KINESIS

DATA COLLECTION INTRODUCTION

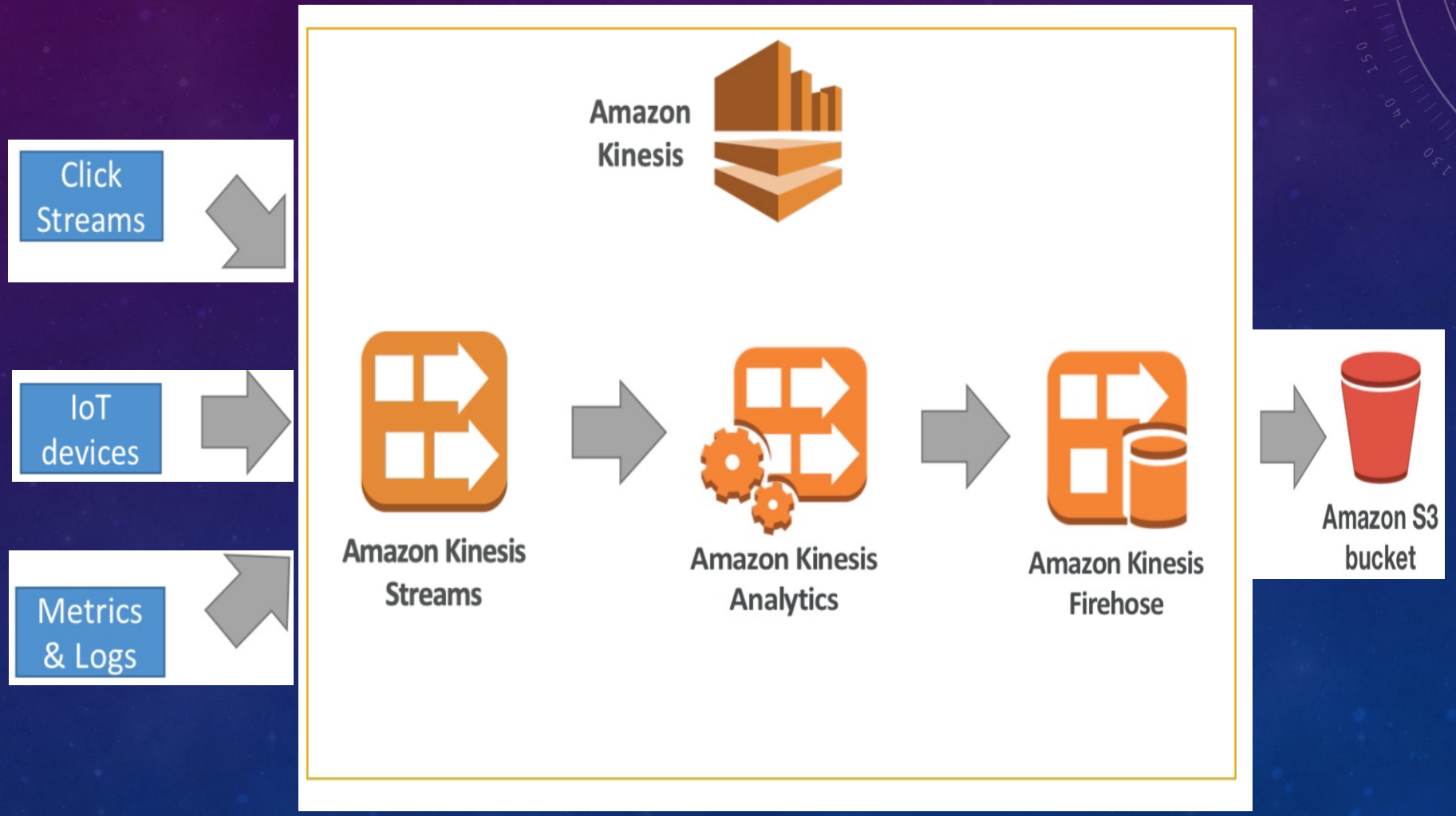
- Real Time - Immediate actions
  - Kinesis Data Streams (KDS)
  - Simple Queue Service (SQS)
  - Internet of Things (IoT)
- Near-real time - Reactive actions
  - Kinesis Data Firehose (KDF)
  - Database Migration Service (DMS)
- Batch - Historical Analysis
  - Snowball
  - Data Pipeline

