Data Lake

Streamline Data Management

Vikas

Compute With Cloud Inc

Data

#	year	university	n_rank	country	region
1	2017	Massachusetts Institute…	1	United States	North America
2	2017	Stanford University	2	United States	North America
3	2017	Harvard University	3	United States	North America
4	2017	University of Cambridge	4	United Kingdom	Europe
5	2017	California Institute of T…	5	United States	North America
6	2018	Massachusetts Institute…	1	United States	North America
7	2018	Stanford University	2	United States	North America
8	2018	Harvard University	3	United States	North America

Relational Databases

```
SELECT year, university, n_rank, country, region
FROM university_ranking
WHERE n_rank < 6
ORDER BY year, n_rank;
```

#	year	university	n_rank	country	region
1	2017	Massachusetts Institute…	1	United States	North America
2	2017	Stanford University	2	United States	North America
3	2017	Harvard University	3	United States	North America
4	2017	University of Cambridge	4	United Kingdom	Europe
5	2017	California Institute of T…	5	United States	North America
6	2018	Massachusetts Institute…	1	United States	North America
7	2018	Stanford University	2	United States	North America
8	2018	Harvard University	3	United States	North America

Relational Databases

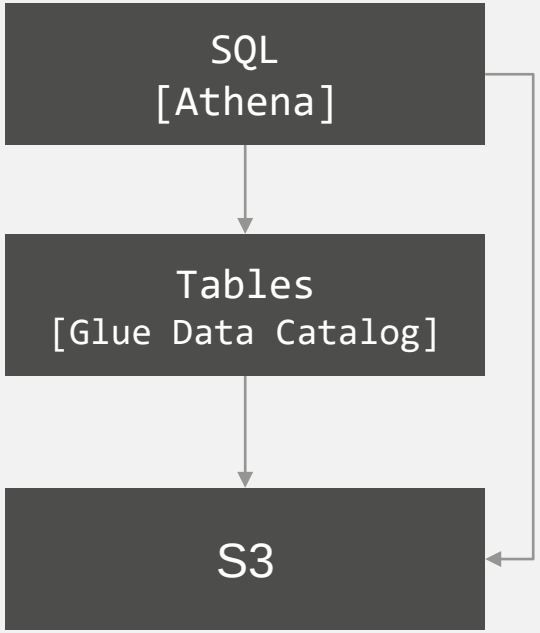
Pros

1. Tabular format
2. SQL
3. Automatic Data types, Integrity checks
4. Consistency
5. Security
6. Ecosystem

Cons

1. Complex
2. Expensive
3. Rigid structure
4. Development and Maintenance effort

Data Lake



1. Store files in S3
2. Glue Data Catalog – Capture record structure in files and represent them as table
3. Query using SQL with Athena service

Data Lake is perfect for Data discovery, Exploratory Data Analysis (EDA) and Machine Learning use cases

Data Discovery

- Finding and identifying data in an organization
- Data source – files, databases, cloud
- Quality assessment – accurate and complete
- Cataloging metadata – structure and format
- Relationship with other data assets

Data discovery is the critical first step in any data driven project: understand data, uncover opportunities, make better decisions

Exploratory Data Analysis

- Cleaning and preparing data
- Handle missing and incorrect data
- Merging data
- Transform data
- Summarizing data characteristics - statistical and visualization techniques

Data Lake gives your data engineers, data analysts and machine learning engineers access to raw data for which purpose is not yet defined

Data Lake Vs Data Warehouse

	Data Warehouse	Data Lake
Data Structure	Well-defined: Tables, Rows, Columns	Native format, Raw data – no defined structure or schema
Purpose	Clearly defined usage scenarios	Exploratory analysis and data discovery, machine learning
Processing	Fast, consistent, reliable results	Ad hoc analysis, mixed data formats, processing time varies
Accessibility	Complex and Expensive	Highly customizable and quick to update

References: https://www.talend.com/resources/data-lake-vs-data-warehouse/
https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

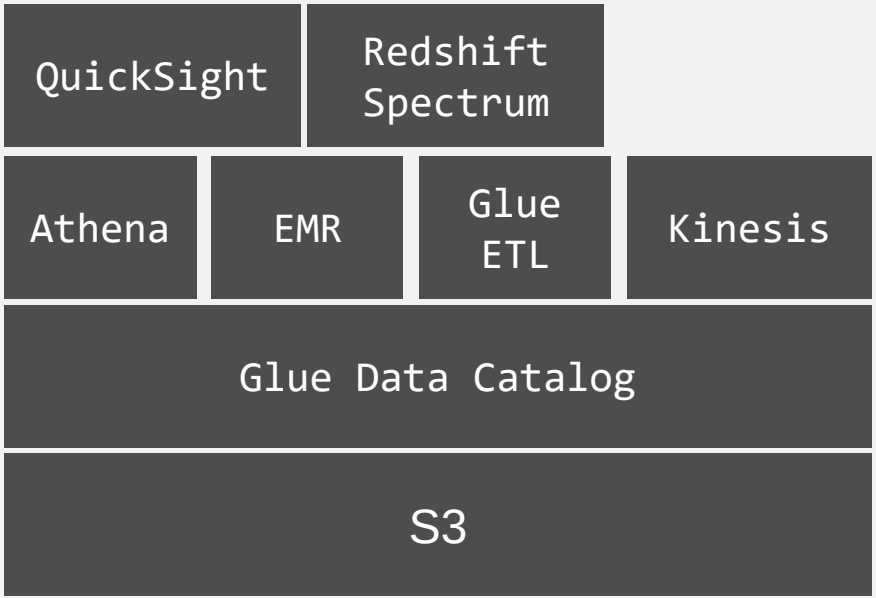
Data Lake - Challenges

Complexity

Data Lake is not a single tool or service

It is a concept or architecture

To build a data lake – you have to use many different services and tools



Lack of Structure

Data Lake – store data in any format

Integrating data from various sources and data formats can be a challenge

Data Quality

Ensuring quality and integrity of data in a data lake is difficult

Large volume of data from diverse sources

Data Governance

Data is stored in raw form – difficult to ensure data is secure, compliant and consistent

Define a strategy on how data is collected, stored, processed and shared in an organization

Data Governance

Area	Responsibility
Data Stewardship	Assign responsibility for managing specific data assets to individuals or teams
Data Quality Management	Monitor and maintain quality of data assets so that they are accurate, complete and consistent
Data Security	Protect against unauthorized access, disclosure or destruction. Establish access control, encryption, audit data access and usage
Data Compliance	Data is managed in compliance with legal, regulatory (GDPR, HIPAA), and contractual requirements on data sharing and usage
Data Consistency	Ensure data is consistent in terms of format, structure and values that it can be trusted

Cost

Data Lake – cost effective for large volume of data

Cost of storage, processing and querying can still add up

S3 Tiered storage and lifecycle management, Glue ETL for transforming data to optimal format, Partition data

Integration Challenges

Integrating data from multiple data sources and setting up a pipeline can be a challenging task

AWS Glue ETL, Storage Gateway, Snowball, Kinesis

Range of analytical tools: Athena, Redshift Spectrum, QuickSight, EMR, SageMaker

Build Proof-of-concept (POC) system

1. Feasibility – Build POC systems to determine feasibility of proposed solution or technology

2. Risks – Identify potential risks and challenges early in the project

3. Performance – Test the performance and ability to store, process and query data

4. Communication – POC helps facility communication and collaboration between teams and stakeholders

Data Lake

1. Ideal for data discovery, exploratory data analysis
2. Streamline data access
3. Build POCs before investing significant time and resources

AWS - Whitepaper

1. Storage
2. Governance
3. Analytics

Data Lake on AWS:
https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/building-data-lake-aws.html

Storage

Service	Purpose	Use
S3	Storage	Object Storage to store and retrieve any amount of data. Cost effective with 99.999999999% (11 9s) of durability Object Life cycle management
Glacier	Backup and Archiving	Backup and Long term archival (multi-year) Extremely low cost and 11 9s durability.

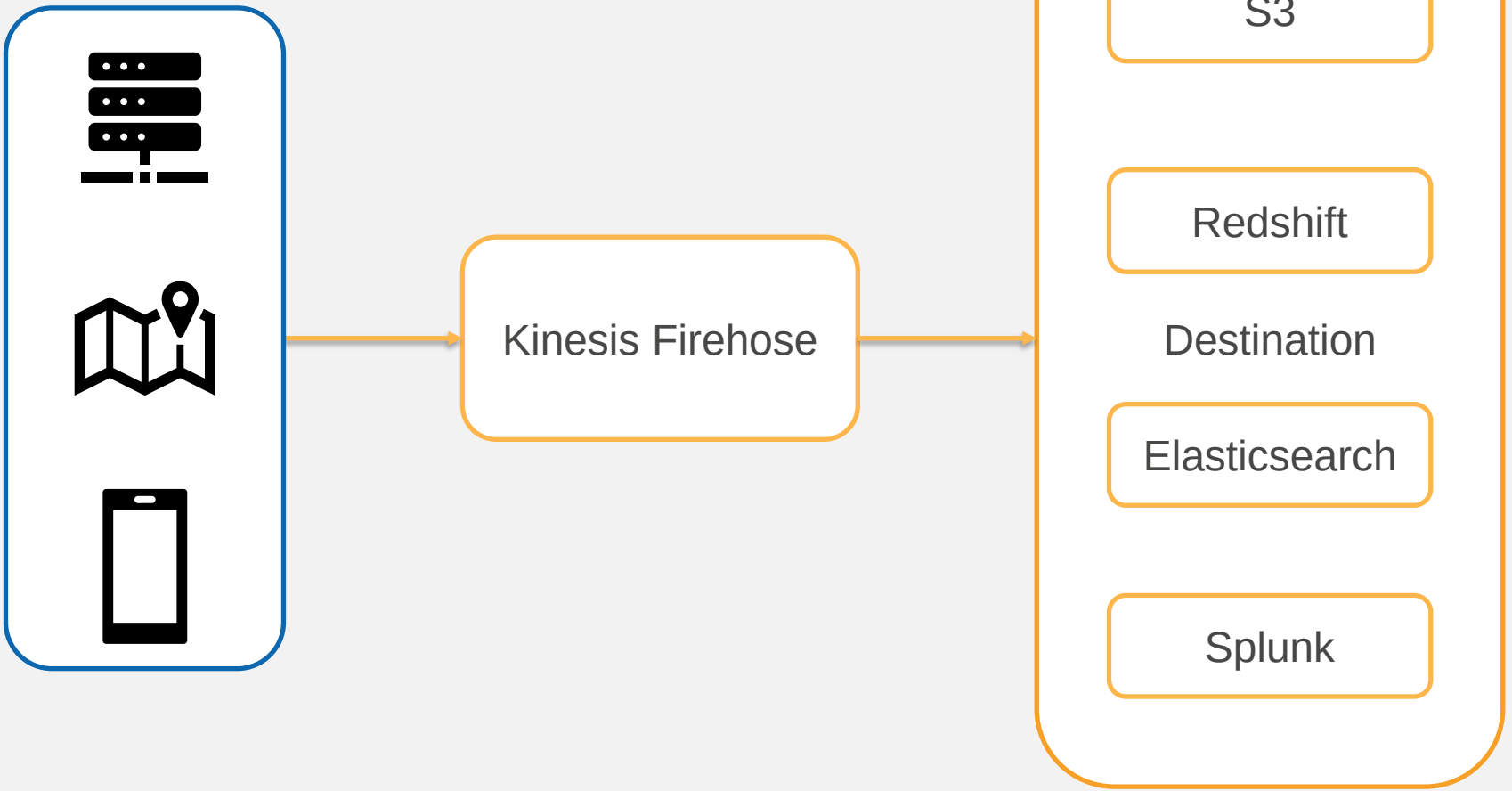
Data Lake Storage

	S3 Standard	S3 Infrequent Access	Glacier
Cost - 500GB per month	USD 11.50	USD 6.25	USD 2.00
Durability	99.999999999% (11 9's)		
Suitable for	Frequently Accessed	Less Frequently Accessed	Long term archival
First byte latency	Immediate	Immediate	Restore can take minutes to hours

