AWS Introduction and Amazon Data Analytics Pipeline

Kapil Dwivedi

Sr Technical Account Manager

Agenda

- AWS 101
- Data Analytics Solution
  - Collection
  - Storage
  - Processing
  - Analysis
  - Visualization

AWS 101

Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud platform, offering over 200 fully featured services from data centers globally. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.

Key facts

- One in three sites you visit on the internet uses AWS services.
- In 2019, Amazon Web Services raked in more than $35 billion in revenue. If AWS were its own company, that would be enough to rank 359th on fortune magazine's Global 500 list

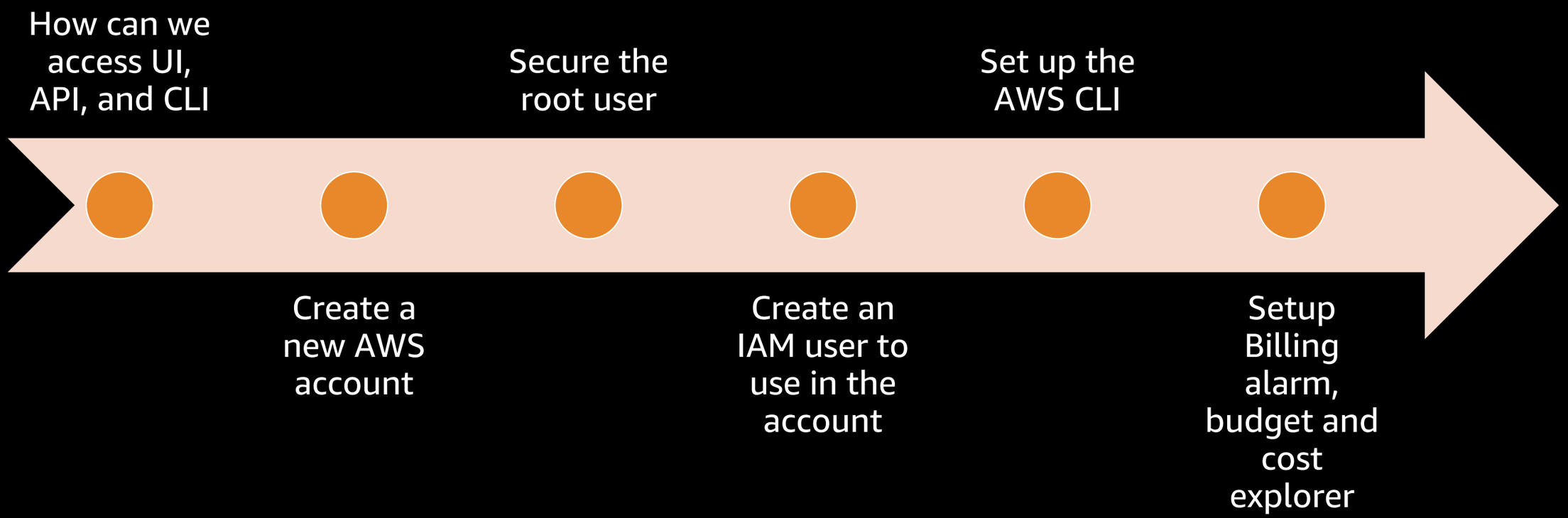
AWS 101 – Global Infrastructure

The Most Secure, Extensive, and Reliable Global Cloud Infrastructure, for all your applications

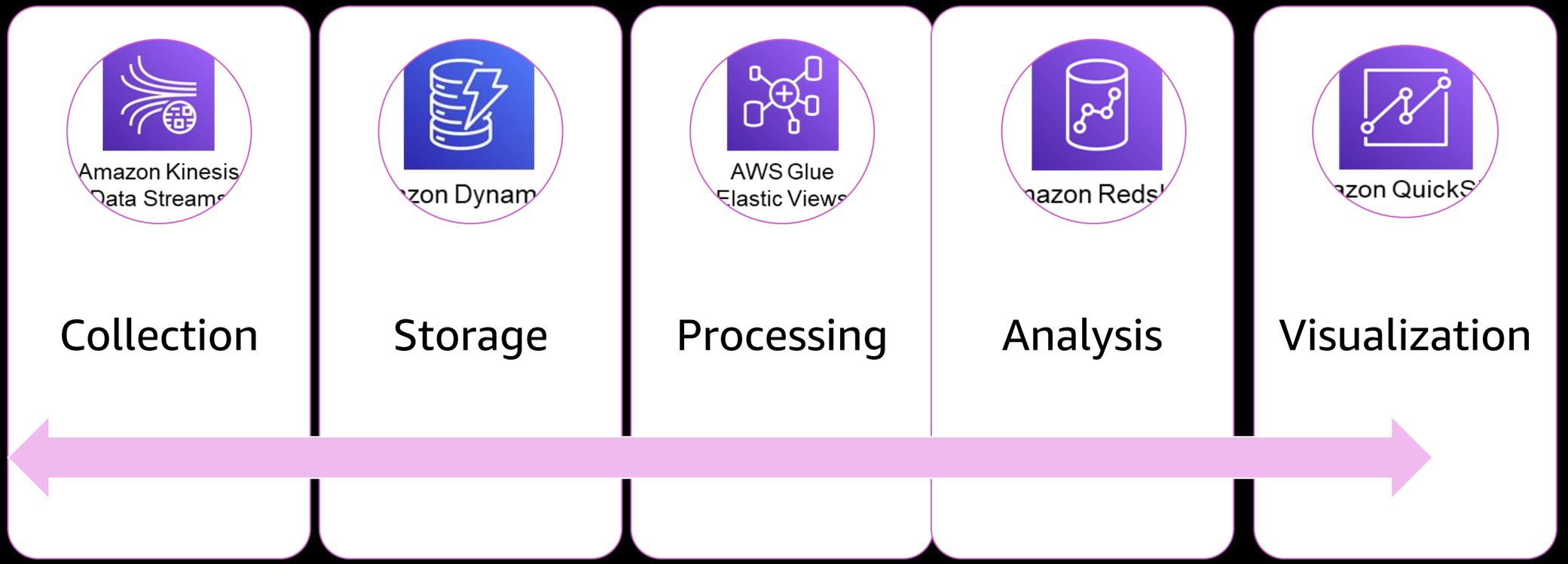
The AWS Cloud spans 81 Availability Zones within 25 geographic regions around the world.

Additionally, there are over 200 edge locations scattered around the world as part of AWS's content delivery network (CDN).