AWS KINESIS

AWS KINESIS OVERVIEW

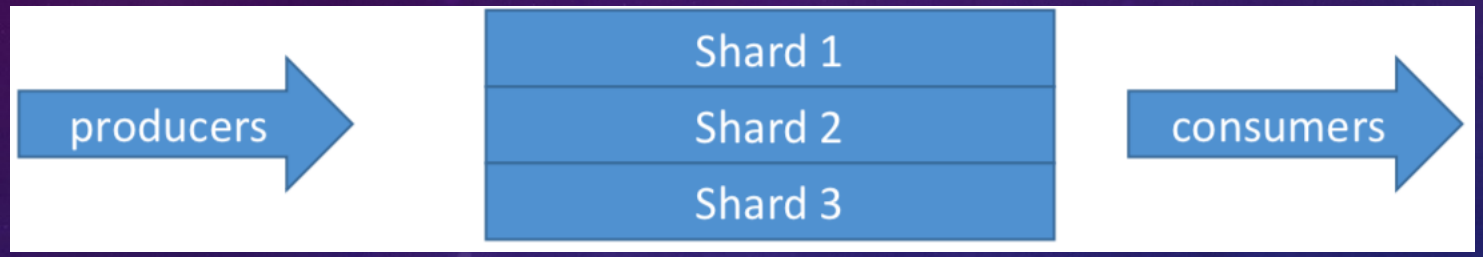
- Kinesis is a managed alternative to Apache Kafka
- Great for application logs, metrics, IoT, clickstreams
- Great for “real-time” big data
- Great for streaming processing frameworks (Spark, NiFi, etc...)
- Data is automatically replicated to 3 AZ
- Kinesis Components
  - Kinesis Streams: low latency streaming ingest at scale
  - Kinesis Analytics: perform real-time analytics on streams using SQL
  - Kinesis Firehose: load streams into S3, Redshift, ElasticSearch …

AWS KINESIS EXAMPLE



AWS KINESIS OVERVIEW

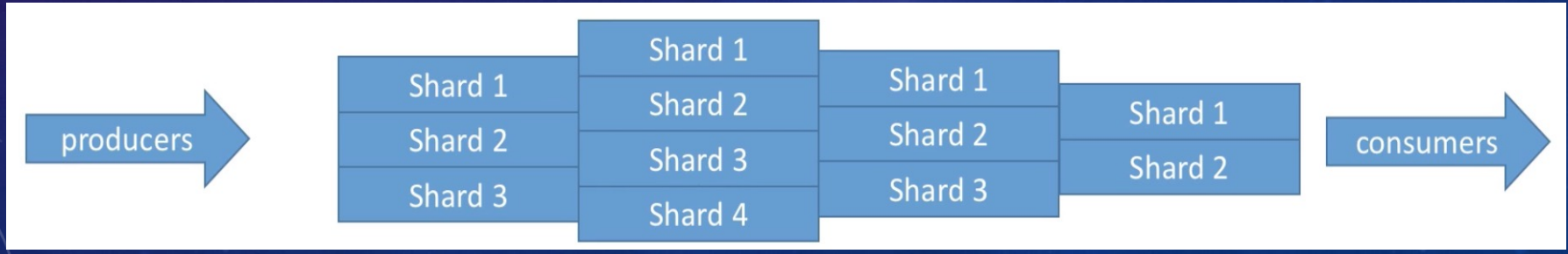
Streams are divided in ordered Shards / Partitions



- Data retention is 1 day by default, can go up to 7 days
- Ability to reprocess / replay data
- Multiple applications can consume the same stream
- Real-time processing with scale of throughput
- Once data is inserted in Kinesis, it can’t be deleted (immutability)

