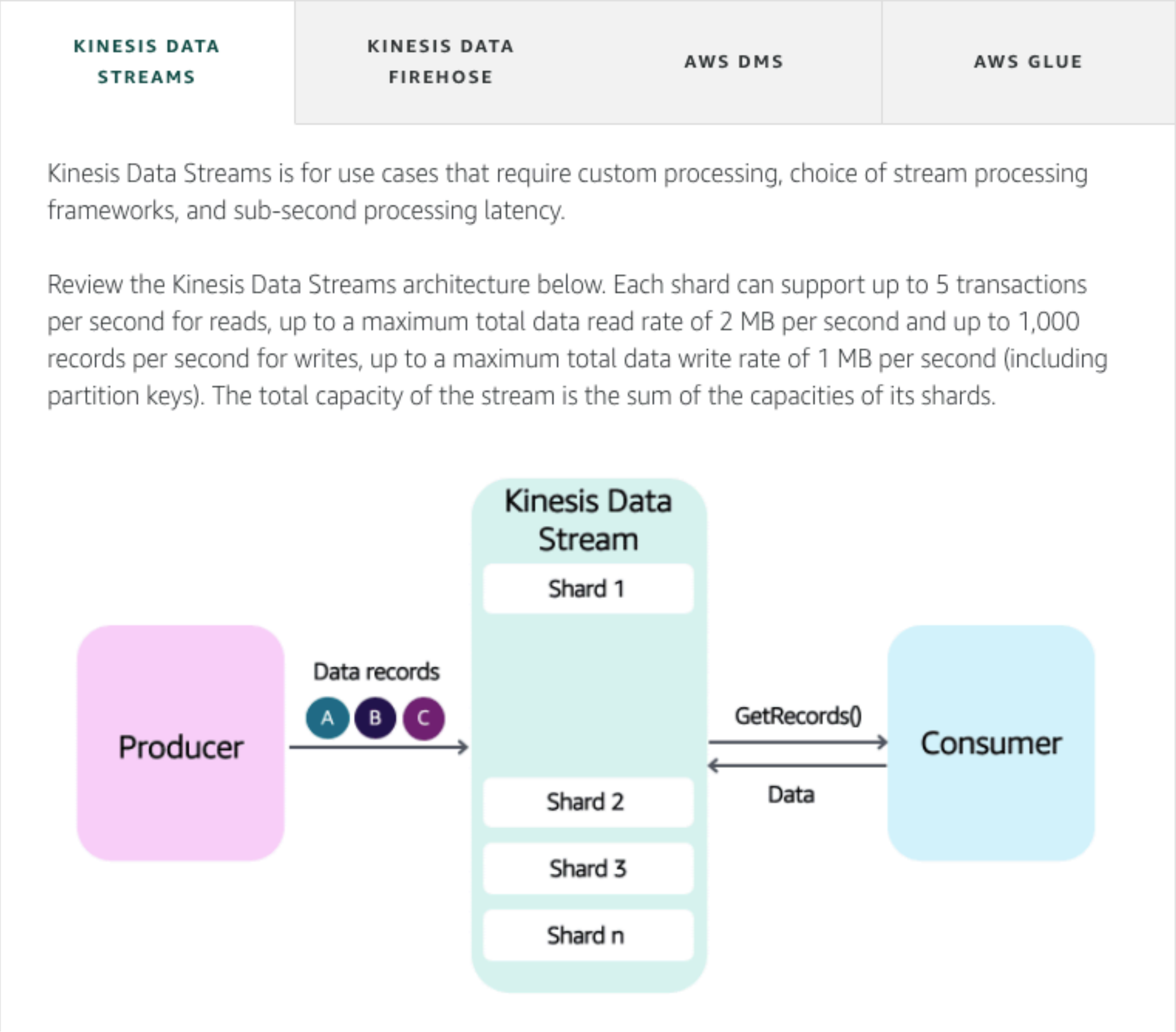
AWS DAS-C01 readiness

Sunday, 23 May 2021 11:42 AM

Domain 1: data collection

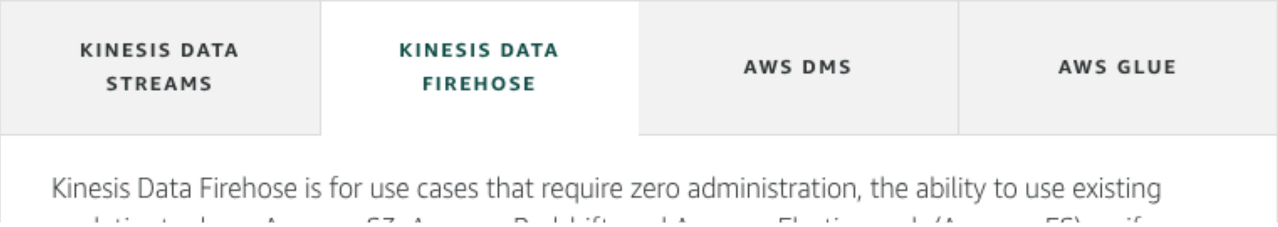
Ingestion method differences:

- Scalability
- Fault tolerance
- Cost



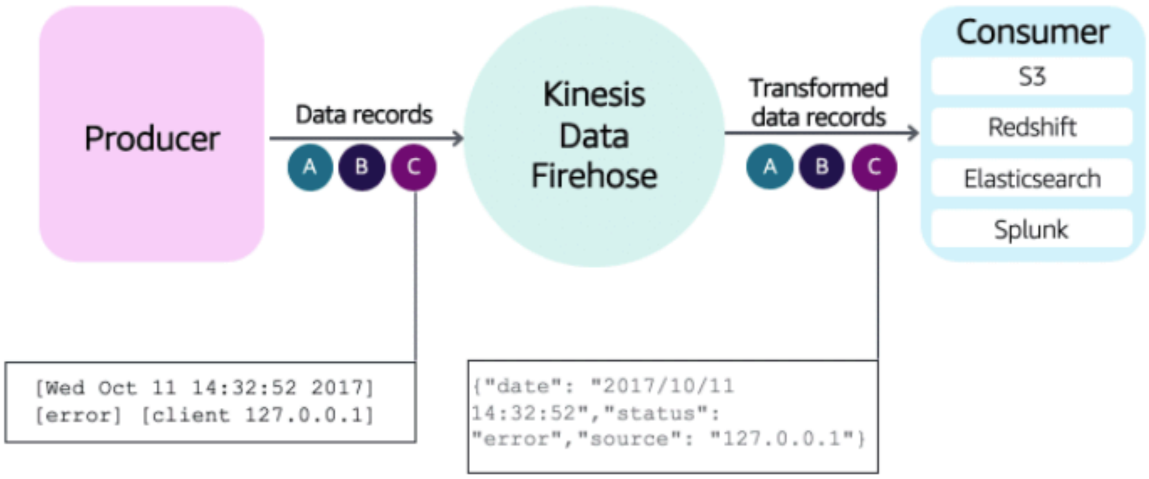
Kinesis Data Streams is for use cases that require custom processing, choice of stream processing frameworks, and sub-second processing latency.

Review the Kinesis Data Streams architecture below. 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 total capacity of the stream is the sum of the capacities of its shards.



Kinesis Data Firehose is for use cases that require zero administration, the ability to use existing analytics tools on Amazon S3, Amazon Redshift, and Amazon Elasticsearch (Amazon ES), or if you require a data latency of 60 seconds or higher.

You may want to take a look at the service architecture as well. Shown in the diagram below, you configure your data producers to send data to Kinesis Data Firehose, and it automatically delivers the data to the destination that you specified. You can also configure Kinesis Data Firehose to transform your data before delivering it.



```
[Wed Oct 11 14:32:52 2017]
[error] [client 127.0.0.1]
```

```
{"date": "2017/10/11
14:32:52","status":
"error","source": "127.0.0.1"}
```