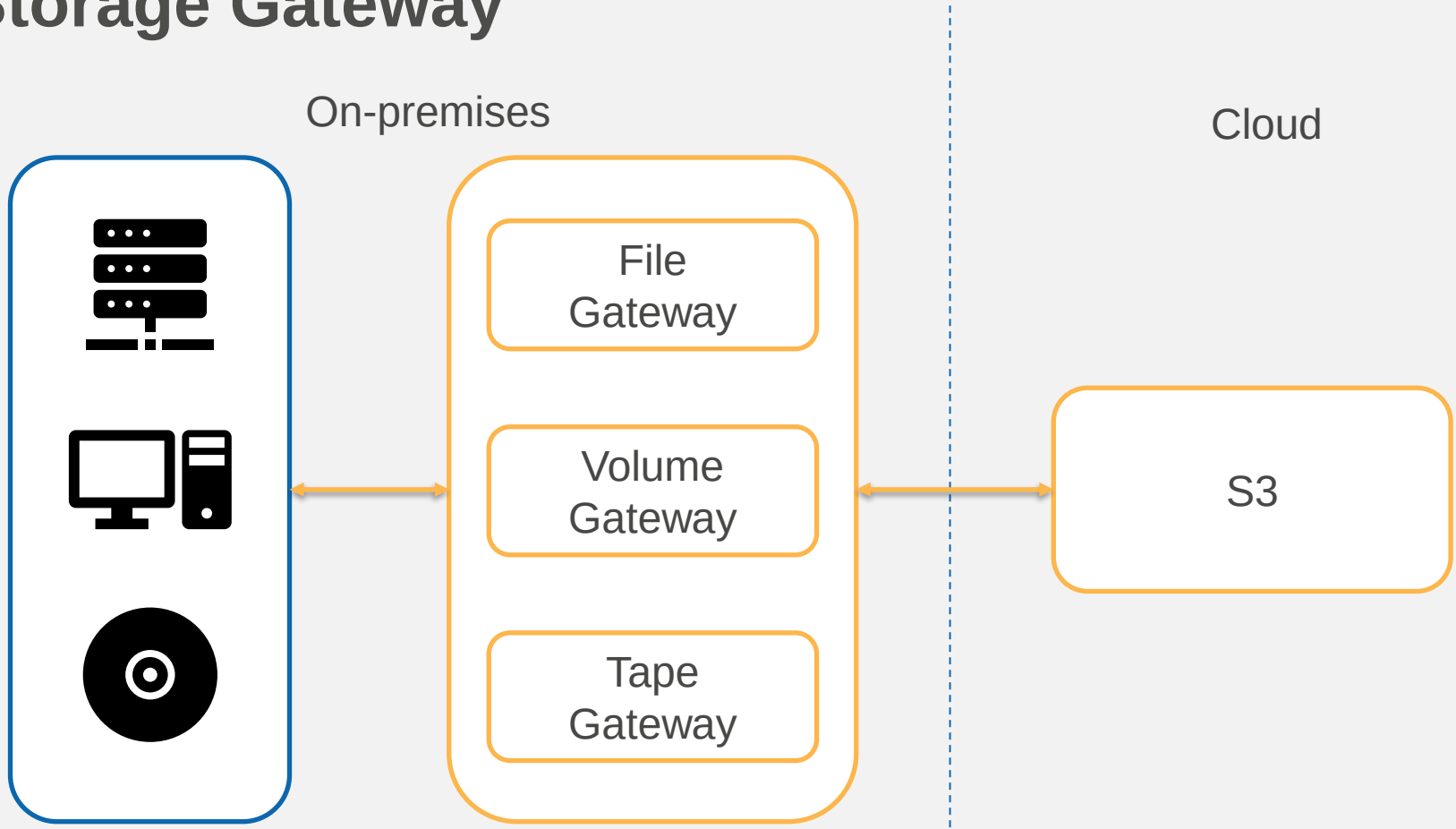
Ingestion

Service	Purpose	Use
Kinesis Firehose	Real-time Streaming Data Ingestion	Capture and deliver real time streaming data directly to S3, Redshift, Elasticsearch, Splunk
Storage Gateway	Hybrid Cloud Storage	Integrate legacy on-premises data processing platforms to S3 Data Lake
Snowball, Snowmobile	Migration (Large scale)	Physically move petabytes to exabytes of data to AWS cloud at 1/5th the cost of internet transfer
SDK, CLI and more	Custom Ingestion	Easy to integrate with variety of tools

Realtime Streaming Data



Storage Gateway



Snowball

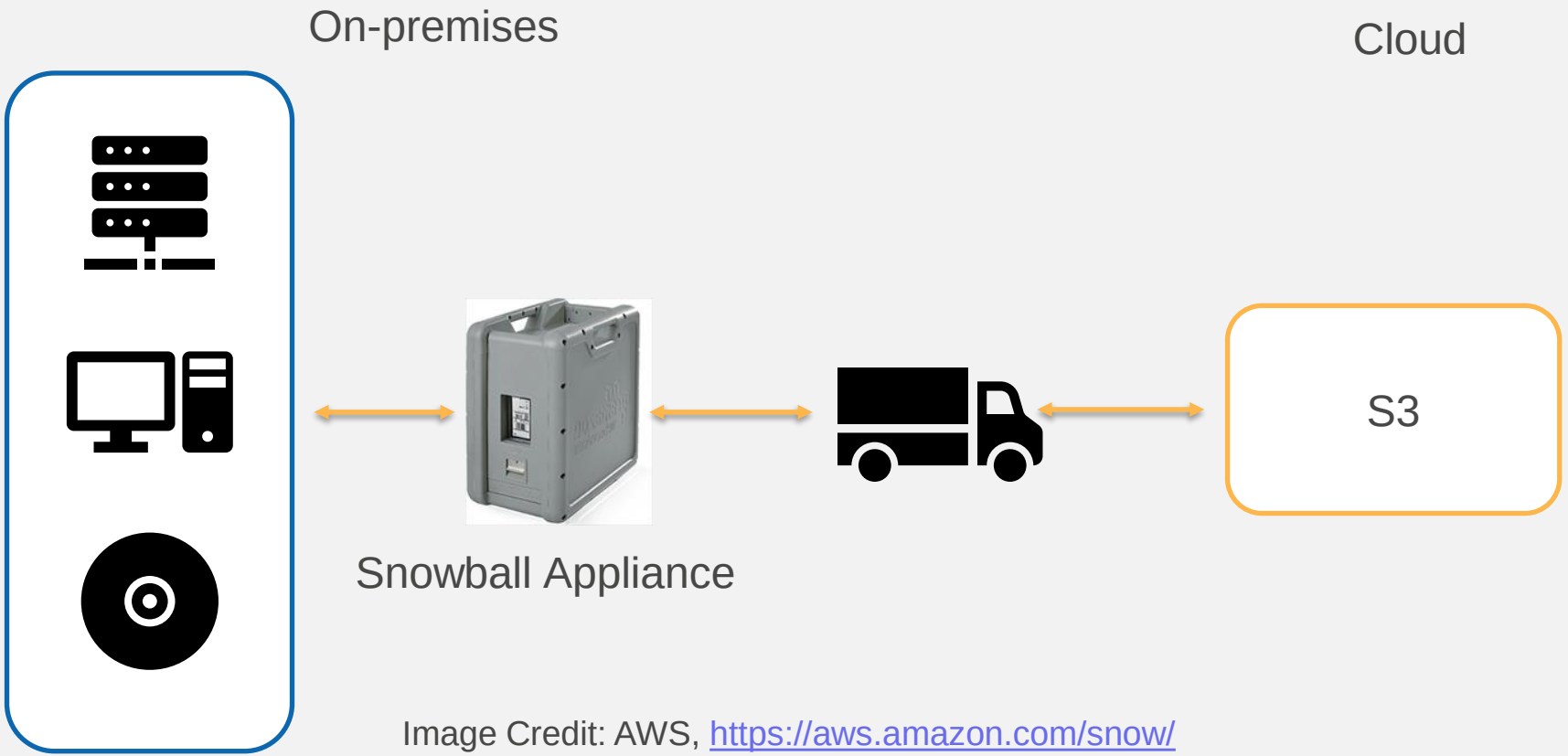
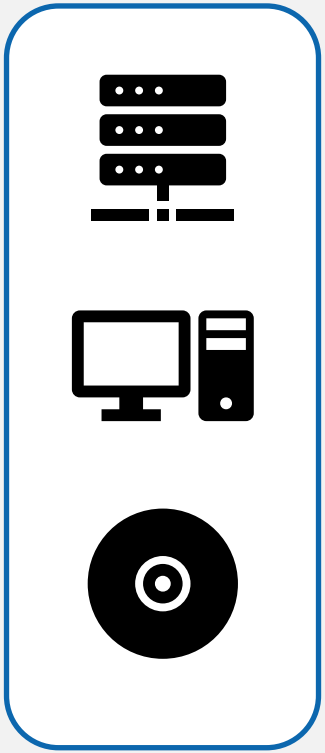


Image Credit: AWS, https://aws.amazon.com/snow/

Snowmobile

On-premises

Cloud



S3

Snowmobile Container

Image Credit: AWS, https://aws.amazon.com/snow/

Ingestion

AWS SDK

AWS Command Line Tool

Third party tools

Data Catalog

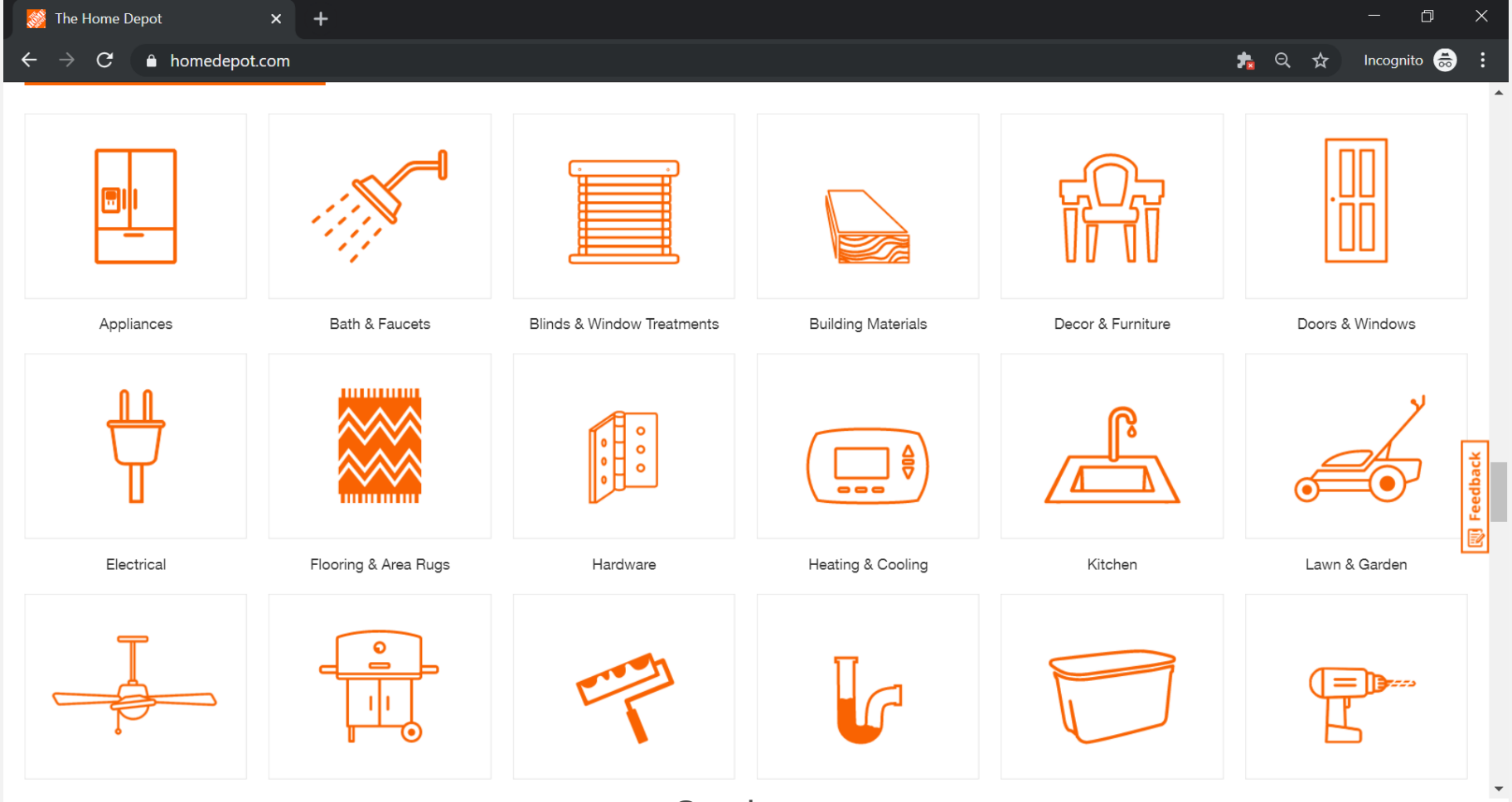
Make data discoverable and usable

Track versions of changes

Queryable interface for all data assets

Data Catalog

Service	Purpose	Use
Do-it-yourself	Comprehensive Data Catalog	Make data discoverable and usable. Use services like Lambda, Elasticsearch, DynamoDB to collect and maintain metadata
Glue	Managed Data Catalog	Make data discoverable and usable. Automatically crawl and collect metadata from S3, DynamoDB and any other databases that supports JDBC connectivity



Catalog

[Image Credit: HomeDepot]

