep Archive joins Amazon S3's portfolio of storage classes, and complements Amazon S3 Glacier, which is ideal for more active archives where some of the data may need to be retrieved in minutes.

A storage class for less frequent access: Amazon S3 Standard-Infrequent Access (S3 Standard-IA) is an Amazon S3 storage class for data that is accessed less frequently but requires rapid access when needed. S3 Standard-IA offers the high durability, throughput, and low latency of the Amazon S3 Standard storage class, with a low per-GB storage price and per-GB retrieval fee.

A storage class limited to a single zone: Amazon S3 One Zone-IA storage class is an Amazon S3 storage class that customers can choose to store objects in a single availability zone. S3 One Zone-IA storage redundantly stores data within that single Availability Zone to deliver storage at 20% less cost than geographically redundant S3 Standard-IA storage, which stores data redundantly across multiple geographically separate Availability Zones.

Automated tiering service: Amazon S3 Intelligent-Tiering (S3 Intelligent-Tiering) is an Amazon S3 storage class for data with unknown access patterns or changing access patterns that are difficult to learn. It is the first cloud storage class that delivers automatic cost savings by moving objects between two access tiers when access patterns change. One tier is optimized for frequent access and the other lower-cost tier is designed for infrequent access.

4 Data security, privacy and compliance

In many organizations, IT security and data governance processes can be complex, as data is stored in multiple IT environments, across hundreds of applications and thousands of users.

Data privacy, including data classification, is a core component of addressing security requirements. Organizations need an easier and pragmatic approach to administering their data assets to mitigate security risks.

Data classification considers confidentiality (controlled and authorized access to data based on permissions); integrity (accuracy and freedom from unauthorized changes); and availability (accessibility and usability of data when required) as the foundations for data security. And in some cases there are mandatory definitions and classifications, for example General Data Protection Regulation (GDPR), United States Munitions List classifications of Controlled Unclassified Information (CUI), Health Insurance Portability and Accountability Act (HIPAA), and Information Security Manual (ISM).

AWS has developed a security assurance program that uses best practices for global privacy and data protection to help you operate securely within AWS, and to make the best use of our security control environment. These security protections and control processes are independently validated by multiple third-party independent assessments.

AWS also meets a variety of compliance requirements with its AWS GovCloud (US) Regions, which offer isolated cloud environments that help customers address the FedRAMP High security control requirements; the DOJ's Criminal Justice Information Systems (CJIS) Security Policy; U.S. International Traffic in Arms Regulations (ITAR); Export Administration Regulations (EAR), the Department of Defense (DoD) Cloud Computing Security Requirements Guide (SRG) for Impact Levels 2, 4 and 5, FIPS 140-2, IRS-1075; and other compliance regimes.

Data security: As an AWS customer, you maintain full control of your content and responsibility for configuring access to AWS services and resources. We provide an advanced set of access, encryption, and logging features to help you do this effectively. We provide APIs for you to configure access control permissions for any of the services you develop or deploy in an AWS environment. We never use your content or derive information from it for marketing or advertising. AWS provides native identity and access management integration across many of its services plus API integration with any of your own applications or services.

Data access: AWS offers you capabilities to define, enforce, and manage user access policies across AWS services. AWS Identity and Access Management (IAM) lets you define individual user accounts with permissions across AWS resources. AWS Multi-Factor Authentication provides secure authentication for privileged accounts, including options for hardware-based authenticators. The AWS Directory Service allows you to integrate and federate with corporate directories to reduce administrative overhead and improve end-user experience.

Data privacy: The AWS infrastructure puts strong safeguards in place to help protect customer privacy. All data is stored in highly secure AWS data centers. AWS offers strong encryption for content in transit and at rest, and also provides you with the option to manage your own encryption keys. These features include the data encryption capabilities available in AWS storage and database services, such as Amazon Elastic Block Store, Amazon Simple Storage Service, Amazon RDS and Amazon Redshift; flexible key management options, including AWS Key Management Service (KMS), allowing customers to choose whether to have AWS manage their encryption keys or enable themselves for control; and Server-Side Encryption (SSE) with Amazon S3-Managed Keys (SSE-S3), SSE with AWS KMS-Managed Keys (SSE-KMS), or SSE with Customer-Provided Encryption Keys (SSE-C).

5 Conclusion

Conclusion

AWS provides a host of services to address an organization's data lifecycle and analytics requirements.

With a broad set of managed services to collect, process, and analyze data, the AWS platform makes it easy to build, deploy, and scale data applications – allowing you to focus on your organization's needs, rather than on updating and managing tools. Users can customize data and analytics solutions using multiple AWS services, to meet their unique business requirements in the most cost-optimized, high performance, and resilient way.

The result is a flexible data architecture that scales with your organization over a lifetime.

