

AWS DATA ANALYTICS SPECIALTY COLLECTIONS

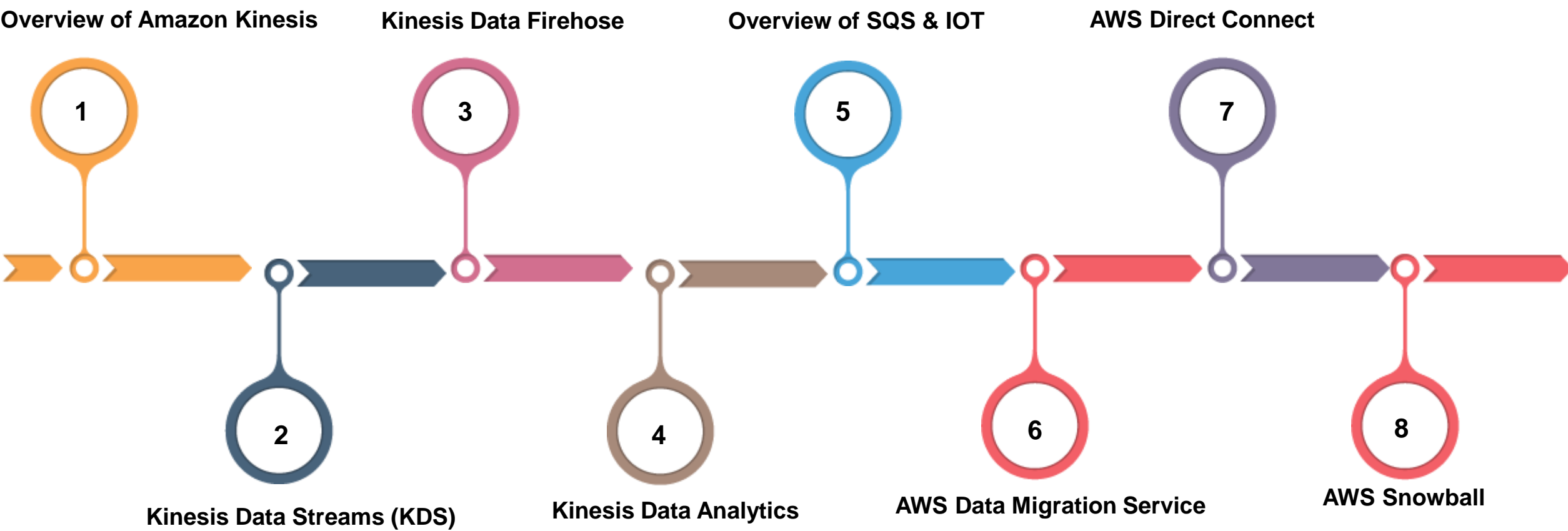
COURSE OBJECTIVES

At the end of this course, you should be able to:

- UNDERSTAND ABOUT AMAZON KINESIS SERVICES
- IMPLEMENT REAL TIME DATA PROCESSING USING KINESIS SERVICES
- UNDERSTAND ABOUT AMAZON SQS SERVICES
- UNDERSTAND ABOUT AMAZON IOT SERVICES
- UNDERSTAND ABOUT AMAZON DATA MIGRATION SERVICE
- UNDERSTAND ABOUT AWS DIRECT CONNECT
- UNDERSTAND ABOUT AWS SNOWBALL



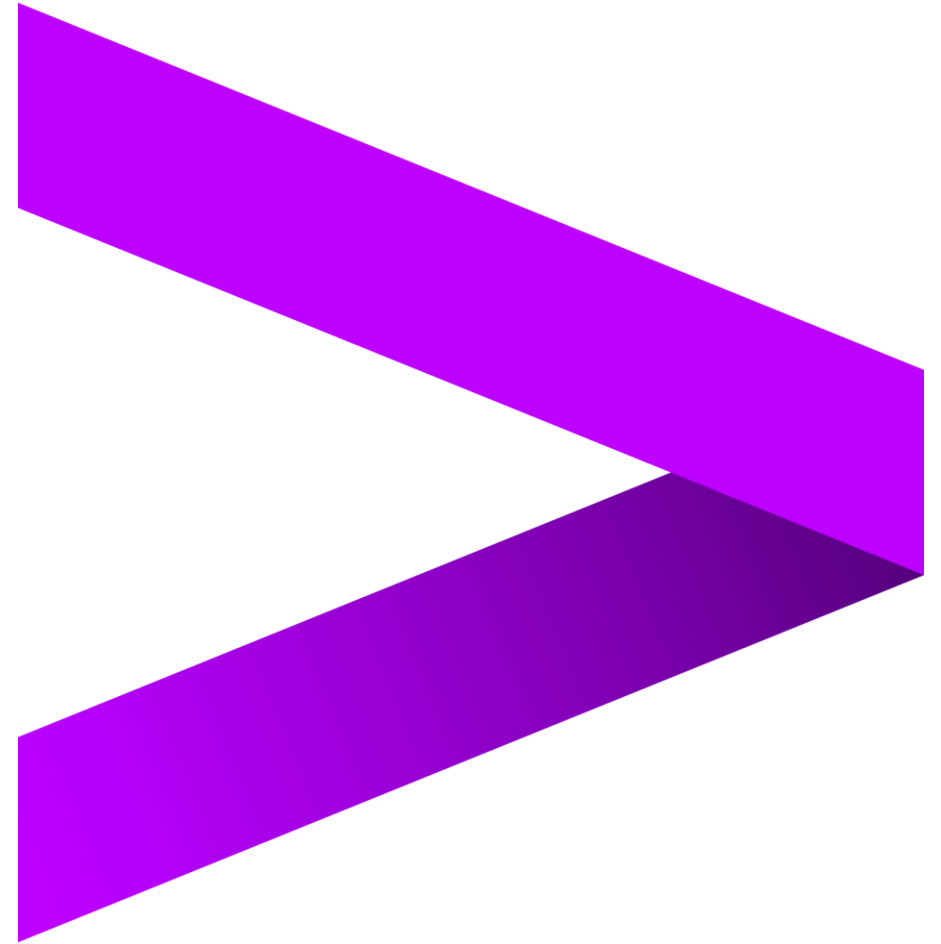
COURSE AGENDA



AWS DATA ANALYTICS SPECIALTY

AMAZON KINESIS

Learning and Knowledge Management



OVERVIEW OF AMAZON KINESIS?



What is Amazon Kinesis ?