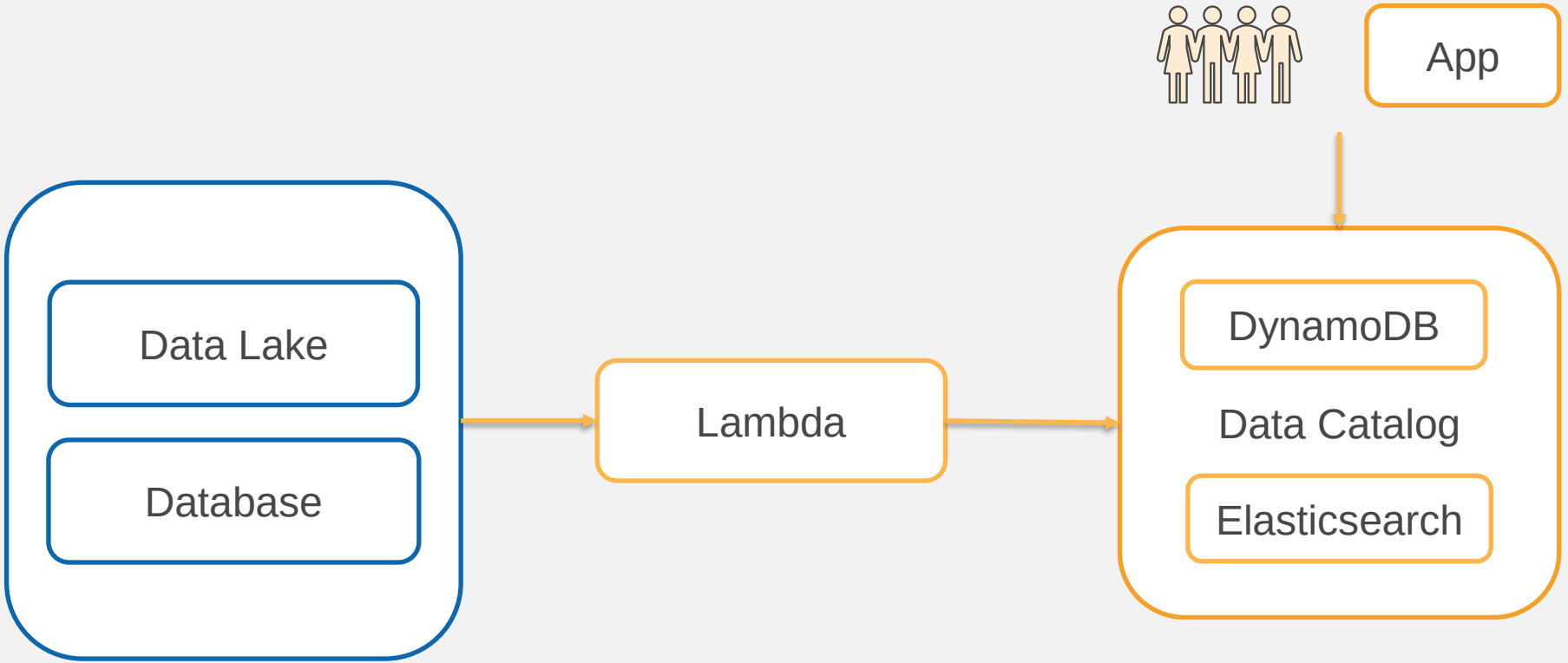
Image Credit: webhamster, flickr

Data Swamp

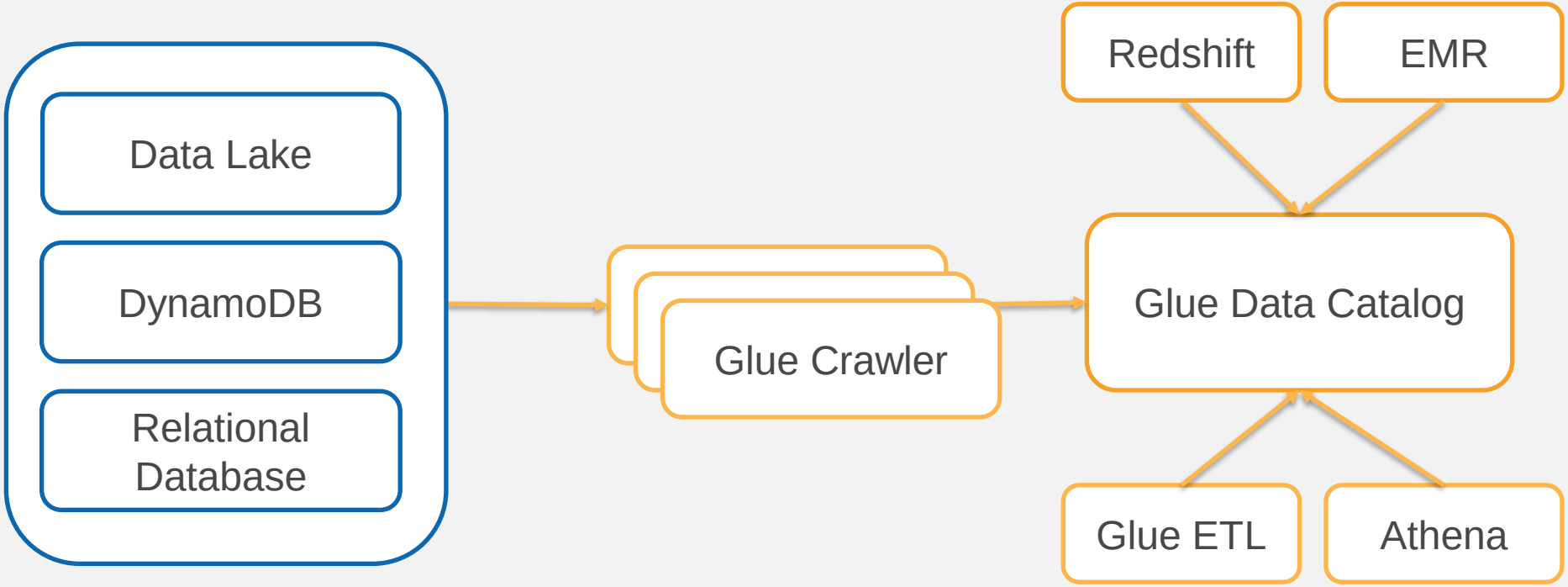
“A **data swamp** is a deteriorated and unmanaged data lake that is either inaccessible to its intended users or is providing little value”

Reference: Data Swamp
https://en.wikipedia.org/wiki/Data_lake

Do-it-yourself Data Catalog



Glue Data Catalog



Data Formats

Popular Formats, Tools for Conversion

Data Formats

Variety of formats

With optimal format, you can:

- Lower storage cost
- Improve query performance

Question: When and where to do the format conversion?

Data Formats

“One of the core values of a data lake is that it is the collection point and repository for all of an organization’s data assets, in whatever their native formats are”

Reference: Data Lake on AWS, https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/building-data-lake-aws.html

Data Formats

Collect data in native format

Transform data in data lake

Organize by row or column

- Row Store – Optimized for reading entire row
- Column Store – Optimized for reading a subset of columns

Text Data Formats

Format	Organization	Use
CSV, TSV	Row	Easy to use No data type support Duplication with hierarchies - For example, in an employee-department CSV file, department information is duplicated for every employee Not optimized for reading only specific columns
JSON, JSON Lines	Row	Format of choice for communication between web services Supports data types Efficiently represent hierarchical data JSON Lines – A record is stored in a line

Binary Data Formats

Format	Organization	Use
Parquet	Columnar	Ideal for use cases that require only subset of columns Efficiently query large amount of data Write Once Read Many (WORM) Compressed Storage Extensive Tool Support Data Type Support Reduce storage footprint, improve query performance and lower query cost

Parquet Performance: https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/monitoring-optimizing-data-lake-environment.html

Binary Data Formats

Format	Organization	Use
ORC	Columnar	Like Parquet
Avro	Row	Ideal for write-heavy use cases Efficiently read the entire record Data Type Support

Data Transformation

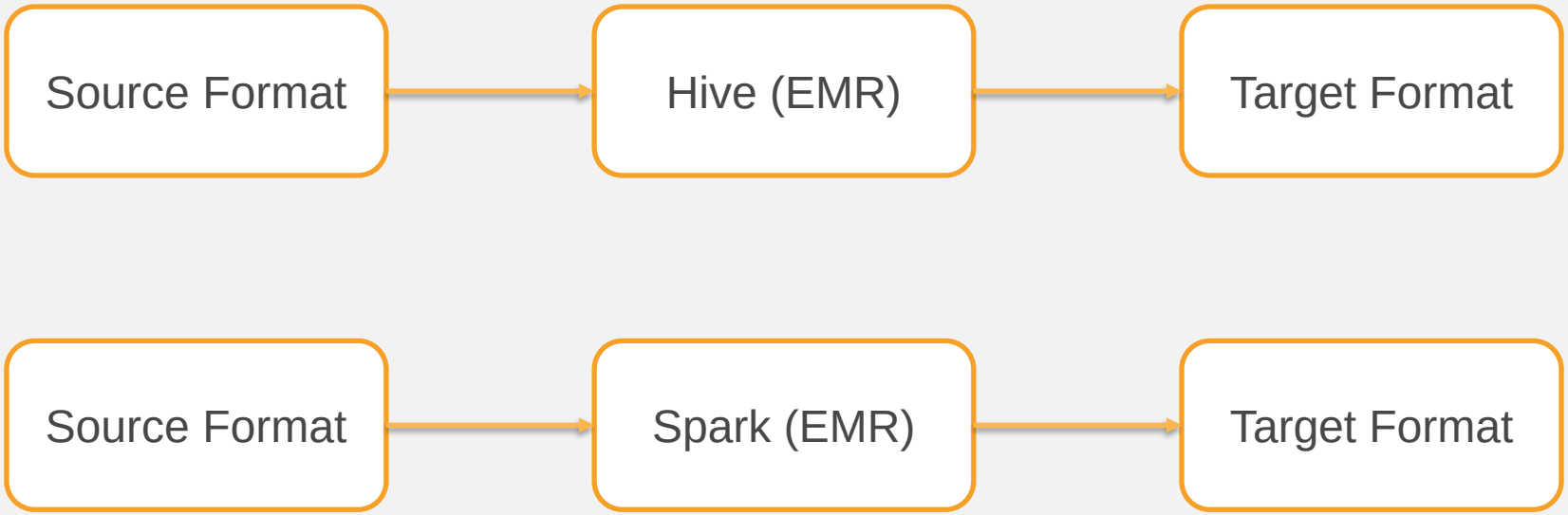
Collect in native format

Transform in data lake

Data Transformation

Service	Purpose	Use
Amazon EMR	Data Processing	Managed Hadoop environment Support for tools like Spark, Hive, HBase Support for ML tools like TensorFlow and MXNet List of tools: https://aws.amazon.com/emr/features/

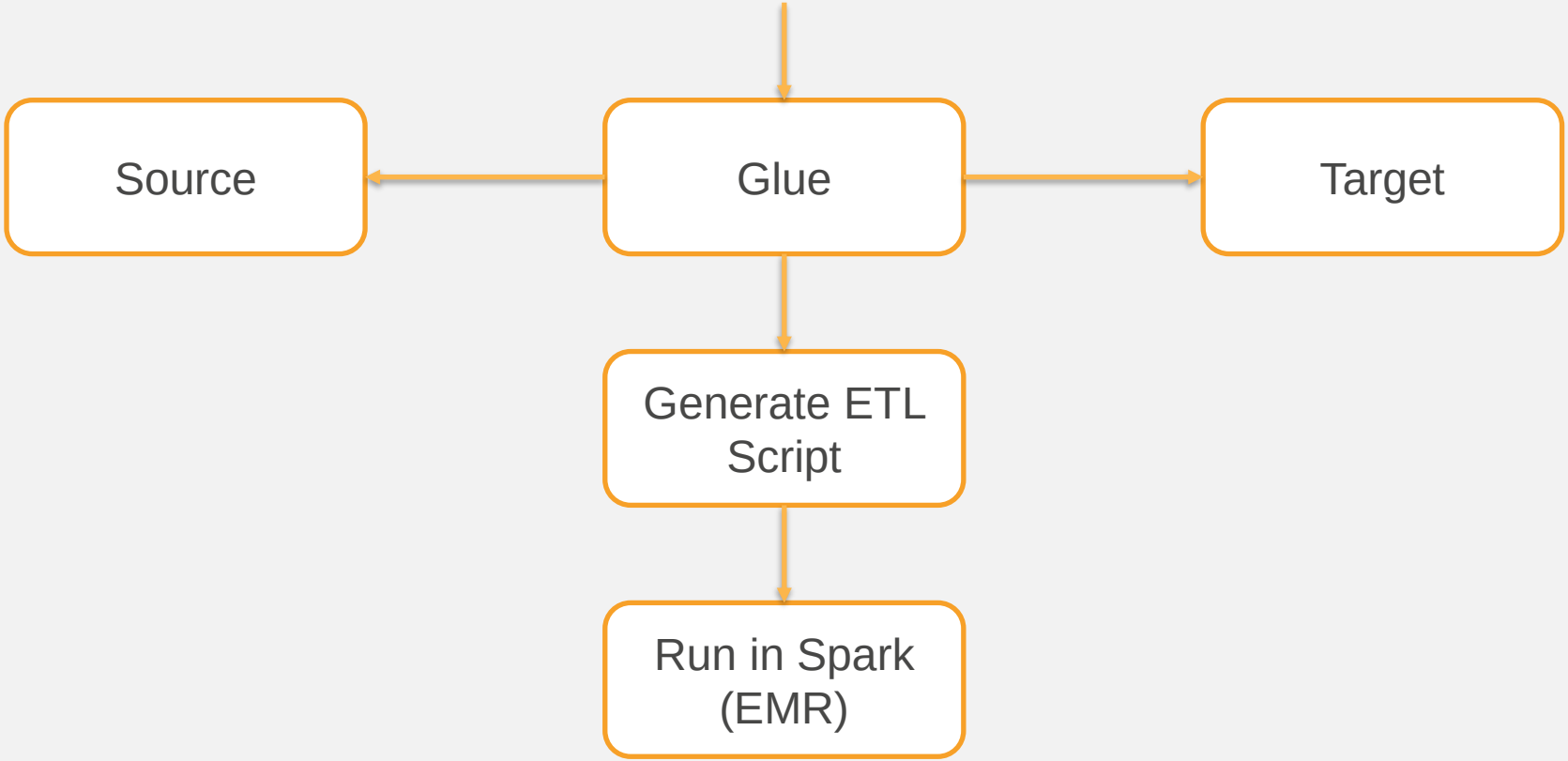
Amazon EMR – Format Conversion



Data Transformation

Service	Purpose	Use
Glue	Managed ETL	Automatically Generate ETL Scripts Schedule and Run in Spark Support for Scala and Python

Glue ETL – Generate and Run Script



Data Transformation

Service	Purpose	Use
Kinesis Firehose	Streaming Data Transformation	Transform streaming data to Parquet, ORC formats Deliver transformed data to AWS Data Stores Optionally, backup original data to S3