6

Further reading & resources

Further reading and resources

Web Resources

AWS in the Public Sector
Big Data on AWS
Cloud Storage with AWS
AWS Databases
Open Data on AWS

Partner Resources

AWS Partner Network
Data & Analytics Partner Solutions
AWS Marketplace
AWS Marketplace: Data Analysis and Visualization guide

Webinars, Case Studies, & Whitepapers:

Big Data Analytics Options Whitepaper
AWS Storage Optimization Whitepaper
Strategies for Migrating Oracle Databases to AWS
Big Data Customer Case Studies
Public Sector Customer Success Stories
Data Lifecycle and Analytics Webinar Recordings
Data Lifecycle Webinar – Slides
Analytics for Data-Driven Decisions Webinar – Slides

Trainings & 10-Minute Tutorials

AWS Training Course on Big data
AWS Training on Storage
Tutorials and Training for Big Data
AWS Tutorial on Storage
AWS Tutorial on Databases

AWS Blogs:

AWS Public Sector Blog
AWS Big Data Blog
AWS Machine Learning Blog
AWS Storage Blog

AWS Quick Start Tools

Get Started with Big Data & Analytics
Big Data Test Drives
Data Lake Foundation on AWS
Customer Ready Solutions
AWS Quick Starts

Appendix 1

AWS GovCloud

AWS GovCloud (US) Regions were built for sensitive data and regulated workloads, including Controlled Unclassified Information, or CUI. AWS GovCloud (US) Regions give government customers and regulated commercial companies the flexibility to architect cloud solutions that comply with: the FedRAMP High baseline, the DOJ's Criminal Justice Information Systems (CJIS) Security Policy, U.S. International Traffic in Arms Regulations (ITAR), Export Administration Regulations (EAR), Department of Defense (DoD) Cloud Computing Security Requirements Guide (SRG) for Impact Levels 2, 4 and 5, and other compliance regimes. AWS GovCloud (US-East) and (US-West) Regions are operated by employees who are U.S. citizens on U.S. soil. AWS GovCloud (US) is only accessible to U.S. entities and root account holders who pass a screening process, where customers must confirm that they will only use a U.S. Person (green card holder or citizen as defined by the U.S. Department of State) to manage and access root account keys to these regions.

Supported Services: The AWS GovCloud (US) Regions currently support AWS services in this list.

For a complete list of all AWS Regions and their supported services, see Products and Services by Region. Also see AWS Services in Scope by Compliance Program

Appendix 2

A Selection of AWS Data and Analytics Partners

C3

C3 provides a comprehensive PaaS for rapidly developing & operating big data, AI, and IoT applications. In addition, we offer a family of configurable and extensible SaaS applications. At the heart of all C3 products is the C3 Type System, which enables programmers and data scientists to develop big data, AI, and IoT applications in 1/10th the time and cost of alternative technologies.

Cloudera (merged with Hortonworks)

Cloudera helps organizations get more value from their data in the cloud. Our modern data management and analytics platform allows the customers the ability to process and explore data wherever it lives to drive greater customer insight, improve products and services and drive efficiencies.

Databricks

Databricks Unified Analytics Platform, unifies data science and engineering across the Machine Learning lifecycle from data preparation, to experimentation and deployment of ML applications.

Digital Reasoning

Digital Reasoning enables automated understanding of human communication. Digital Reasoning's award-winning machine learning platform, Synthesys, identifies threats, risks and opportunities by transforming information into a private Knowledge Graph.

Informatica

Informatica is a leading provider of enterprise cloud data management software that empowers organizations to access, integrate, manage and govern all its data, giving organizations a competitive advantage in today's global information economy.

Marklogic

MarkLogic's operational and transactional enterprise NoSQL database platform empowers our customers to build next generation applications on a unified, 360-degree view of their data.

Microstrategy

MicroStrategy is a leading worldwide provider of enterprise software platforms. We provide enterprise analytics, mobility, and security platforms that are flexible, powerful, scalable, and user-friendly. MicroStrategy enables users to conduct ad hoc analysis and share their insights anywhere, anytime with reports, documents and dashboards.

Recorded Future

Recorded Future delivers a complete threat intelligence solution powered by patented machine learning to lower risk. We empower organizations to reveal unknown threats before they impact business, and enable teams to respond to alerts 10 times faster. To supercharge the efforts of security teams, our technology automatically collects and analyzes intelligence from technical, open web, and dark web sources and aggregates customer-proprietary data.

Tableau

Tableau Software helps people see and understand data. Tableau helps anyone quickly analyze, visualize and share information. More than 10,000 organizations get rapid results with Tableau in the office and on-the-go. And tens of thousands of people use Tableau Public to share data in their blogs and websites.

Teradata

Teradata transforms how businesses work and people live through the power of data. Teradata leverages all of the data, all of the time, so you can analyze anything, deploy anywhere, and deliver analytics that matter. We call this pervasive data intelligence. And it's the answer to the complexity, cost, and inadequacy of today's approach to analytics.

TIBCO

TIBCO fuels digital business by enabling better decisions and faster, smarter actions through the TIBCO Connected Intelligence Cloud. From APIs and systems to devices and people, we interconnect everything, capture data in real time wherever it is, and augment the intelligence of your business through analytical insights.

Contributors

The following individuals from the World Wide Public Sector (WWPS) organization at Amazon Web Services (AWS) contributed to the Data Lifecycle and Analytics in the AWS Cloud reference guide:

- Asif Moinuddin, Business Development Manager – Storage
- Ben Snively, Solutions Architect – Data and Analytics, AI/ML
- Edgar Haren, Senior Product Marketing Manager
- Leila Nouri, Senior Product Marketing Manager
- Nader Nanjiani, Senior Product Marketing Manager
- Paul Macey, Specialist Solutions Architect – Public Sector, Big Data and Insights
- Peter Schmiedeskamp, Technical Business Development Data Analyst