It is a service in AWS to easily collect, process, and analyze data streams in real time

WITH AMAZON KINESIS,

Ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications

Enables to process and analyze data as it arrives and respond instantly

makes it easy to collect, process, and analyze real-time, streaming data to get timely insights and react quickly to new information.

Cost-effectively process streaming data at any scale, along with the flexibility to choose the tools that best suit the requirements of your application.

BENEFITS

- **REAL-TIME**
 - Amazon Kinesis enables you to ingest, buffer, and process streaming data in real-time, so you can derive insights in seconds or minutes instead of hours or days.
- **FULLY MANAGED**
 - Amazon Kinesis is fully managed and runs your streaming applications without requiring you to manage any infrastructure.
- **SCALABLE**
 - Amazon Kinesis can handle any amount of streaming data and process data from hundreds of thousands of sources with very low latencies.

AMAZON KINESIS SERVICES

KINESIS DATA STREAMS

- Capture and store data streams
- Amazon Kinesis Data Streams is a scalable and durable real-time data streaming service that can continuously capture gigabytes of data per second from hundreds of thousands of sources.

KINESIS DATA FIREHOSE

- Load data streams into AWS data stores
- Amazon Kinesis Data Firehose is the easiest way to capture, transform, and load data streams into AWS data stores for near real-time analytics with existing business intelligence tools.

AMAZON KINESIS SERVICES

KINESIS DATA ANALYTICS

- Analyze data streams with SQL or Apache Flink
- Amazon Kinesis Data Analytics is the easiest way to process data streams in real time with SQL or Apache Flink without having to learn new programming languages or processing frameworks.



AWS KINESIS

USE CASES



CLICKSTREAM ANALYTICS



Input

Websites send clickstream data to Amazon Kinesis Data Firehose



Amazon Kinesis Data Firehose

Collects the data and sends it to Amazon Kinesis Data Analytics



Amazon Kinesis Data Analytics

Processes data in real-time



Amazon Kinesis Data Firehose

Loads processed data into Amazon Redshift



Amazon Redshift

Run analytics models that come up with content recommendations



Output

Readers see personalized content suggestions and engage more

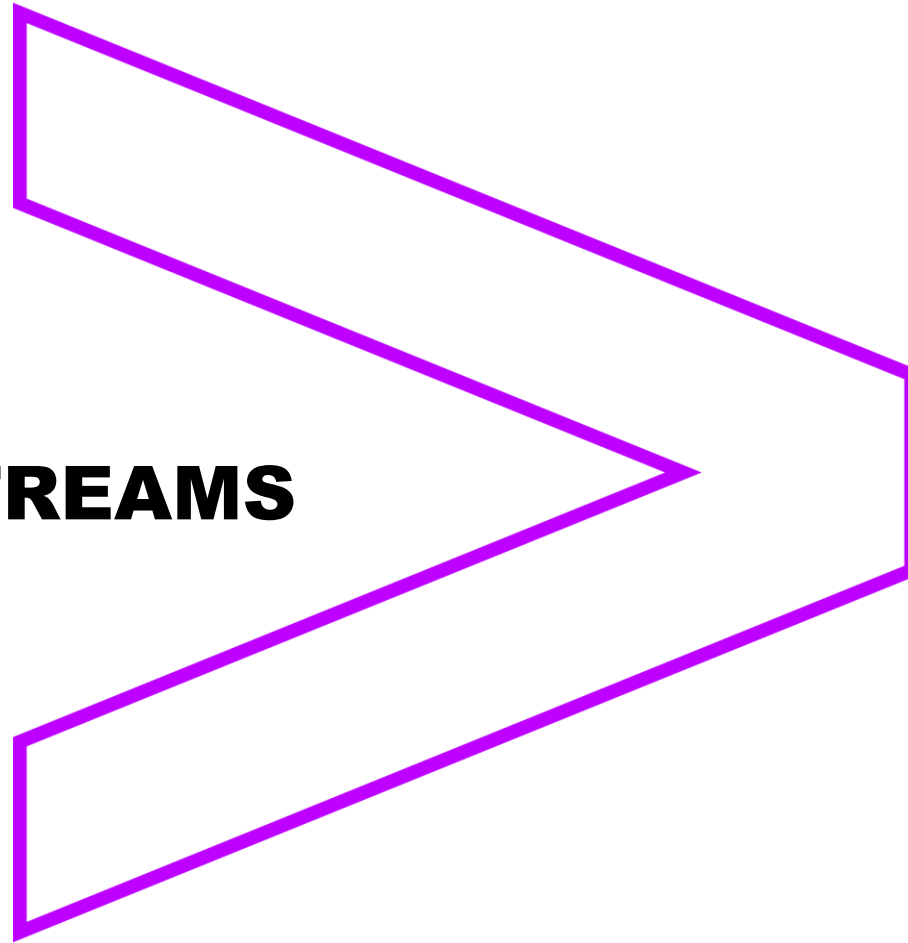
ANALYSIS OF STREAMING SOCIAL MEDIA DATA

Example: Analysis of streaming social media data



AWS

KINESIS DATA STREAMS



AMAZON KINESIS DATA STREAMS

AMAZON KINESIS DATA STREAMS (KDS)

- A massively scalable and durable real-time data streaming service.
- 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, dynamic pricing, and more.

