AWS DAS-C01 readiness domain4

Wednesday, 26 May 2021 3:19 PM

The whitepaper Big Data Analytics Options on AWS is a good reference for starting to understand the operational characteristics of these services including:

- Cost
- Performance
- Durability and availability
- Scalability and elasticity
- Interfaces

Cost

Amazon Athena:

Pay for the resources you consume. Priced **per query, per TB of data scanned**, and charges based on the amount of data scanned by the query. You can **save significantly on per-query costs** and get better performance by compressing, partitioning, and converting your data into **columnar formats.** This allows Athena to read only the columns it needs for the query.

Amazon Elasticsearch Service

Pay only for what you use. You are charged for Amazon ES instance hours, Amazon EBS storage (if you choose this option), and standard data transfer fees.

There is a free tier. Storage for automated snapshots that are stored for 14 days is free of charge. Manual snapshots are charged per Amazon S3 storage rates.

Amazon EMR

you only **pay for the hours the cluster is up**. You can launch a **persistent** cluster that stays up **indefinitely** or a **temporary cluster** that terminates after the analysis is complete.

Amazon EMR supports a variety of Amazon EC2 instance types and all Amazon EC2 pricing options.

When you launch an Amazon EMR cluster (also called a "job flow"), you choose how many and what type of Amazon EC2 instances to provision. The

Amazon EMR price is in addition to the Amazon EC2 price.

Amazon Kinesis

Amazon Kinesis services are **serverless**. You pay for the resources you consume.

Amazon Redshift

No long-term commitments or upfront costs. Charges are based on the **size** and **number of nodes** of your cluster.

No additional charge for backup storage up to 100% of your provisioned storage.

Backup storage beyond the provisioned storage size, and backups stored after your cluster is terminated, are billed at standard Amazon S3 rates.

No data transfer charge for communication between Amazon S3 and Amazon Redshift.

Performance

Amazon Athena

You can improve the performance of your query by **compressing**, **partitioning**, and **converting** your data into **columnar** formats. This means Athena can scan less data from Amazon S3 when executing your query.

Amazon Athena supports open source columnar data formats such as **Apache Parquet and Apache ORC**.

Amazon Elasticsearch Service (Amazon ES)

Performance depends on multiple factors including:

1. instance type
2. workload
3. index
4. number of shards used
5. read replica configuration
6. storage configuration (instance or EBS)

A search engine makes heavy use of storage devices. Making disks faster will result in faster query and search performance. Amazon ES can use either the fast SSD instance storage for storing indexes or multiple EBS either the fast SSD instance storage for storing indexes or multiple EBS volumes.

Amazon EMR

Amazon EMR performance is driven by the type and number of EC2 instances you choose to run your analytics. Consider processing requirements, sufficient memory, storage, and processing power.

- For best performance, you should launch the cluster in the same region as your data and use the same region for all of your AWS resources that will be used with the cluster.
- For low latency workloads that need to run in close proximity to on-premises resources, consider Amazon EMR on AWS Outposts.
- Consider scaling back debugging when you've finished development and put your data processing application into full production to save on log costs and reduce the processing load on the cluster.

Amazon Kinesis

Kinesis Data Streams performance: Choose throughput capacity in terms of shards. The **enhanced fan-out option** can improve performance by increasing the throughput available to each individual consumer.

Kinesis Data Firehose performance: Specify a batch size or batch interval and data compression to control how quickly data is uploaded to destinations.

Amazon Redshift

Amazon Redshift uses a variety of innovations to obtain **very high performance** on data sets ranging in size from hundreds of gigabytes to a petabyte or more.

- It uses **columnar** storage, data compression, and zone maps to reduce the amount of I/O needed to perform queries.
- Amazon Redshift has a **massively parallel processing (MPP) architecture**, parallelizing and distributing SQL operations to take advantage of all available resources.
- The underlying hardware is designed for high-performance data

- The underlying hardware is designed for high-performance data processing, using locally attached storage to maximize throughput between the CPUs and drives, and a 10 GigE mesh network to maximize throughput between nodes.
- An Amazon Redshift cluster can use either dense storage or dense compute nodes. Performance can be tuned based on your data warehousing needs: AWS offers Dense Compute (DC) with SSD drives as well as Dense Storage (DS) options.