Stream and Batch Processing

Streaming Data

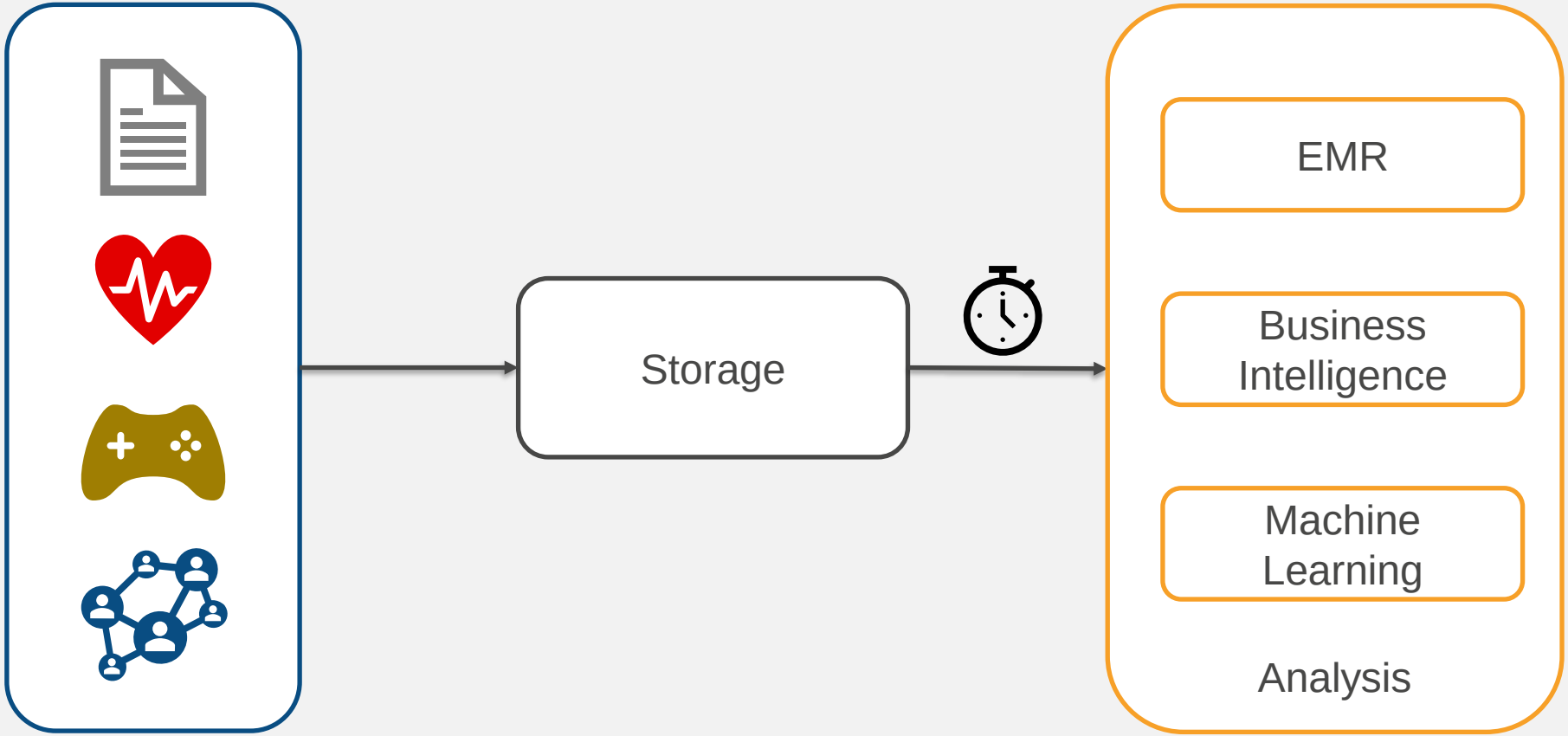
Generated Continuously

Thousands of sources

Small Payloads




Batch Processing

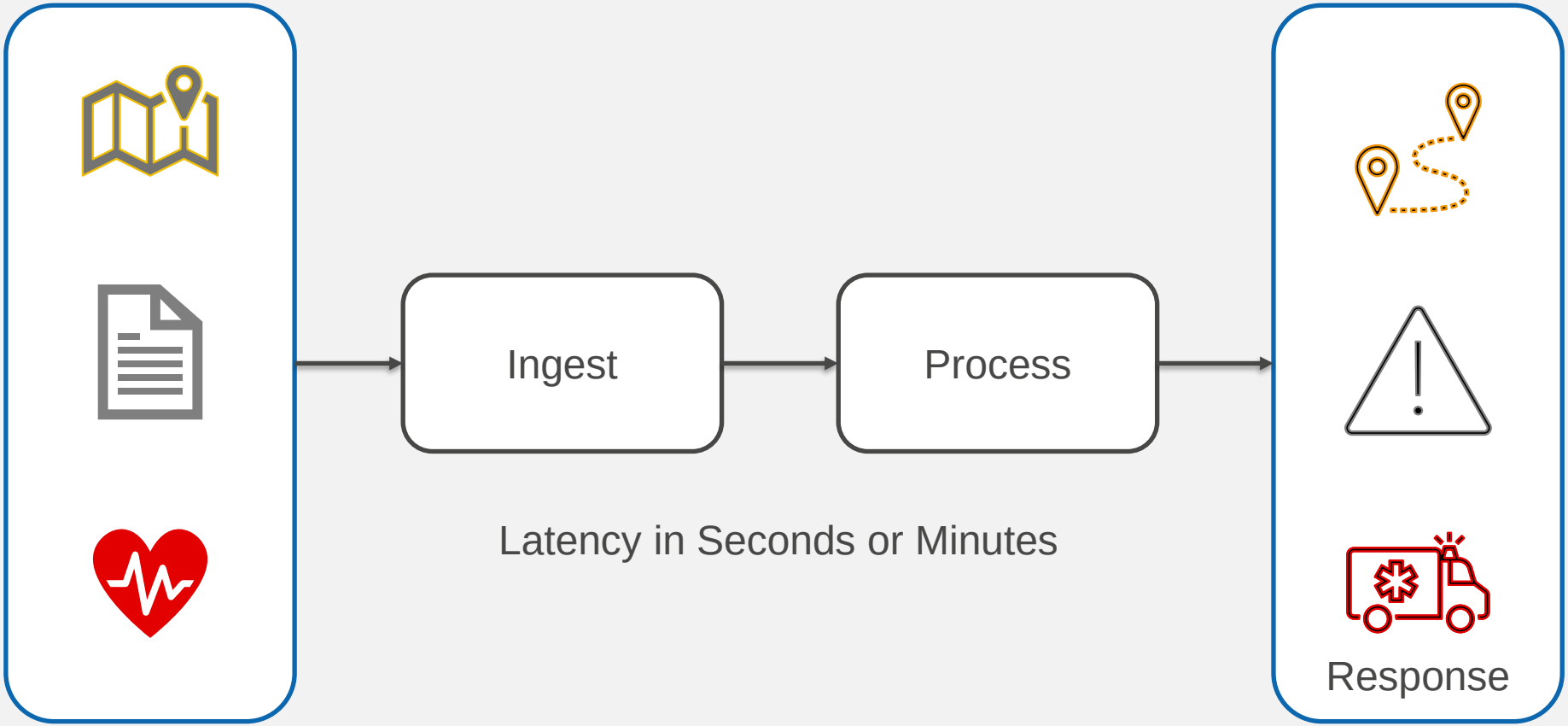


Batch Processing Use Cases

Utility bill generation

Daily, monthly manufacturing reports

Stream Processing



Amazon Kinesis

Collect, Process, Analyze Streaming Data

Amazon Kinesis

“Amazon Kinesis enables you to ingest, buffer and process streaming data in real-time”

“you can derive insights in seconds or minutes.”

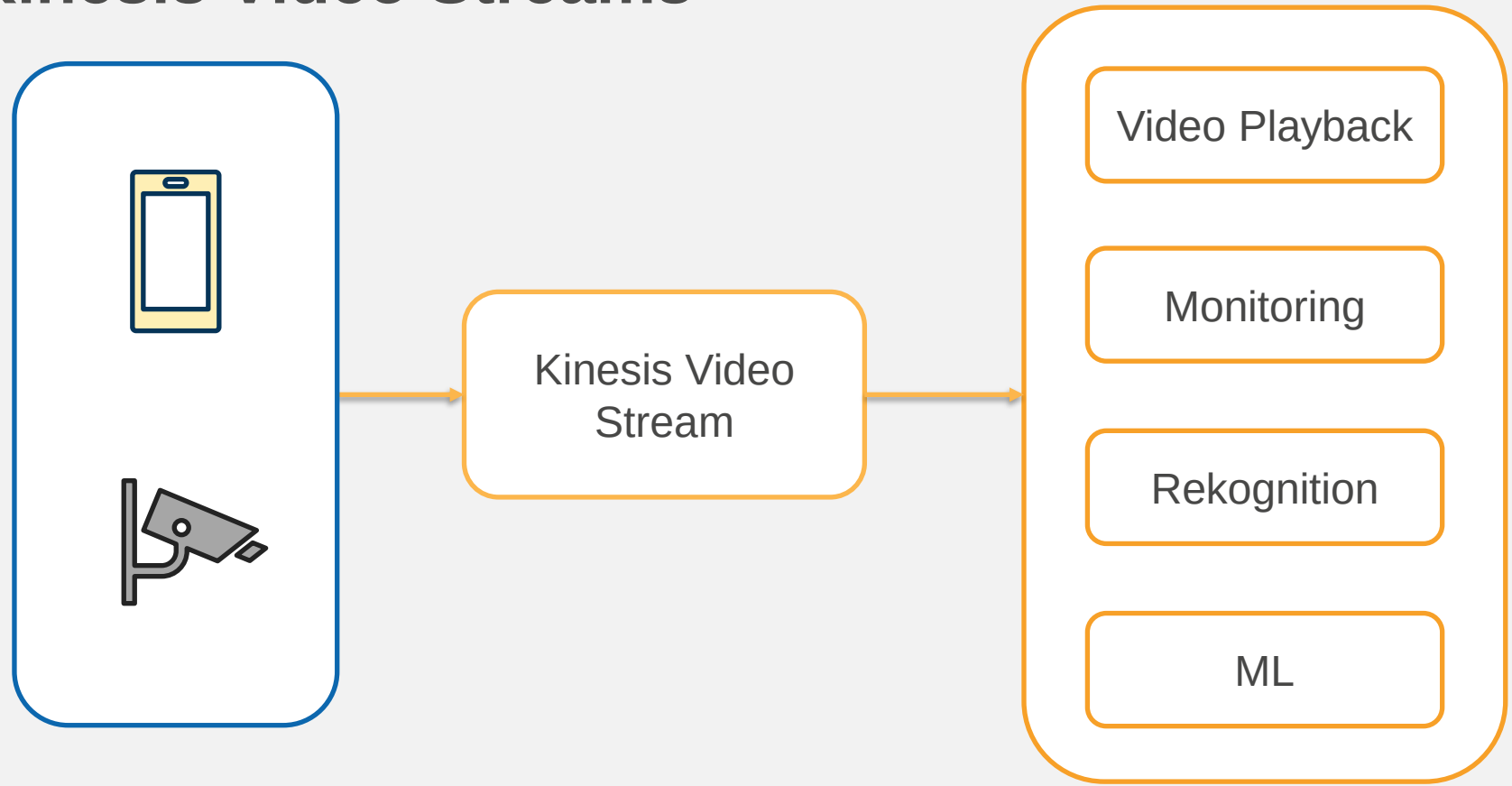
“Handle any amount of streaming data from hundreds of thousands of sources with very low latencies”