KDS WITH OTHER SERVICES

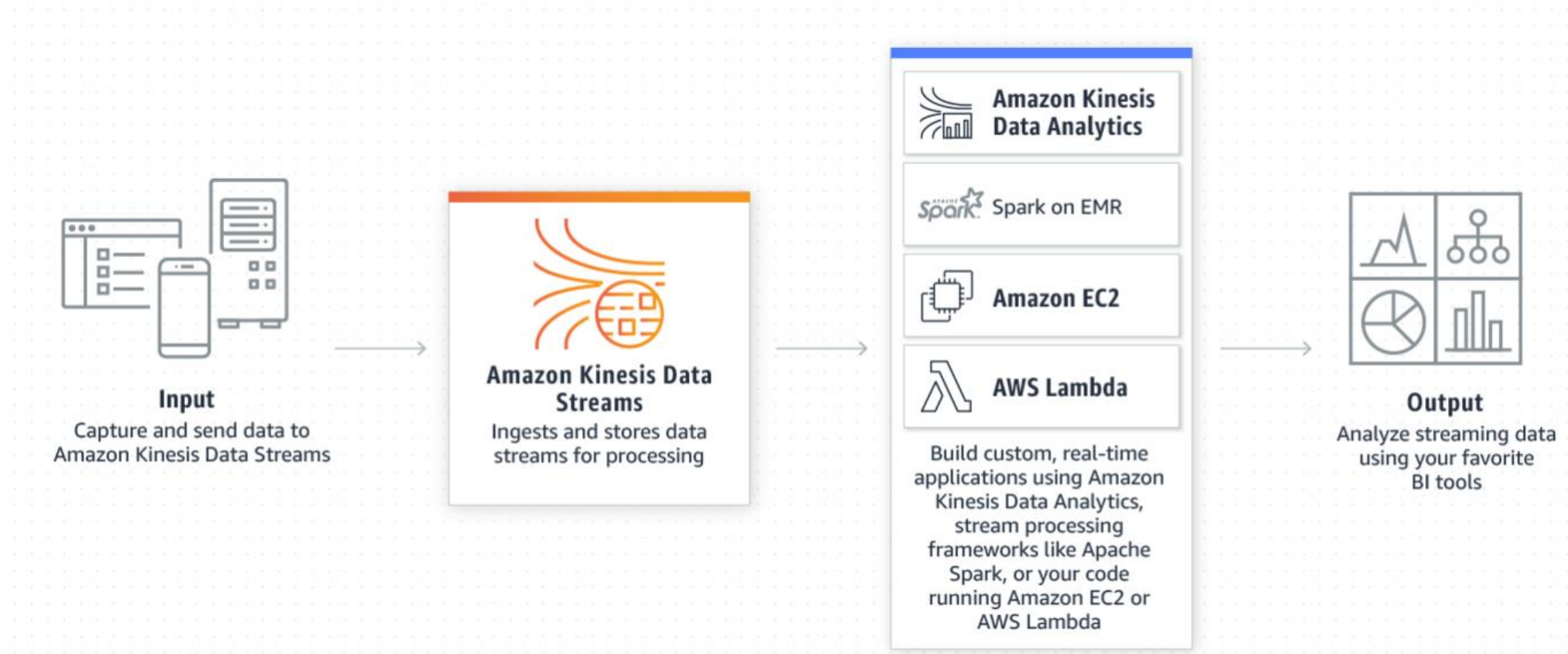
- Amazon Kinesis Data Streams is integrated with a number of AWS services,
 - **Amazon Kinesis Data Firehose**
 - for near real-time transformation and delivery of streaming data into an AWS data lake like **Amazon S3**, **Kinesis Data Analytics** for managed stream processing,
 - **AWS Lambda**
 - for event or record processing,
 - **AWS PrivateLink**
 - for private connectivity,
 -
 - **Amazon Cloudwatch**
 - for metrics and log processing, and
 - **AWS KMS**
 - for server-side encryption.

KDS BENEFITS

- REAL-TIME PERFORMANCE
- DURABLE
- SECURE
- EASY TO USE
- ELASTIC
- LOW COST

HOW KDS WORKS

Collect streaming data, at scale, for real-time analytics



Source :- <https://aws.amazon.com/kinesis/>

KDS KEY CONCEPTS

DATA PRODUCER

- A data producer is an application that typically **emits data records** as they are generated to a Kinesis data stream.
- Data producers assign partition keys to records.
- Partition keys ultimately determine which shard ingests the data record for a data stream.

DATA CONSUMER

- A data consumer is a distributed Kinesis application or AWS service retrieving data from all shards in a stream as it is generated.
- Most data consumers are retrieving the most recent data in a shard, enabling real-time analytics or handling of data.

KDS KEY CONCEPTS

DATA STREAM

- A data stream is a logical grouping of shards.
- There are no bounds on the number of shards within a data stream.
- A data stream will retain data for 24 hours, or up to 7 days when extended retention is enabled.

KDS KEY CONCEPTS

SHARD

- A shard is the base throughput unit of an Amazon Kinesis data stream.
 - A shard is an append-only log and a unit of streaming capability. A shard contains an ordered sequence of records ordered by arrival time.
 - One shard can ingest up to 1000 data records per second, or 1MB/sec. Add more shards to increase your ingestion capability.
 - Add or remove shards from your stream dynamically as your data throughput changes using the AWS console, **UpdateShardCount API**, trigger automatic scaling via AWS Lambda, or using an auto scaling utility.
 - When consumers use enhanced fan-out, one shard provides 1MB/sec data input and 2MB/sec data output for each data consumer registered to use enhanced fan-out.

KDS KEY CONCEPTS

DATA RECORD

- A record is the unit of data stored in an Amazon Kinesis stream.
- A record is composed of a sequence number, partition key, and data blob.
- A data blob is the data of interest your data producer adds to a stream.
- The maximum size of a data blob (the data payload after Base64-decoding) is 1 megabyte (MB).

KDS KEY CONCEPTS

PARTITION KEY

- A partition key is typically a meaningful identifier, such as a user ID or timestamp.
 - It is specified by your data producer while putting data into an Amazon Kinesis data stream, and useful for consumers as they can use the partition key to replay or build a history associated with the partition key.
- The partition key is also used to segregate and route data records to different shards of a stream.
 - For example, assuming you have an Amazon Kinesis data stream with two shards (Shard 1 and Shard 2).
 - You can configure your data producer to use two partition keys (Key A and Key B) so that all data records with Key A are added to Shard 1 and all data records with Key B are added to Shard 2.