AWS 101


How can we access UI, API, and CLI
Create a new AWS account
Secure the root user
Create an IAM user to use in the account
Set up the AWS CLI
Setup Billing alarm, budget and cost explorer

Amazon Data Analytics Pipeline

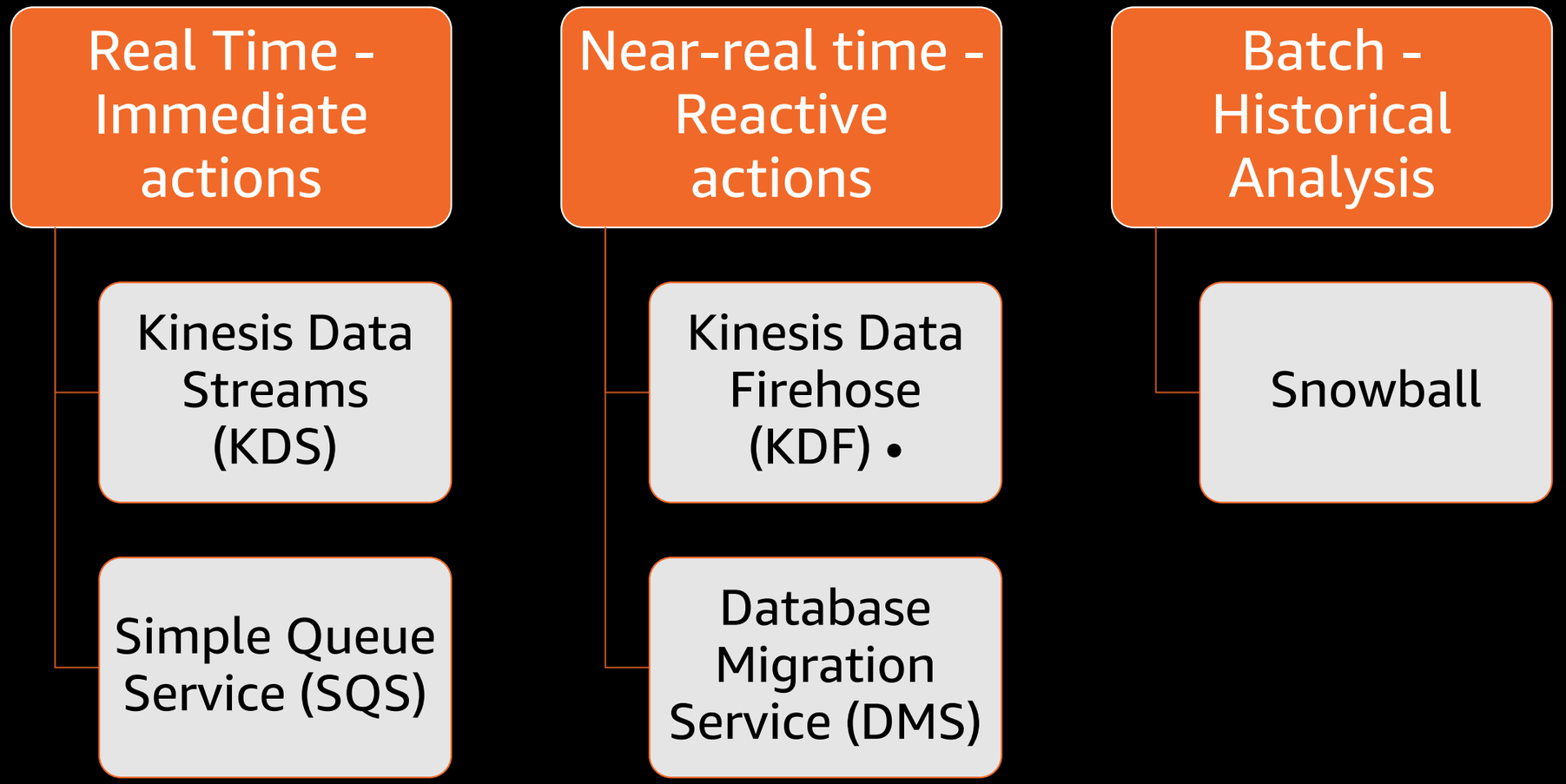


Amazon Data Analytics Pipeline

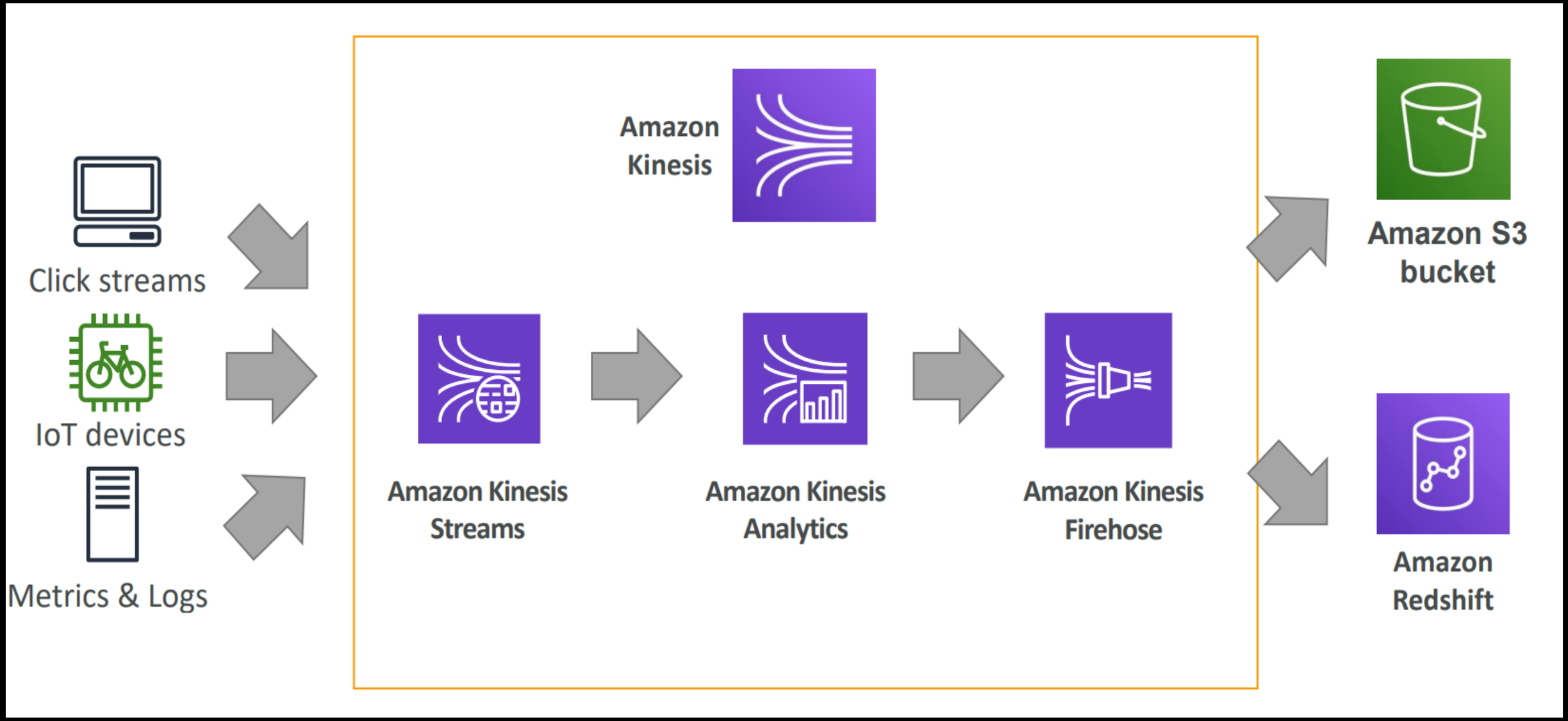
COLLECTION

Amazon Data Analytics Pipeline

COLLECTION - MOVING DATA INTO AWS

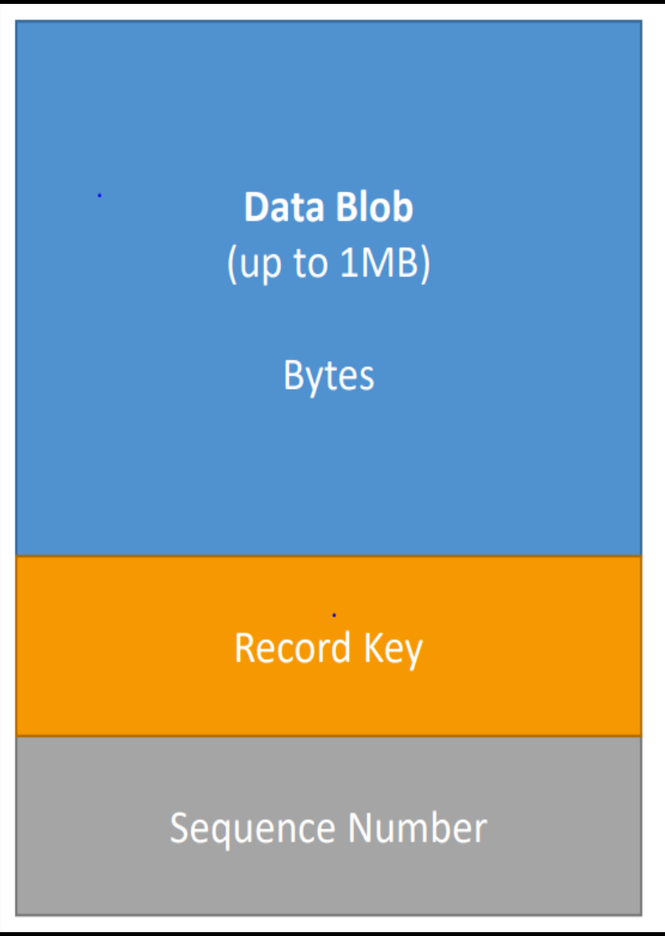


Kinesis


Click streams
IoT devices
Metrics & Logs
Amazon Kinesis
Amazon Kinesis Streams
Amazon Kinesis Analytics
Amazon Kinesis Firehose
Amazon S3 bucket
Amazon Redshift

Kinesis over view

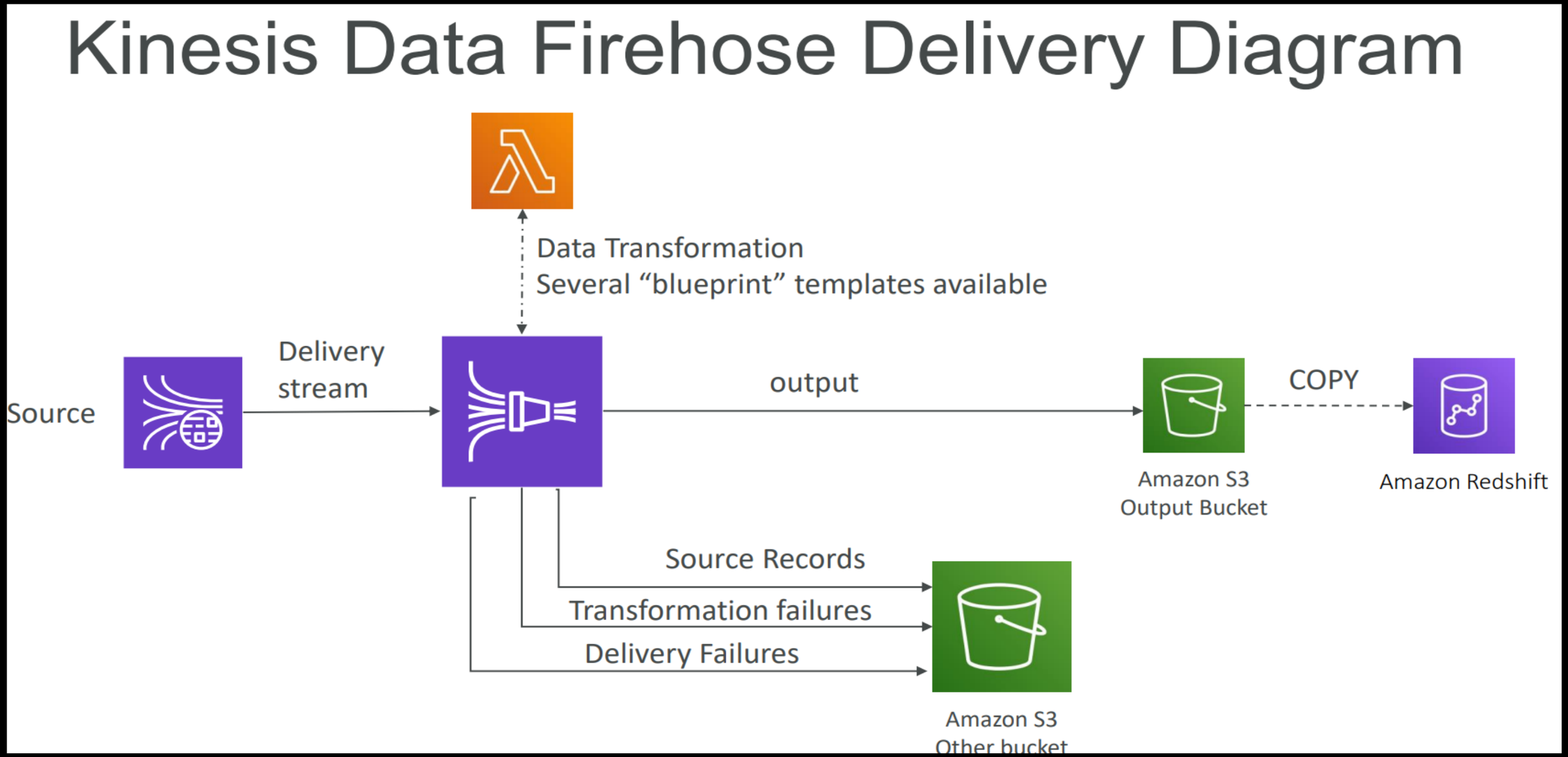
- Kinesis is a managed alternative to Apache Kafka
- Great for application logs, metrics, IoT, clickstreams
- Data is automatically replicated synchronously to 3 AZ
- Data retention is 24 hours by default, can go up to 7 days
- Ability to reprocess / replay data
- Multiple applications can consume the same stream
- Once data is inserted in Kinesis, it can't be deleted (immutability)
- Kinesis Streams Records
  - Data Blob – 1MB
  - Record Key - Same key = Same shard
  - Sequence number - Unique identifier for each records put in shards
- Kinesis Data Streams Limits
  - 1MB/s or 1000 messages/s at write PER SHARD
  - 2MB/s at read PER SHARD across all consumers
- Producers - kinesis SDK, Kinesis Producer Library (KPL) , Kinesis Agent, 3 rd party libraries: Spark, Kafka Connect
- Kinesis Consumers - Kinesis Firehose, AWS Lambda



