AWS Data Engineering Course by Johnny Chivers

https://www.youtube.com/c/JohnnyChivers

What is Data Enginering?

Data Engineering is the process of collecting, analysing and transforming data from numerous sources. Data can be transient, or persisted to a repository.

AWS Data Streaming

AWS Data Engineering

What is AWS Kinesis?

Realtime

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.

AWS Managed Service

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.

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.

Kinesis Data Firehose

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.

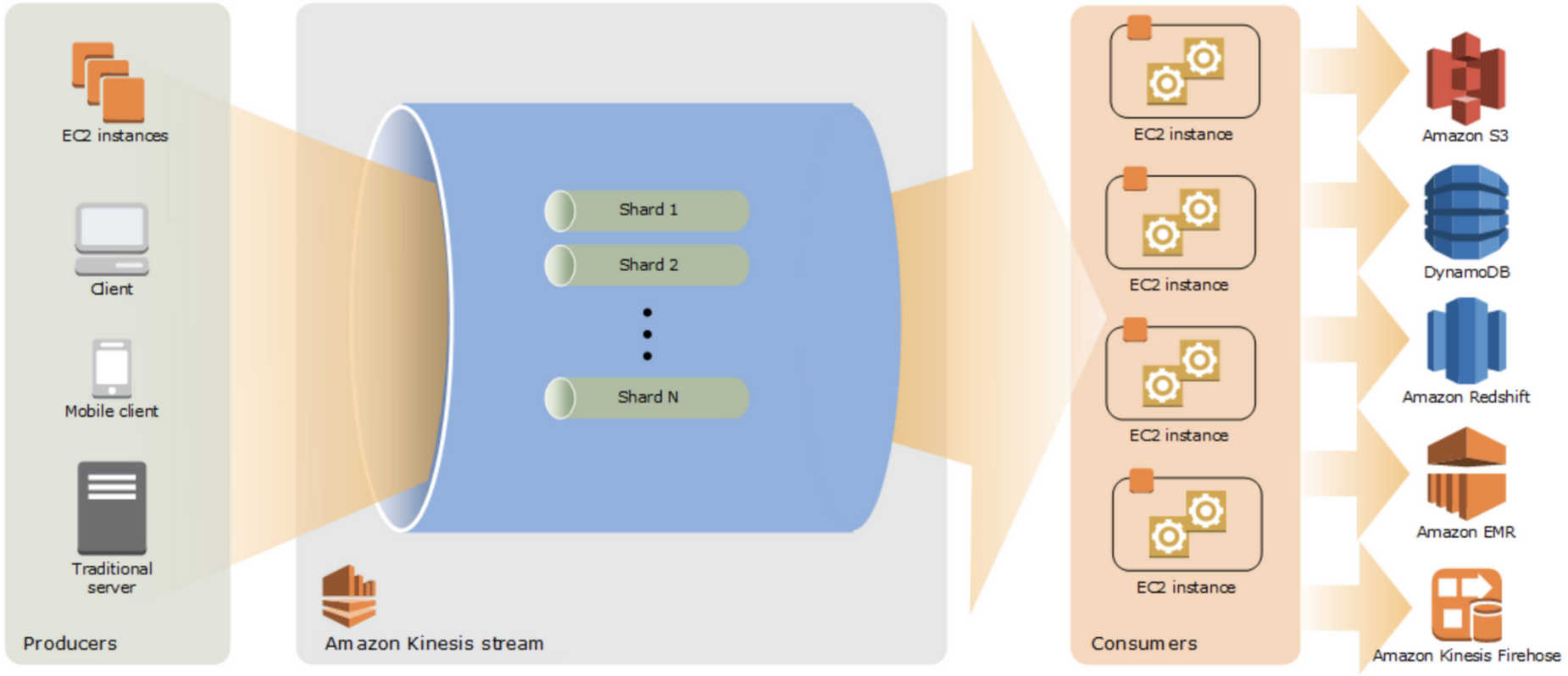
Kinesis Data Streams

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 Analytics

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.

Kinesis Data Streams High-Level Architecture



https://docs.aws.amazon.com/streams/latest/dev/key-concepts.html

Producer

Producers put records into Amazon Kinesis Data Streams. For example, a web server sending log data to a stream is a producer.