KDS KEY CONCEPTS

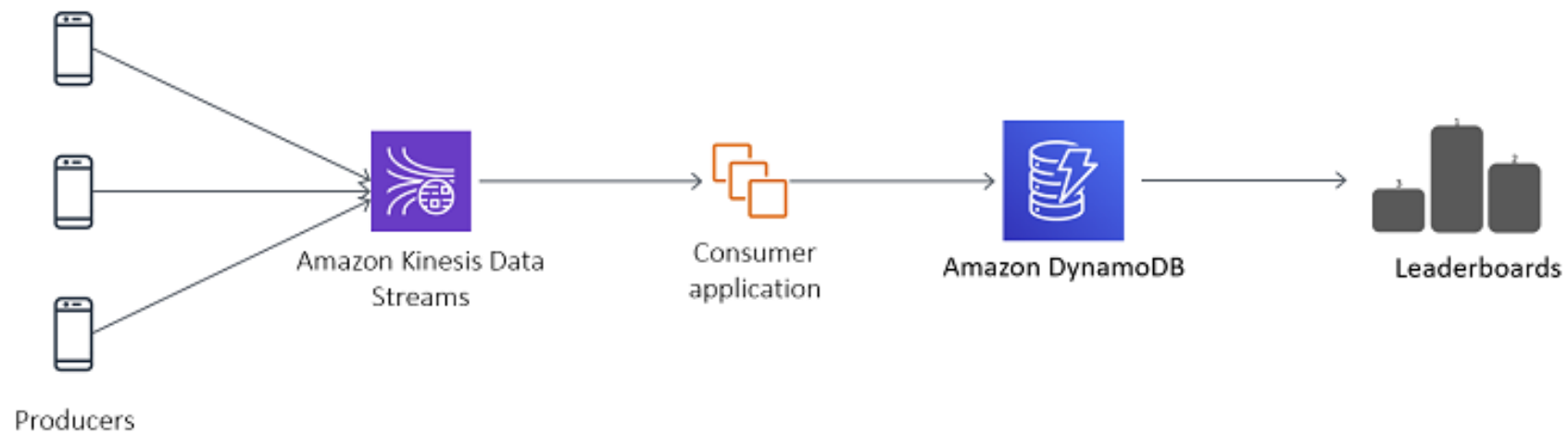
SEQUENCE NUMBER

- A sequence number is a unique identifier for each data record.
- Sequence number is assigned by Amazon Kinesis Data Streams when a data producer calls PutRecord or PutRecords API to add data to an Amazon Kinesis data stream.
- Sequence numbers for the same partition key generally increase over time; the longer the time period between PutRecord or PutRecords requests, the larger the sequence numbers become.

USE CASES

- LOG AND EVENT DATA COLLECTION
- MOBILE DATA CAPTURE
- REAL-TIME ANALYTICS
- GAMING DATA FEED

DATA FLOW



AMAZON KINESIS DATA FIREHOSE



accenture

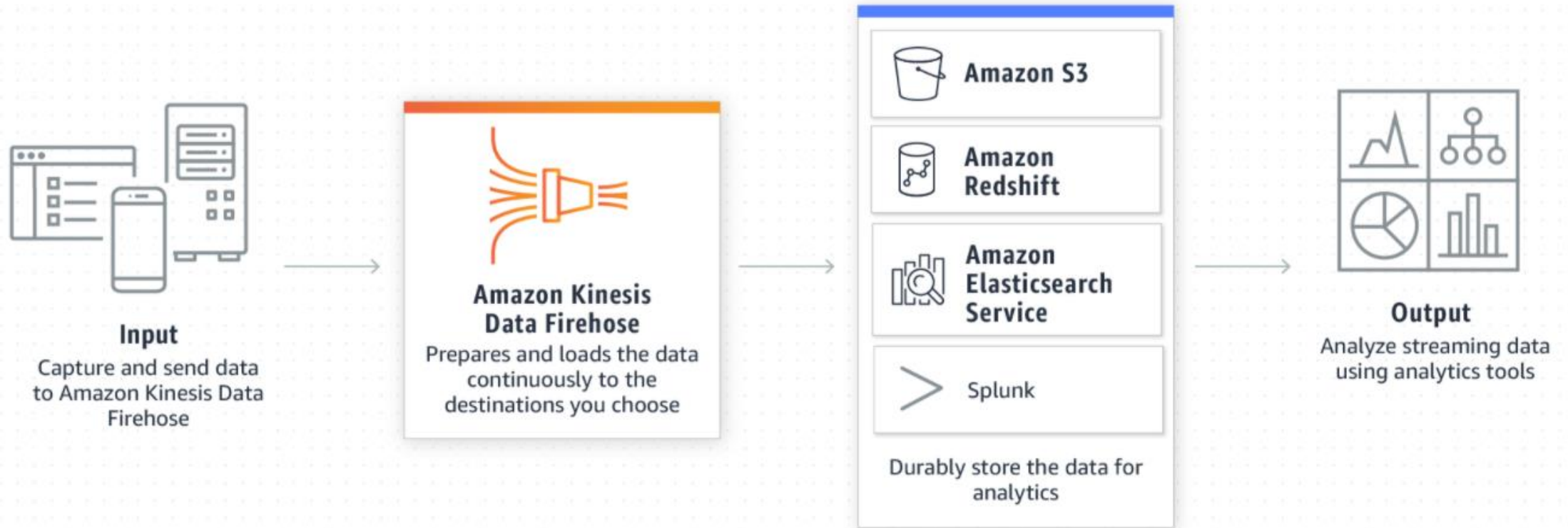
WHAT IS AMAZON KINESIS DATA FIREHOSE?

- Amazon Kinesis Data Firehose is a fully managed service for delivering real-time streaming data to destinations such as
 - Amazon Simple Storage Service (Amazon S3),
 - Amazon Redshift,
 - Amazon Elasticsearch Service (Amazon ES),
 - Splunk, and
 - any custom HTTP endpoint or HTTP endpoints owned by supported third-party service providers, including Datadog, MongoDB, and New Relic.

WHAT IS AMAZON KINESIS DATA FIREHOSE?

- Configure the data producers to send data to Kinesis Data Firehose, and it **automatically delivers** the data to the destination that you specified.
- Can also configure Kinesis Data Firehose to transform your data before delivering it.

HOW IT WORKS



KINESIS DATA FIREHOSE KEY CONCEPTS

Kinesis Data Firehose Delivery Stream

- The underlying entity of Kinesis Data Firehose.
- Use Kinesis Data Firehose by creating a **Kinesis Data Firehose delivery stream** and then sending data to it.

Record

- The data of interest that your data producer sends to a Kinesis Data Firehose delivery stream.
- A record can be as large as 1,000 KB.



AMAZON KINESIS DATA ANALYTICS

AMAZON KINESIS DATA ANALYTICS

- Amazon Kinesis Data Analytics is the easiest way to transform and analyze streaming data in real time with SQL or Apache Flink.
 - **Apache Flink** is an open source framework and engine for processing data streams. Amazon Kinesis Data Analytics reduces the complexity of building, managing, and integrating Apache Flink applications with other AWS services.
- Amazon Kinesis Data Analytics takes care of everything required to run streaming applications continuously, and scales automatically to match the volume and throughput of your incoming data.
- With Amazon Kinesis Data Analytics,
 - there are no servers to manage,
 - no minimum fee or setup cost, and
 - pay for the resources your streaming applications consume.

HOW KDA WORKS



USE CASES

- **STREAMING ETL**

- You can develop streaming extract-transform-load (ETL) applications with Amazon Kinesis Data Analytics built-in operators to transform, aggregate, and filter streaming data.
- You can easily deliver your data in seconds to Amazon Kinesis Data Streams, Amazon Managed Streaming for Apache Kafka (Amazon MSK), Amazon Elasticsearch Service, Amazon S3, custom integrations, and more using built-in connectors.