Kinesis Firehose

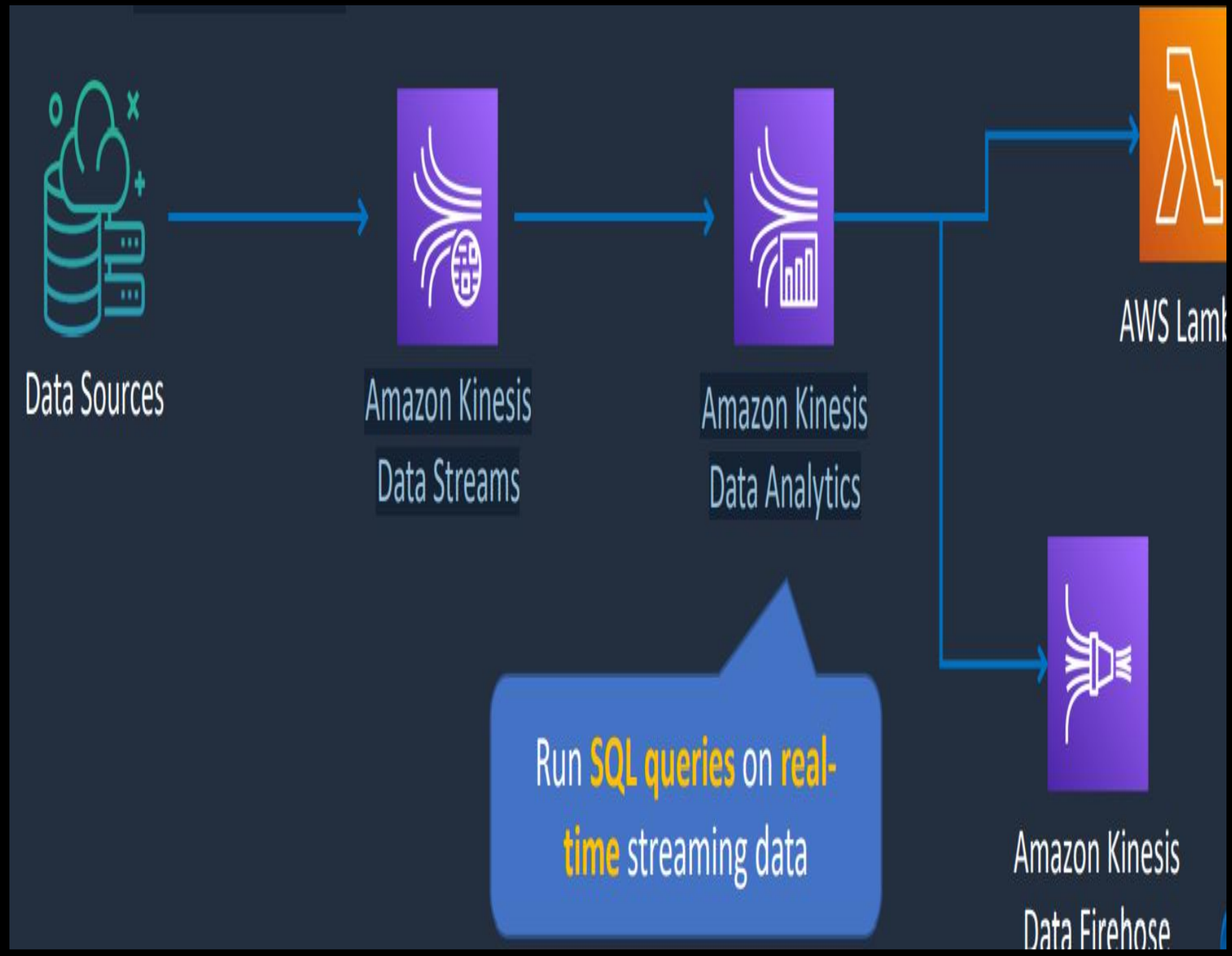
- Producers send data to Firehose, There are no Shards, completely automated (scalability is elastic).
- Firehose data is sent to another AWS service for storing, data can be optionally processed/transformed using AWS Lambda
- Near real-time delivery (~60 seconds latency)
- Kinesis Data Firehose destinations
  a. RedShift (via an intermediate S3 bucket)
  b. Elasticsearch,
  c. Splunk
  d. Datadog
  e. MongoDB
  f. HTTP Endpoint
  g. Amazon S3

Kinesis Firehose



Kinesis Analytics

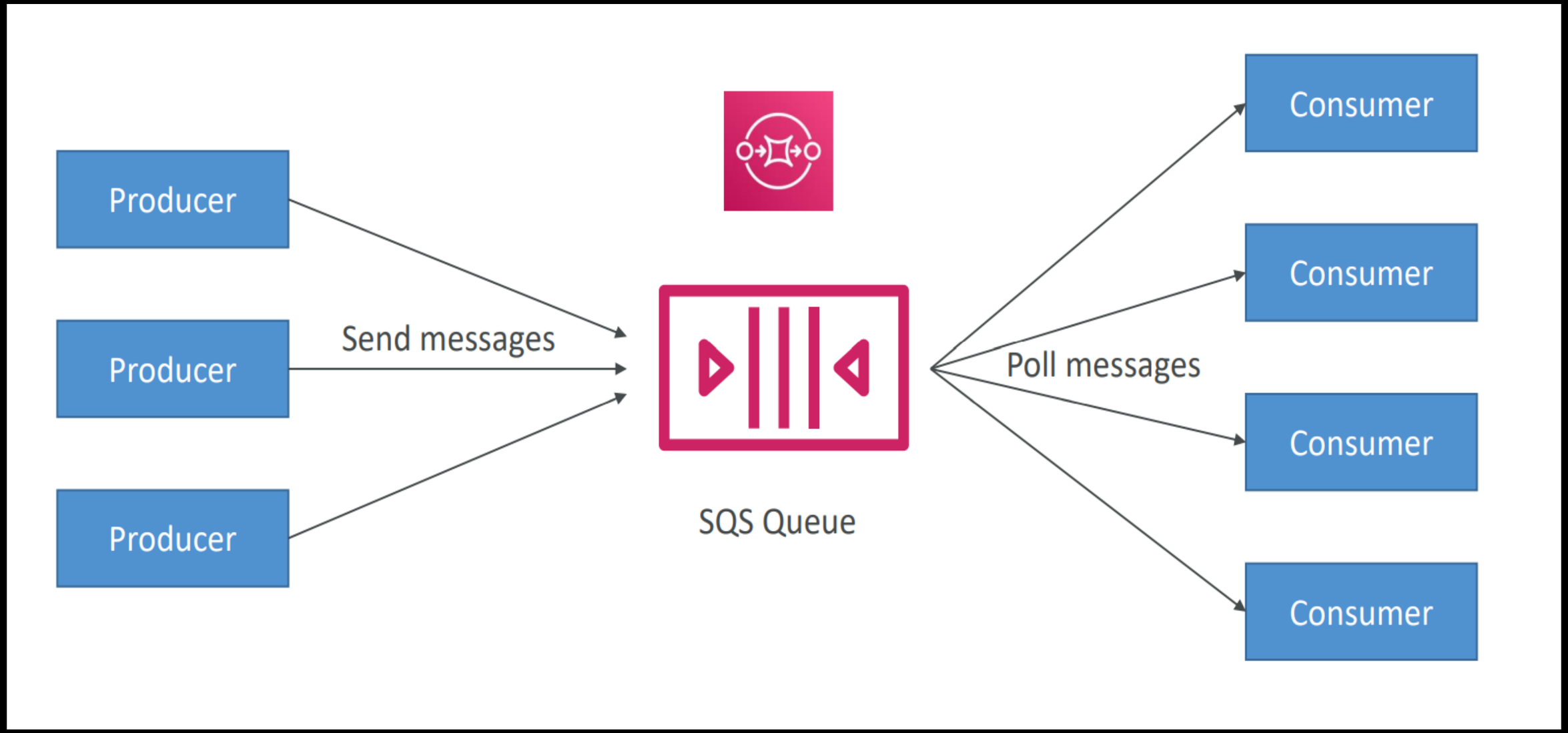
- Provides real-time SQL processing for streaming data
- Provides analytics for data coming in from Kinesis Data Streams and Kinesis Data Firehose
- Destinations can be Kinesis Data Streams, Kinesis Data Firehose, or AWS Lambda Amazon Kinesis Data Streams Amazon Kinesis Data Analytics



Amazon Data Analytics Pipeline

COLLECTION - SQS

SQS – Decouple applications



AWS SQS – Standard Queue and FIFO

Standard Queue

- Unlimited Throughput - Unlimited TPS
- Best-Effort Ordering
- At-Least-Once Delivery