Retention Period

The length of time that data records are accessible after they are added to the stream. A stream's retention period is set to a default of 24 hours after creation. You can increase the retention period up to 8760 hours (365 days)

Shard

A shard is a uniquely identified sequence of data records in a stream. A stream is composed of one or more shards, each of which provides a fixed unit of capacity. Each shard can support up to 5 transactions per second for reads, up to a maximum total data read rate of 2 MB per second and up to 1,000 records per second for writes, up to a maximum total data write rate of 1 MB per second (including partition keys). The data capacity of your stream is a function of the number of shards that you specify for the stream. The total capacity of the stream is the sum of the capacities of its shards.
If your data rate increases, you can increase or decrease the number of shards allocated to your stream

Partition Key

A partition key is used to group data by shard within a stream.

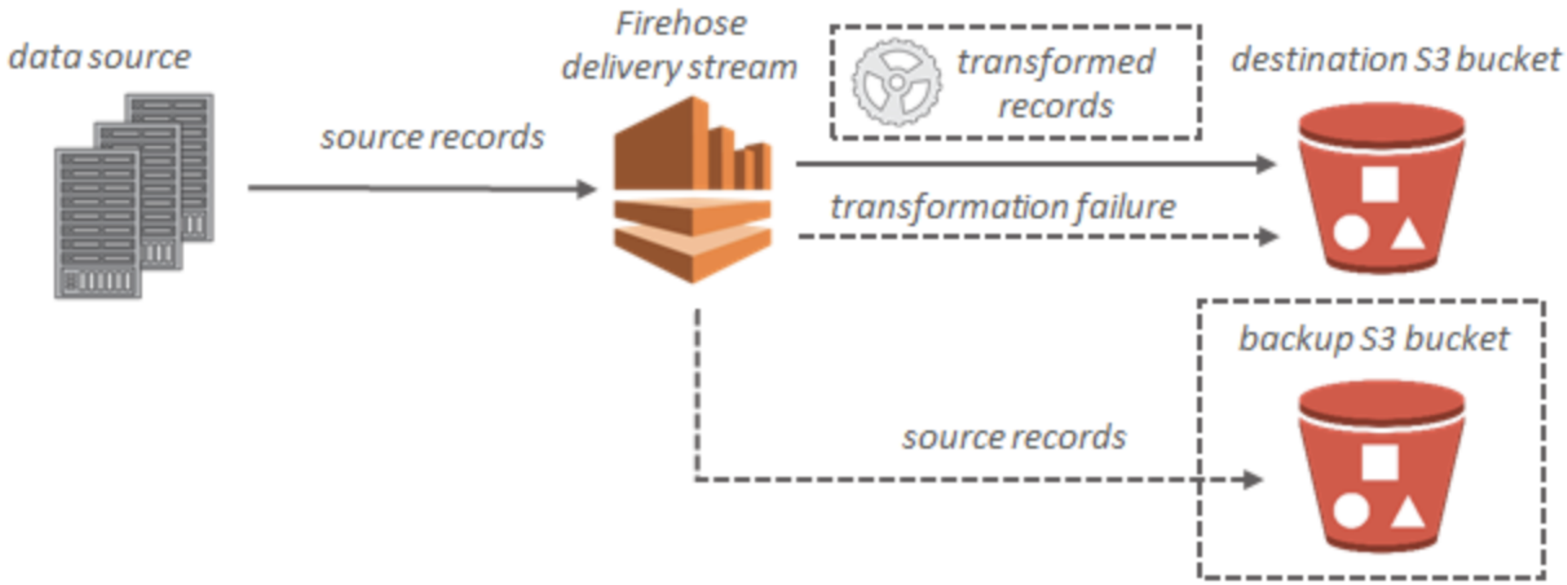
Consumer

Consumers get records from Amazon Kinesis Data Streams and process them. These consumers are known as Amazon Kinesis Data Streams Application.

Sequence Number

Each data record has a sequence number that is unique per partition-key within its shard.

Kinesis Data Firehose High-Level Architecture



https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html

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.