Amazon sagemaker

Amazon SageMaker hosting automatically scales to the performance needed for your application using **Application Auto Scaling**.

By using Amazon **SageMaker Elastic Inference (EI)**, you can speed up the throughput and decrease the latency of getting real-time inferences from your deep learning models that are deployed as Amazon SageMaker hosted models.

Durability and availability

Amazon Athena

Amazon Athena is highly available and executes queries using compute resources across multiple facilities, automatically routing queries appropriately if a particular facility is unreachable.

Because Athena uses Amazon S3 as its underlying data store, you get **Amazon S3 durability (99.999999999%)** of objects. Data is redundantly stored across multiple facilities and multiple devices in each facility.

Amazon Elasticsearch service (ES)

Enable zone awareness for high availability. When Zone Awareness is enabled, Amazon ES distributes the instances supporting the domain across two different Availability Zones.

Then, if you enable replicas in Amazon ES, the instances are automatically distributed to deliver cross-zone replication. Use **snapshots** (automated and manual) to build data durability for your Amazon ES domain.

You can use snapshots to recover your domain with preloaded data or to create a new domain with preloaded data. Snapshots are stored in Amazon S3, which is a secure, durable, highly-scalable object storage.

Amazon ES automatically creates **daily snapshots of each domain by default**.

Use the Amazon ES snapshot APIs to create additional manual snapshots. Manual snapshots can be used for cross-region disaster recovery and to provide additional durability.

Amazon kinesis

All Kinesis services are **fully managed.**

Kinesis Data Streams is the underlying entity for Kinesis Data Firehose and Kinesis Data Analytics.

- Kinesis Data Streams synchronously replicates data across three Availability Zones in an AWS Region, and stores that data for up to seven days.
- You can store a cursor in DynamoDB to durably track what has been read from an Amazon Kinesis Data Streams.
- For Kinesis Data Analytics applications, you can create and delete durable application backups through a simple API call.
- Amazon Kinesis Video Streams uses Amazon S3 as the underlying data store, which means your data is stored durably and reliably.

Amazon EMR

By default, Amazon EMR is fault tolerant for core node failures and continues job execution if a slave node goes down.

- When a core node fails, Amazon EMR will provision a new node.
- If all nodes in the cluster are lost, Amazon EMR will not replace them.

Amazon EMR provides configuration options that control how your cluster is terminated By default, clusters that you create using the console or the AWS CLI continue to run until you shut them down.

- You can configure the cluster to continue running after processing completes so that you can choose to terminate it manually when you no longer need it. Or, you can create a cluster, interact with the installed applications directly, and then manually terminate the cluster when you no longer need it. The clusters in these examples are referred to as long-running clusters.

- If you configure your cluster to be automatically terminated, it is terminated after all the steps complete. This is referred to as a transient cluster.

Amazon Redshift

Amazon Redshift **automatically detects and replaces a failed node in your data warehouse cluster**. The data warehouse cluster is read-only until a replacement node is provisioned and added to the DB, which typically only takes a few minutes.

Amazon Redshift makes your replacement node available immediately and streams your most frequently accessed data from Amazon S3 first to allow you to resume querying your data as quickly as possible.

Additionally, your data warehouse cluster remains available in the event of a drive failure; because Amazon Redshift mirrors your data across the cluster, it uses the data from another node to rebuild failed drives.

Amazon Redshift clusters reside within one Availability Zone, but if you wish to have a multi-AZ set up for Amazon Redshift, you can set up a mirror and then self-manage replication and failover.

Scalability and elasticity

Amazon Athena

Athena is serverless, so there is no infrastructure to set up or manage, and it can scale automatically, as needed.

Amazon ES

You can **add or remove instances, and modify Amazon EBS volumes** to accommodate data growth.

You can write code to monitor your domain using CloudWatch metrics and call the Amazon ES Service API to scale up and down based on thresholds call the Amazon ES service API to scale up and down based on thresholds you set. The service will execute the scaling without any downtime.

Amazon EMR

You can **resize your cluster** to add instances for peak workloads and remove instances to control costs when peak workloads subside.