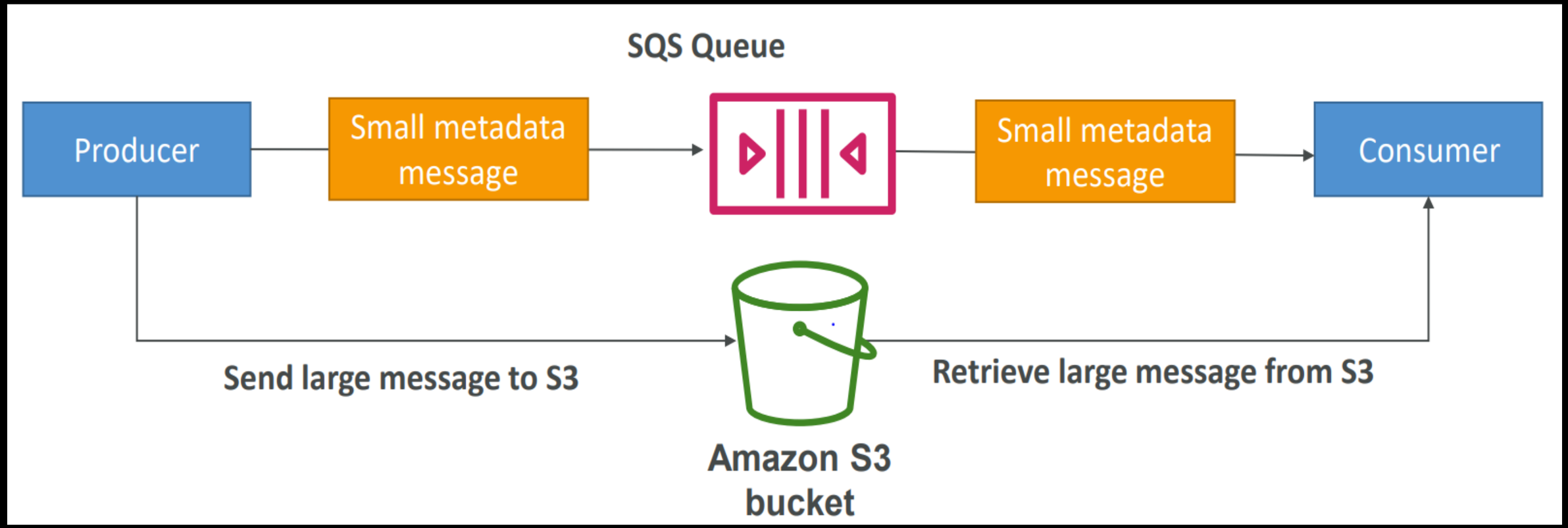
FIFO Queue

- High Throughput: 300 message/sec
- First-In-First-out Delivery
- Exactly-Once Processing

- FIFO queues require the Message Group ID and Message Deduplication ID parameters to be added to messages.
- Dead Letter Queue - Dead-letter queue is designed for handling message failure.
- Amazon SQS delay queues
- SQS Long Polling vs Short Polling - Long polling can lower costs, can be enabled at queue level of API level **WaitTimeSeconds**.
- Max message size is 256KB (or use Extended Client)
- Data retention from 1 minute to 14 days

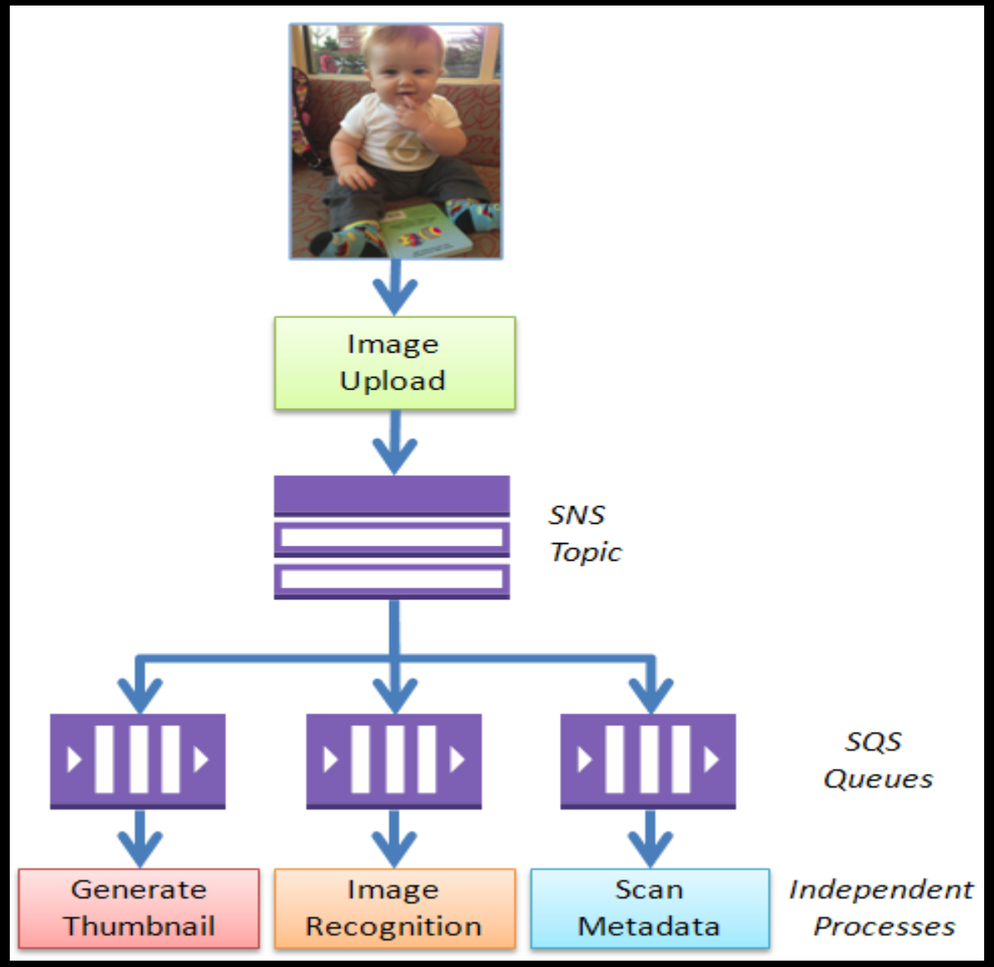
AWS SQS – SQS Extended Client

Message size limit is 256KB, how to send large messages?
Using the SQS Extended Client (Java Library)



SQS Queues and SNS Notifications – Now Best Friends

One common design pattern is called “fanout.” In this pattern, a message published to an SNS topic is distributed to a number of SQS queues in parallel. By using this pattern, you can build applications that take advantage parallel, asynchronous processing. For example, you could publish a message to a topic every time a new image is uploaded. Independent processes, each reading from a separate SQS queue, could generate thumbnails, perform image recognition, and store metadata about the image:



Anatomy of JSON policy

Resource-based policies are JSON policy documents that you attach to a resource such as an Amazon S3 bucket, or SQS.

sp-conversions	{ "Version": "2008-10-17", "Id": "{REPLACE_WITH_ANY_VALUE}", "Statement": [{

Amazon.com Confidential 11

```
    "Sid": "{REPLACE_WITH_ANY_VALUE}",
    "Effect": "Allow",
    "Principal": {
     "Service": "sns.amazonaws.com"
    },
    "Action": "SQS:SendMessage",
    "Resource":
"{REPLACE WITH ARN Of SQS DESTINATION QUEUE}",
    "Condition": {
     "ArnEquals": {
      "aws:SourceArn": "arn:aws:sns:us-east-
1:802324068763:*"
     }
    }
   }
  ]
}
```

Streams vs Firehose and Amazon SQS

	Kinesis Data Streams	Kinesis Data Firehose	Amazon SQS Standard	Amazon SQS FIFO
Managed by AWS	yes	yes	yes	yes
Ordering	Shard / Key	No	No	Specify Group ID
Delivery	At least once	At least once	At least once	Exactly Once
Replay	Yes	No	No	No
Max Data Retention	7 days	No	14 days	14 days
Scaling	Provision Shards: 1MB/s producer 2MB/s consumer	No limit	No limit	~3000 messages per second with batching (soft limit)
Max Object Size	1MB	128 MB at destination	256KB (more if using extended lib)	256KB (more if using extended lib)

Amazon Data Analytics Pipeline

STORAGE

Simple Storage Service

Amazon S3 provides a simple web service interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. The service gives any developer access to the same highly scalable, reliable, secure, fast, inexpensive infrastructure that Amazon uses to run its own global network of websites. The service aims to maximize benefits of scale and pass those benefits on to developers.