Buffer Size and Buffer Interval

Kinesis Data Firehose buffers incoming streaming data to a certain size or for a certain period of time before delivering it to destinations. Buffer Size is in MBs and Buffer Interval is in seconds.

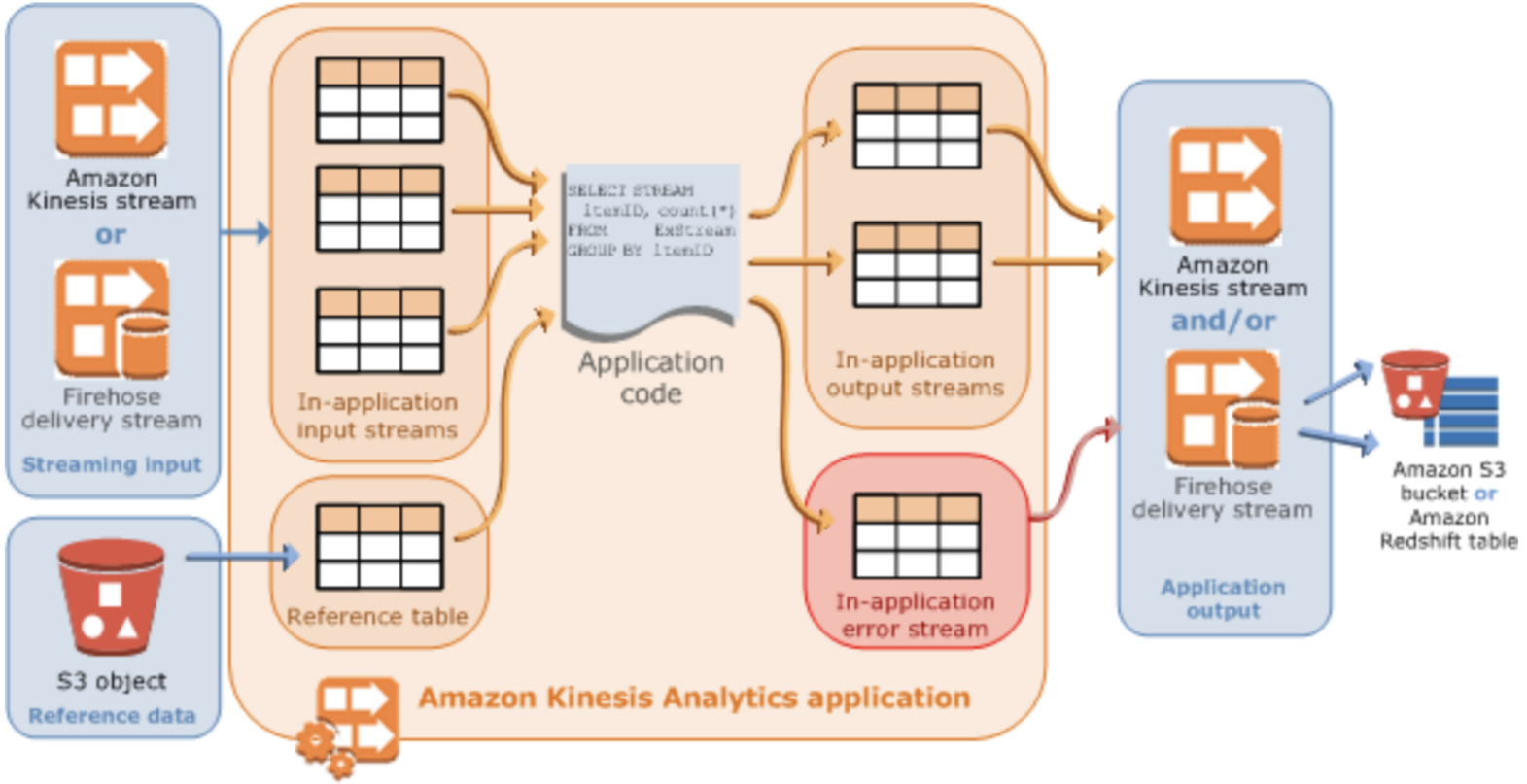
Data Producer

Producers send records to Kinesis Data Firehose delivery streams. For example, a web server that sends log data to a delivery stream is a data producer. You can also configure your Kinesis Data Firehose delivery stream to automatically read data from an existing Kinesis data stream, and load it into destinations. For more information, see Sending Data to an Amazon Kinesis Data Firehose Delivery Stream.