- You can **add core nodes** that hold the Hadoop Distributed File System (HDFS) at any time to increase processing power and HDFS storage capacity (and throughput).
- You can also **add and remove task nodes** at any time which can process Hadoop jobs but do not maintain HDFS.
- You can **decouple memory and compute** from storage by **using Amazon S3 on EMRFS** along with or instead of local HDFS. This provides greater flexibility and cost-efficiency.

Amazon Kinesis

All of the Amazon Kinesis services are designed to handle any amount of streaming data and process data from hundreds of thousands of sources with very low latencies.

Kinesis Data Streams:

- The initial scale is based on the number of shards you select for the stream.
- You can increase or decrease the capacity of the stream at any time.
- Use API calls or development tools to automate scaling.

Kinesis Data Firehose

- Streams automatically scale up and down based on the data rate you specify for the stream.

Kinesis Data Analytics

- Set up your application for your future scaling needs by proactively increasing the number of input in-application streams from the default (one).
- Use multiple streams and Kinesis Data Analytics for SQL applications if your application has scaling needs beyond 100 MB/second.
- Use Kinesis Data Analytics for Java Applications if you want to use a single stream and application.

Kinesis Video Streams

- Automatically provisions and elastically scales to millions of devices and scales down when the devices are not transmitting video.

Amazon Redshift

You can easily **change the number or type of nodes** in your data warehouse as your performance or capacity needs change.
You can use **elastic resize** to scale your cluster by changing the number of nodes. Or, you can use **classic resize** to scale the cluster by specifying a different node type.

- While resizing, Amazon Redshift places your existing cluster into read-only mode, provisions a new cluster of your chosen size, and then copies data from your old cluster to your new one in parallel.
- During this process, you pay only for the active Amazon Redshift cluster.
- You can continue running queries against your old cluster while the new one is being provisioned. After your data has been copied to your new cluster, Amazon Redshift automatically redirects queries to your new cluster and removes the old cluster.

Interfaces

Amazon Athena

Querying can be done using the Athena Console. You can connect to Athena using the **CLI, API via SDK and JDBC**.

Athena integrates with Amazon Quick Sight to create visualizations based on the Athena queries.
Athena natively supports querying datasets and data sources that are registered with the A**WS Glue Data Catalog**.

Amazon ES

Amazon ES supports many of the commonly used open-source Elasticsearch APIs. The developer guide includes a list of supported Elasticsearch operations.

The **AWS CLI, API, or the AWS Management Console** can be used for creating and managing your domains. Amazon ES also integrates with **Amazon CloudWatch** for monitoring Amazon ES domain metrics and CloudTrail for auditing configuration API calls to Amazon ES domains.

Amazon ES includes built-in integration with Kibana, an open-source analytics, and visualization platform and supports integration with Logstash, an open-source data pipeline that helps you process logs and other event data. You can set up your Amazon ES domain as the backend store for all logs coming through your Logstash implementation to easily ingest structured and unstructured data from a variety of sources.

Amazon EMR

Amazon EMR supports many tools on top of **Hadoop** that can be used for big data analytics and each has its own interfaces.

Apache Hive - open source **data warehouse** and **analytics package** that runs on Hadoop.

- Hive is operated by Hive QL. It abstracts programming models and supports typical data warehouse interactions.
- Values in Hive tables are structured elements such as JSON objects, any user-defined data type, or any function written in Java.
- Amazon EMR improvements to Hive include direct integration with DynamoDB and Amazon S3.

Apache Pig - open source analytics package that runs on top of Hadoop.

- Pig is operated by Pig Latin. It provides a scripting language that lets you transform large data sets without having to write complex code in a lower level computer language.
- Pig works with structured and unstructured data in a variety of formats.
- Amazon EMR improvements to Pig include the ability to use multiple file systems, the ability to load customer JARs and scripts from Amazon S3, and additional functionality for String and DateTime processing.