AWS KINESIS STREAMS SHARDS

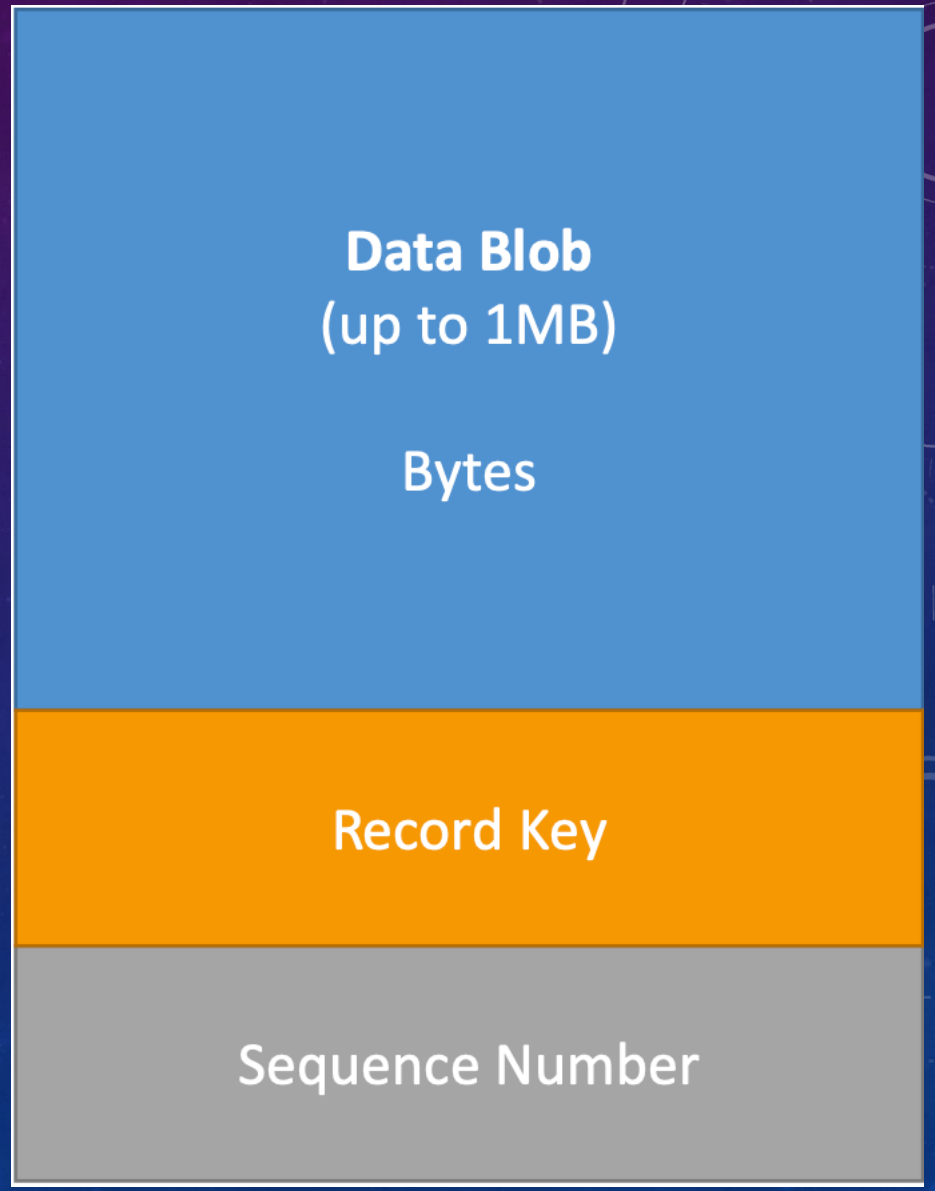
- One stream is made of many different shards
- Billing is per shard provisioned, can have as many shards as you want
- Batching available or per message calls.
- The number of shards can evolve over time (reshard / merge)
- Records are ordered per shard



AWS KINESIS STREAMS - SHARDS

- AWS Kinesis Streams Records
  - Data Blob: data being sent, serialized as bytes. Up to 1 MB. Can represent anything
  - Record Key:

    sent alongside a record, helps to group records in Shards. Same key = Same shard.
    Use a highly distributed key to avoid the “hot partition” problem
  - Sequence number: Unique identifier for each records put in shards.
    Added by Kinesis after ingestion