Destinations

- Amazon Simple Storage Service (Amazon S3)
- Amazon Redshift
- Amazon Elasticsearch Service (Amazon ES)
- Splunk
- Datadog
- Dynatrace
- LogicMonitor
- MongoDB
- New Relic
- Sumo Logic

Kinesis Data Analytics



https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works.html

AWS Glue

AWS Data Engineering

What is AWS Glue?

Managed ETL Service

AWS Glue is a fully managed ETL (extract, transform, and load) service that makes it simple and cost-effective to categorize your data, clean it, enrich it, and move it reliably between various data stores and data streams.

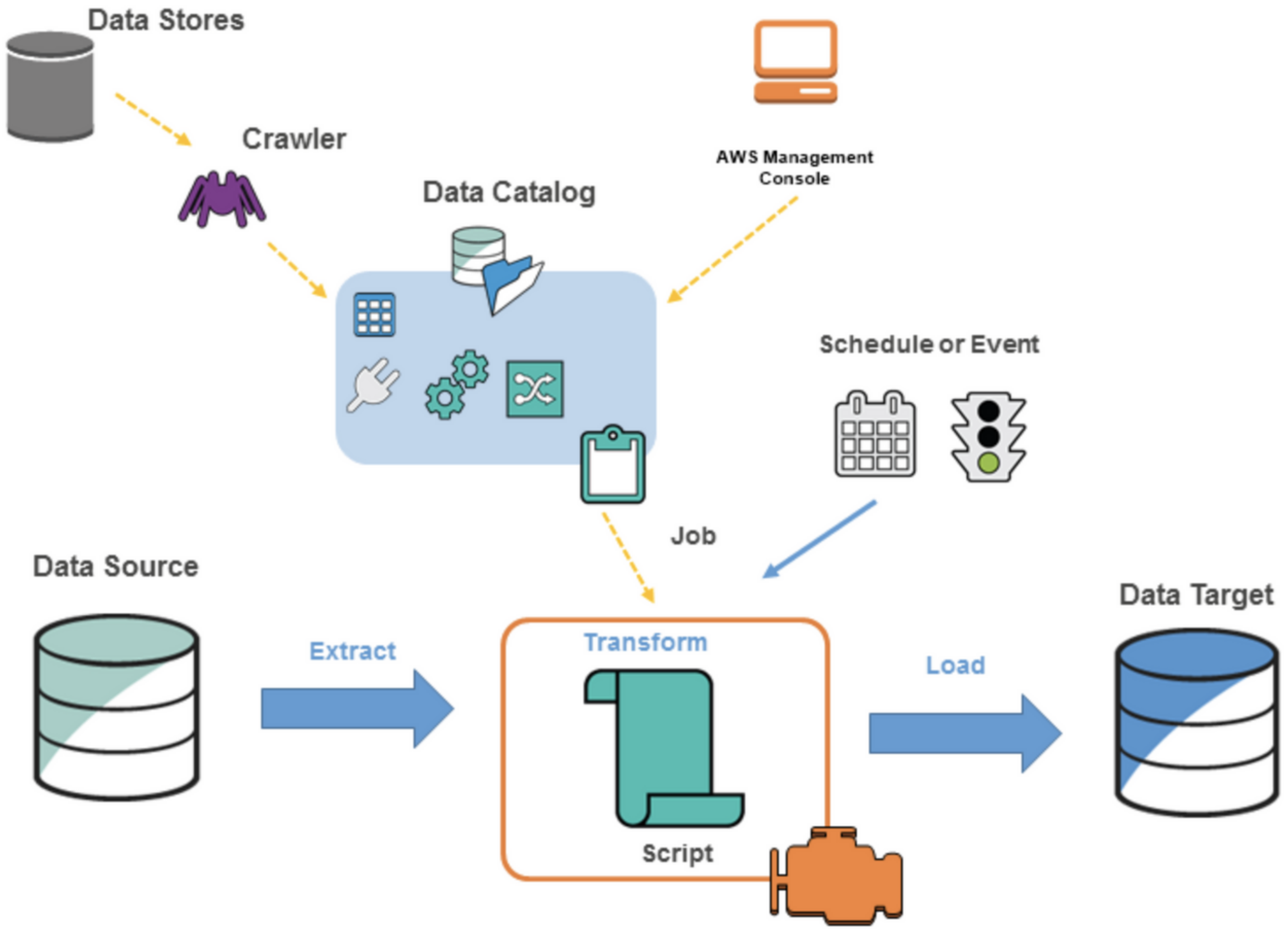
Collection of Components

AWS Glue consists of a central metadata repository known as the AWS Glue Data Catalog, an ETL engine that automatically generates Python or Scala code, and a flexible scheduler that handles dependency resolution, job monitoring, and retries.

Serverless

AWS Glue is serverless, so there's no infrastructure to set up or manage.

AWS Glue Key Concepts



https://docs.aws.amazon.com/glue/latest/dg/components-key-concepts.html

AWS Glue Data Catalog

The persistent metadata store in AWS Glue. It contains table definitions, job definitions, and other control information to manage your AWS Glue environment. Each AWS account has one AWS Glue Data Catalog per region.

Crawler

A program that connects to a data store (source or target), progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in the AWS Glue Data Catalog.

