# Data Lake in AWS

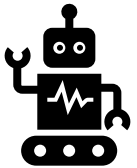
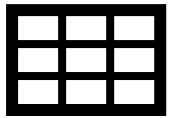
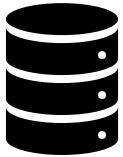
## Storage, Data Governance, Analytics

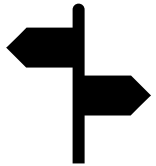
Chandra Lingam

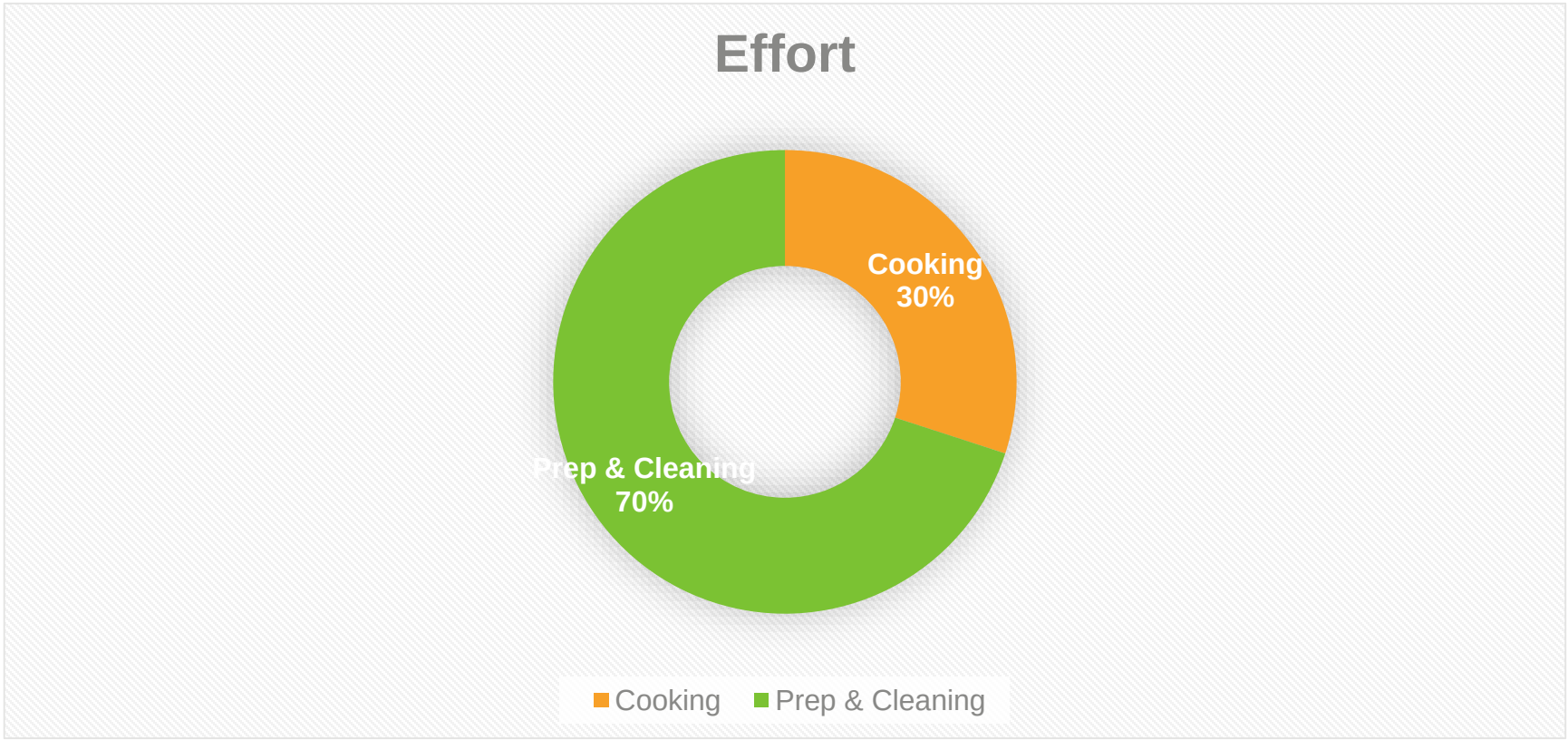
Cloud Wave LLC

# Data Lake Motivation



# Data Lake

*Streamline Data Management*

# Date Lake Vs Data Warehouse

"A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose."

Reference: Talend, https://www.talend.com/resources/data-lake-vs-data-warehouse/

# Data Lake

“A data lake is a centralized repository that allows you to migrate and store all structured and unstructured data at unlimited scale…”

Reference: AWS, https://aws.amazon.com/products/storage/data-lake-storage/infographic/

# 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

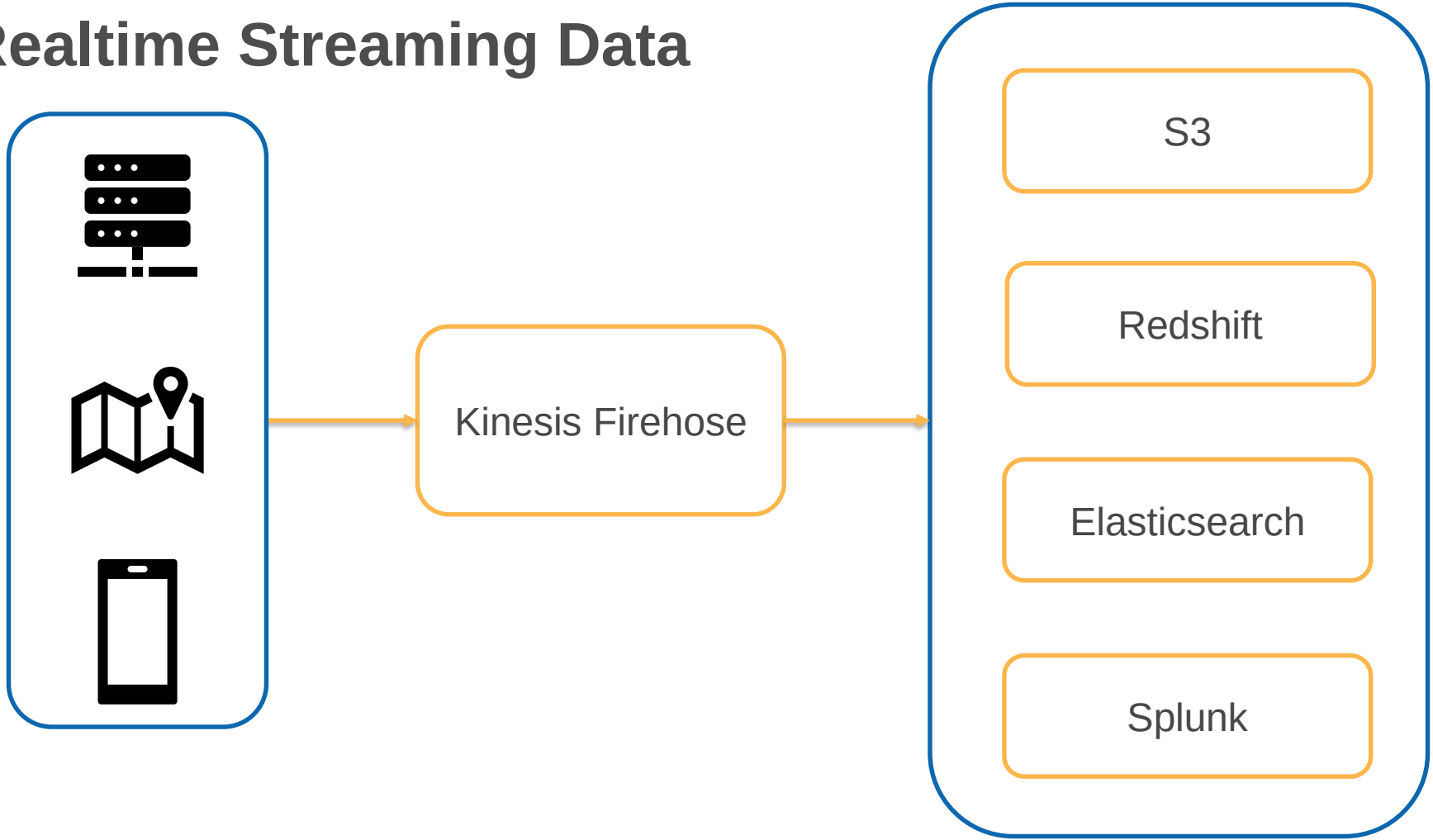
# Storage

| Service | Purpose | Use |
| --- | --- | --- |
| S3 | Storage (Exabyte scale) | Object Storage to store and retrieve any amount of data.<br>Cost effective with 99.999999999% (11 9s) of durability<br>Object Life cycle management and Tiered storage based on access patterns |
| Glacier | Backup and Archiving (Exabyte scale) | Backup and Long term archival (multi-year) at extremely low cost and 11 9s durability. |

