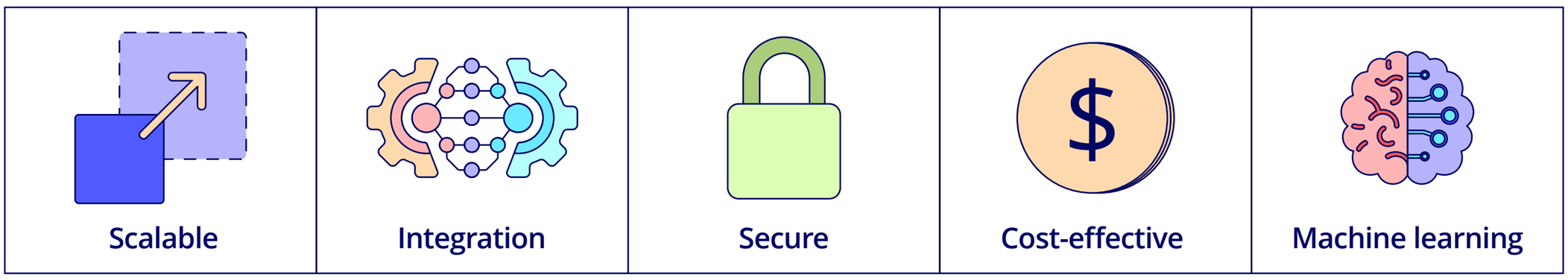
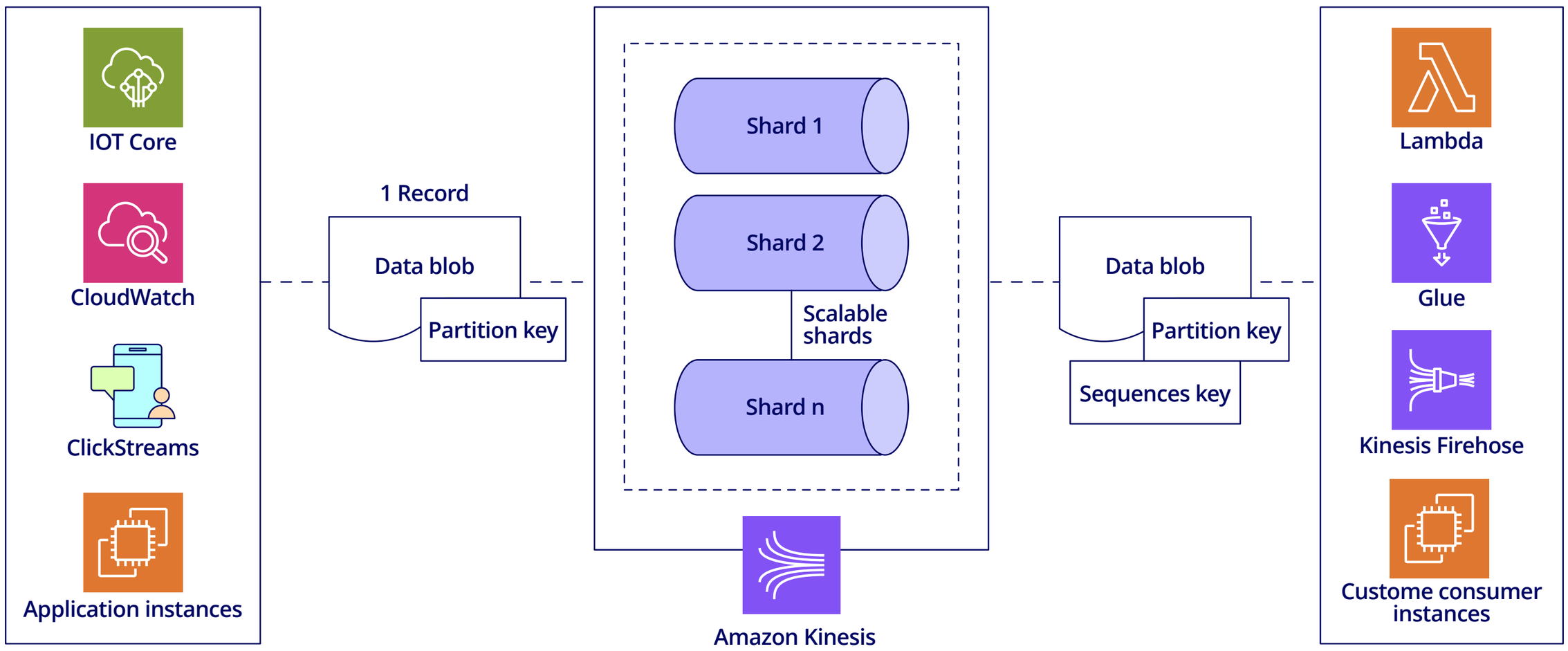
AWS Analytical Services