- **REAL-TIME ANALYTICS**

- You can interactively query streaming data using standard SQL and build Apache Flink applications using Java and Scala to analyze data streams.
- You can calculate key business and operational metrics, refresh content performance dashboards, and analyze customer experiences, at scale, in real time.

- **STATEFUL EVENT PROCESSING**

- You can develop applications that process events from one or more data streams and trigger conditional processing and external actions.
- You can identify patterns like anomaly detection in your data streams using standard SQL and Apache Flink libraries for complex event processing.

SUMMARY

Amazon Kinesis Data Streams

- Data is captured from multiple sources and is sent to Kinesis data streams.
- Amazon Kinesis Data Streams stores the data for processing.
- This data can be utilized in many ways, like building customized and real-time applications or performing stream processing frameworks like Apache Spark.
- Users can then use any of their favourite business intelligence tools to understand the outputs.

Amazon Kinesis Data Firehose

- Data captured is sent to Amazon Kinesis Data Firehouse.
- Kinesis data firehouse is continuously loaded and prepared for the chosen destinations.
- Streamed data hereafter is stored for analytics with the help of tools like Splunk and Amazon Elasticsearch service for analysis.

SUMMARY


Amazon Kinesis Data Analytics

- Streaming data is collected with the help of Kinesis data firehouse and Kinesis data streams.
- Amazon Kinesis Data Analytics is used for query purposes and for analyzing streaming data.
- For allowing users to create alerts and respond quickly, Amazon Kinesis Data Analytics sends processed data to analytics tools.

Kinesis Video Streams

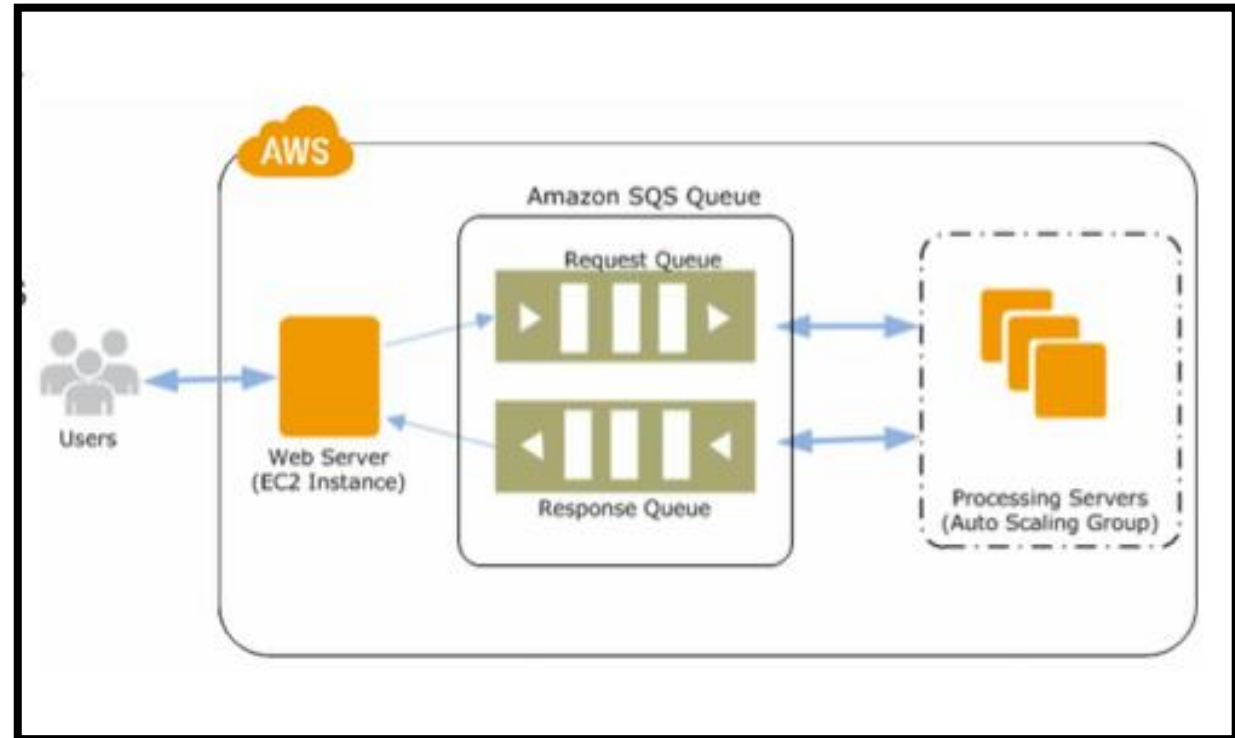
- Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), and other processing.

OTHER IMPORTANT SERVICES



AMAZON SQS

- Amazon **Simple Queue Service (SQS)** is a fully managed message queuing service that enables you to decouple and scale microservices, distributed systems, and serverless applications.
- SQS eliminates the complexity and overhead associated with managing and operating message oriented middleware, and empowers developers to focus on differentiating work.





AMAZON SQS

- Using SQS, you can send, store, and receive messages between software components at any volume, without losing messages or requiring other services to be available.
- Can implement SQS using the AWS console, Command Line Interface or SDK of your choice, and three simple commands.
- SQS offers two types of message queues.
 - **Standard queues** offer maximum throughput, best-effort ordering, and at-least-once delivery.
 - **SQS FIFO queues** are designed to guarantee that messages are processed exactly once, in the exact order that they are sent.



AMAZON SQS

Standard Queue	FIFO Queue
	
Unlimited Throughput Supports a nearly unlimited number of transactions per second (TPS) per API action.	High Throughput By default, FIFO queues support up to 3,000 messages per second (TPS), per API action through batching
At-Least-Once Delivery A message is delivered at least once. Occasionally more than one copy of a message is delivered.	Exactly-Once Processing A message is delivered once and remains available until a consumer processes and deletes it. Duplicates are not introduced into the queue.
Best-Effort Ordering Occasionally, messages might be delivered in an order different from which they were sent.	First-In-First-Out Delivery The order in which messages are sent and received is strictly preserved.

You can configure the Amazon **SQS** message **retention period** to a value from 1 minute to 14 days. The default is 4 days.

KINESIS DATA STREAMS VS SQS

What is the difference between Kinesis vs SQS?

- Amazon **Kinesis** is differentiated from Amazon's Simple Queue Service (SQS) in that Kinesis is used to **enable real-time processing of streaming big data**.
- **SQS**, on the other hand, is used as a message queue **to store messages transmitted between distributed application components**.