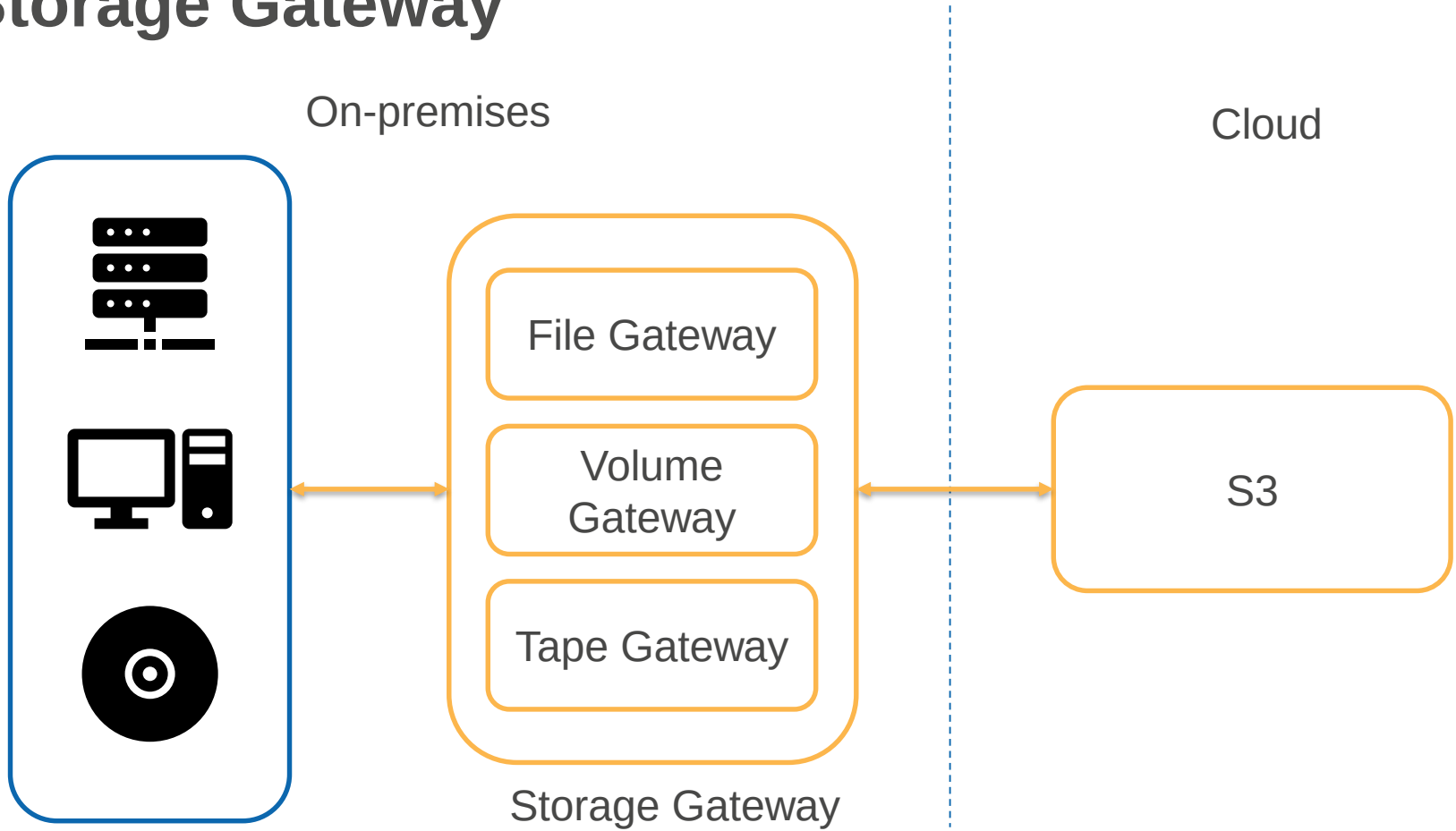
# Ingestion

| Service | Purpose | Use |
|---|---|---|
| Kinesis Firehose | Real-time Streaming Data Ingestion | Capture and deliver real time streaming data directly to S3 and other destinations like Redshift, Elasticsearch, Splunk |
| Storage Gateway | Hybrid Cloud Storage | Integrate legacy on-premises data processing platforms to S3 Data Lake. Files, volumes and tape backups |
| Snowball, Snowmobile | Very Large Data Transfer | Appliance to move petabytes to exabytes of data to AWS cloud at one-fifth the cost of moving over internet |
| SDK, CLI and more | Transfer data to S3 | Software driven infrastructure - easy to integrate with variety of tools |

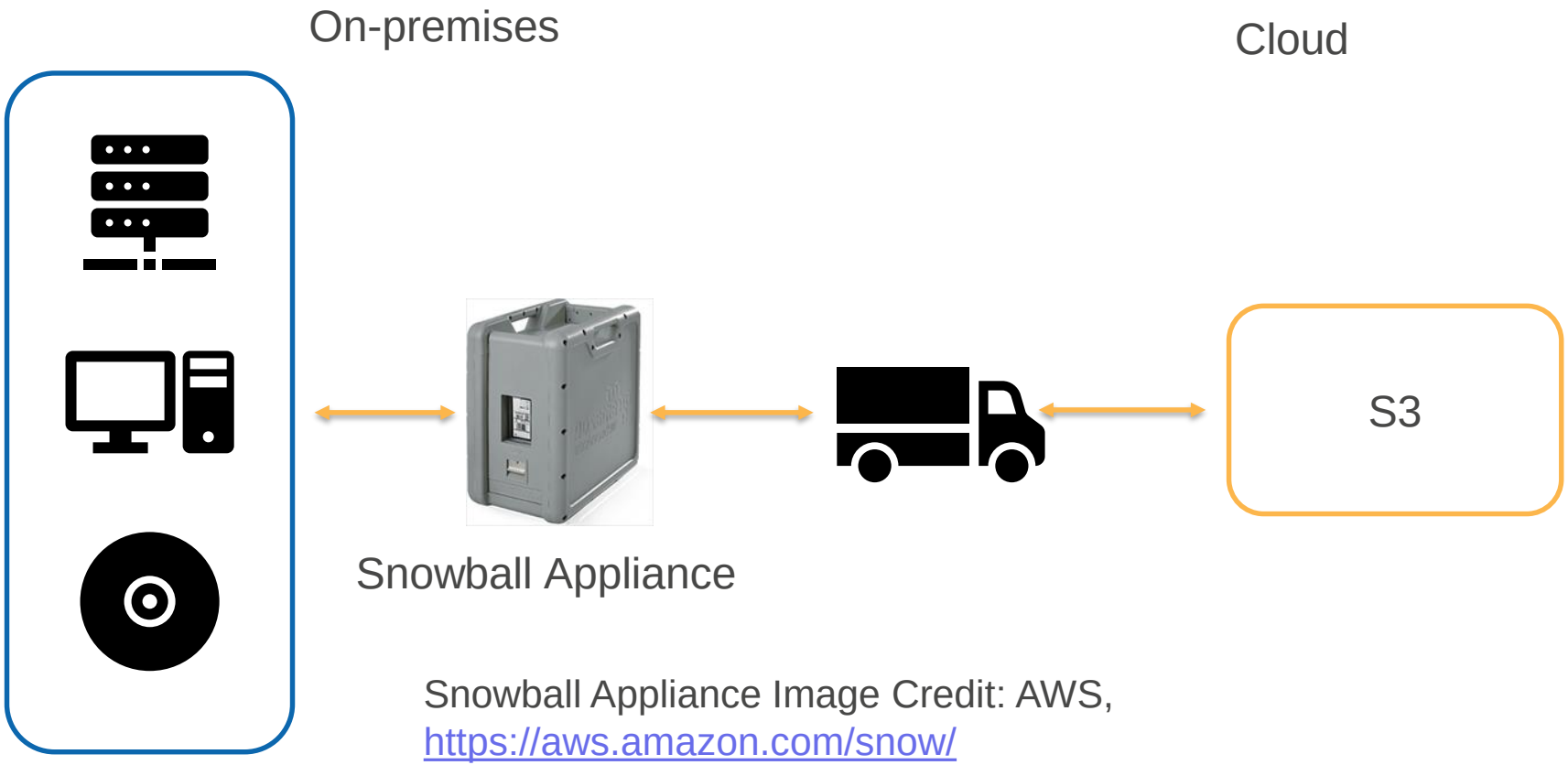
# Realtime Streaming Data



# Storage Gateway



# Snowball

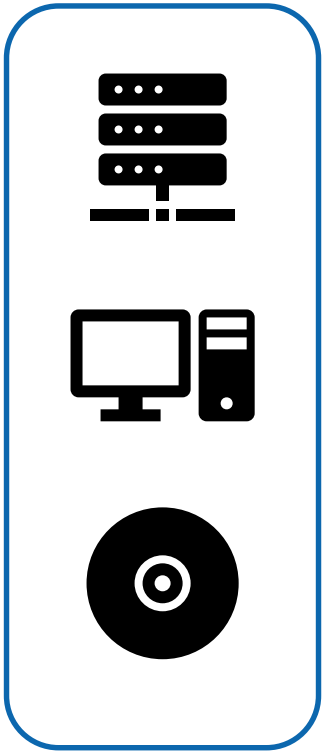



Snowball Appliance Image Credit: AWS, https://aws.amazon.com/snow/

# Snowmobile

On-premises

Cloud



S3

Snowmobile Container

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

# Data Catalog

| Service | Purpose | Use |
|---|---|---|
| Do-it-yourself | Comprehensive Data Catalog | Make data discoverable and usable.<br><br>Use services like S3, Lambda, Elasticsearch, DynamoDB to maintain metadata |
| Glue | Data Catalog (Metadata repository) | Make data discoverable and usable.<br><br>Automatically crawl and collect metadata from S3, DynamoDB and any other database that supports JDBC connectivity |

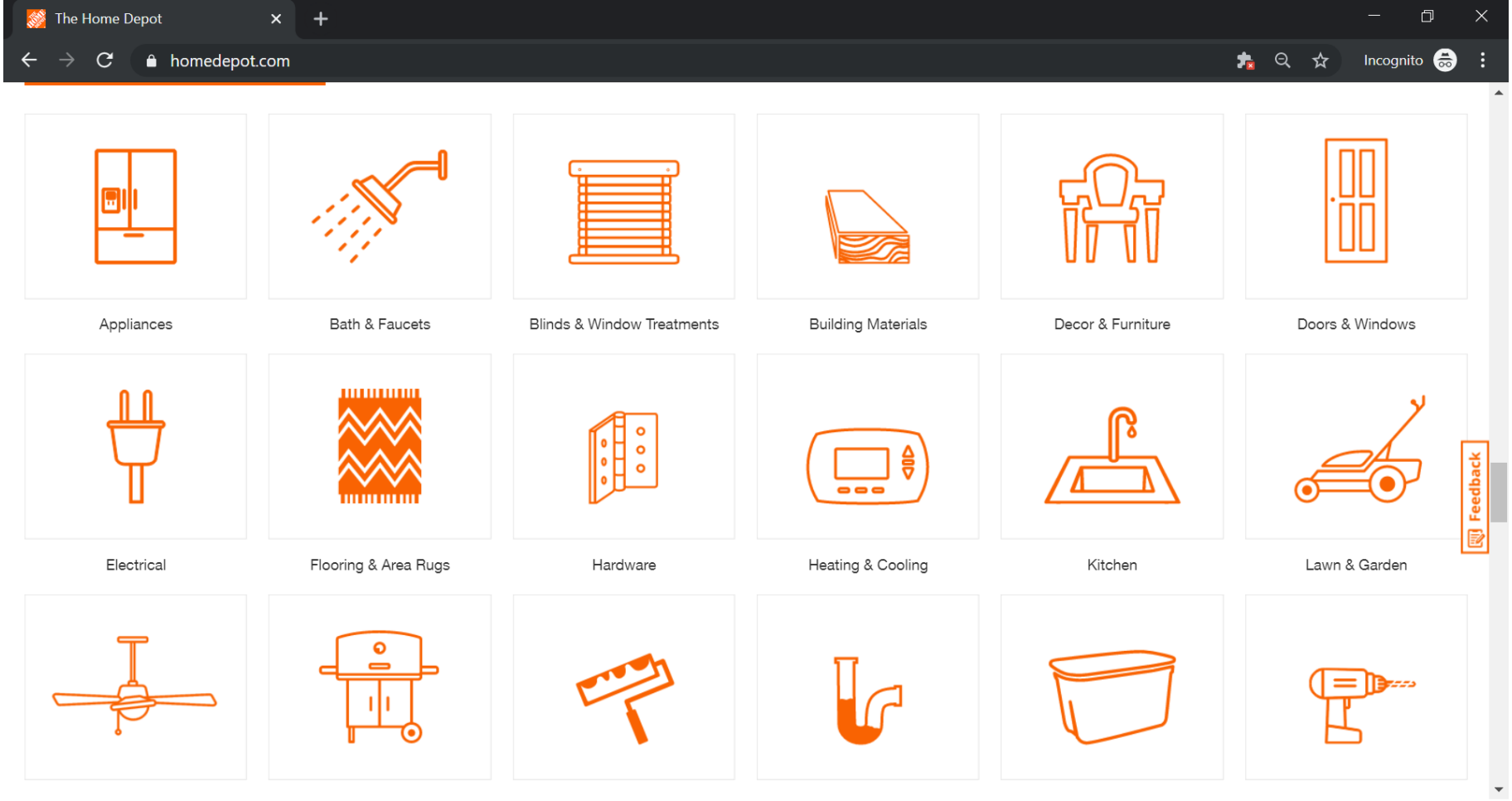



Image Credit: HomeDepot

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

# Amazon Kinesis

*Collect, Process, Analyze Data Streams*

# 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/

# Stream Vs. Batch Processing

What is stream processing?