KINESIS DATA STREAMS

KINESIS DATA FIREHOSE

AWS DMS

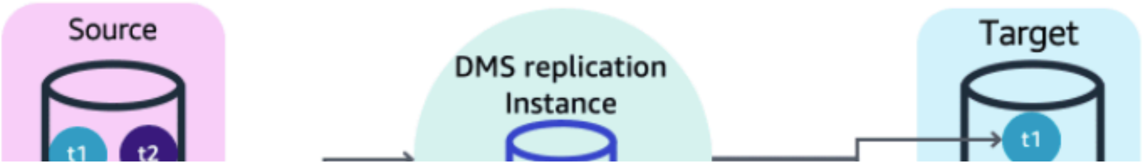
AWS GLUE

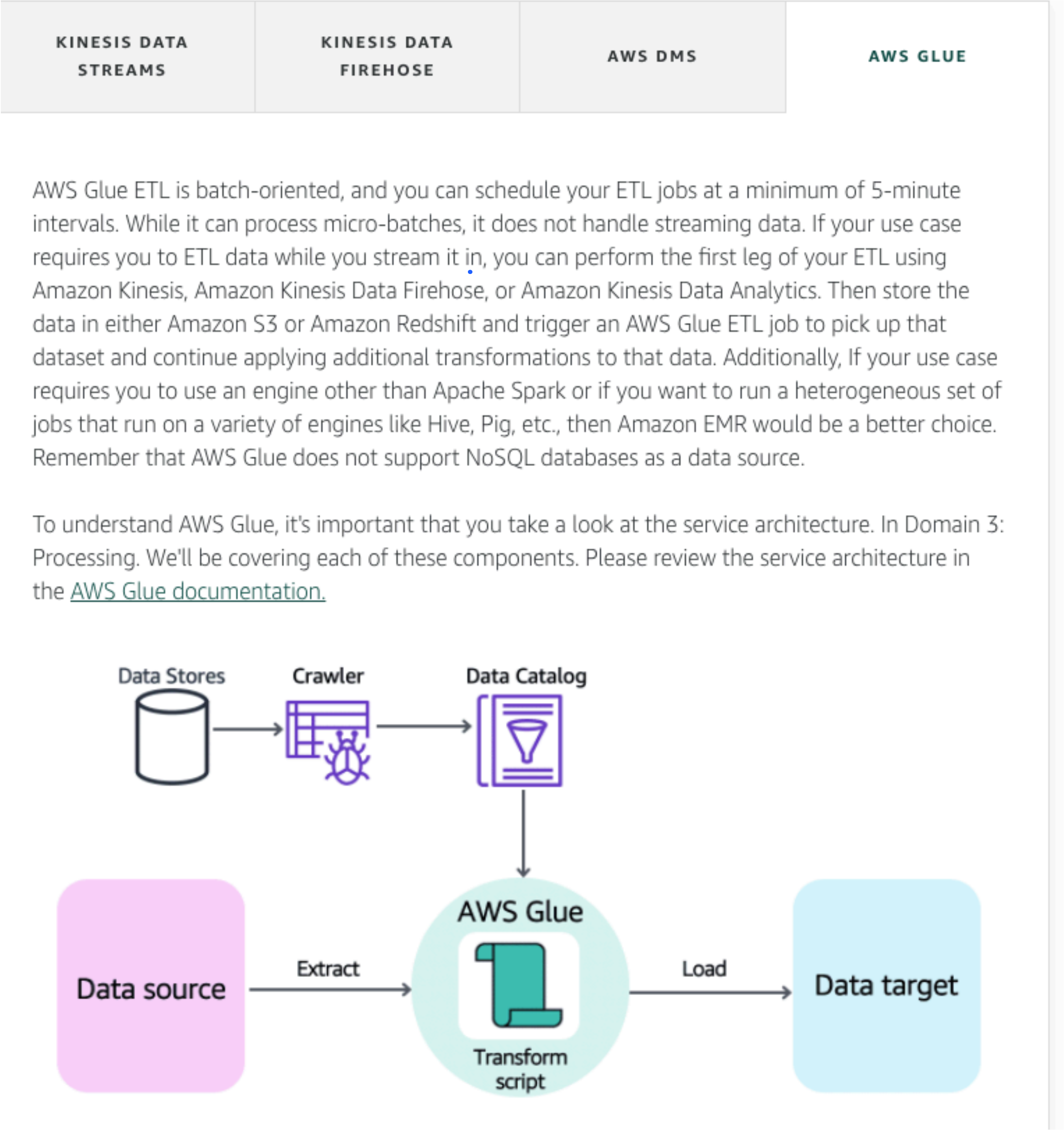
AWS DMS provides one-time migration and continuous replication of your database records and data structures to AWS. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. AWS DMS supports the most popular data engines as a source and target for data replication.