Classifers

Determines the schema of your data. AWS Glue provides classifiers for common file types, such as CSV, JSON, AVRO, XML, and others. Plus Common relational database management systems using a JDBC connection. You can write your own classifier

Data Store

A data store is a repository for persistently storing your data. Examples include Amazon S3 buckets and relational databases. A data source is a data store that is used as input to a process or transform. A data target is a data store that a process or transform writes to.

https://docs.aws.amazon.com/glue/latest/dg/components-overview.html

AWS Glue Data Catalog

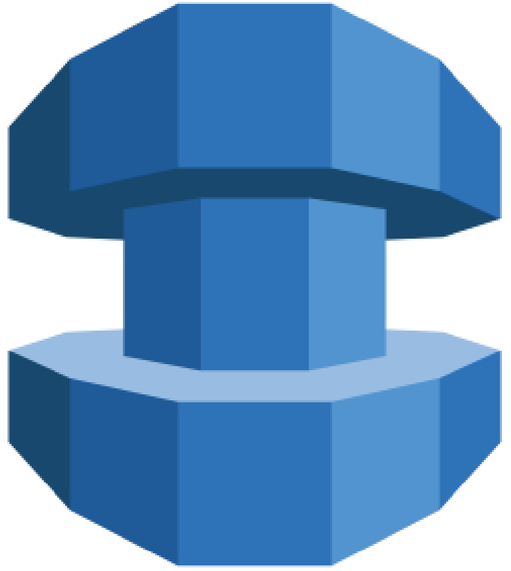
The AWS Glue Data Catalog is your persistent metadata store. It is a managed service that lets you store, annotate, and share metadata in the AWS Cloud in the same way you would in an Apache Hive metastore.

The Data Catalog also provides comprehensive audit and governance capabilities, with schema change tracking and data access controls. You can audit changes to data schemas. This helps ensure that data is not inappropriately modified or inadvertently shared.

Each AWS account has one AWS Glue Data Catalog per AWS region.

AWS Glue Data Catalog consists of a hierarchy of databases and tables.Tables are the metadata definition that represents your data and databases are logically grouped tables.

AWS Database Migration Service (DMS)