Amazon Simple Storage Service
Amazon S3

S3 Storage Classes

S3 Standard –

- General purpose storage for active, frequently accessed data with millisecond access

S3 Intelligent-Tiering –

- Only cloud storage class with automatic cost savings by moving objects between two tiers.

S3 Standard-Infrequent Access (S3 Standard-IA) –

- Long-term storage, backups, and disaster recovery

S3 One Zone-Infrequent Access (S3 One Zone-IA) –

- Ideal for secondary backups and workloads with easily re-creatable data

S3 Glacier –

- Archive or backup data with secure, durable, and low-cost storage

S3 Glacier Deep Archive –

- Lowest-cost cloud storage for long-term archives at about

Simple Storage Service

- Life cycle policies
- S3 Versioning
- S3 Replication
- S3 Encryption
- S3 Event Notifications

- Store anything
- Secure object storage
- Natively online, HTTP access
- Unlimited scalability
- 99.999999999% durability

Dynamo DB

- Fully managed NoSQL database service
- It is a non-relational, key-value type of database good for when data is not well structured or unpredictable
- Fully serverless service, Highly available, fault tolerant, service
- Horizontal, Push button scaling
- DynamoDB Time to Live (TTL)
- DynamoDB Streams - A Lambda function can be triggered
- DynamoDB Accelerator (DAX) -Fully managed in-memory cache for DynamoDB that increases performance (microsecond latency)

Amazon Data Analytics Pipeline

PROCESSING - AWS LAMBDA, AMAZON EMR, AMAZON ML, AMAZON SAGEMAKER

AWS Lambda - Serverless data processing

A way to run code snippets “in the cloud”

- Serverless
- Continuous scaling

Main uses of Lambda

- Real-time file processing
- Real-time stream processing
- ETL
- Cron replacement
- Process AWS events

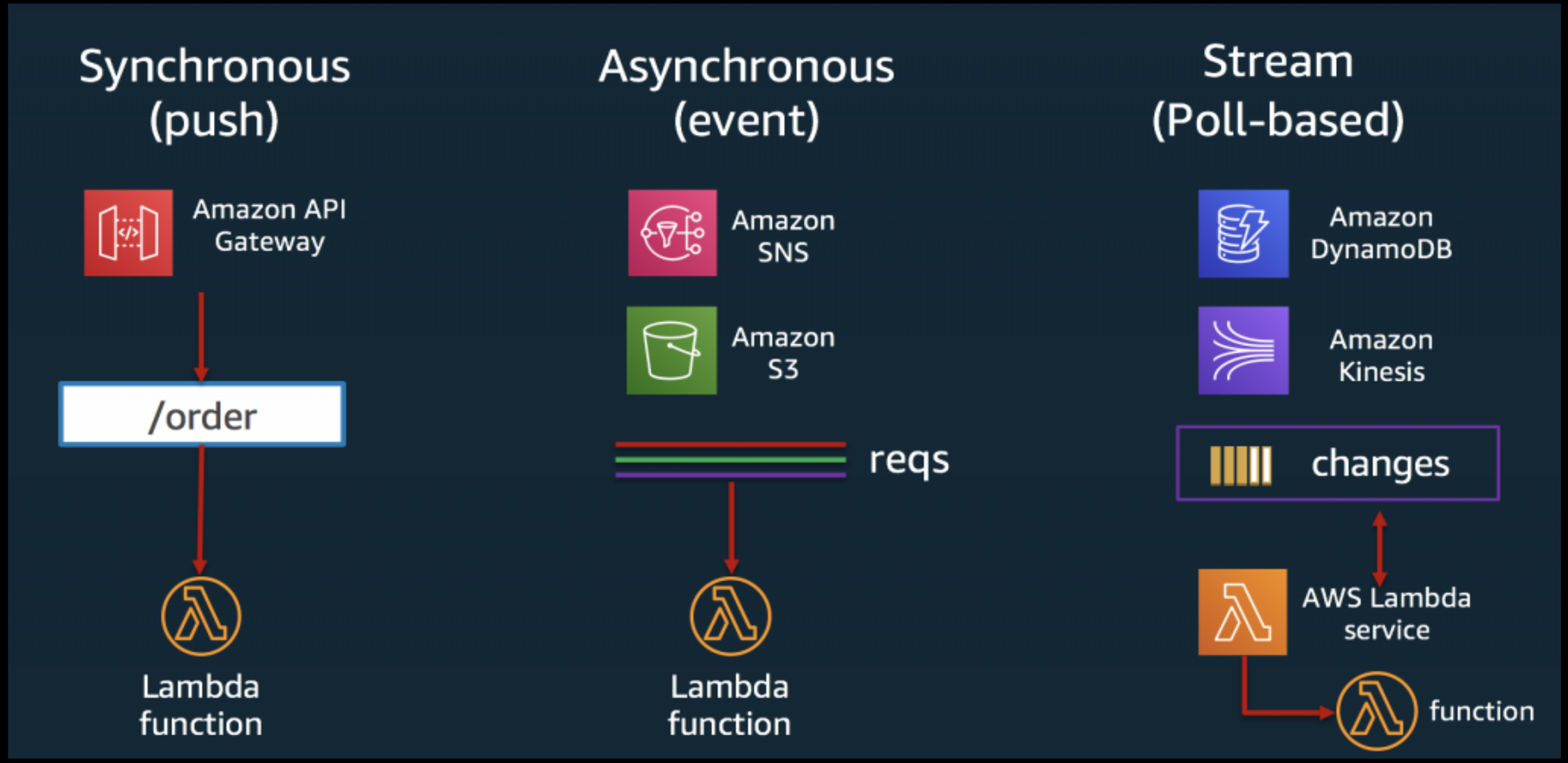
Supported language –

- Node.js • Python • Java • C# • Go • PowerShell • Ruby

Lambda Triggers –

- **Synchronous Invocation method with Request Response** - API Gateway, Load Balancer, AWS Step Functions CloudFront (Lambda@Edge), Amazon Kinesis Data Firehose, Amazon Cognito, Amazon Alexa
- **Asynchronous Invocation method with Event** – S3,SNS, CF, Cloud Watch, Events, IOT, Developer Tools
- **Event Source Mapping method with Poll Based** - Kinesis, DynamoDB, Simple Queue Service

AWS Lambda - Serverless data processing



AWS Lambda

Cost -

- Pay for what you use” Generous free tier (1M requests / month, 400K GB -seconds compute time)
- $0.20 / million requests
- $.00001667 per GB/second

Antipatterns -

- Stateful applications
- Long running Apps – 15 min timeout limit

Amazon Data Analytics Pipeline

ANALYSIS - ES, ATHENA, REDSHIFT

What is Athena - Serverless

- Interactive query service for S3 (SQL)
- No need to load data, it stays in S3
- Unstructured, semi-structured, or structured
- Supports many data formats
  - CSV (human readable)
  - JSON (human readable)
  - ORC (columnar, split table)
  - Parquet (columnar, split table)
  - Avro (split table)

Athena - Use cases

- Querying staging data before loading to Redshift
- Integration with Quick Sight
- Integration via ODBC / JDBC with other visualization tools

Data warehouse

A data warehouse is a central repo of structured data from many data sources, this data is transformed, aggregated, and prepared for business reporting and analysis.