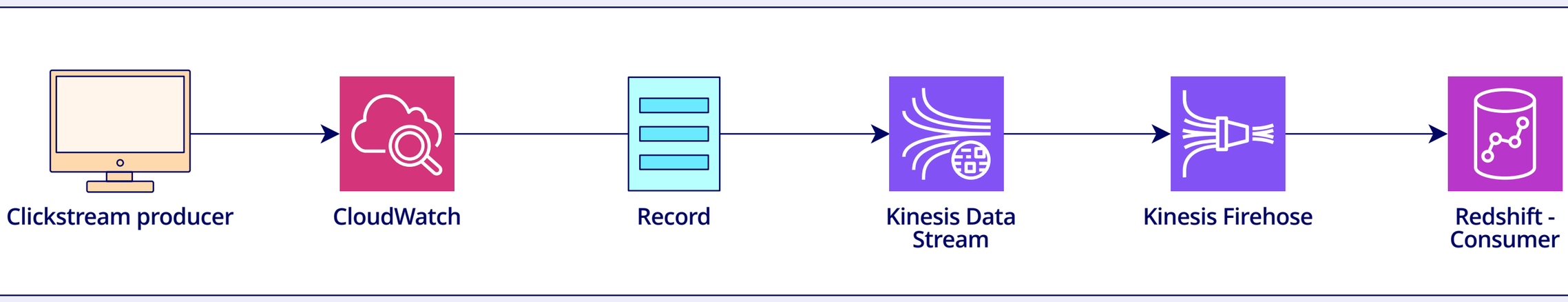


Data Ingestion

- **Amazon Kinesis**
  - It’s a fully managed service to ingest real-time streaming data such as video, audio, logs, clickStream, and IoT telemetry data.
  - The fundamental unit of data capacity is called a shard. Each shard can ingest 1000 records per second.
  - It’s a data record consisting of a partition key, sequence number, and immutable data blob. It is composed of 1 or more shards.



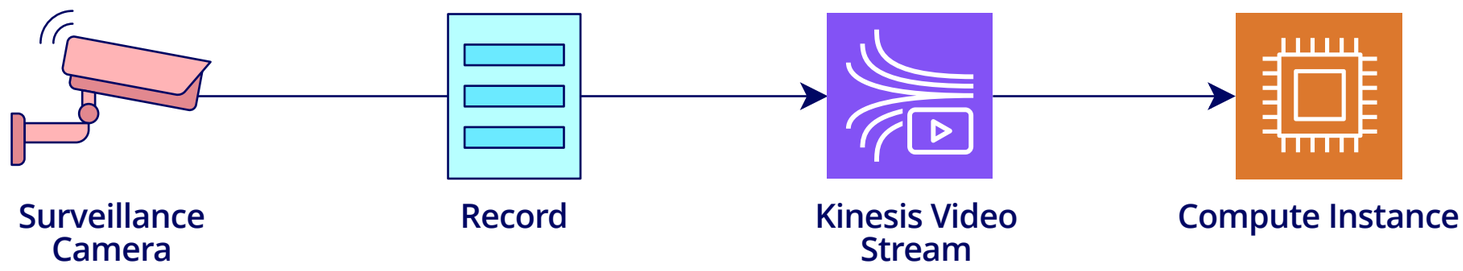
- **Kinesis Data Streams**
  - Processes large-scale streaming data in real time.
  - Supports resharding to adjust the number of shards to change the data flow rate through the stream.
  - Data is encrypted in the stream using AWS KMS.