AWS IOT

- IoT services for industrial, consumer, and commercial solutions

Device software

Connect your devices and operate them at the edge.



FreeRTOS

FreeRTOS is an operating system for microcontrollers that makes small, low-power edge devices easy to program, deploy, secure, connect, and manage.



AWS IoT Greengrass

AWS IoT Greengrass is software that lets you run local compute, messaging, data caching, sync, and machine learning inference capabilities on connected devices in a secure way.

AWS IOT

Connectivity & control services

Secure, control, and manage your devices from the cloud.



AWS IoT Core

AWS IoT Core lets connected devices easily and securely interact with cloud applications and other devices.



AWS IoT Device Defender

AWS IoT Device Defender continuously monitors and audits your IoT configurations to make sure that they aren't deviating from security best practices.



AWS IoT Device Management

AWS IoT Device Management makes it easy to securely register, organize, monitor, and remotely manage IoT devices at scale.

AWS IOT

Analytics services

Work with IoT data faster to extract value from your IoT data.



AWS IoT Analytics

AWS IoT Analytics makes it easy to run sophisticated analytics on massive volumes of IoT data.



AWS IoT SiteWise

AWS IoT SiteWise makes it easy to collect, organize and analyze industrial data at scale.



AWS IoT Events

AWS IoT Events makes it easy to detect and respond to events from large numbers of IoT sensors and applications.



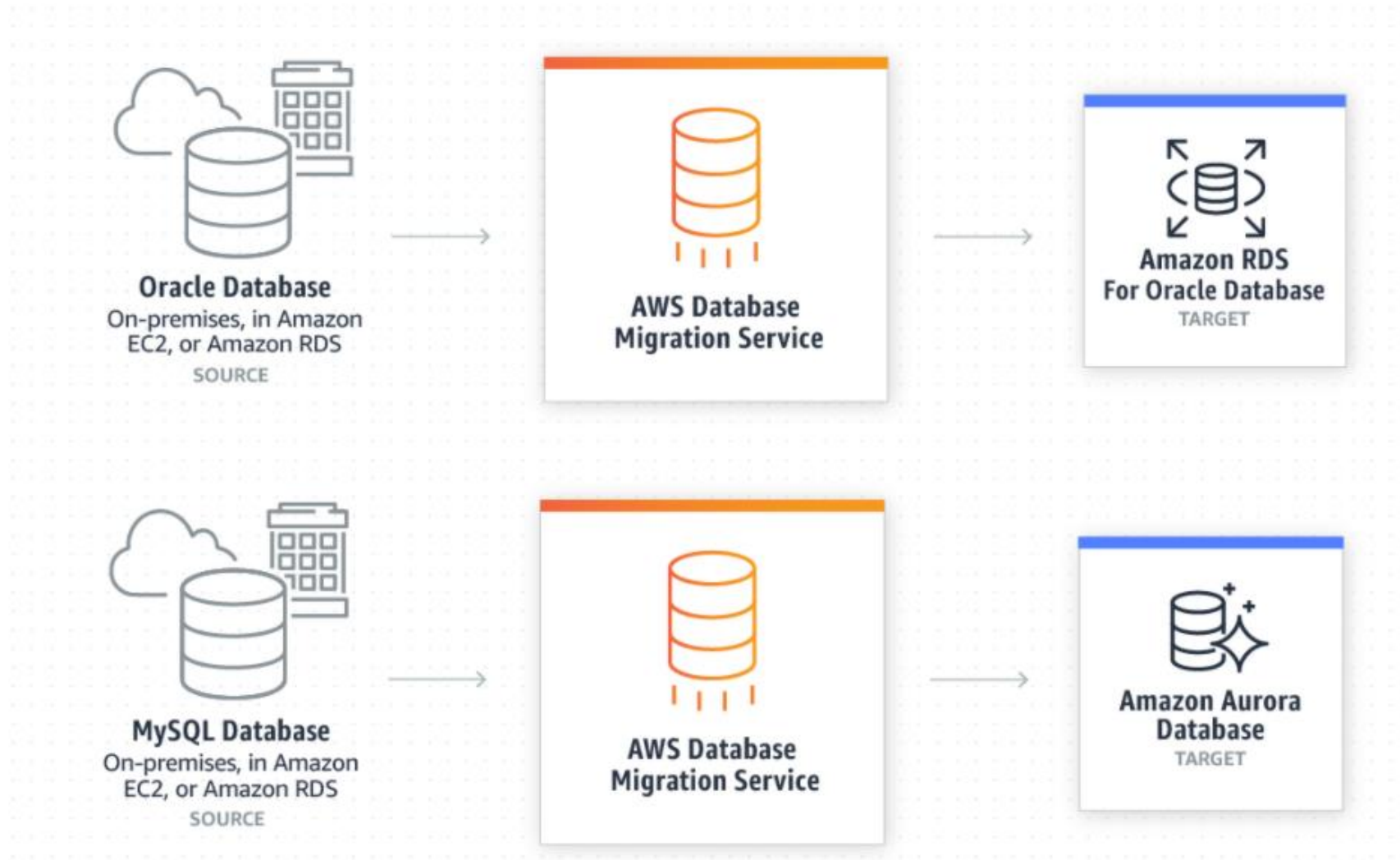
AWS IoT Things Graph

AWS IoT Things Graph makes it easy to connect different devices and cloud services to build IoT applications.

AWS DATA MIGRATION SERVICE

- AWS Database Migration Service helps you 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 most widely used commercial and open-source databases.
- AWS Database Migration Service supports **homogeneous 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.
- When migrating databases to Amazon Aurora, Amazon Redshift, Amazon DynamoDB or Amazon DocumentDB (with MongoDB compatibility) we can use DMS. It is freely available for 6 months.