Apache Spark - open source distributed processing framework and programming model that helps you do machine learning, stream processing, or graph analytics using Amazon EMR clusters. Spark provides additional speed for certain analytics and is the foundation for other power additional speed for certain analytics and is the foundation for other power tools such as:

- Shark (SQL driven data warehousing),
- Spark Streaming (streaming applications),
- GraphX (graph systems)
- MLlib (machine learning)

Apache HBase - an open source, non-relational, distributed database modeled after Google's BigTable that runs on top of Hadoop.

- It provides a fault-tolerant, efficient way of storing large quantities of sparse data using column-based compression and storage.
- It provides for fast lookup of data because data is stored in-memory instead of on disk.
- Optimized for sequential write operations, and highly efficient for batch inserts, updates, and deletes.
- Integrates with Apache Hive
- With Amazon EMR, you can back up HBase to Amazon S3 (full or incremental, manual or automated) and you can restore from a previously created backup.

Presto - open-source distributed SQL query engine optimized for low-latency, ad-hoc analysis of data

- Supports the **ANSI SQL standard**, including complex queries, aggregations, joins, and window functions.
- Process data from **multiple data sources** including the Hadoop Distributed File System (HDFS) and Amazon S3.

Kinesis Connector- enables EMR to directly read and query data from Kinesis Data Streams.

- Perform batch processing of Kinesis streams using existing Hadoop ecosystem tools such as Hive, Pig, MapReduce, Hadoop Streaming, and Cascading.
- Useful for streaming log analysis, complex data processing workflows, and ad-hoc queries.
- AWS provides the ability to quickly move large amounts of data from Amazon S3 to HDFS, from HDFS to Amazon S3, and between Amazon S3 buckets using Amazon EMR's S3DistCp.

Amazon Kinesis

Kinesis Data Streams:

- **Producers** write data using the Amazon **Kinesis PUT API**, an AWS

- **Producers** write data using the Amazon **Kinesis PUT API**, an AWS Software Development Kit (SDK) or toolkit abstraction, the Amazon Kinesis Producer Library (KPL), or the Amazon Kinesis Agent.
- For consumers, client libraries are provided to build and operate real-time streaming data processing applications. The Kinesis Client Library (KCL) acts as an intermediary between Amazon Kinesis Data Streams and your business applications.
- AWS services that be consumers include **Kinesis Data Analytics, Kinesis Data Firehose, and AWS Lambda**

Kinesis Data Firehose

- You can use a Kinesis data stream, the Kinesis Agent, or the Kinesis Data Firehose API using the AWS SDK to write to a Kinesis Data Firehose stream.
- You can also use Amazon CloudWatch Logs, CloudWatch Events, or AWS IoT Core as your data source.
- Kinesis Data Firehose streams can deliver data to one of four destinations: Amazon S3, Amazon Elasticsearch Service, Amazon Redshift, or Splunk.
- You can configure your Kinesis Data Firehose delivery stream to invoke a Lambda function to transform your data prior to its delivery to your selected destination.

Kinesis Data Analytics

- Input options for your Kinesis Data Analytics application are a Kinesis data stream or a Kinesis Data Firehose delivery stream.
- In your application code, you write the output of SQL statements to **one or more in-application streams**. You can optionally add an output configuration to your application to persist everything written to an in-application stream to an external destination such as an Amazon Kinesis data stream, a Kinesis Data Firehose delivery stream, or an AWS Lambda function.
- If the data in your stream requires format conversion, transformation, enrichment, or filtering, you can pre-process the data using an AWS Lambda function.

Kinesis Video Streams

- Kinesis Video Streams provides APIs for creating and managing Kinesis video streams.
- It also provides APIs for reading and writing media data to a stream.

- The Amazon Kinesis Video Streams Producer libraries are a set of easy-to-use libraries that are part of the Kinesis Video Streams Producer SDK.