How does it differ from batch processing?

# Streaming Data

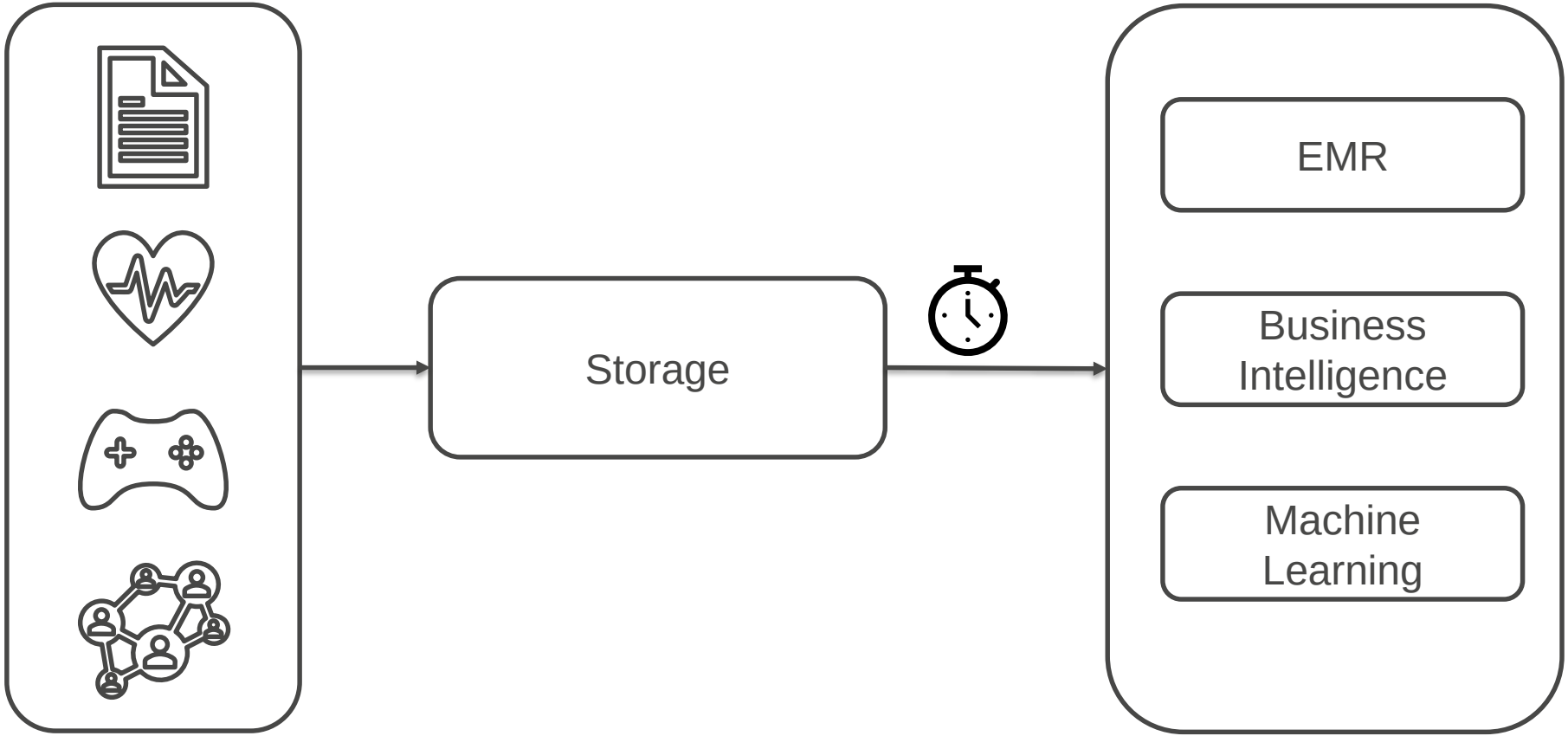
Thousands of sources

Generated Continuously

Small Payloads




# Batch Processing

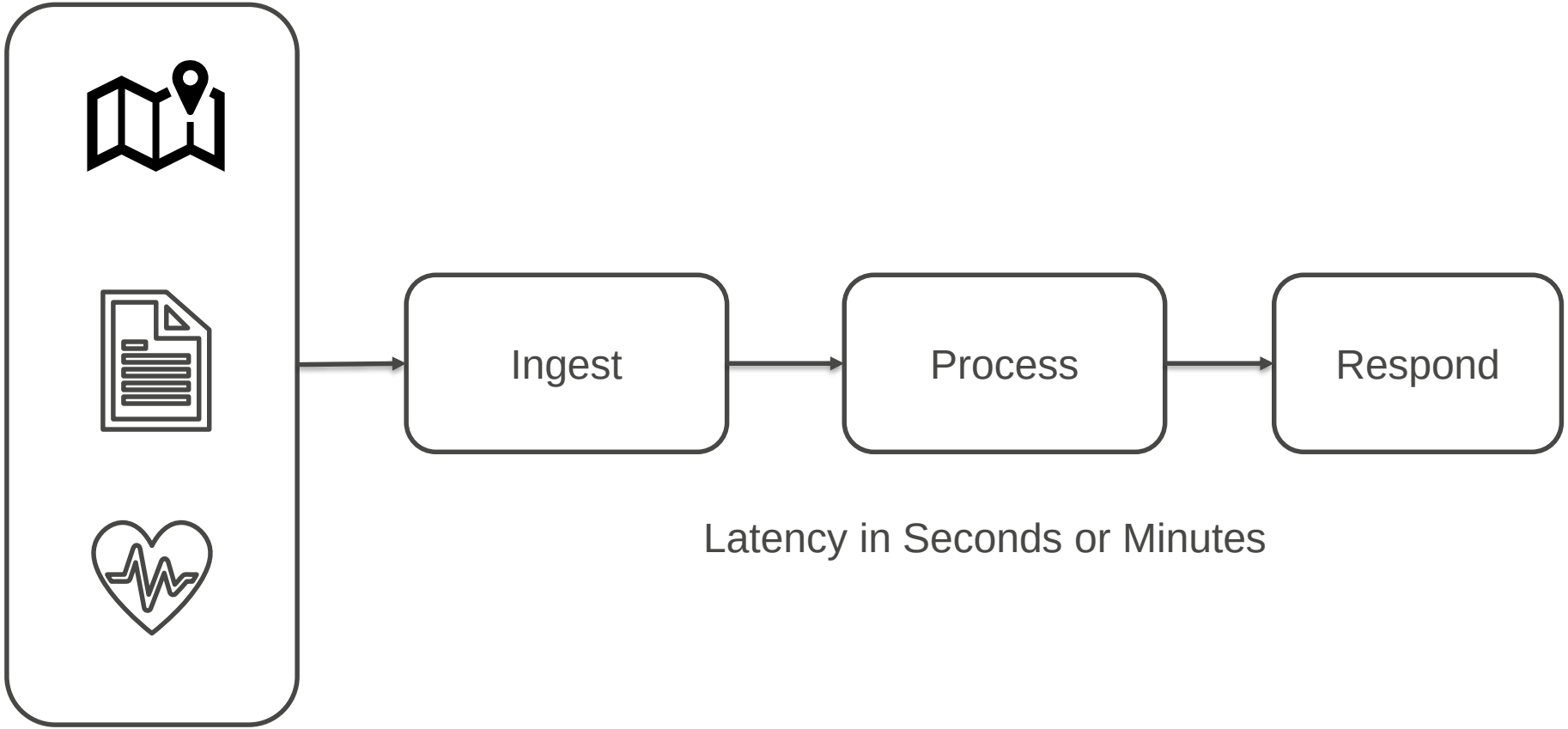


# Batch Processing Use Cases

Utility bill generation

Daily, Monthly Manufacturing Reports

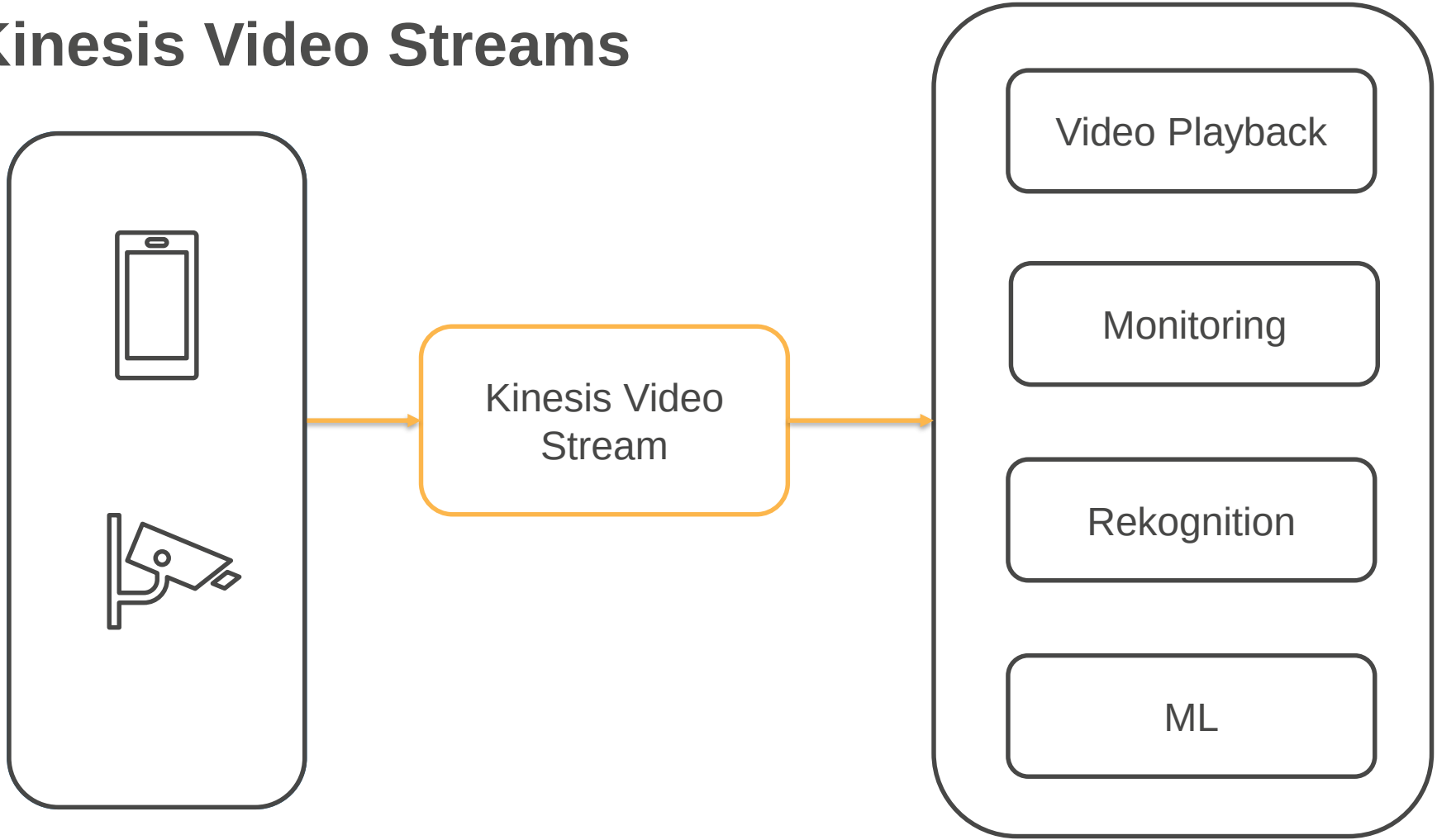
# Stream Processing



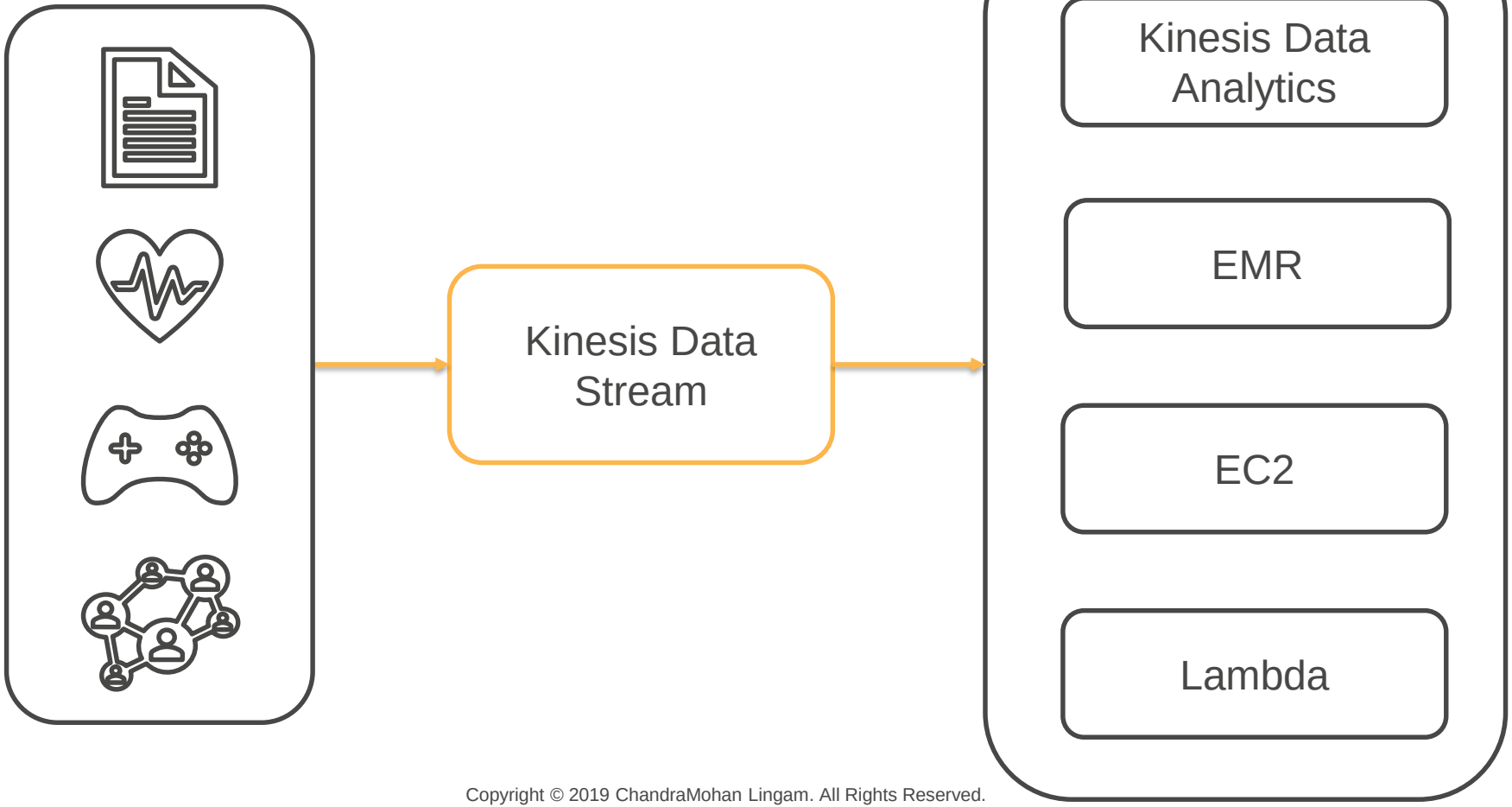
# Kinesis

| 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 streams to AWS Data Stores | Use Existing BI tools for Streaming Data: S3, Redshift, ElasticSearch, Splunk. |