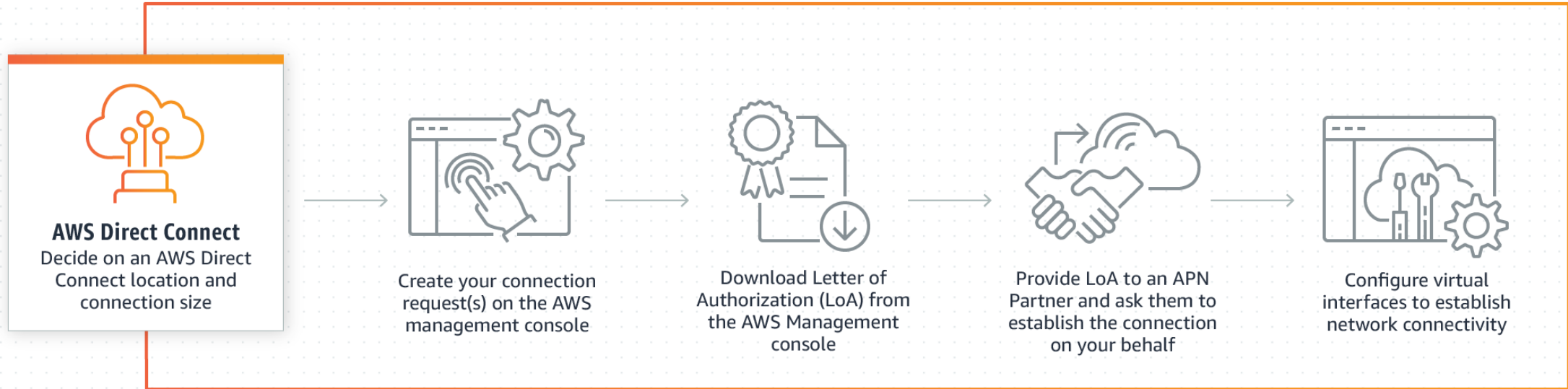
DATABASE MIGRATIONS USE CASE



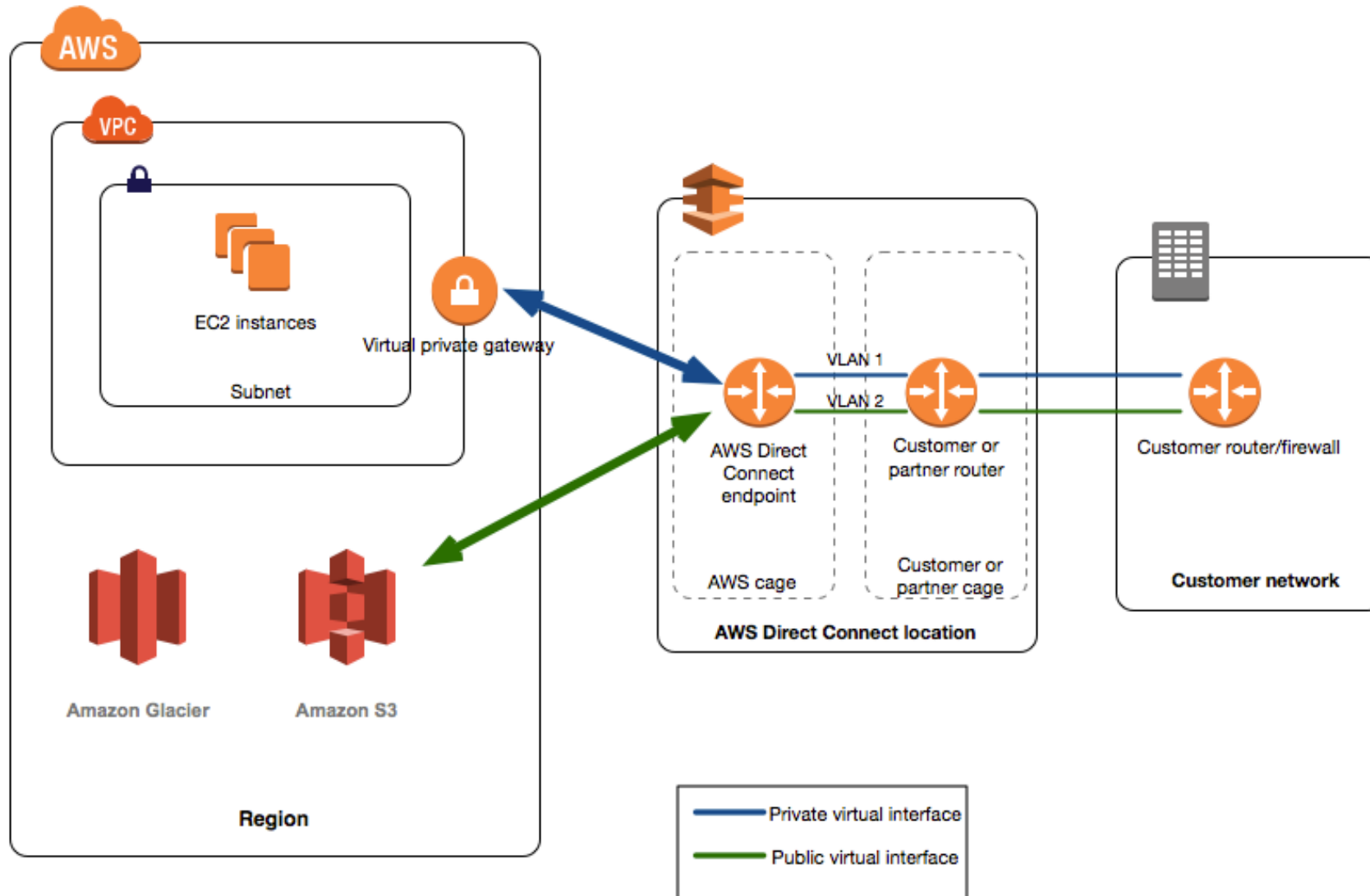
AWS DIRECT CONNECT

- AWS Direct Connect is a cloud service solution that makes it easy **to establish a dedicated network connection from your on-premise to AWS.**
- Using AWS Direct Connect, you **can establish private connectivity between AWS and your datacenter, office, or colocation environment,**
 - can reduce your network costs,
 - increase bandwidth throughput, and
 - provide a more consistent network experience than Internet-based connections.