Reference: Amazon Kinesis, https://aws.amazon.com/kinesis/

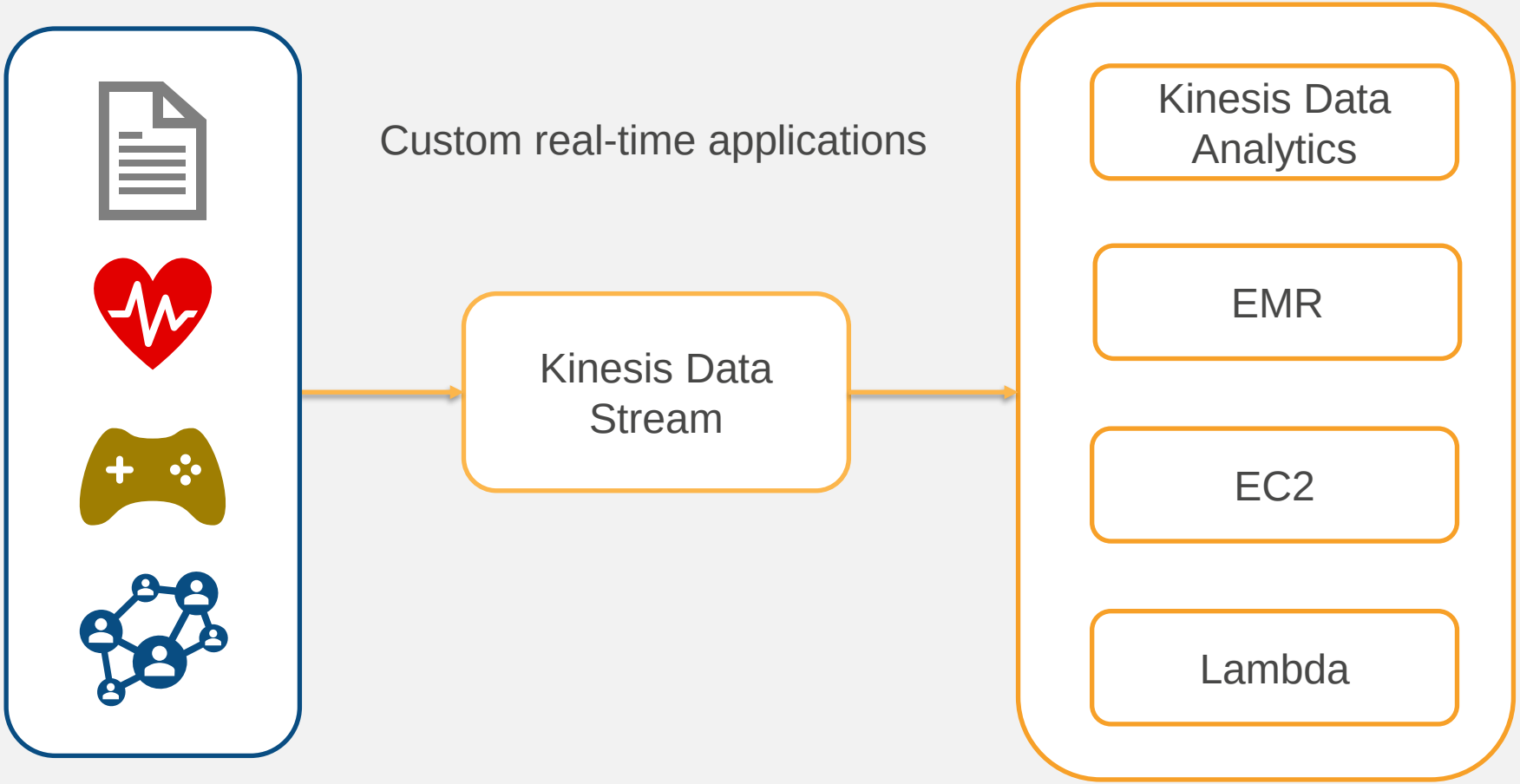
Kinesis Family

Service	Purpose	Use
Video Streams	Capture and Analyze Video Stream	Security Monitoring, Video Playback, Face detection
Data Streams	Capture and Analyze Data Stream	Custom real-time application
Firehose	Capture and Deliver Data Stream to AWS Data Stores	Use Existing BI tools for Streaming Data: S3, Redshift, ElasticSearch, Splunk
Data Analytics	Analyze Data Stream with SQL and Java	Real-time analytics, Anomaly detection

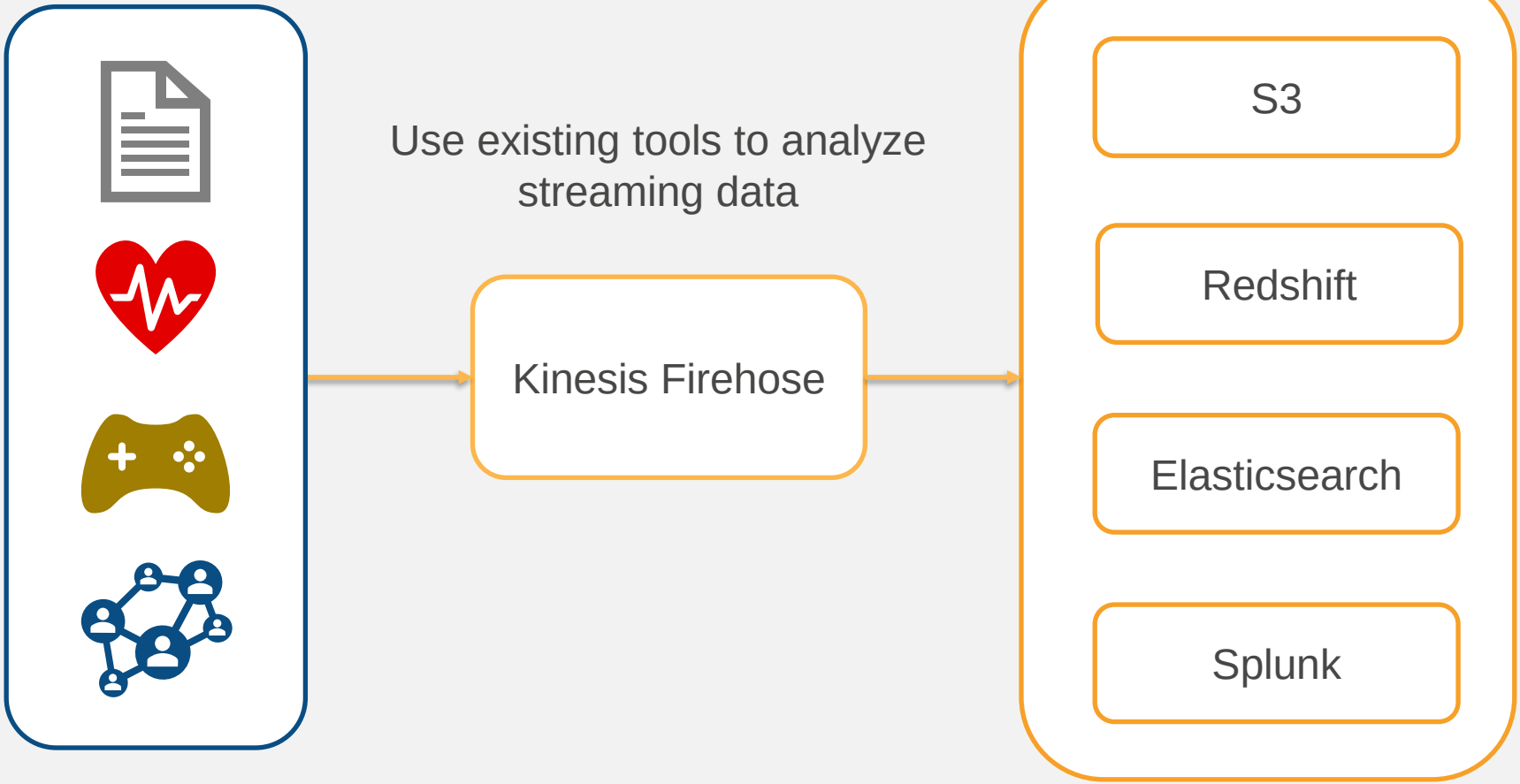
Kinesis Video Streams



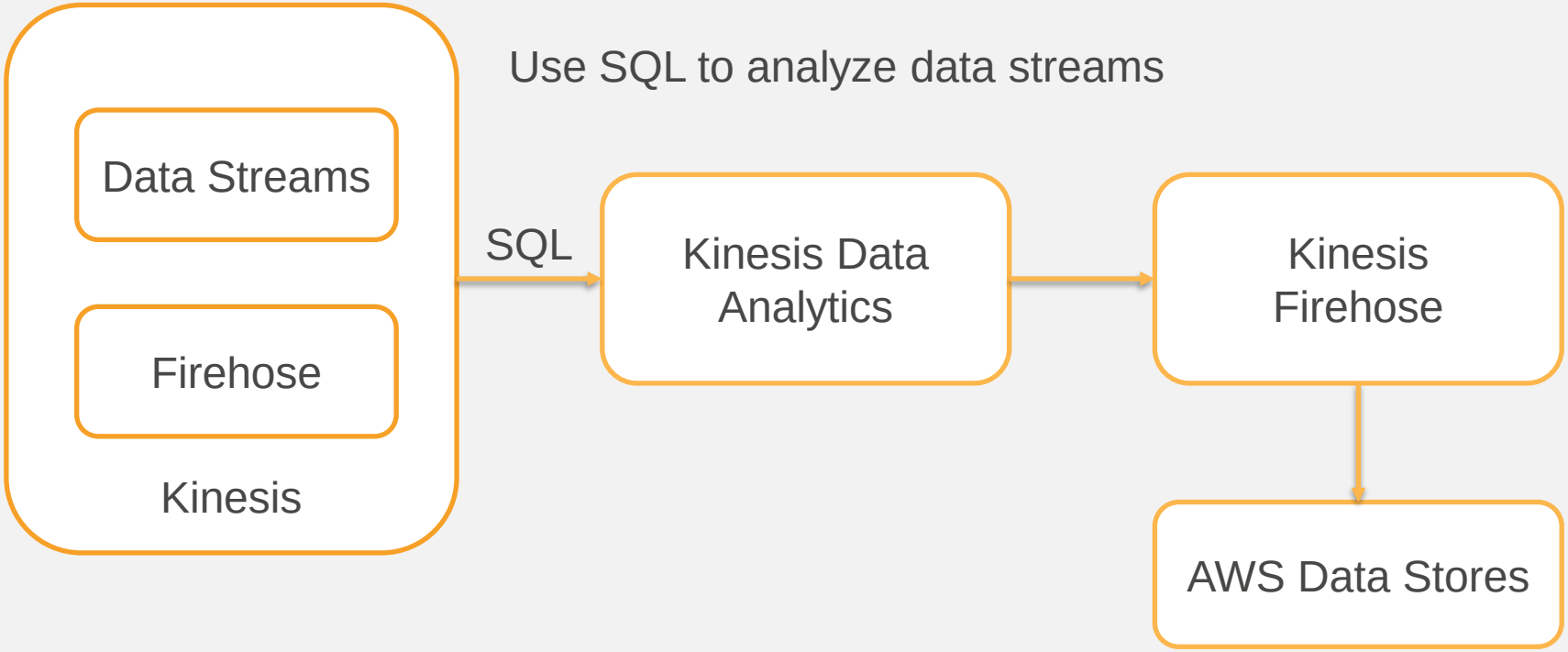
Kinesis Data Streams



Kinesis Data Firehose



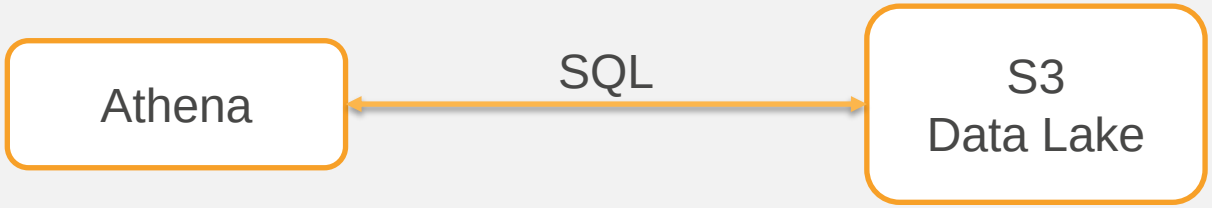
Kinesis Data Analytics



In-place Querying

- Directly query data in S3 using SQL
- Athena, Redshift Spectrum

Athena In-place Query

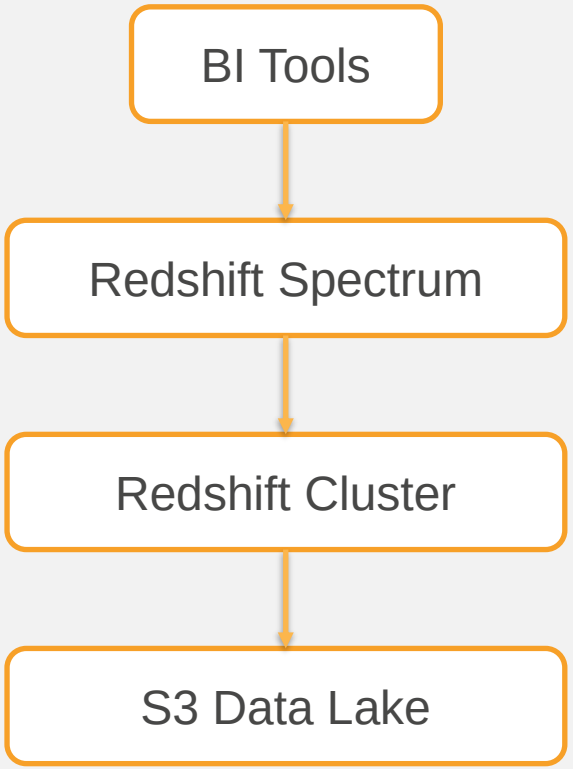


- Directly run SQL query against files in S3
- No need to provision servers (serverless)
- Charges based on amount of data scanned
- Support for popular file formats: CSV, JSON, Parquet, ORC, Avro

“This makes vast amount of unstructured data accessible to any data lake user who can use SQL.”