| Data Analytics | Analyze data streams with SQL and Java | Real-time analytics, Anomaly detection |

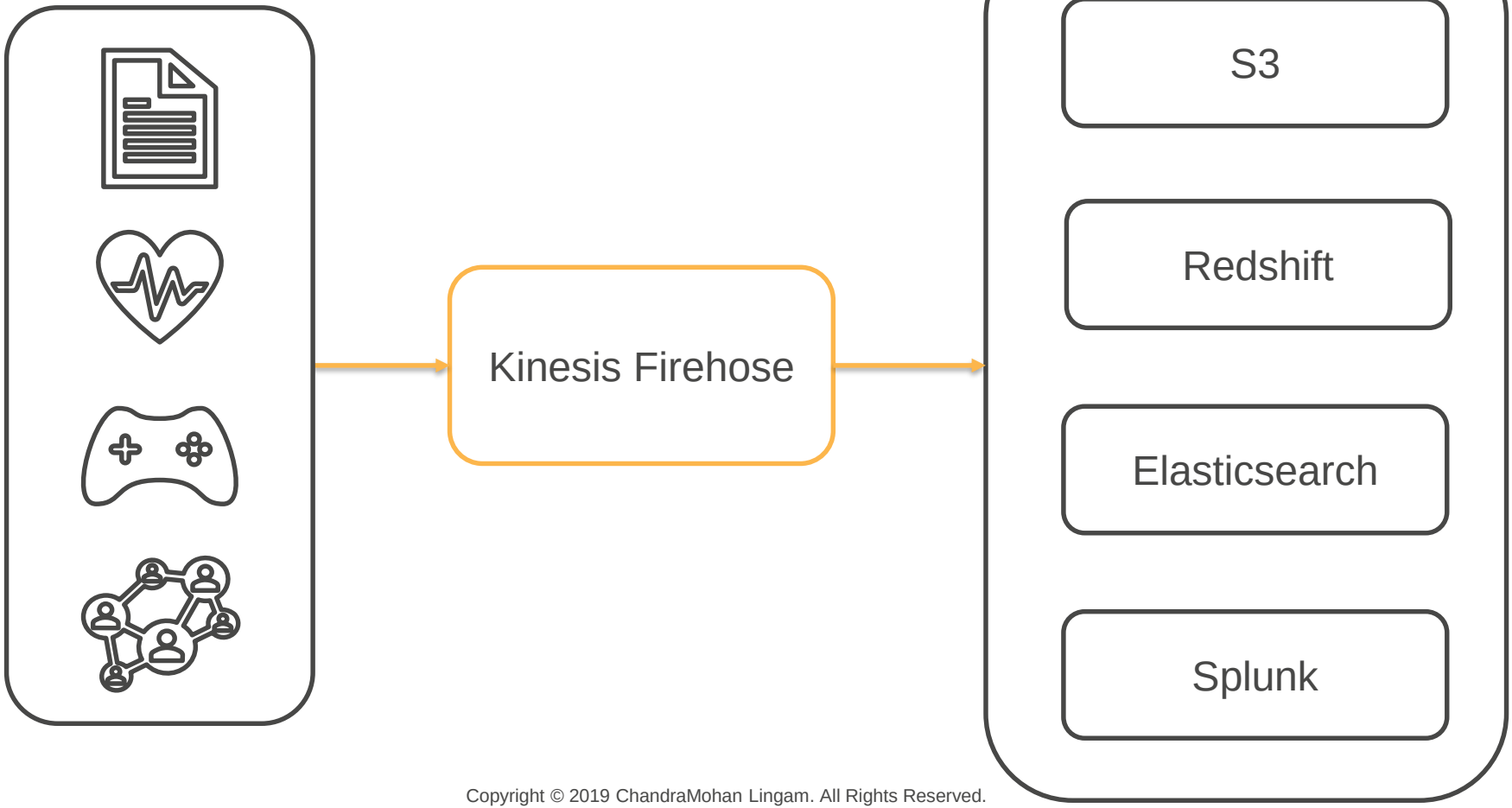
# Kinesis Video Streams



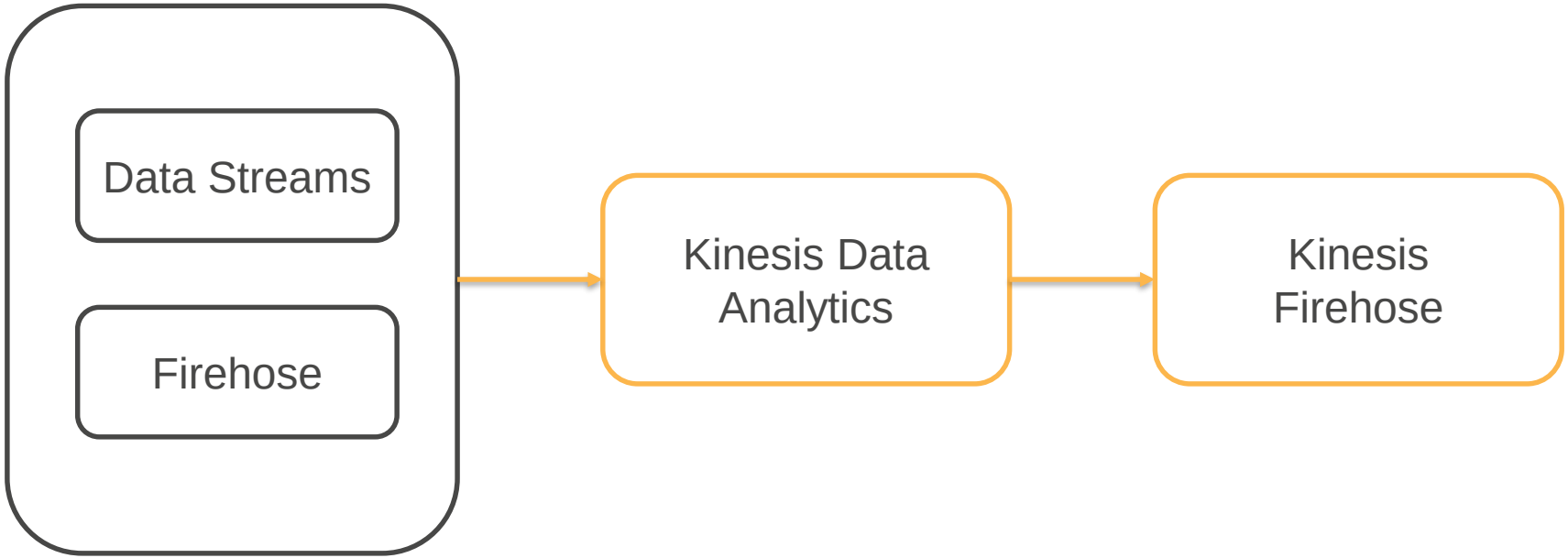
# Kinesis Data Streams



# Kinesis Data Firehose



# Kinesis Data Analytics



# Data Formats

Popular Formats, Tools for Conversion

# Data Formats

Variety of Formats

Optimal Format 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

Data Organization:

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

# Data Formats (Text)

| Format | Organization | Use |
|---|---|---|
| CSV, TSV | Row | Easy to use<br>No data type support<br>Duplication when used for hierarchical data:<br>For example, in an employee-department CSV file, department information is duplicated for every employee<br>Not optimized for reading only specific columns |
| JSON | Row | Format of choice for communication between web services<br>Supports data types<br>Efficiently represent hierarchical data |
| JSON Lines | Row | New Line Delimited JSON<br>Convenient for processing one record at a time |

# Data Formats (Binary)

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

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

# Data Formats (Binary)

| Format | Organization | Use |
|---|---|---|
| ORC | Columnar | Like Parquet |
| Avro | Row | Ideal for write-heavy use cases<br>Ideal for scenarios where you need to read the entire record<br>Data Type Support |

# Data Transformation

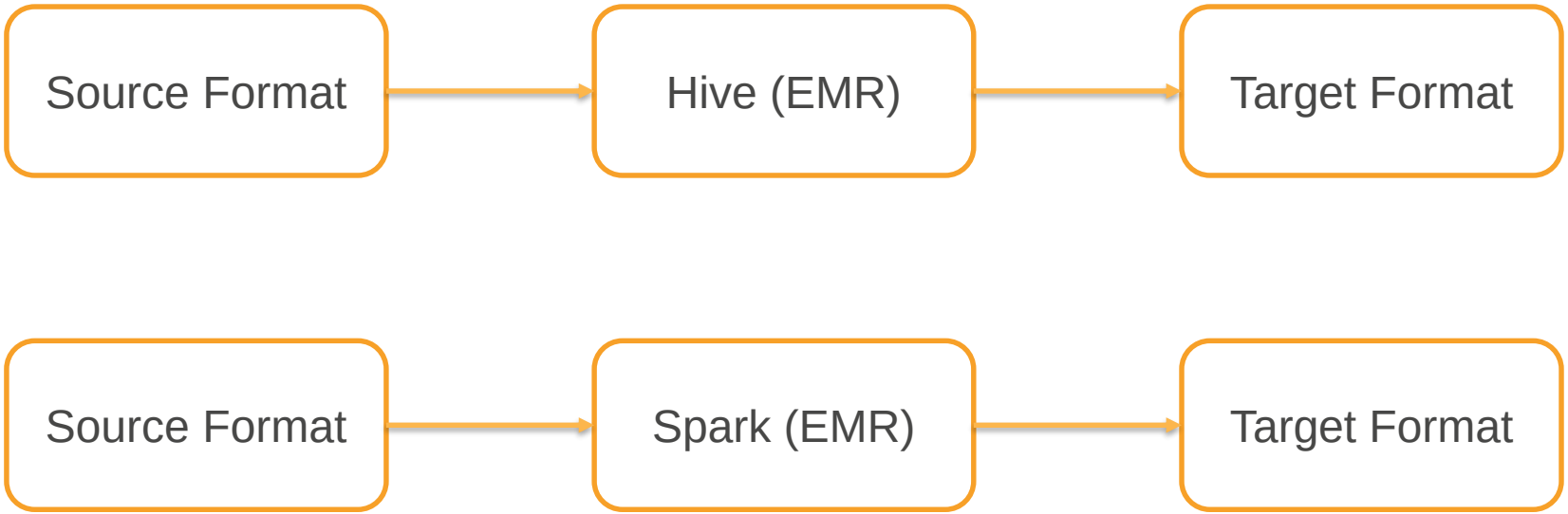
When and Where to do the format conversion?

- Collect in Native Format
- Transform in Data Lake

# Data Transformation

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

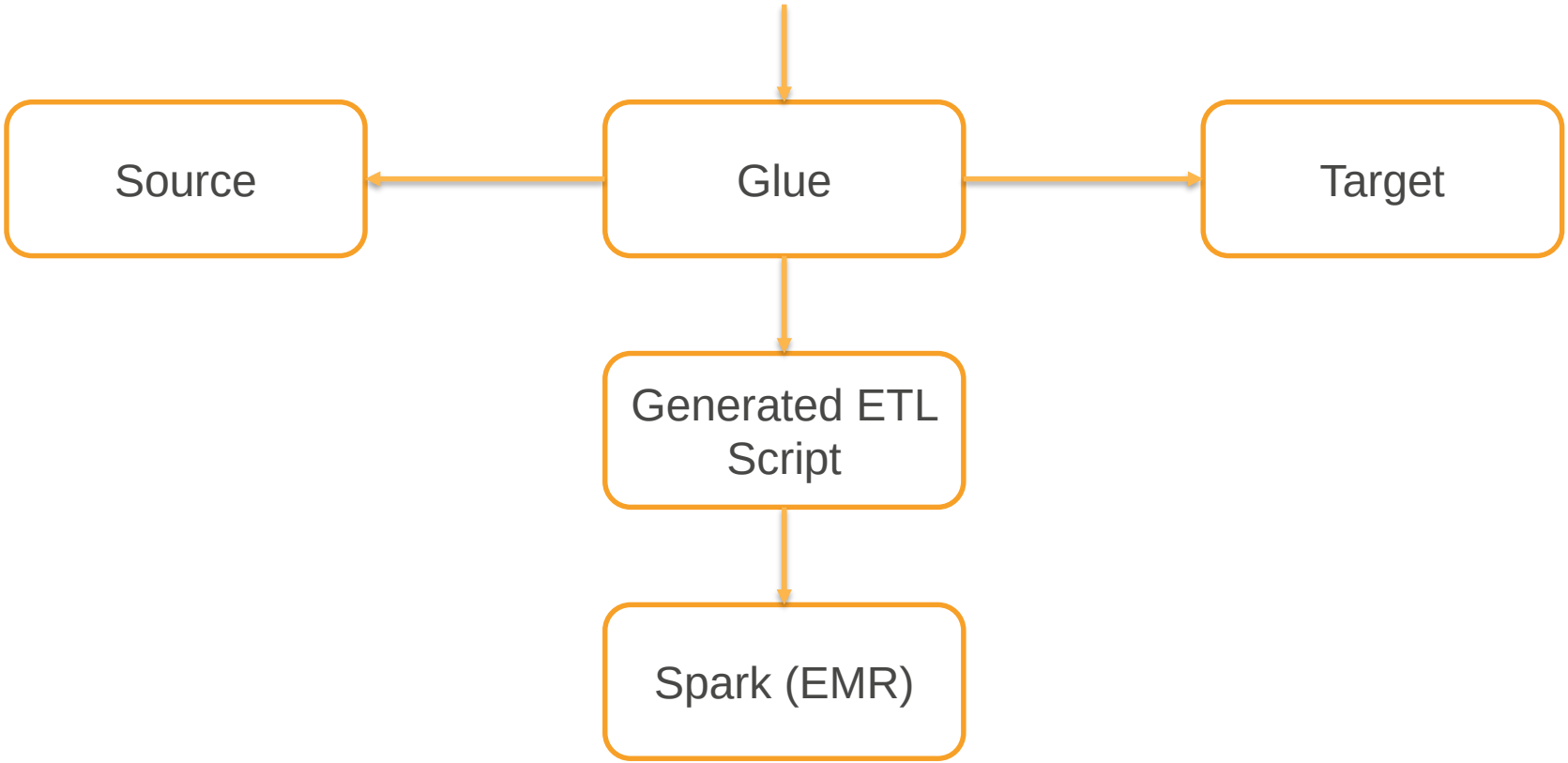
# Amazon EMR – Format Conversion



# Data Transformation

| Service | Purpose | Use |
| --- | --- | --- |
| Glue | ETL | Automatically Generate ETL Scripts<br><br>Schedule and Run on Managed Spark Environment |

# Glue ETL – Generate and Run Script



# Data Transformation

| Service | Purpose | Use |
| --- | --- | --- |
| Kinesis Firehose | Streaming Data Transformation | Transform streaming data to Parquet, ORC<br>Deliver transformed data to AWS Data Stores<br>Backup original data to S3 |

# In-Place Querying

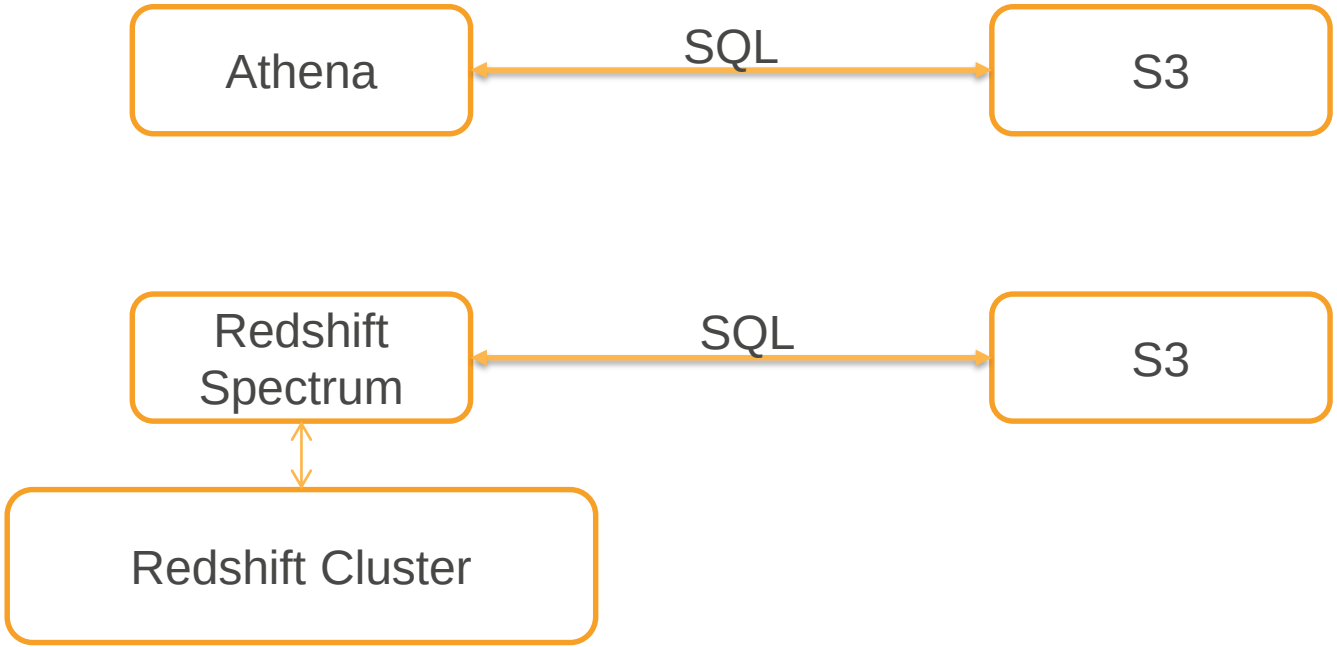
Directly query data in S3 using SQL

# In-place Query

“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

# In-Place Query



# In-Place Query

| Service | Purpose | Use |
| --- | --- | --- |
| Athena | In-place SQL Query | Query data in S3 without needing to extract, load into a separate service or platform.<br><br>Charged based on amount of data scanned<br>https://aws.amazon.com/s3/features/ |
| Redshift Spectrum | In-place SQL Query (Redshift Compatible SQL) | Query data in S3 without needing to extract, load into a separate service or platform.<br><br>More suitable for complex queries and large datasets (up to Exabytes).<br>https://aws.amazon.com/s3/features/ |