HOW IT WORKS



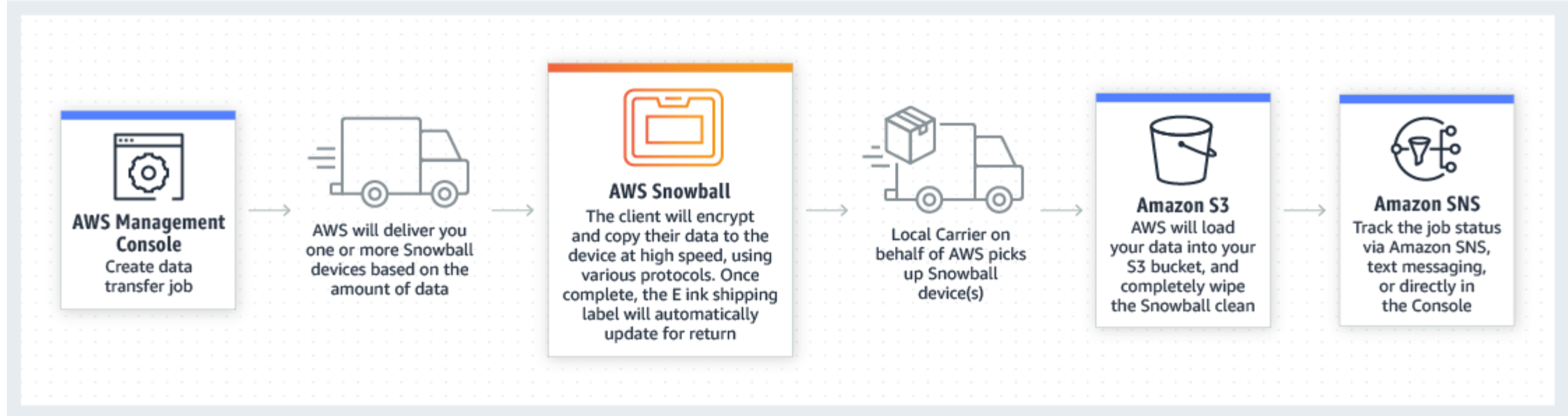
AWS DIRECT CONNECT



AWS SNOWBALL

- **AWS Snowball**, a part of the **AWS Snow Family**, is an edge computing, data migration, and edge storage device that comes in two options.
 - Compute Optimized and
 - Storage Optimized.
- Snowball Edge Storage Optimized devices
 - provide both block storage and Amazon S3-compatible object storage, and 40 vCPUs.
 - Well suited for local storage and large scale-data transfer.
- Snowball Edge Compute Optimized devices
 - provide 52 vCPUs, block and object storage, and an optional GPU for use cases like advanced machine learning and full motion video analysis in disconnected environments.
 - Can be used for data collection, machine learning and processing, and storage in environments with intermittent connectivity (like manufacturing, industrial, and transportation) or in extremely remote locations (like military or maritime operations) before shipping them back to AWS.

AWS SNOWBALL



AWS SNOWMOBILE



AWS Snowmobile moves up to 100 PB of data in a 45-foot long ruggedized shipping container and is ideal for multi-petabyte or Exabyte-scale digital media migrations and data center shutdowns.

QUIZ : QUESTION 1

If Kinesis Firehose experiences data delivery issues to S3, it will retry delivery to S3 for a period of

- A. 3 hours
- B. 7 days**
- C. 7 hours
- D. 24 hours

QUIZ : QUESTION 2

What are the main uses of Kinesis Data streams? Choose 2

- A.** They can carry out real-time reporting and analysis of streaming data
- B.** They can accept data as soon as it is produced, without need for batching
- C.** They can provide long term data storage
- D.** They can undertake loading of streamed data directly into data stores

QUIZ : QUESTION 3

You currently work for a company that is specialised in baggage management. GPS devices installed on all the baggages, deliver the coordinates of the unit every 10 seconds. You need to process these coordinates in real-time from multiple sources. Which tool should you use to process the data?

- A.** Amazon EMR
- B.** Amazon SQS
- C.** AWS Data Pipeline
- D.** Amazon Kinesis

QUIZ : QUESTION 4

Your present log analysis application takes more than four hours to generate a report of the top 10 users of your web application. You have been asked to implement a system that can report this information in real time, ensure that the report is always up to date, and handle increases in the number of requests to your web application. Choose the option that is cost-effective and can fulfill the requirements.

- A.** Publish your data to CloudWatch Logs, and configure your application to Auto Scale to handle the load on demand.
- B.** Publish your log data to an Amazon S3 bucket. Use AWS CloudFormation to create an AutoScaling group to scale your post-processing application which is configured to pull down your log files stored in Amazon S3.
- C.** Post your log data to an Amazon Kinesis data stream, and subscribe your log-processing application so that is configured to process your logging data.
- D.** Configure an Auto Scaling group to increase the size of your Amazon EMR cluster

QUIZ : QUESTION 5

IOT sensors monitor the number of bags that are handled at an airport. The data gets sent back to a Kinesis stream with default settings. Every alternate day, the data from the stream is sent to S3 for processing. But it is noticed that S3 is not receiving all of the data that is being sent to the Kinesis stream. What could be the reason for this?

- A.** The sensors probably stopped working on some days, hence data is not sent to the stream.
- B.** S3 can only store data for a day.
- C.** Data records are only accessible for a default of 24 hours from the time they are added to a stream.
- D.** Kinesis streams are not meant to handle IoT related data.

QUIZ : QUESTION 6

Which service does Kinesis Firehose not load streaming data into ?

- A. Dynamodb
- B. S3
- C. Redshift
- D. Elasticsearch

QUIZ : QUESTION 7

Kinesis Firehose buffers incoming data before delivering the data into your s3 bucket. What are the buffer size ranges?

- A.** 1-128 mb
- B.** 8-64 mb
- C.** 2-128 mb
- D.** 4-256 mb

QUIZ : QUESTION 8

You require the ability to analyze a customer's clickstream data on a website so they can do behavioral analysis. Your customer needs to know what sequence of pages and ads their customer clicked on. This data will be used in real time to modify the page layouts as customers click through the site to increase stickiness and advertising click-through. Which option meets the requirements for capturing and analyzing this data?

- A.** Log clicks in weblogs by URL store to Amazon S3, and then analyze with Elastic MapReduce.
- B.** Push web clicks by session to Amazon Kinesis and analyze behavior using Kinesis workers.
- C.** Write click events directly to Amazon Redshift and then analyze with SQL.
- D.** Publish web clicks by session to an Amazon SQS queue. Then send the events to AWS RDS for further processing.

QUIZ : QUESTION 9

A company has an infrastructure that consists of machines which keep sending log information every 5 minutes. The number of these machines can run into thousands and it is required to ensure that the data can be analyzed at a later stage. Which of the following would help in fulfilling this requirement?

- A.** Use Kinesis Firehose with S3 to take the logs and store them in S3 for further processing.
- B.** Launch an Elastic Beanstalk application to take the processing job of the logs.
- C.** Launch an EC2 instance with enough EBS volumes to consume the logs which can be used for further processing.
- D.** Use CloudTrail to store all the logs which can be analyzed at a later stage.

QUIZ : QUESTION 10

Which of the following statements are True about SQS ?

- A.** Messages can be sent and read simultaneously.
- B.** Messages can be retained in queues for up to 14 days.
- C.** Messages can be retained in queues for up to 7 days.
- D.** A queue can only be created in limited regions, and you should check the SQS website to see which are supported.
- E.** A queue can be created in any region.

QUESTIONS