KINESIS AGENT

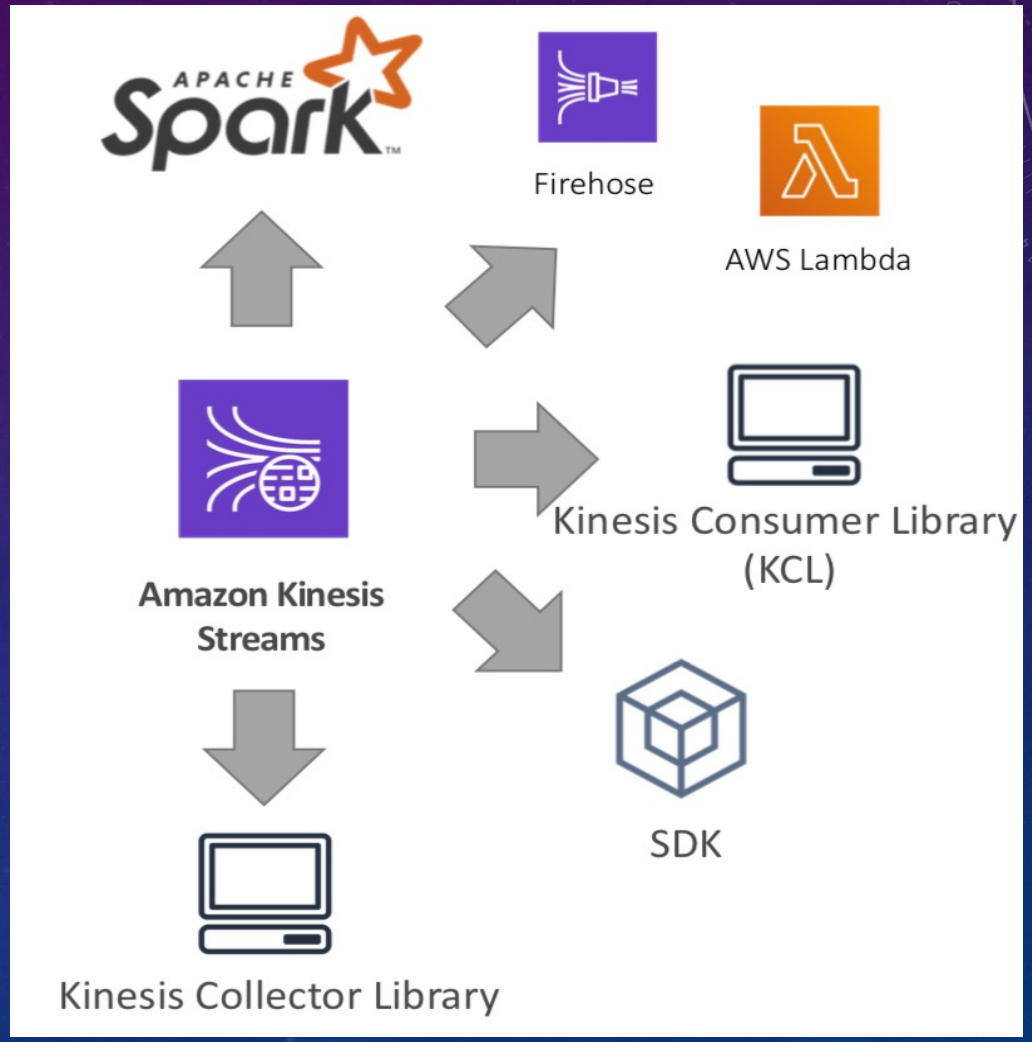
- Monitor Log files and sends them to Kinesis Data Streams
- Java-based agent, built on top of KPL
- Install in Linux-based server environments

Features:

- Write from multiple directories and write to multiple streams
- Routing feature based on directory / log file
- Pre-process data before sending to streams (single line, csv to json, log to json...)
- The agent handles file rotation, checkpointing, and retry upon failures
- Emits metrics to CloudWatch for monitoring