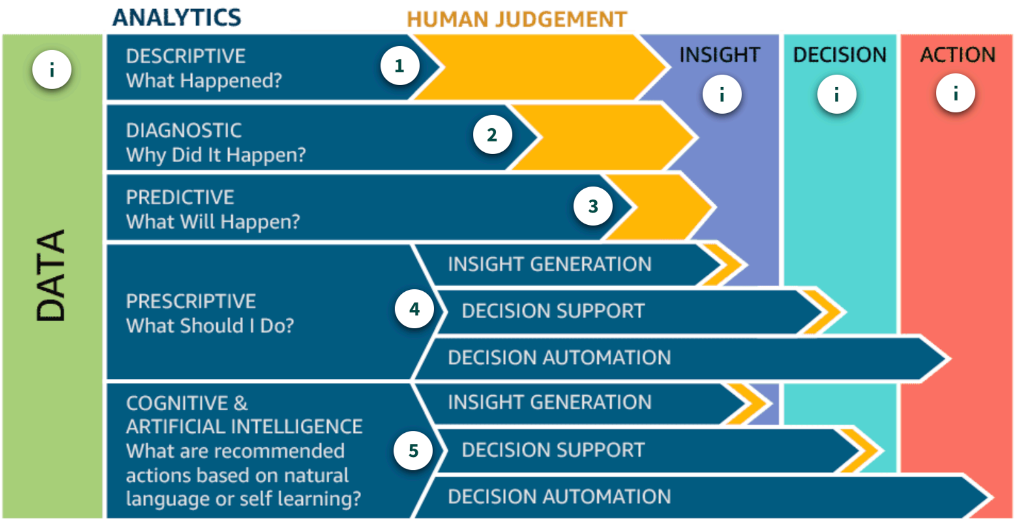
Amazon Redshift

Amazon Redshift has custom JDBC and ODBC drivers that you can download from the Connect Client tab of the console, allowing you to use a wide range of familiar SQL clients.
You can use standard PostgreSQL JDBC and ODBC drivers.

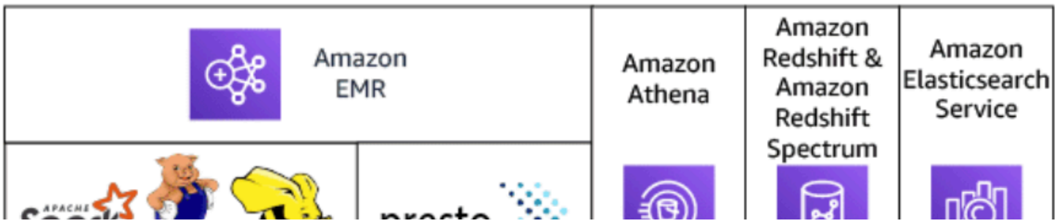
There are validated integrations with many popular BI and ETL vendors. Loads and unloads are attempted in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster as well as to and from Amazon S3 and DynamoDB.

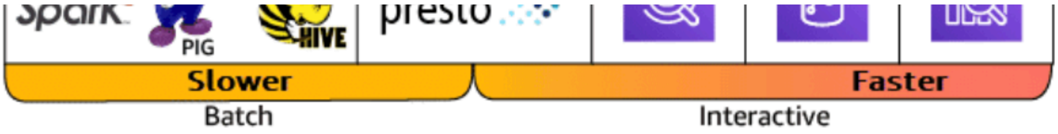
You can **load streaming data into Amazon Redshift using Amazon Kinesis Data Firehose**, enabling near real-time analytics with existing business intelligence tools and dashboards.

Selecting the right type of analysis



Selecting the right type of processing


Amazon EMR
Amazon Athena
Amazon Redshift & Amazon Redshift Spectrum
Amazon Elasticsearch Service
APACHE Spark
presto



Interactive analytics

Interactive analytics typically involves running complex queries across complex data sets at high speeds. This type of analytics is interactive in that it allows a user to query and see results right away.

Amazon Athena makes it easy to analyze data directly in Amazon S3 and Amazon S3 Glacier using standard SQL queries.

Amazon Elasticsearch Service allows you to search, explore, filter, aggregate, and visualize your data in near real-time.

Amazon Redshift provides the ability to run complex, analytic queries against petabytes of structured data and includes Redshift Spectrum, which runs **SQL queries directly** against exabytes of structured or unstructured data in **Amazon S3** without the need for unnecessary data movement.

Stream analytics

Amazon Kinesis services are designed for streaming data (Kinesis Data Streams, Kinesis Data Firehose, and Kinesis Data Analytics).