With AWS DMS, you can create a task that captures ongoing changes after you complete your initial migration to a supported target data store. This process is called *ongoing replication* or *change data capture* (CDC).

Here's how it works (as shown in the diagram):

1. A database has a transaction log. Changes to the database (such as changes t1 and t2) are written to that log.
2. AWS uses the native APIs of the database engine to read the changes from that log and replicate them to the target database.





AWS Glue ETL is batch-oriented, and you can schedule your ETL jobs at a minimum of 5-minute intervals. While it can process micro-batches, it does not handle streaming data. If your use case requires you to ETL data while you stream it in, you can perform the first leg of your ETL using Amazon Kinesis, Amazon Kinesis Data Firehose, or Amazon Kinesis Data Analytics. Then store the data in either Amazon S3 or Amazon Redshift and trigger an AWS Glue ETL job to pick up that dataset and continue applying additional transformations to that data. Additionally, If your use case requires you to use an engine other than Apache Spark or if you want to run a heterogeneous set of jobs that run on a variety of engines like Hive, Pig, etc., then Amazon EMR would be a better choice. Remember that AWS Glue does not support NoSQL databases as a data source.

To understand AWS Glue, it's important that you take a look at the service architecture. In Domain 3: Processing. We'll be covering each of these components. Please review the service architecture in the AWS Glue documentation.

Kinesis Data Streams:

Each shard can support up to **1,000 PUT** records per second. However, you can increase the number of shards limitlessly. One shard provides a capacity of **1 MB/sec data input** and **2 MB/sec data output.**

Kinesis Data Streams synchronously replicates data across three Availability zones

Kinesis Data Streams synchronously replicates data across three Availability Zones, providing high availability and data durability.

Kinesis Data Firehose:
Kinesis Data Firehose will automatically scale to match the throughput of your data, without any manual intervention or developer overhead.
Amazon Kinesis Data Firehose synchronously replicates data across three Availability Zones, providing high availability and data durability.

AWS DMS:
AWS DMS uses Amazon EC2 instances as the replication instance. You can scale up or down your replication instance, depending on utilization.
You have the option of enabling Multi-AZ which provides a replication stream that is fault-tolerant through redundant replication servers.

AWS Glue:
AWS Glue uses a scale-out Apache Spark environment to load your data into its destination. To scale out, you specify the number of DPUs (data processing units) that you want to allocate to your ETL jobs.
AWS Glue connects to the data source of your preference, whether it is in an Amazon S3 file, Amazon RDS table, or another set of data. As a result, all your data is stored and available as it pertains to that data stores durability characteristics. AWS Glue also provides default retry behavior that will retry all failures **three times** before sending out an error notification. To be informed of job failures or completions, you can set up Amazon **SNS notifications** via **Amazon CloudWatch** actions.

Issues with incoming data:

- Out of order data
- Duplicated data

How to fix:

- Delivery functionality(de-duping)
- Guaranteed ordering

Ingestion tools categorizes data delivery using **three** phrases:

- At least once: can duplicate
- At most once: can lost
- Exactly once: no message loss, no duplicate

Attribute	Amazon DynamoDB Streams	Amazon Kinesis Data Streams	Amazon Kinesis Data Firehose	Amazon SQS (Standard)	Amazon SQS (FIFO)	Apache Kafka/ Amazon MSK
Guaranteed ordering	✓	✓	✗	✗	✓	✓
Delivery (deduping)	Exactly-once	At-least-once	At-least-once	At-least-once	Exactly-once	At-least-once

All services listed, except for Kinesis Data Firehose and Amazon SQS (Standard), support guaranteed delivery. The delivery mechanism is exactly-once for Amazon SQS (FIFO) and Amazon DynamoDB Streams, and at-least-once for all the other services listed.

	Amazon DynamoDB Streams	Amazon Kinesis Data Streams	Amazon Kinesis Data Firehose	Amazon SQS Standard	Amazon SQS FIFO	Apache Kafka
Row/object size	400 KB	1 MB	Destination row/object size	256 KB	256 KB	Varies
Parallel Consumption	✓	✓	✗	✗	✗	✓
Stream MapReduce	✓	✓	N/A	N/A	N/A	✓