AWS KINESIS CONSUMERS

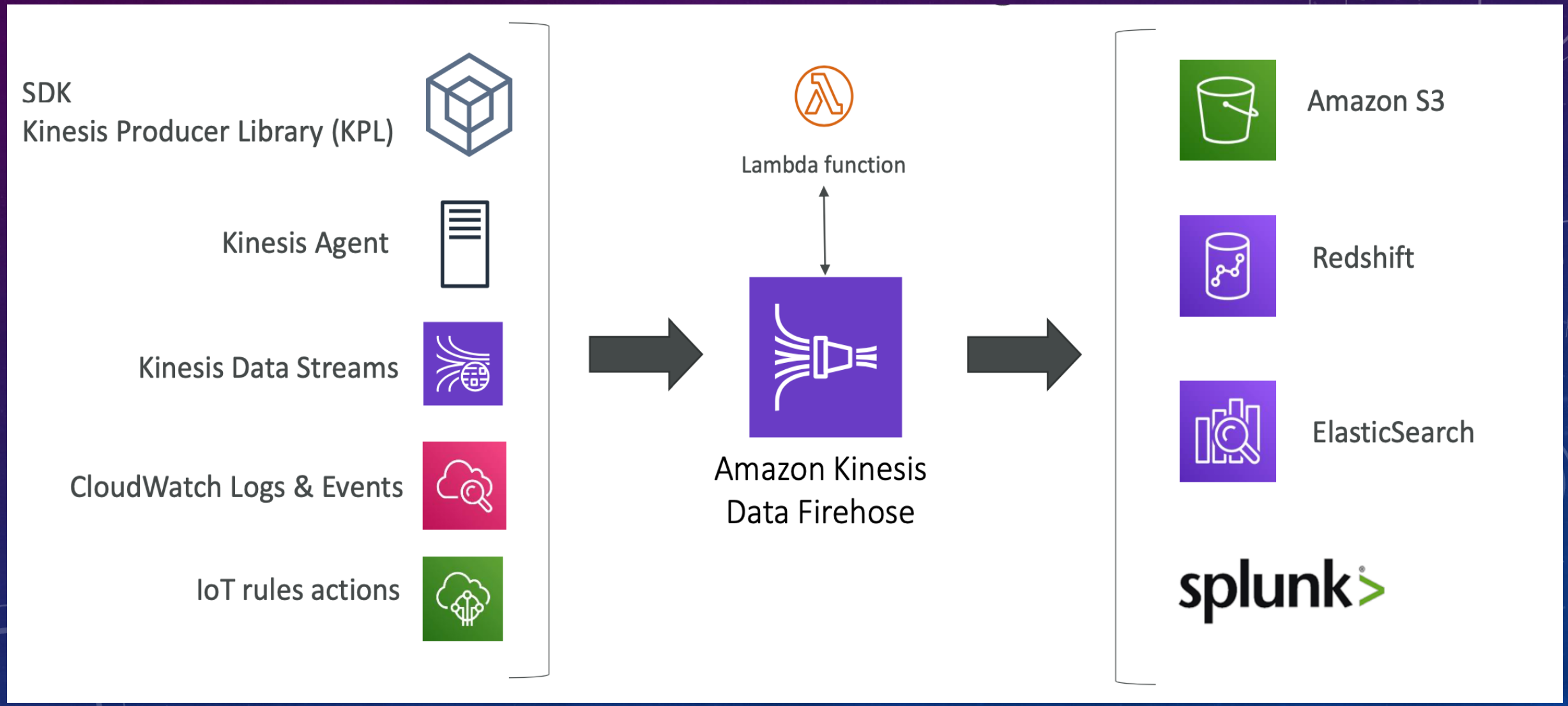
- Kinesis SDK
- Kinesis Client Library (KCL)
- Kinesis Connector Library
- Kinesis Firehose
- AWS Lambda
- 3rd party libraries: Spark, Log4J Appenders, Flume, Kafka Connect...
- Kinesis Consumer Enhanced Fan