Amazon Athena patterns and anti-patterns

Ideal patterns

1. *Interactive ad hoc querying for weblogs* – **Athena is a good tool for interactive one-time SQL queries against data on Amazon S3**. For example, you could use Athena to run a query on web and application logs to troubleshoot a performance issue. Athena integrates with Amazon QuickSight for easy visualization.
2. *Interactive Analytical Solutions with notebook-based solutions* - Data scientists and Analysts are often concerned about managing the infrastructure behind big data platforms while running notebook-based solutions such as RStudio, Jupyter, and Zeppelin. Amazon Athena **makes it easy to analyze data using standard**

Amazon Athena **makes it easy to analyze data using standard SQL without the need to manage infrastructure.**

3. *Analyze AWS service logs* – AWS CloudTrail, Amazon CloudFront, Elastic Load Balancing and Amazon VPC flow logs can be analyzed with Athena. The logs allow you to investigate network traffic patterns and identify threats and risks across your Amazon VPC estate.
4. *Query staging data before loading into Amazon Redshift* – You can stage your raw data in Amazon S3 before processing and loading it into Amazon Redshift, and then use Athena to query that data.

Anti-patterns

1. *Enterprise Reporting and Business Intelligence Workloads* – **Amazon Redshift** is a **better tool for Enterprise Reporting and Business Intelligence** Workloads involving iceberg queries or cached data at the nodes. Data warehouses pull data from many sources, format and organize it, store it, and support complex, high speed queries that produce business reports. The query engine in Amazon Redshift has been optimized to perform especially well on data warehouse workloads.
2. ETL Workloads – You should use **Amazon EMR/AWS Glue** if you are looking for an **ETL tool t**o process extremely large datasets and analyze them with the latest big data processing frameworks such as Spark, Hadoop, Presto, or Hbase.
3. RDBMS – Athena is not a relational/transactional database. It is not meant to be a replacement for SQL engines like MySQL.

Amazon ES patterns and anti-patterns

Ideal patterns

1. Analyze activity logs, e.g., logs for customer facing applications or websites
2. Analyze CloudWatch logs
3. Analyze product usage data coming from various services and systems
4. Analyze social media sentiments, CRM data and find trends for your brand and products
5. Analyze data stream updates from other AWS services, e.g.

5. Analyze data stream updates from other AWS services, e.g., Amazon Kinesis Data Streams and Amazon DynamoDB
6. Provide customers a rich search and navigation experience.
7. Usage monitoring for mobile applications

<u>Anti-patterns</u>

1. Online transaction processing (OLTP) - Amazon ES is a real-time distributed search and analytics engine. There is **no support for transactions or processing** on data manipulation. If your requirement is for a **fast transactional system**, then a relational database system built on **Amazon RDS**, or a non-relational database offering functionality such as **DynamoDB, is a better choice.**
2. Ad hoc data querying – While Amazon ES takes care of the operational overhead of building a highly scalable Elasticsearch cluster if running Ad hoc queries or one-off queries against your data set is your use-case, Amazon Athena is a better choice.

Amazon EMR patterns and anti-patterns

<u>Ideal pattern</u>

Amazon EMR's flexible framework **reduces large processing problems and data sets into smaller jobs and distributes them across many compute nodes in a Hadoop cluster**. This capability lends itself to many usage patterns with big data analytics including:

1. Log processing and analytics
2. Large extract, transform, and load (ETL) data movement
3. Risk modeling and threat analytics
4. Ad targeting and clickstream analytics
5. Genomics
6. Predictive analytics
7. Ad hoc data mining and analytics

<u>Anti-pattern</u>

1. Small data sets – Amazon EMR is built for massively parallel processing; if your data set is small enough to run quickly on a single machine, in a single thread, the added overhead to map and reduce jobs may not be worth it for small data sets that can easily be processed in memory on a single system.

2. **ACID transaction requirements** – While there are ways to achieve ACID (atomicity, consistency, isolation, durability) or limited ACID on Hadoop, using another database, such as **Amazon RDS** or a relational database running on Amazon EC2 may be a better option for workloads with stringent requirements.