# Recommendations

Athena

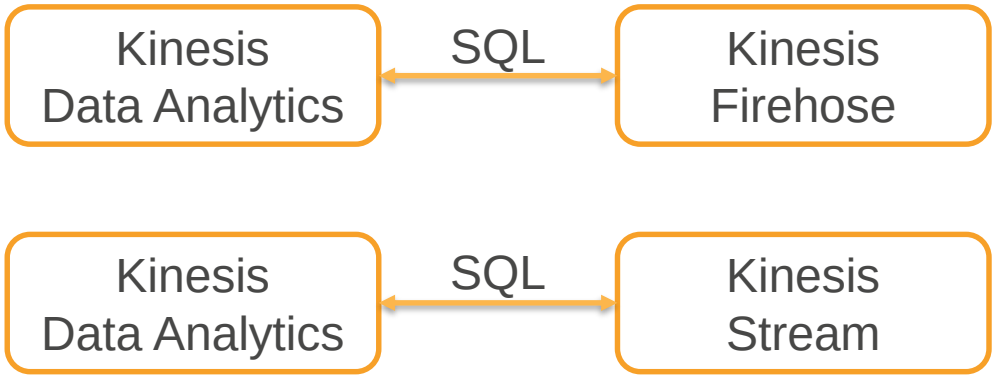
- Ad-hoc data discovery and SQL querying

Redshift Spectrum

- More complex queries
- Large number of users

# Streaming Query

| Service | Purpose | Use |
| --- | --- | --- |
| Kinesis Data Analytics | Streaming Data SQL Querying | Query and analyze Streaming data with SQL<br><br>https://aws.amazon.com/kinesis/data-analytics/ |



# Broader Analytics Portfolio

| Service | Purpose | Use |
|---|---|---|
| Amazon EMR | Hadoop Ecosystem tools | You can run variety of workloads using Hadoop tools: Spark, Hive, Pig, Hbase, TensorFlow, MxNet and so forth |
| SageMaker | Machine Learning | Managed Machine Learning service with wide selection of algorithms |
| Artificial Intelligence | Video, Image, Natural language | Pre-trained, ready-to-use AI service for video analysis, speech and natural language processing |

# Broader Analytics Portfolio

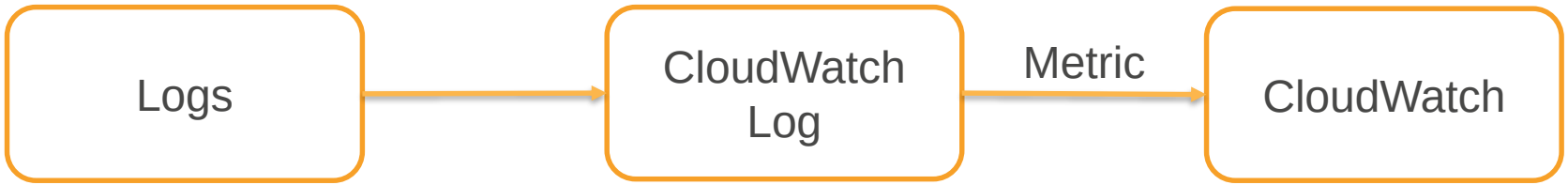
| Service | Purpose | Use |
|---|---|---|
| Quicksight | Business Intelligence | Managed BI tool to create interactive dashboards |
| Redshift | Data warehouse (Columnar Storage) | Managed Petabyte scale data warehouse. SQL based querying and easily integrates with your existing Business Intelligence tools |
| Lambda | Business Logic (Function as a service) | Serverless Backend processing logic with trigger-based code execution |

# Monitoring and Optimization

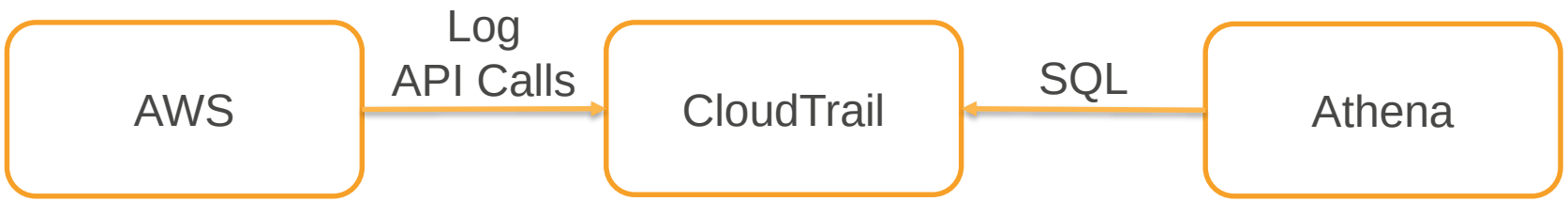
# Monitoring

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

# CloudWatch Log – Consolidate Logs and Monitor



# CloudTrail – Audit Trail of all API Activities



# 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

1. S3 Lifecycle Management
2. S3 Storage Class Analysis
3. Intelligent Tiering
4. Amazon Glacier and Glacier Deep Archive
5. Data Formats

# Lifecycle Storage Tiering and Expiration

1. Object Age
2. Name and Folder Structure
3. S3 Object Tags

# S3 Lifecycle Management


| | Standard | 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 | Rarely accessed |
| 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

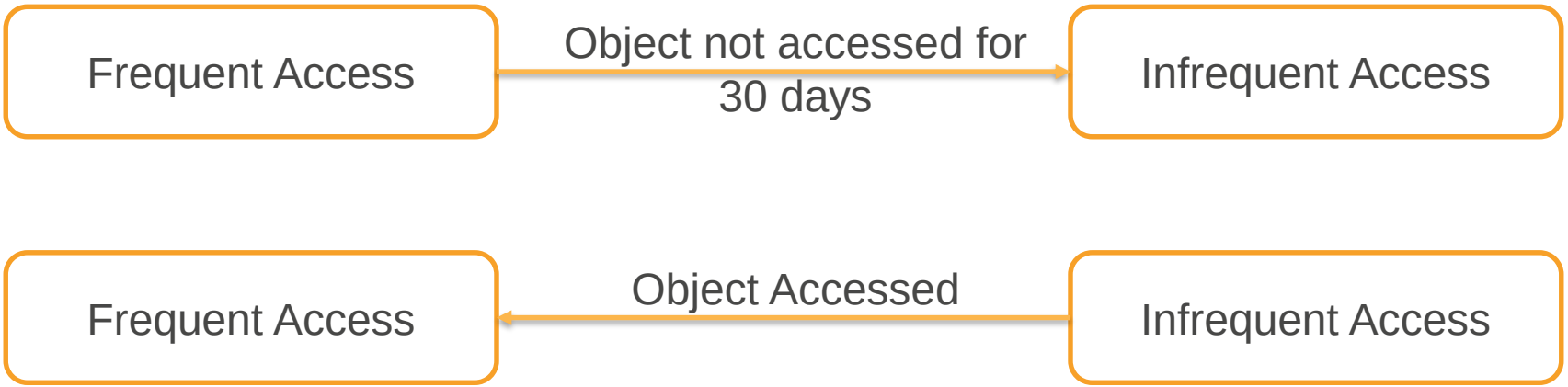
# Storage Class Analysis

“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<br>Durability: 11 9's<br>Retrieval Time: Minutes to Hours<br>Vault Lock to prevent future edits |
| Glacier<br>Deep Archive | Archive and Backup | Cost: USD 0.50 for 500 GB/Month<br>Durability: 11 9's<br>Retrieval Time: 12 to 48 hours<br>Vault Lock to prevent future edits |

# Security and Protection

# Data Lake Security

- Data Lake is Centralized
- Consolidates all data in one place

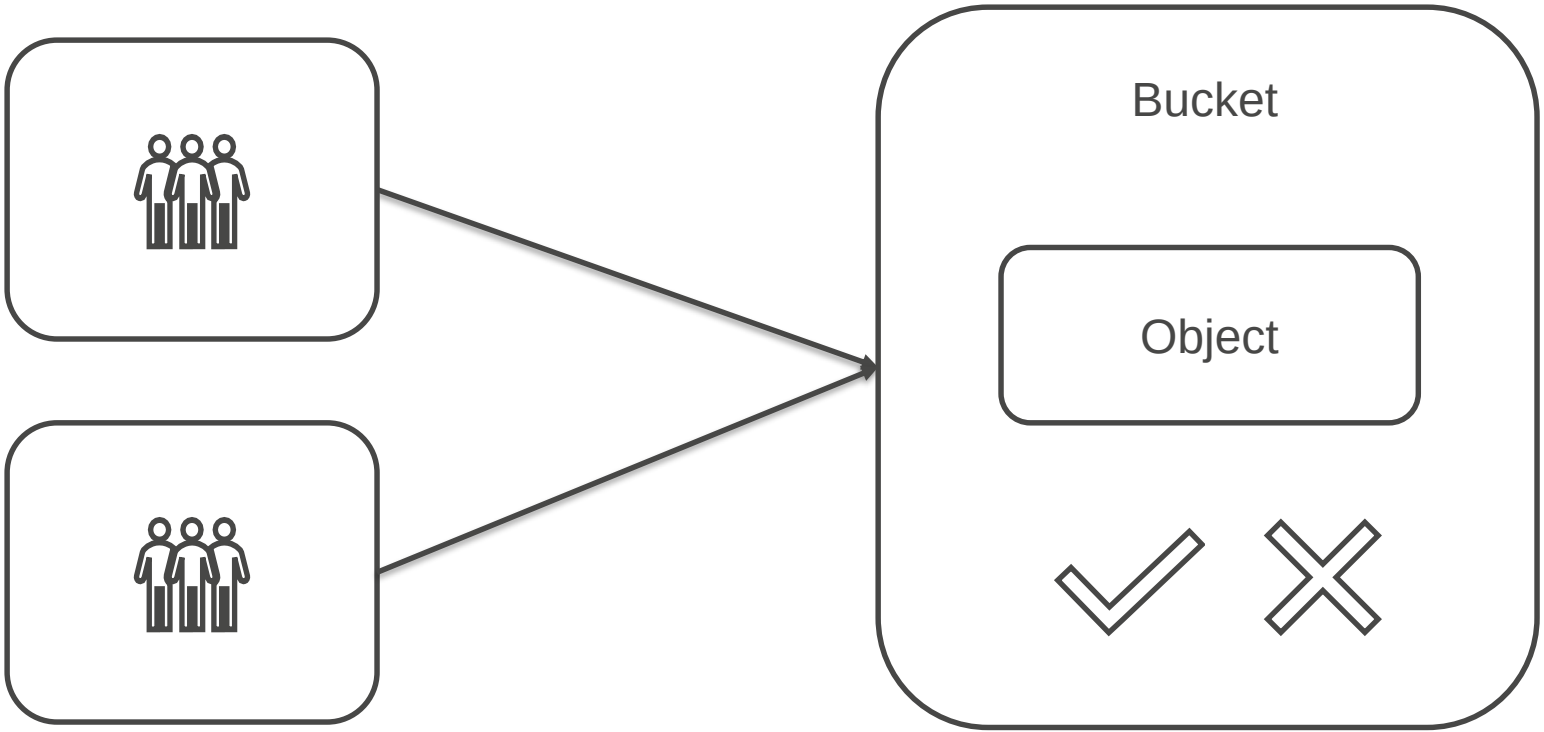
**Protecting and Managing Data is very important**