Reference: Data Lake on AWS,
https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/building-data-lake-aws.html

Redshift Spectrum In-place Query



- Sophisticated Query Optimization
- Distribute query across multiple nodes
- Redshift Data Warehouse SQL Syntax
- Use with existing BI tools
- Query can span Redshift Tables and S3 Data Lake

AWS Recommendations

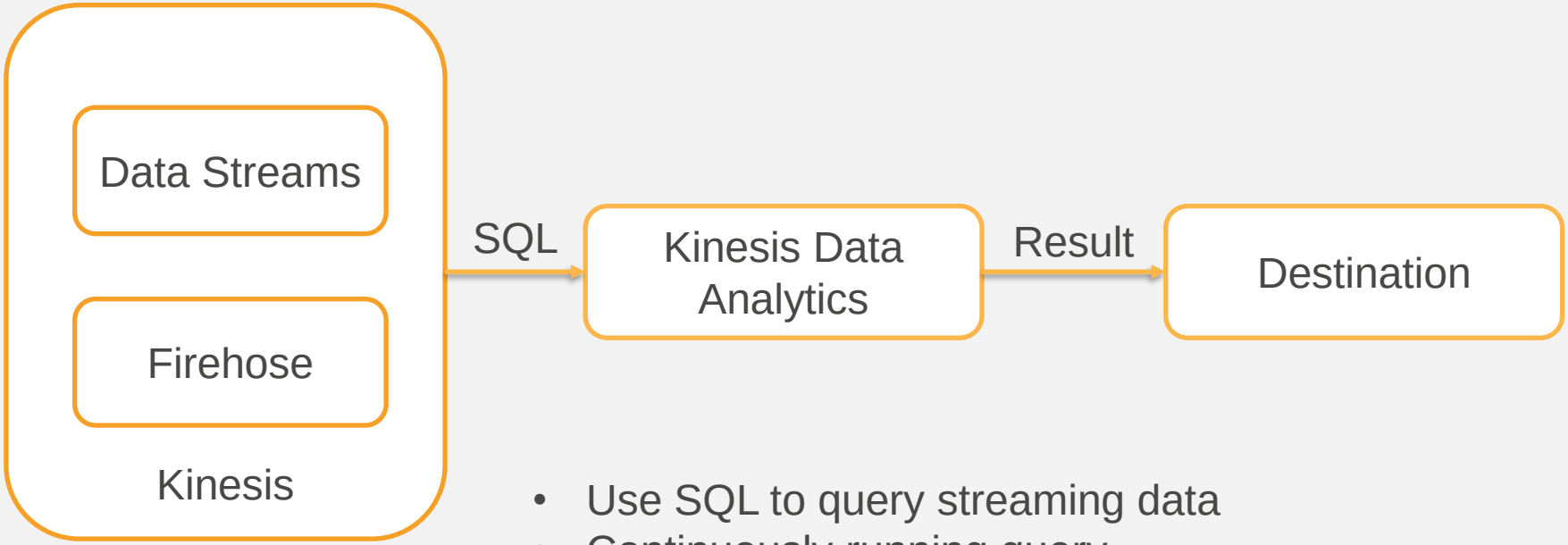
Athena

- Ad-hoc data discovery and SQL querying

Redshift Spectrum

- More complex queries
- Large number of users

Streaming Query Kinesis Data Analytics



- Use SQL to query streaming data
- Continuously running query
- Sends matching results to configured destination

Analytics Tools

- Data Lake needs to support current and future tools
- S3 is a popular cloud service
- Several third-party tools natively support S3

Broader Analytics Portfolio

Service	Purpose	Use
Amazon EMR	Hadoop Ecosystem tools	Process data in S3 using Spark, Hive, Pig, Hbase, TensorFlow, MxNet and so forth
SageMaker	Machine Learning	Train models with data in S3 Generate real-time and Batch predictions
Artificial Intelligence	Video, Image, Natural language processing	Analyze audio, video, image, text data in S3

Broader Analytics Portfolio

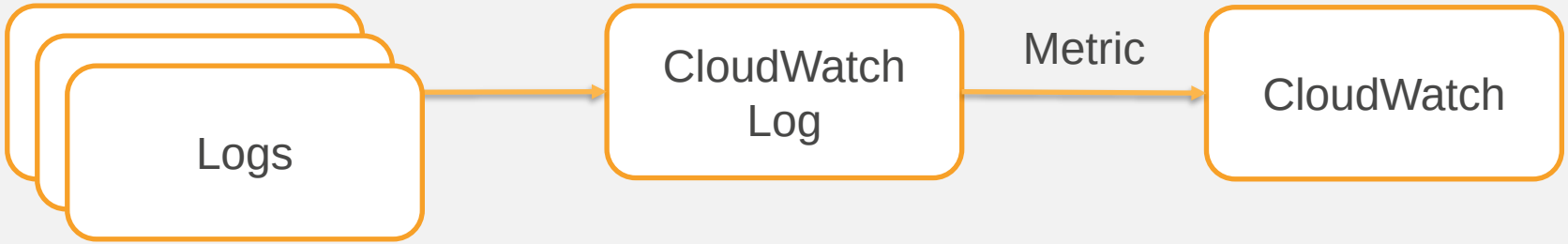
Service	Purpose	Use
QuickSight	Business Intelligence	Create Interactive Dashboard Supports Athena, Redshift, Relational database
Redshift	Petabyte Scale Data warehouse (Columnar Storage)	Load data to tables from S3 - local querying Query S3 directly using Redshift Spectrum
Lambda	Business Logic (Function as a service)	Serverless code execution Trigger-based function invocation

Monitoring and Optimization

Monitoring

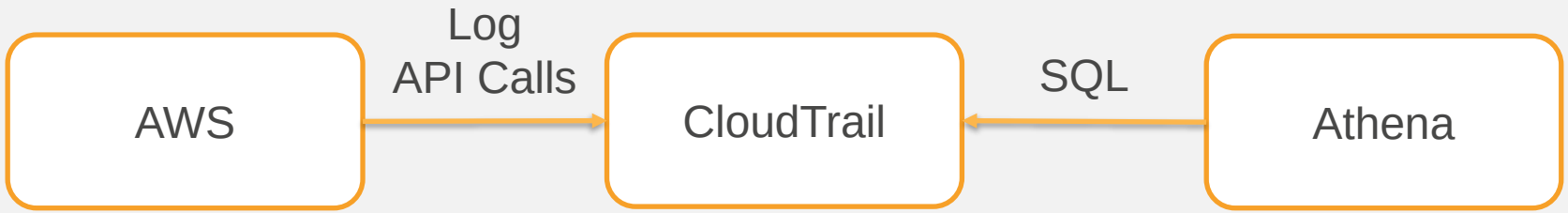
Service	Purpose	Use
CloudWatch	Monitoring	Monitor your resources Configure Alarms to alert Take automated action
CloudWatch Log	Log Monitoring	Consolidate log files and monitor
CloudTrail	Audit Trail	Log all activities and who performed those actions Useful for investigation, compliance monitoring

CloudWatch Log



- Consolidate Logs
- Monitor

CloudTrail



- Log all activities and who performed those actions
- Useful for investigation, compliance monitoring

Optimization

“Data storage is often a significant portion of the costs associated with a data lake.”

Reference: Data Lake on AWS, https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/building-data-lake-aws.html

Cost Optimization

S3 Lifecycle Management

S3 Storage Class Analysis

Intelligent Tiering

Glacier and Glacier Deep Archive

Data Formats

S3 Lifecycle Management

	S3 Standard	S3 Infrequent Access	Glacier
Cost - 500GB per month	USD 11.50	USD 6.25	USD 2.00
Retrieval Fee	-	Per GB	Per GB
Suitable for	Frequently Accessed	Rarely Accessed	Archival and Backup
First byte latency	Immediate	Immediate	Restore can take minutes to hours

Lifecycle Storage Tiering and Expiration

Object Age

Name and Folder Structure

S3 Object Tags

S3 Lifecycle Management

	S3 Standard	S3 Infrequent Access	Glacier
Cost - 500GB per month	USD 11.50	USD 6.25	USD 2.00
Retrieval Fee	-	Per GB	Per GB
Suitable for	Frequently Accessed	Rarely Accessed	Archival and Backup
First byte latency	Immediate	Immediate	Restore can take minutes to hours

Storage Class Analysis

“One of the challenges of developing and configuring lifecycle rules for the data lake is gaining an understanding of how data assets are accessed over time.”

Reference: Data Lake on AWS, https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/building-data-lake-aws.html

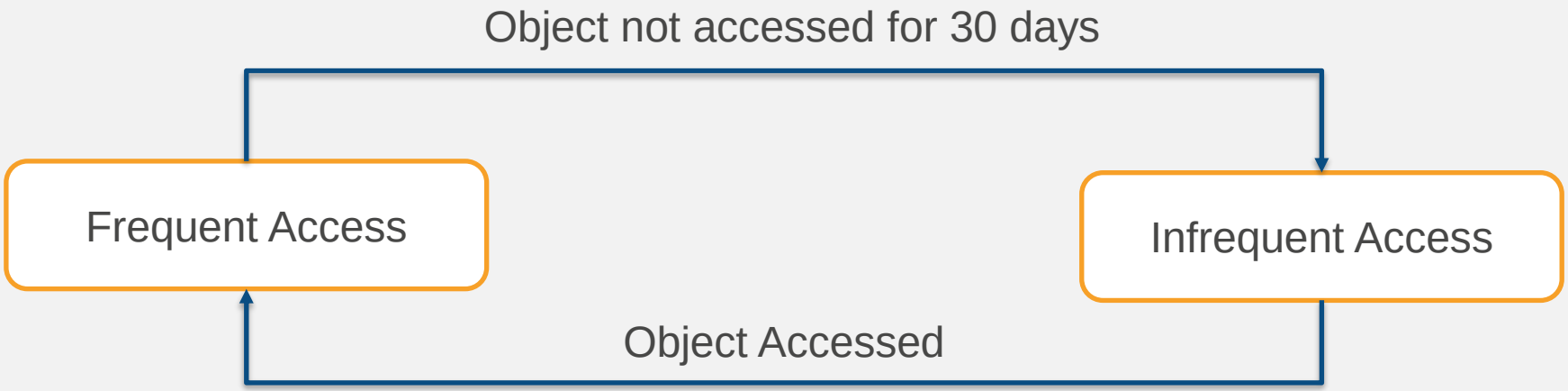
Storage Class Analysis

“This new Amazon S3 analytics feature observes data access patterns to help you determine when to transition less frequently accessed STANDARD storage to the STANDARD_IA storage class”

Reference: S3, https://docs.aws.amazon.com/AmazonS3/latest/dev/analytics-storage-class.html

S3 Intelligent Tiering

Objects are automatically moved between frequent access and infrequent access storage class



Glacier, Glacier Deep Archive

Service	Purpose	Use
Glacier	Archive and Backup	Cost: USD 2.00 for 500 GB/Month Durability: 11 9's Retrieval Time: Minutes to Hours Vault Lock to prevent future edits
Glacier Deep Archive	Archive and Backup	Cost: USD 0.50 for 500 GB/Month Durability: 11 9's Retrieval Time: 12 to 48 hours Vault Lock to prevent future edits

Cost Optimization

S3 Lifecycle Management

S3 Storage Class Analysis

Intelligent Tiering

Glacier and Glacier Deep Archive

Data Formats

Security and Protection

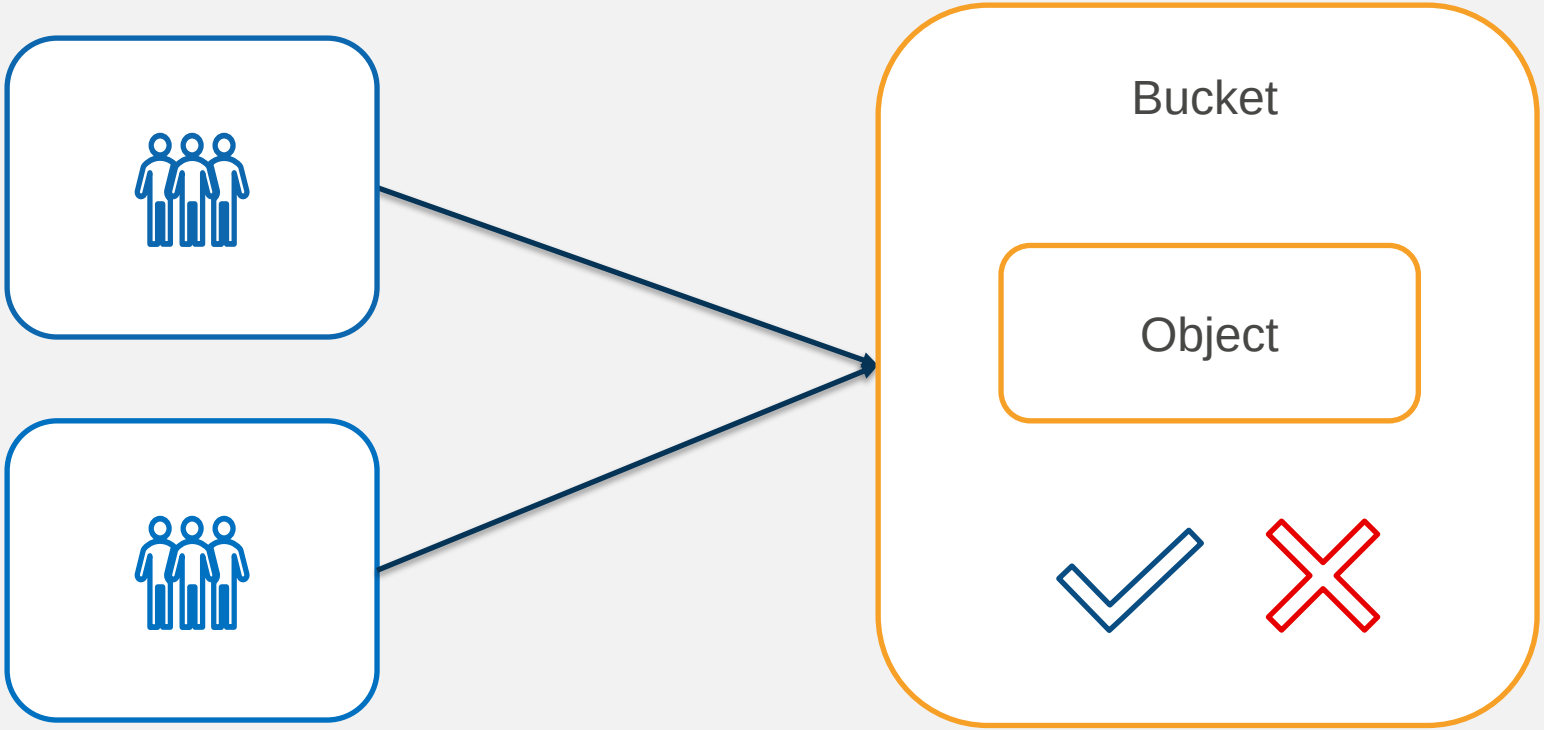
- Data Lake is centralized
- Consolidates all data in one place
- Protecting and managing data is very important