Challenges –

- Costly to implement
- Maintenance can be challenging.
- Security concerns
- Hard to scale to meet to demand

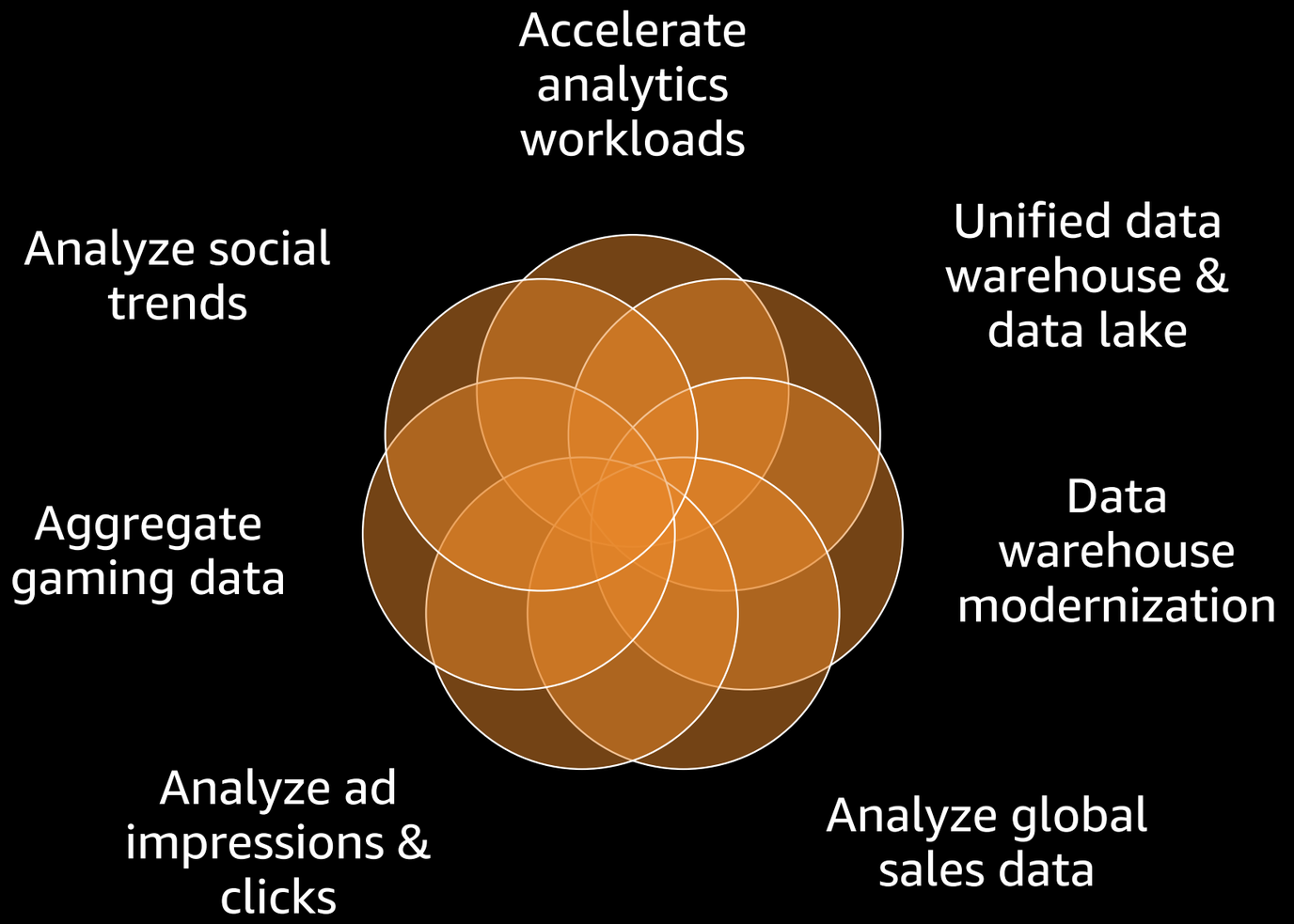
Amazon Redshift - Fully -managed, petabyte - scale data warehouse

Redshift provides a cloud-based, scalable, secure environment for your data warehouse. Amazon Redshift is easy to set up, deploy, and manage and provides up to 10 times faster performance than other data warehousing solutions.

Benefits –

- Designed for OLAP, not OLTP
- Cost effective
- SQL, ODBC, JDBC interfaces
- Scale up or down on demand
- Built-in replication & backups • Monitoring via CloudWatch / CloudTrail

Amazon Redshift – Use cases

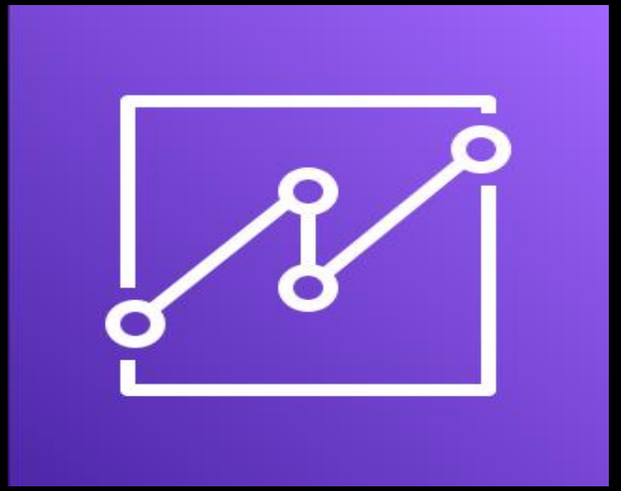


Amazon Kinesis Data Analytics

VISUALIZATION

Quick Sight

- Fast, easy, cloud -powered business analytics service
- Allows all employees in an organization to:
  - Build visualizations
  - Perform ad-hoc analysis
  - Quickly get business insights from data
  - Anytime, on any device (browsers, mobile)
- Serverless