# S3 Access Control

Resource-based Policy and Access Control

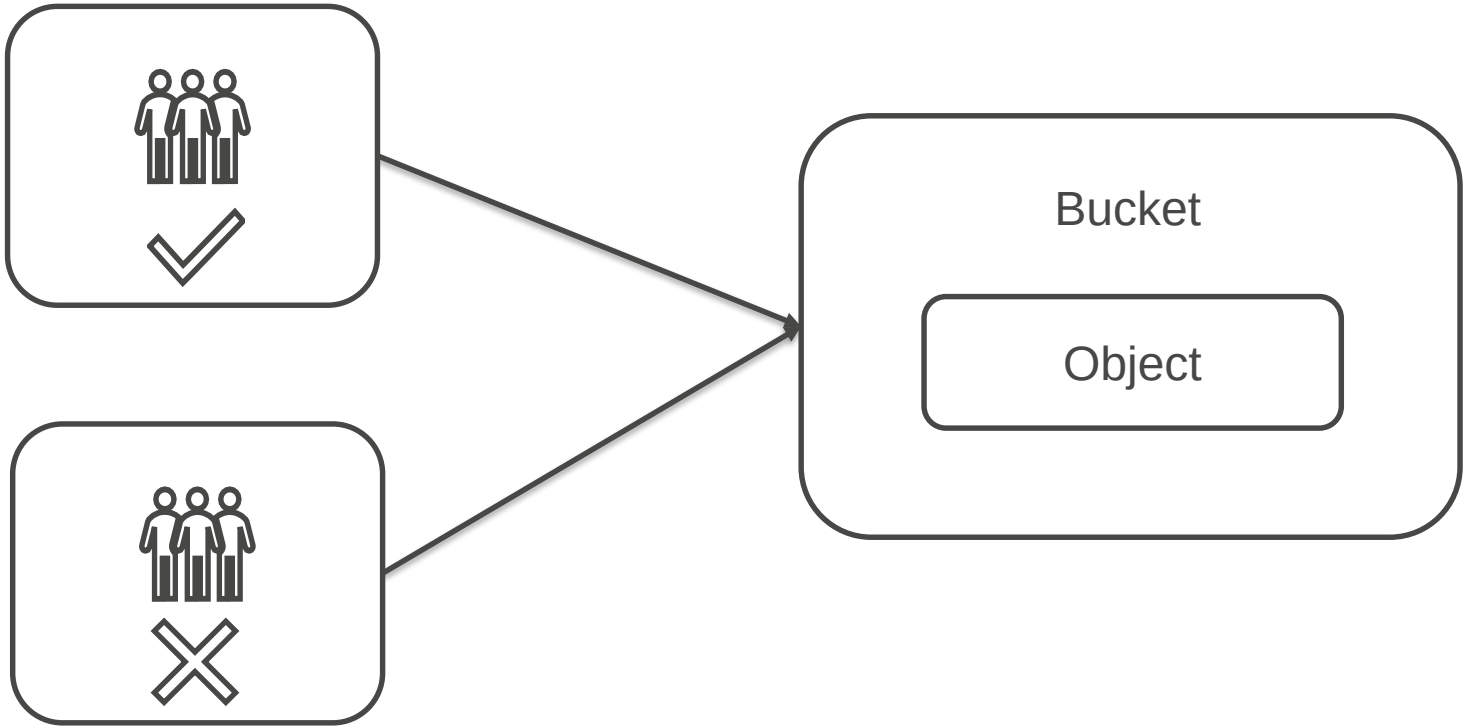
User-based Policy

# S3 Resource Based Policy



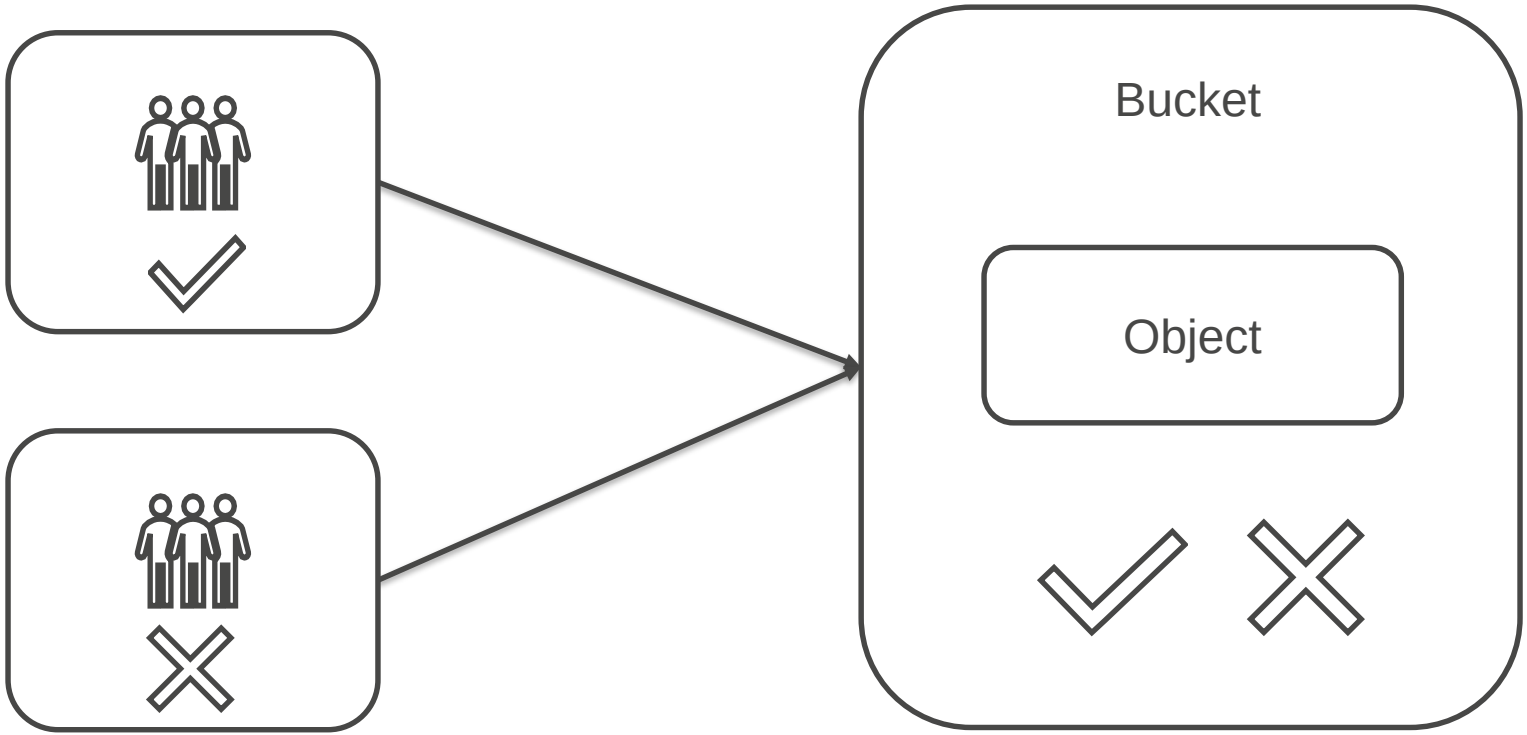

Permissions are embedded as part of Bucket and Object

# S3 User Based Policy



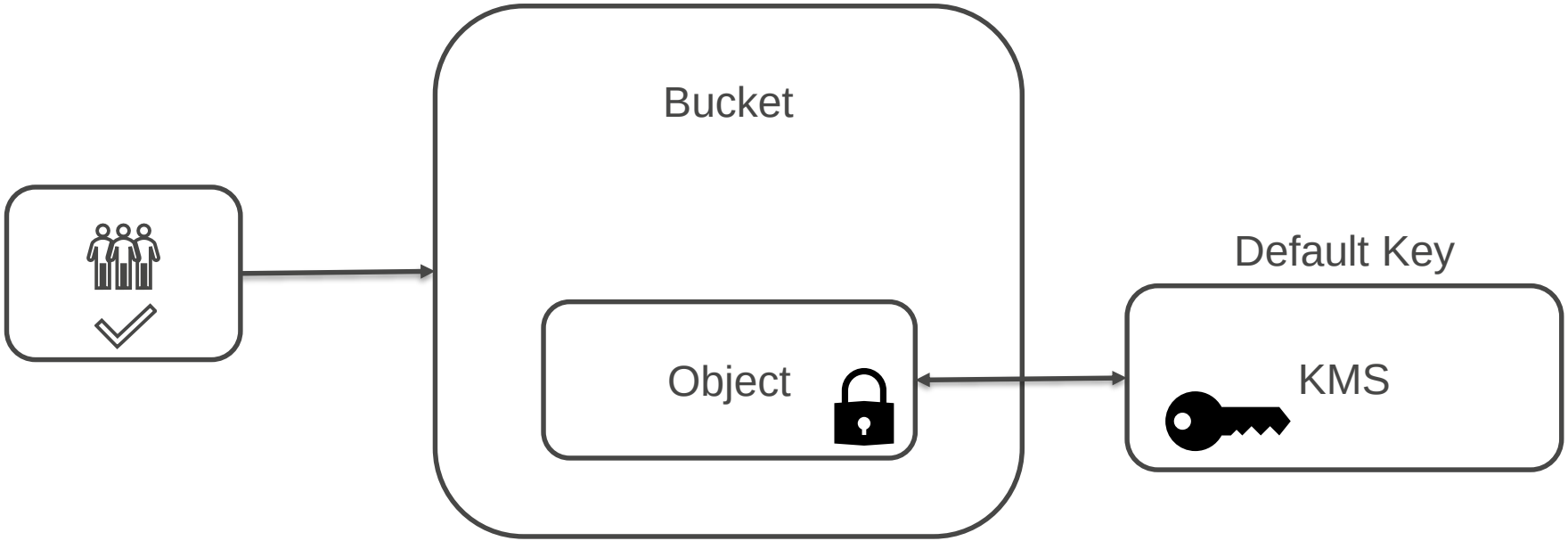

Permissions are granted to Users and Groups

# S3 User and Resource Based Policy



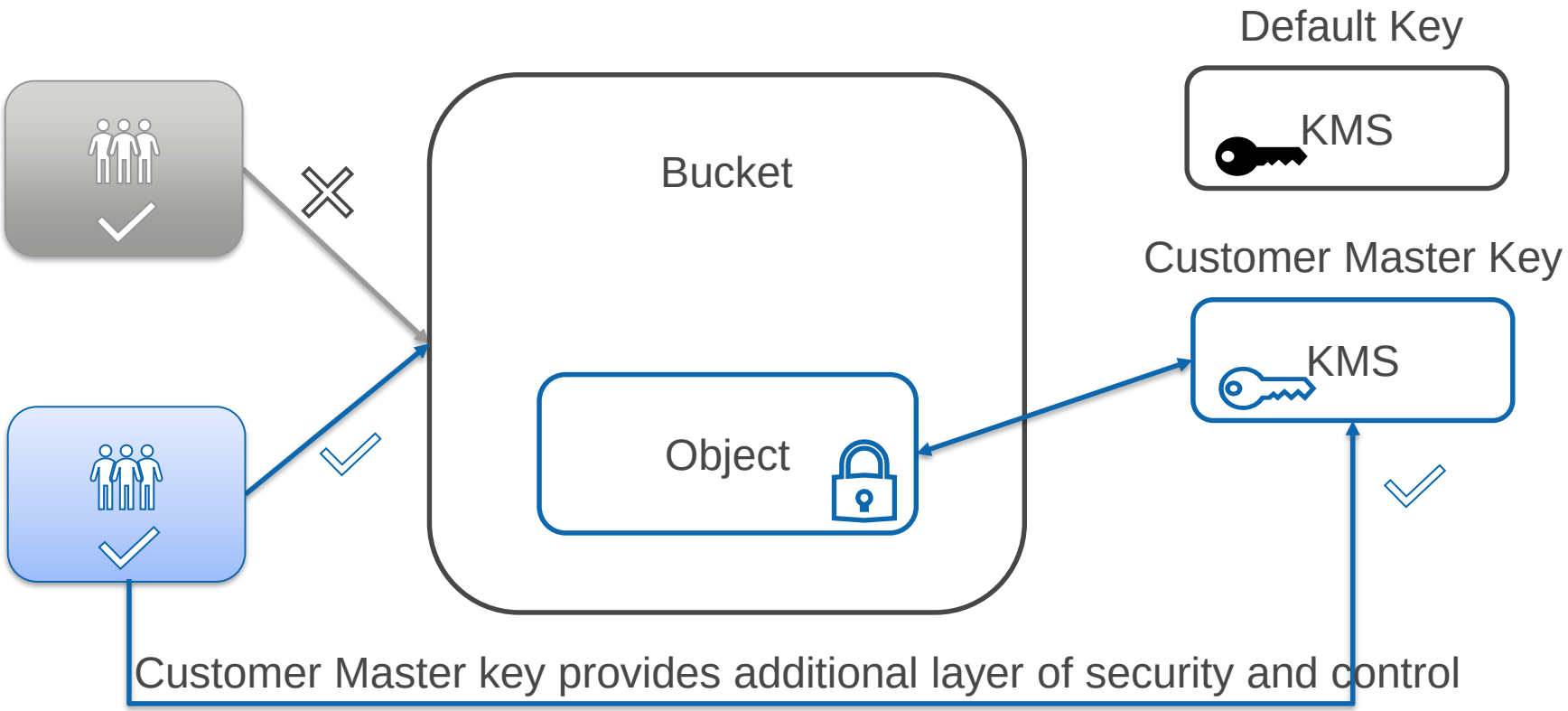

Deny all access that do not originate from on-premises

# S3 Data Encryption – Default Key



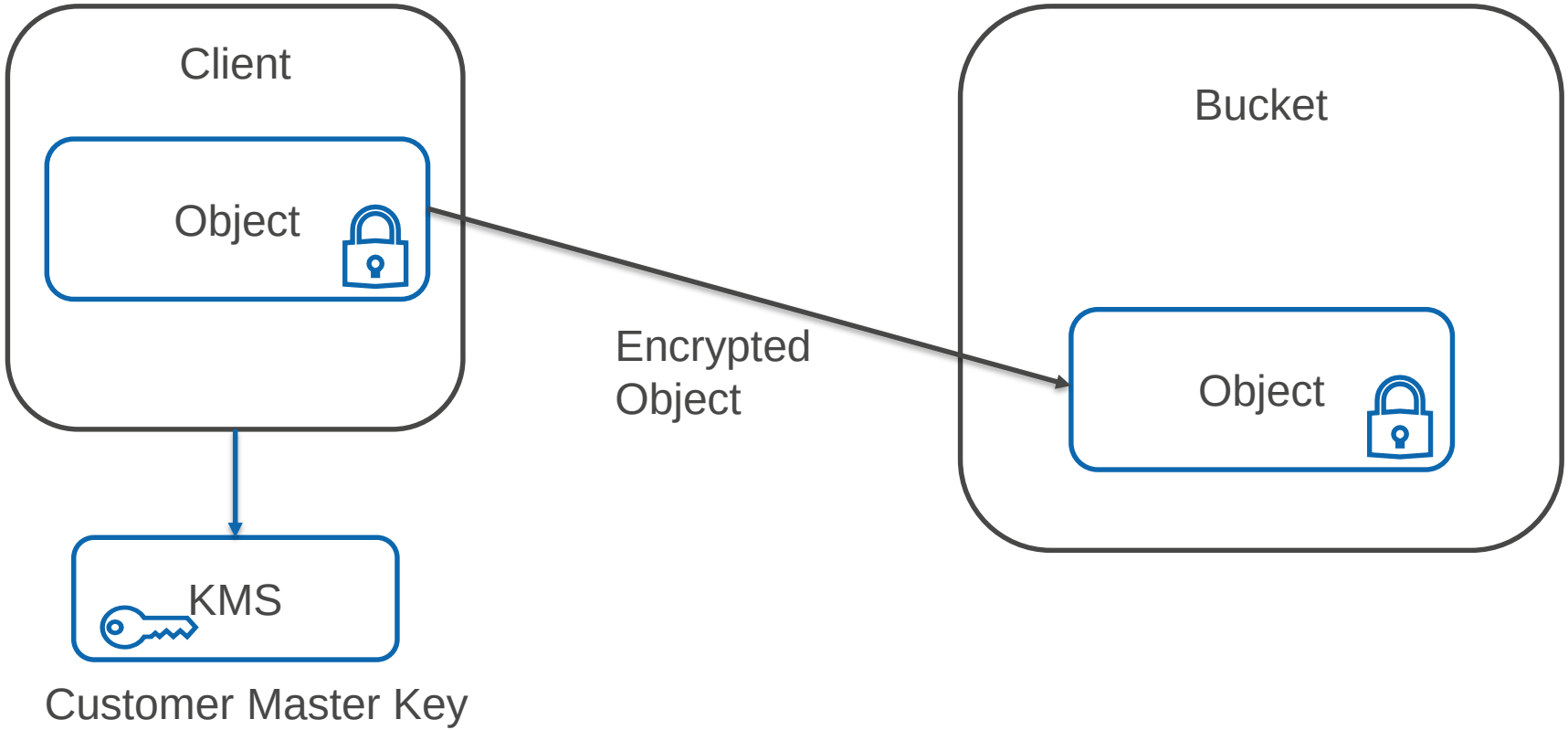

With default key, S3 automatically decrypts object for any user who is allowed access to the bucket or object

# S3 Data Encryption – Customer Master Key (CMK)



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

Object encryption and decryption is client responsibility



# Protection

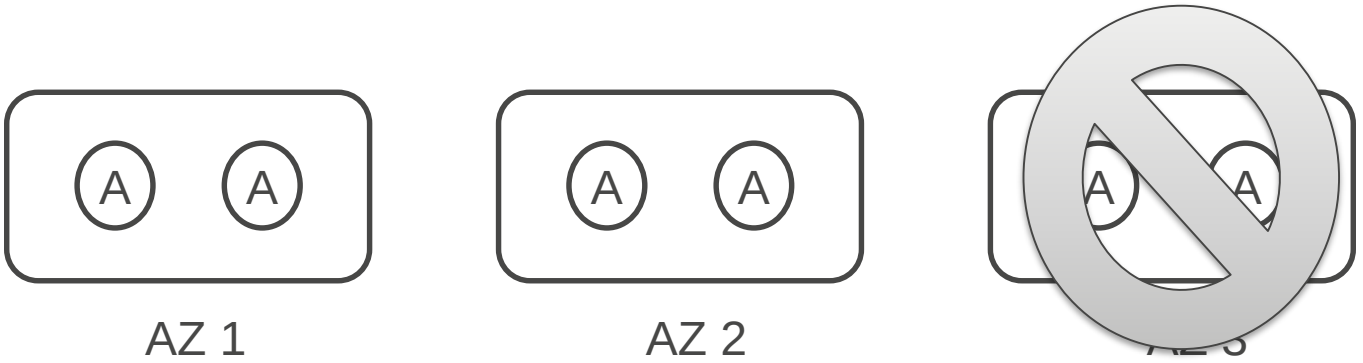