S3 Access Management

Resource-based Policy

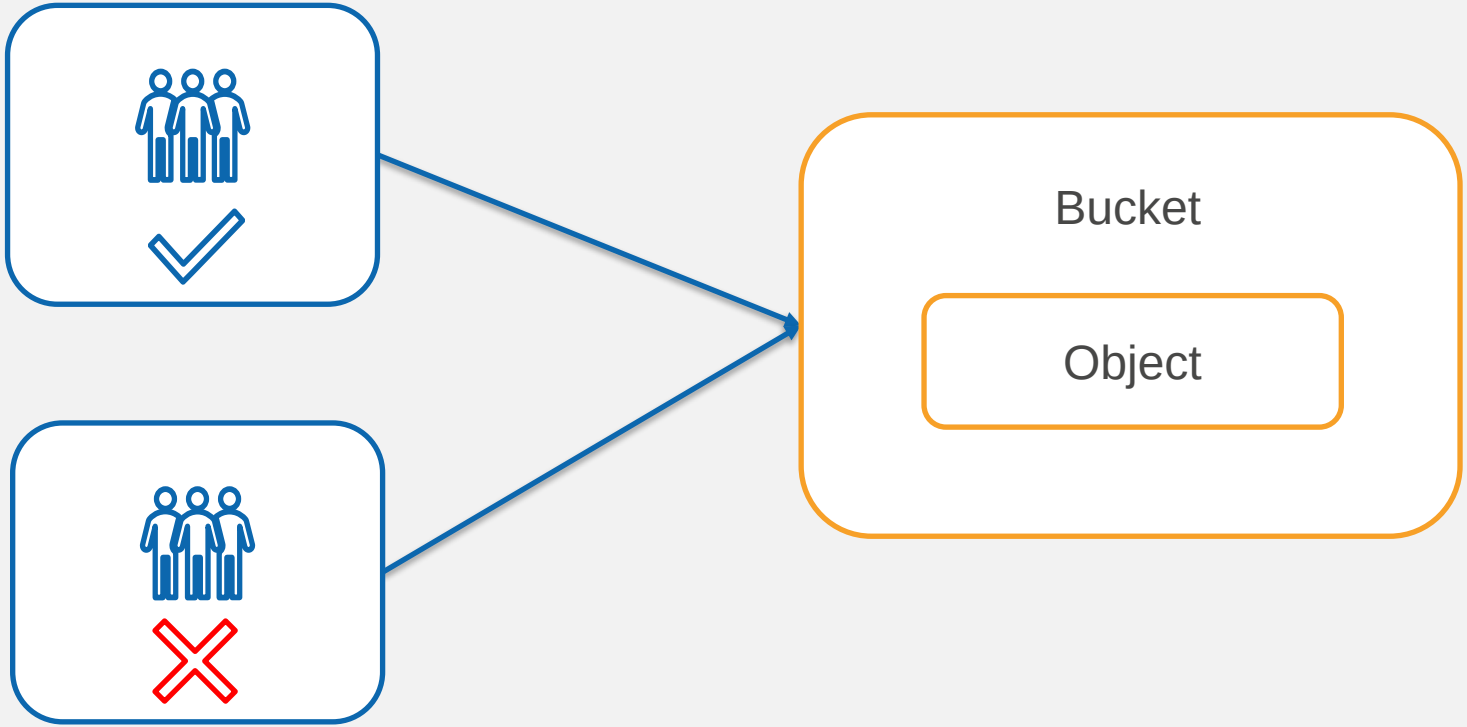
User-based Policy

S3 Resource Based Policy



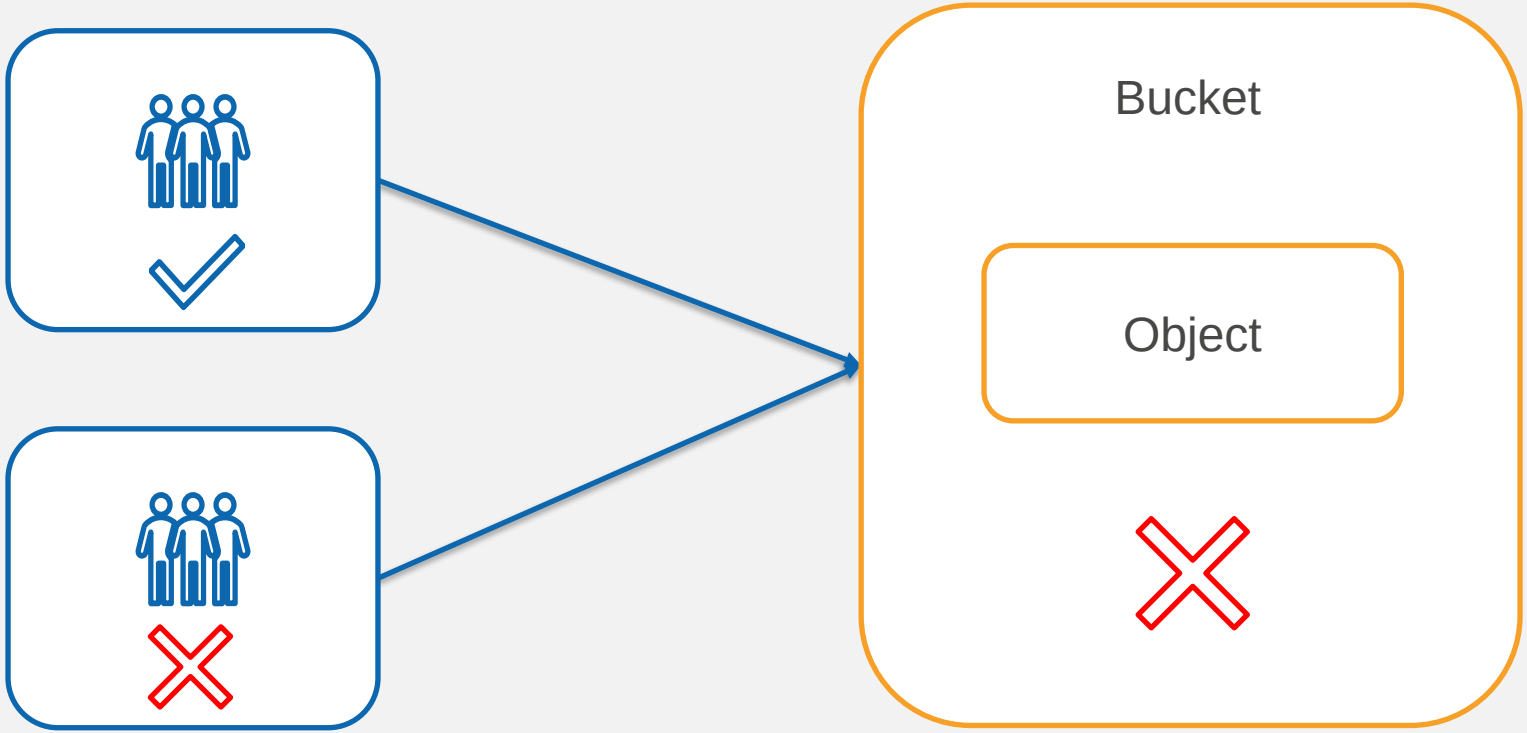
Permissions are embedded as part of Bucket and Object

S3 User Based Policy



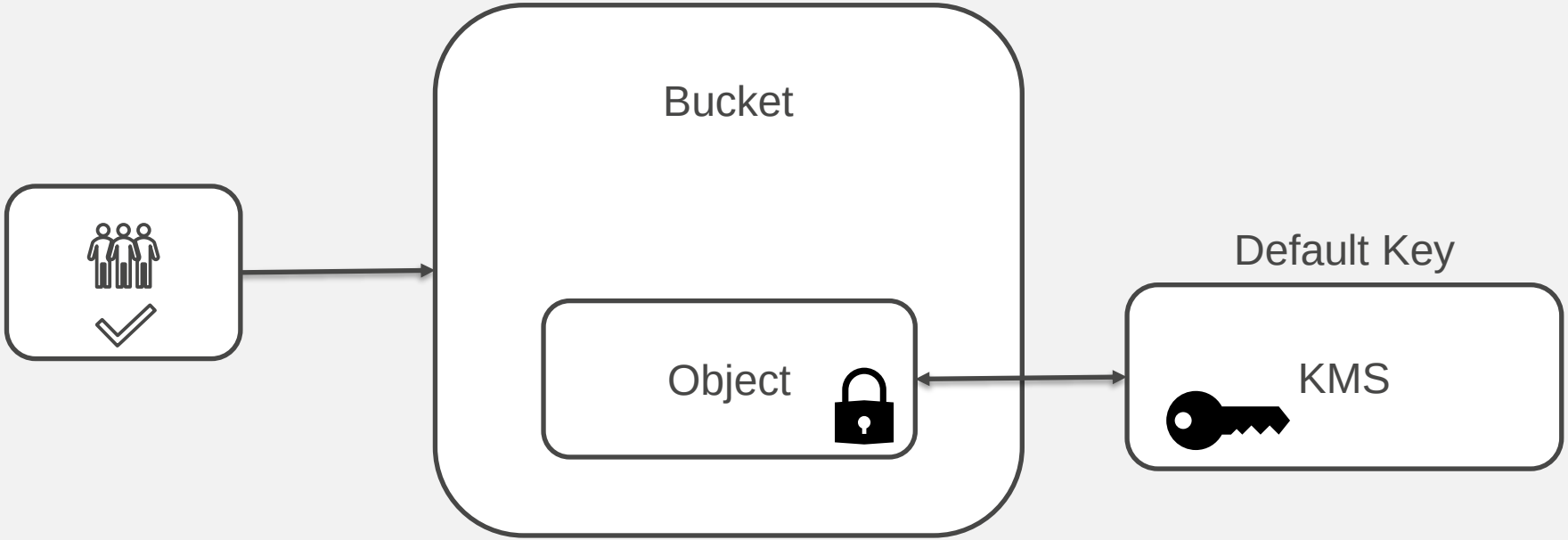
Permissions are granted to Users and Groups

S3 User and Resource Based Policy



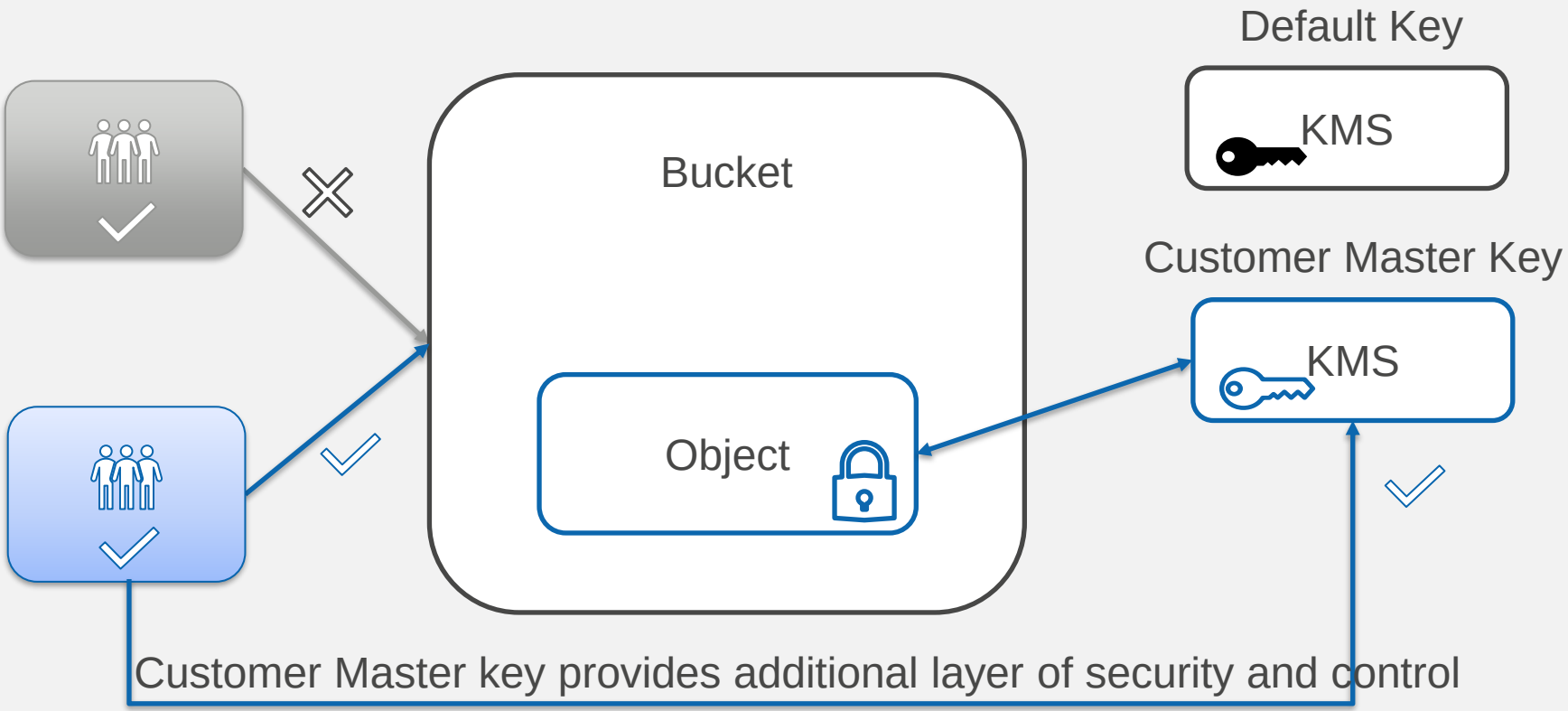
Deny all access that do not originate from on-premises