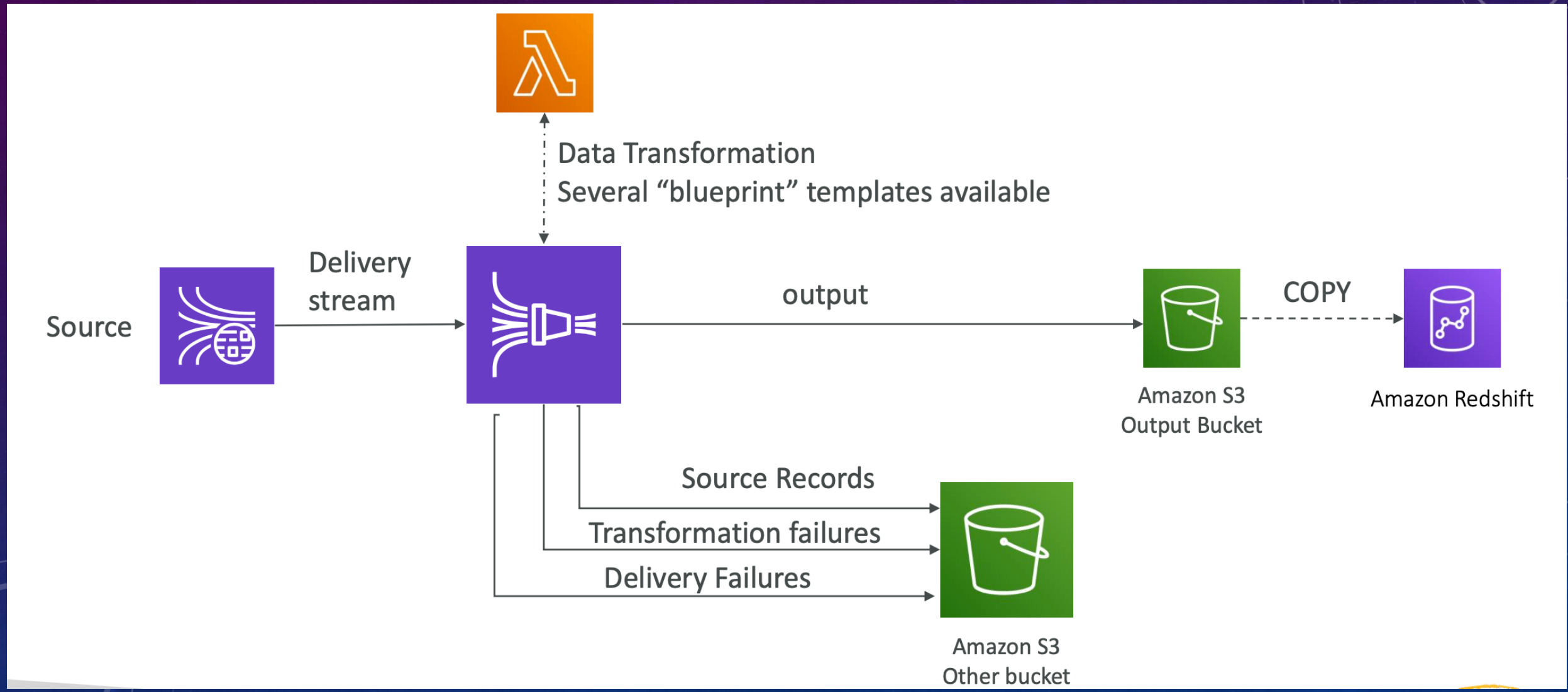
AWS KINESIS DATA FIREHOSE

- Fully Managed Service, no administration
- Near Real Time (60 seconds latency minimum for non full batches)
- Load data into Redshift / Amazon S3 / ElasticSearch / Splunk
- Automatic scaling
- Supports many data formats
- Data Conversions from JSON to Parquet / ORC (only for S3)
- Data Transformation through AWS Lambda (ex: CSV => JSON)
- Supports compression when target is Amazon S3 (GZIP, ZIP, and SNAPPY)
- Only GZIP is the data is further loaded into Redshift
- Spark / KCL do *not* read from KDF
- Pay for the amount of data going through Firehose

AWS KINESIS DATA FIREHOSE DIAGRAM



KINESIS DATA FIREHOSE DELIVERY DIAGRAM



FIREHOSE BUFFER SIZING

- Firehose accumulates records in a buffer
- The buffer is flushed based on time and size rules

- Buffer Size (ex: 32MB): if that buffer size is reached, it's flushed
- Buffer Time (ex: 2 minutes): if that time is reached, it's flushed
- Firehose can automatically increase the buffer size to increase throughput

- High throughput => Buffer Size will be hit
- Low throughput => Buffer Time will be hit