“A data lake must protect data against corruption, loss, accidental or malicious overwrites, modifications, and deletions.”

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)

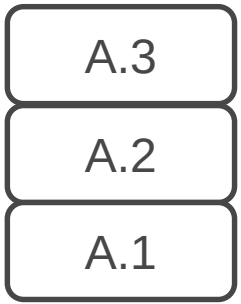
Measure of protection against data loss and corruption



# 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

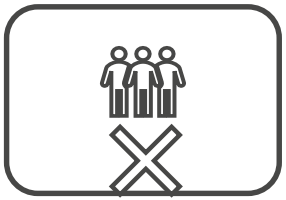
Deletes are not replicated

# 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

# Data Lake 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

# 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

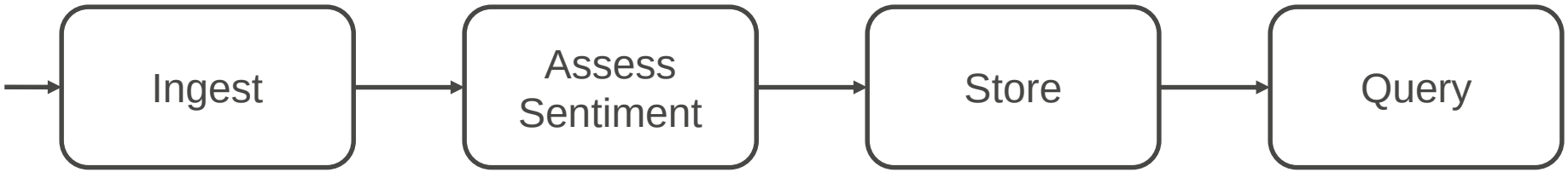
```
SELECT product_title, star_rating,review_body
FROM  "demo_db"."amazon_reviews_parquet"
WHERE product_category = 'Books'
and product_title like 'Harry Potter%'
and star_rating > 3
limit 100;
```

# 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