Amazon QuickSight

Quick Sight Data Sources

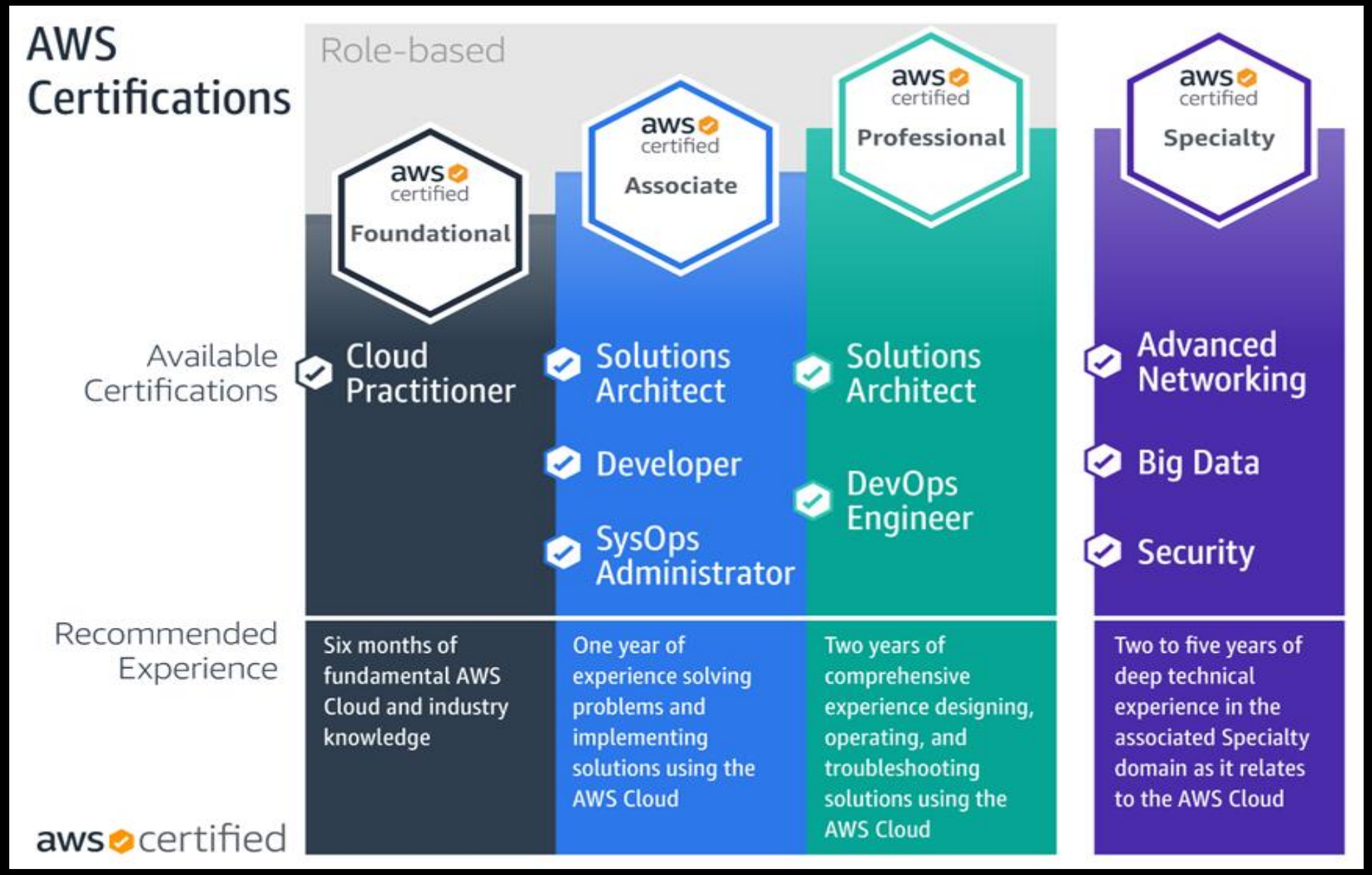
- Redshift
- Aurora / RDS
- Athena
- EC2-hosted databases
- Files (S3 or on-premises)
  - Excel
  - CSV, TSV
  - Common or extended log format
- Data preparation allows limited ETL

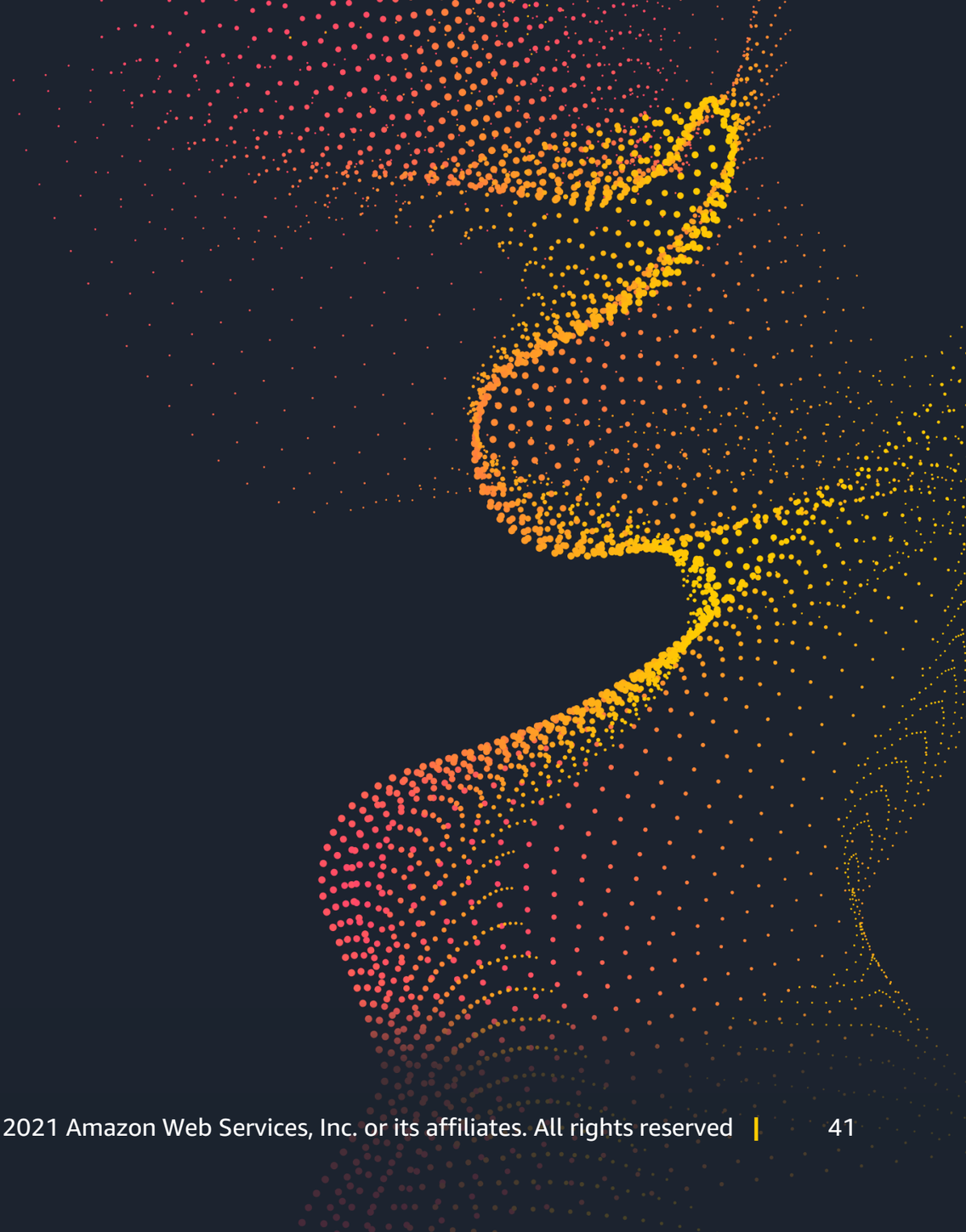
SPICE

- Data sets are imported into SPICE
  - Super-fast, Parallel, In-memory Calculation Engine
  - Uses columnar storage, in -memory, machine code generation
  - Accelerates interactive queries on large datasets
- Each user gets 10GB of SPICE
- Highly available / durable
- Scales to hundreds of thousands of users

Certification Path

- How to pass the AWS Associate Certification exams





aws machine learning