Amazon Kinesis patterns and anti-patterns

Ideal patterns

1. **Real-time data analytics** – Kinesis Data Streams enables real-time data analytics on streaming data, such as analyzing website clickstream data and customer engagement analytics.
2. **Log and data feed intake and processing** – With Kinesis Data Streams, you can have producers push data directly into an Amazon Kinesis data stream. For example, you can submit system and application logs to Kinesis Data Streams and access the stream for processing within seconds. This prevents the log data from being lost if the front-end or application server fails, and reduces local log storage on the source. Kinesis Data Streams provides accelerated data intake because you are not batching up the data on the servers before you submit it for intake.
3. **Real-time** metrics and reporting – You can use data ingested into Kinesis Data Streams for extracting metrics and generating KPIs to power reports and dashboards at real-time speeds. This enables data-processing application logic to work on data as it is streaming in continuously, rather than wait for data batches to arrive.

Anti-pattern

1. **Small scale** consistent throughput – Even though Kinesis Data Streams works for streaming data at 200 KB/sec or less, it is designed and optimized for larger data throughputs.
2. **Long-term data storage and analytics** –Kinesis Data Streams is not suited for long-term data storage. By default, data is retained for 24 hours, and you can extend the retention period by up to 7 days. You can move any data that needs to be stored for longer than 7 days into another durable storage service such as Amazon

than 7 days into another durable storage service such as Amazon S3, Amazon S3 Glacier, Amazon Redshift, or DynamoDB.

Amazon Redshift patterns and anti-patterns

Ideal patterns

Amazon Redshift is ideal for **online analytical processing (OLAP)** using your existing business intelligence tools. Organizations are using Amazon Redshift to:

1. Analyze global sales data for multiple products
2. Store historical stock trade data
3. Analyze ad impressions and clicks
4. Aggregate gaming data
5. Analyze social trends
6. Measure clinical quality, operation efficiency, and financial performance in health care

Anti-pattern

1. Small data sets – Amazon Redshift is built for parallel processing across a cluster. If your data set is less than a hundred gigabytes, you are not going to get all the benefits that Amazon Redshift has to offer and Amazon RDS may be a better solution.
2. On-line transaction processing (OLTP) – Amazon Redshift is designed for data warehouse workloads producing extremely fast and inexpensive analytic capabilities. If you require a fast transactional system, you may want to choose a traditional relational database system built on Amazon RDS or a Non-relational database offering, such as DynamoDB.

Amazon QuickSight patterns and anti-patterns

Ideal pattern

Amazon QuickSight is an ideal Business Intelligence tool allowing end-users to create visualizations that provide insight into their data to help them make better business decisions.

1. Quick interactive ad-hoc exploration and optimized visualization of data
2. Create and share dashboards and KPI's to provide insight into your data
3. Create Stories which are guided tours through specific views of an analysis and allow you to share insights and collaborate with

an analysis and allow you to share insights and collaborate with others. They are used to convey key points, a thought process, or the evolution of analysis for collaboration.

4. Analyze and visualize data coming from logs and stored in Amazon S3
5. Analyze and visualize data from on-premises databases like SQL Server, Oracle, PostgreSQL, and MySQL
6. Analyze and visualize data in various AWS resources, e.g., Amazon RDS databases, Amazon Redshift, Amazon Athena, and Amazon S3.
7. Analyze and visualize data in software as a service (SaaS) applications like Salesforce.
8. Analyze and visualize data in data sources that can be connected to using JDBC/ODBC connection.

Anti-pattern

1. Highly formatted canned Reports – Amazon QuickSight is much more suited for ad hoc query, analysis and visualization of data. For highly formatted reports e.g. formatted financial statements consider using a different tool.
2. ETL - While Amazon QuickSight can perform some transformations it is not a full-fledged ETL tool. AWS offers AWS Glue, which is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics.

Amazon SageMaker patterns and anti-patterns

Ideal pattern

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

that potentially impact business decisions.

Anti-pattern

1. Very large data sets – Terabyte-scale ingestion of data is not currently supported. Using Amazon EMR to run Spark’s Machine Learning Library (MLlib) is a common tool for such a use case.
2. Cases, where you need full control over your ML environment, are not ideal for Amazon SageMaker since it is a managed service.