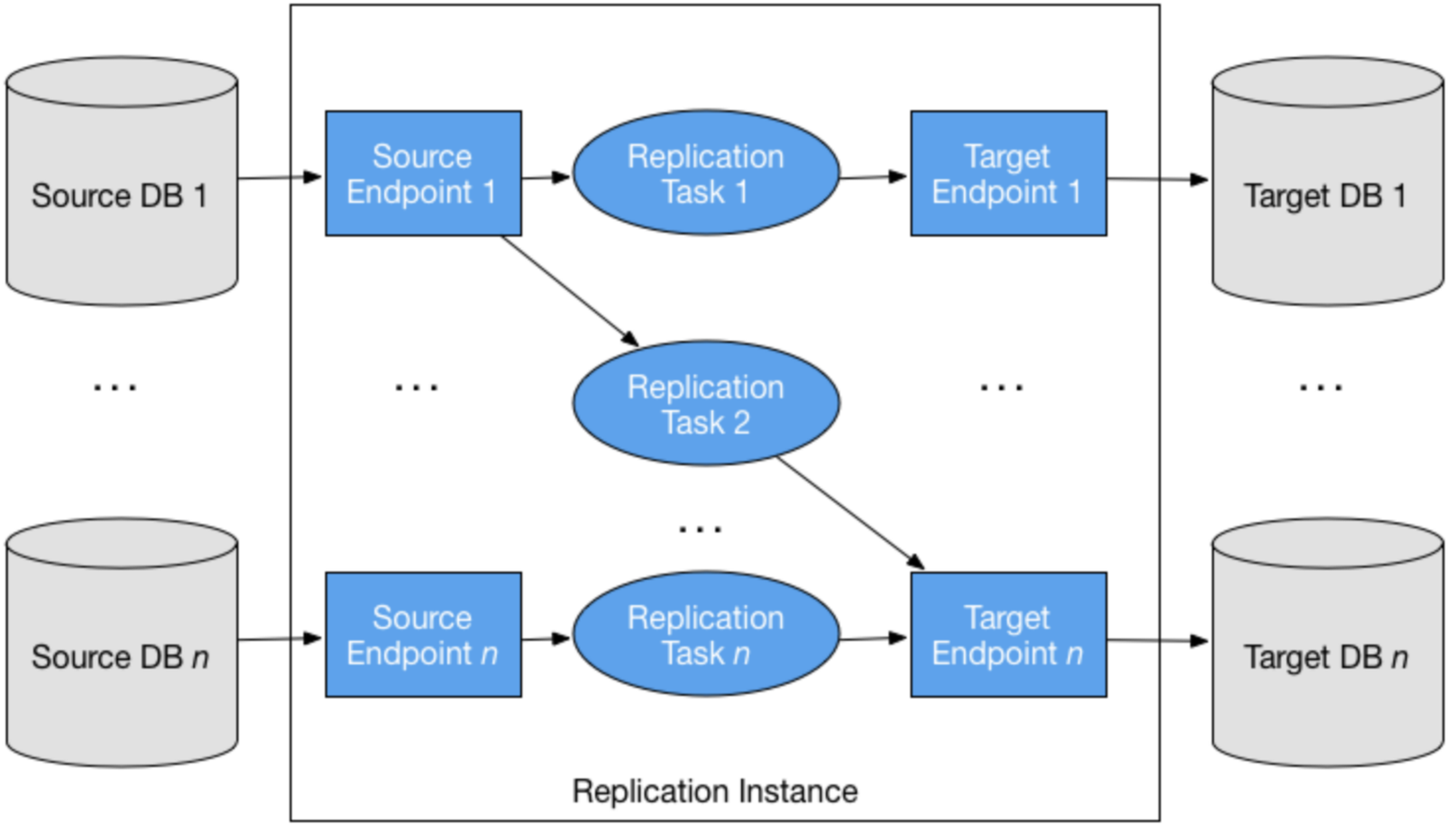


AWS Data Engineering

What is AWS DMS?

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.

https://aws.amazon.com/dms/



https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Introduction.Components.html

Replication instance

At a high level, an AWS DMS replication instance is simply a managed Amazon Elastic Compute Cloud (Amazon EC2) instance that hosts one or more replication tasks.

Replication tasks

You use an AWS DMS replication task to move a set of data from the source endpoint to the target endpoint. Creating a replication task is the last step you need to take before you start a migration.

Endpoint

AWS DMS uses an endpoint to access your source or target data store. The specific connection information is different, depending on your data store, but in general you supply the following information when you create an endpoint:

Schema and Code Migration

AWS DMS doesn't perform schema or code conversion

https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Introduction.Components.html

Sources for AWS DMS

- Oracle
- Microsoft SQL Server
- MySQL
- MariaDB
- PostgreSQL
- MongoDB
- SAP Adaptive Server Enterprise (ASE)
- Microsoft Azure
- Amazon RDS instance databases, and Amazon Simple Storage Service

https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Introduction.Sources.html