S3 Data Encryption – Default Key



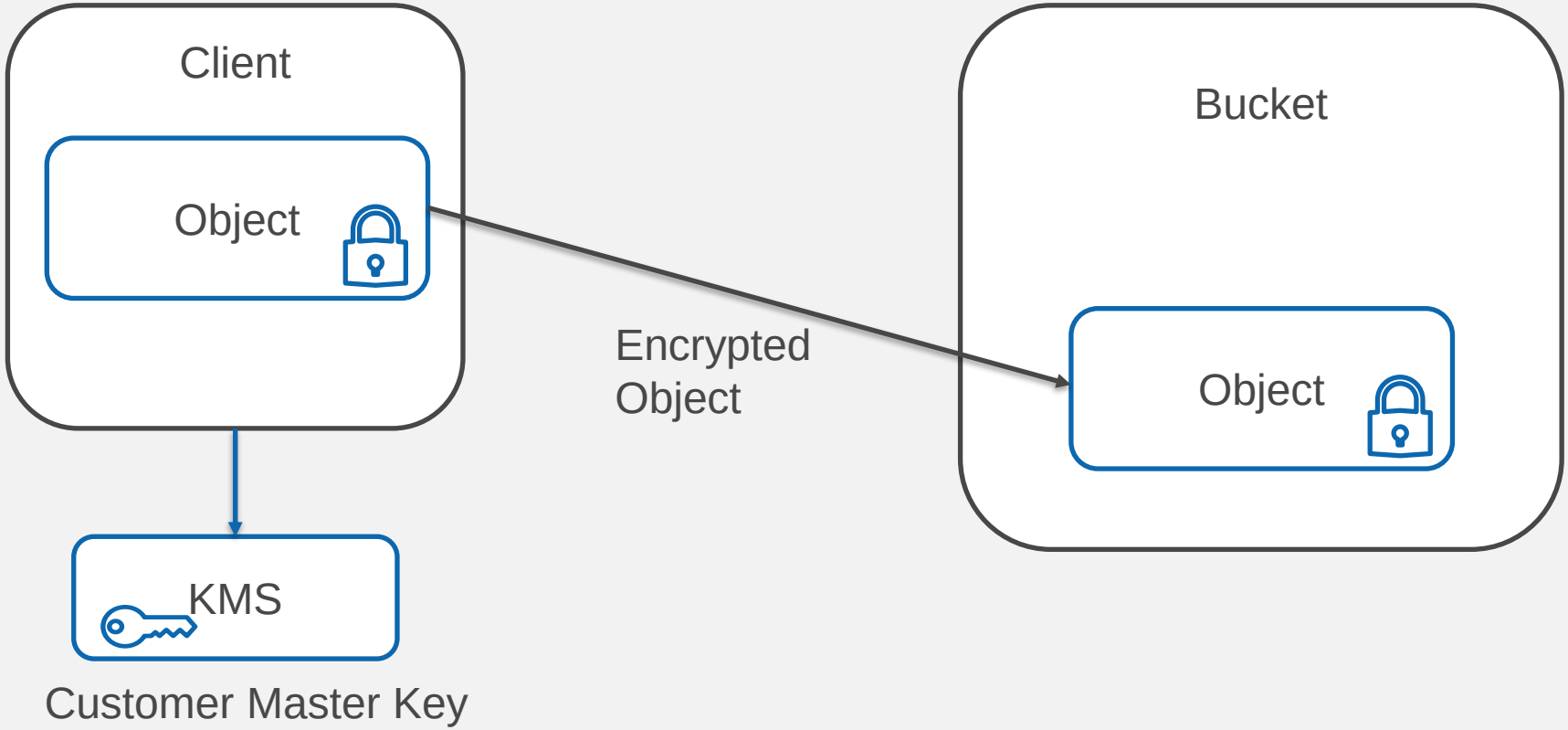
With default key, S3 automatically decrypts object for any user who is allowed access to the bucket or object

S3 Data Encryption – Customer Master Key (CMK)



S3 Client-Side Encryption – Customer Master Key (CMK)

Object encryption and decryption is client responsibility



Protection

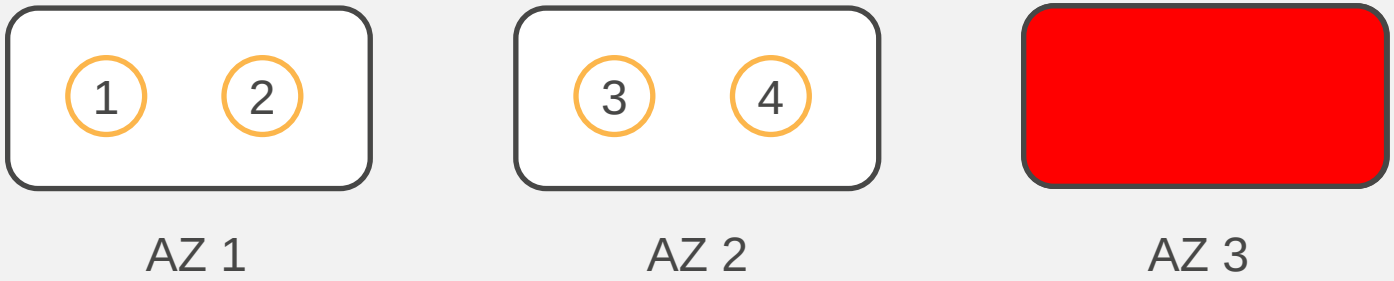
“A data lake must protect data against corruption, loss, accidental or malicious overwrites, modifications, and deletions.”

Reference: Data Lake on AWS, https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/building-data-lake-aws.html

S3 Durability

S3 Durability 99.999999999% (11 9’s)

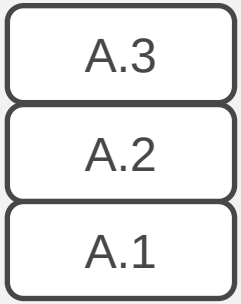
Measure of protection against data loss and corruption



S3 Versioning

Protection against accidental and malicious deletes

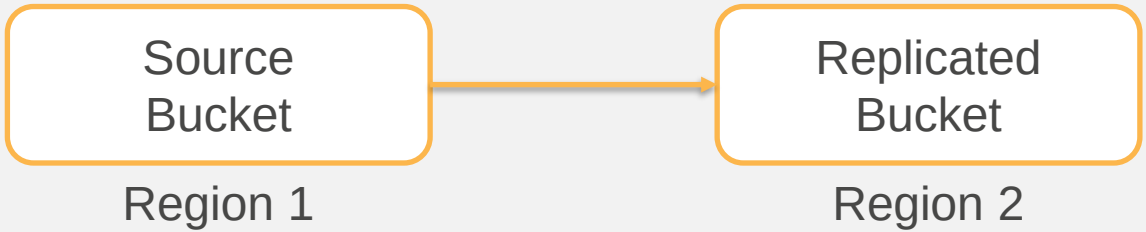
S3 maintains versions of objects



Configure Lifecycle rules for current and previous versions

Multi-Factor Authentication (MFA) for additional layer of authentication

S3 Cross Region Replication (CRR)



Replicate S3 bucket in another region for Disaster Recovery

Automatic and continuous replication

Deletes are not replicated

S3 Object Tagging

Tags are additional meta-data that you can add to Object

Define access control policies based on tags

ALLOW Classification=PHI

DENY Classification=PHI

Object

Classification=PHI

Security and Protection

AWS and S3 provides several features to secure and protect your data

As part of Shared Responsibility Model, Customers are responsible for configuring these security features according to their organization needs

Summary

S3 Data Lake Architecture provides a template on how to design and run a data lake for your organization

- Ingest and Store Data
- Discover and Make data usable
- Transform data
- Analyze data in-place
- Future proofing
- Monitor
- Optimize
- Security and Protection

AWS Housekeeping

Account Setup, Support

Chandra Lingam

Cloud Wave LLC

Hands-on Experience

“Gain free, hands-on experience with the AWS platform, products, and services.”

https://aws.amazon.com

Three Types of Offers

- Always Free
- 12 months free and
- Trials

https://aws.amazon.com/free

https://aws.amazon.com/free/free-tier-faqs/

Billing

You are billed standard pay-as-you-go rates when -

- Usage exceeds free tier limits or
- Term expires

AWS requires a Credit or Debit card to sign-up for an account

Billing

Billing Alerts

Dashboard

- Free-Tier usage
- Monthly Charge Summary
- Itemized charges
- Past Bills and Usage

Free Support Center

- Account Issues
- Billing Enquires
- Service Limit Changes

- Technical Support – Part of paid plans

Service Quotas

Amazon SageMaker quotas for new accounts might be different from the default quotas listed here. If you receive an error that you've exceeded your quota, contact customer service to request a quota increase for the resources you want to use.

On-demand and Spot instance quotas are tracked and modified separately. For example, with the default quotas, you could run up to 20 training jobs with on-demand ml.m4.xlarge instances and up to 20 training jobs with Managed Spot ml.m4.xlarge instances simultaneously. Request quota increases for on-demand and spot instances separately.

Amazon SageMaker Notebooks

Resource	Default
ml.t2.medium instances	20
ml.t2.large instances	20

https://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html
https://docs.aws.amazon.com/general/latest/gr/sagemaker.html

Billing Alert - Best Practices

- Enable billing access to authorized users in your account
- Configure Free Tier Alerts
- Enable billing data collection for CloudWatch monitoring
- Configure Billing Alarms with CloudWatch
- Configure AWS Budget

User Accounts

Account/User	Purpose
Root Account (Highest Privilege)	Responsible for paying bills. Sign-in at https://aws.amazon.com/ Enable MFA
my_admin	IAM User with administrative access Sign-in Link https://<AccountId>.signin.aws.amazon.com/console https://<Alias>.signin.aws.amazon.com/console

MFA Setup

Recommended for root account

Login credentials + one-time passwords

- Google Authenticator App or similar

Summary

Account Setup

Types of free offers

Billing Dashboard and Alerts

Support

Lab – AWS Account Setup

Create AWS Account

Configure user and permission

Lab – S3

- Storage Class
- S3 Versioning
- Age Based Retention
- Storage Tiering
- Replication
- Encryption with SS3-S3 and KMS

Lab – Glue Data Catalog and Athena

In-place Querying of files stored in S3

- Store file in S3
- Collect metadata with Glue Crawler
- Run Query using Athena

Example Queries (Lab)

- Query first 10 rows

```
SELECT * FROM "demo_db"."iris_csv" limit 10;
```

- Query for a specific class

```
SELECT * FROM "demo_db"."iris_csv"
WHERE class = 'Iris-setosa';
```

- Query by wildcard

```
SELECT * FROM "demo_db"."iris_csv"
where class like '%setosa%';
```

- Get a count

```
SELECT count(*) AS COUNT FROM "demo_db"."iris_csv"
```

- Compute new columns

```
SELECT sepal_length, sepal_width,
       sepal_length * sepal_width as sepal_area
FROM "demo_db"."iris_csv";
```

Lab – Glue ETL

Use Glue ETL to convert files to Parquet format

- Glue automates process of ETL script generation, scheduling and execution
- Glue ETL provisions required Apache Spark infrastructure to run the job

Example Queries - Parquet (Lab)

- Query Iris Parquet Table

```
SELECT sepal_length, sepal_width,
       sepal_length * sepal_width as sepal_area
FROM "demo_db"."iris_parquet" limit 10;
```

Lab – Customer Review

Query Amazon Customer Reviews Public Dataset using Athena

- Create table definition (instead of using Glue Crawler)
- Update catalog with partition
- Query using Athena

Reference:

https://s3.amazonaws.com/amazon-reviews-pds/readme.html

https://registry.opendata.aws/

Example Queries – Customer Review

- Highly Rated Books

```
SELECT product_title, star_rating,review_body
FROM  "demo_db"."amazon_reviews_parquet"
WHERE product_category = 'Books'
and star_rating > 3
limit 10;
```

- Book Reviews for specified book title pattern

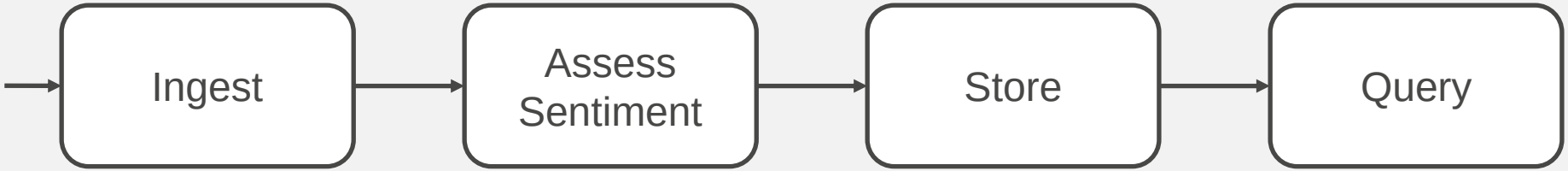
```
SELECT product_title, star_rating,review_body
FROM  "demo_db"."amazon_reviews_parquet"
WHERE product_category = 'Books'
and product_title like 'Harry Potter%'
and star_rating > 3
limit 100;
```

Lab – Sentiment of the Customer Review

Find Sentiment of the customer review using Comprehend AI Service

With Athena, Query the reviews using sentiment

Lab – Serverless Customer Review Solution



Lab – Serverless Customer Review Solution