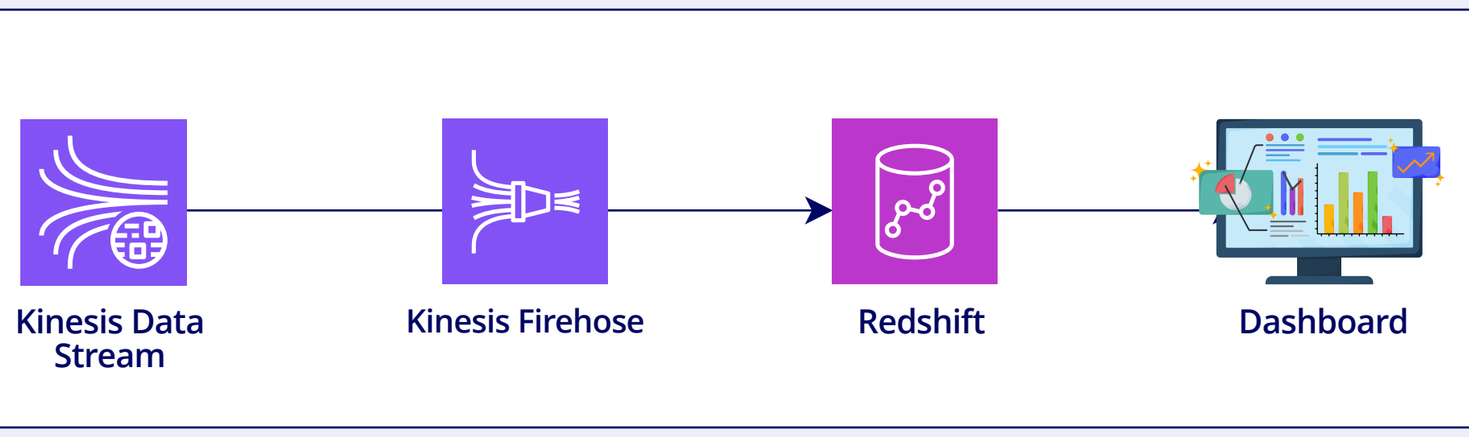
AWS Analytics Services

educative

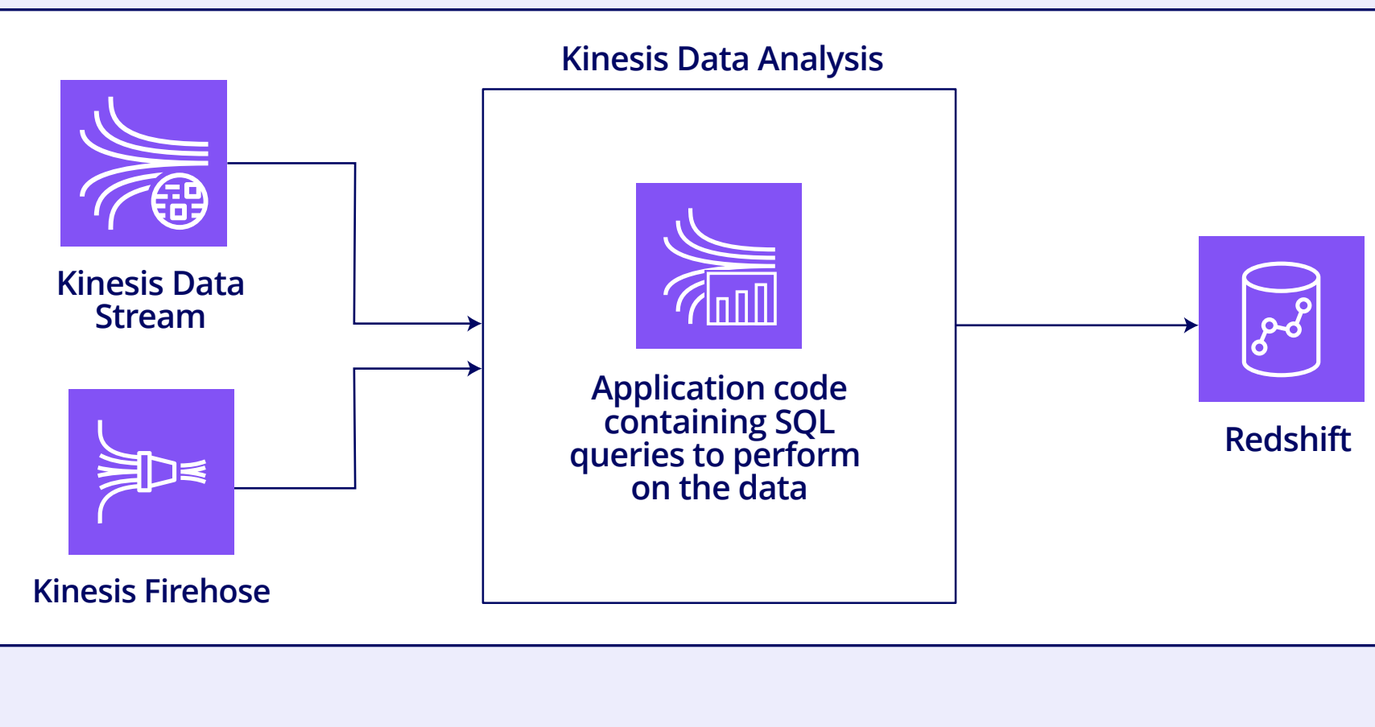
- **Kinesis Video Streams**
  - A serverless service to process large-scale data and video streams for analysis, playback, and ML.
  - Uses TLS-based encryption on data streaming and AWW KMS for data at rest.
  - Stores data for a custom retention period.
  - Time-indexes the stored data based on both producer and ingestion timestamps.



- **Kinesis Data Firehose**
  - Streams data directly into warehouses, data lakes, and other analytical services.
  - Allows transformation of data between source and target.
  - Automatically scales up and down to match the input data rate.
  - Can convert the data format from JSON to Parquet or ORC formats before storing it in S3.
  - Supports AWS S3, Redshift, ElasticSearch, Splunk, and custom HTTP endpoints.

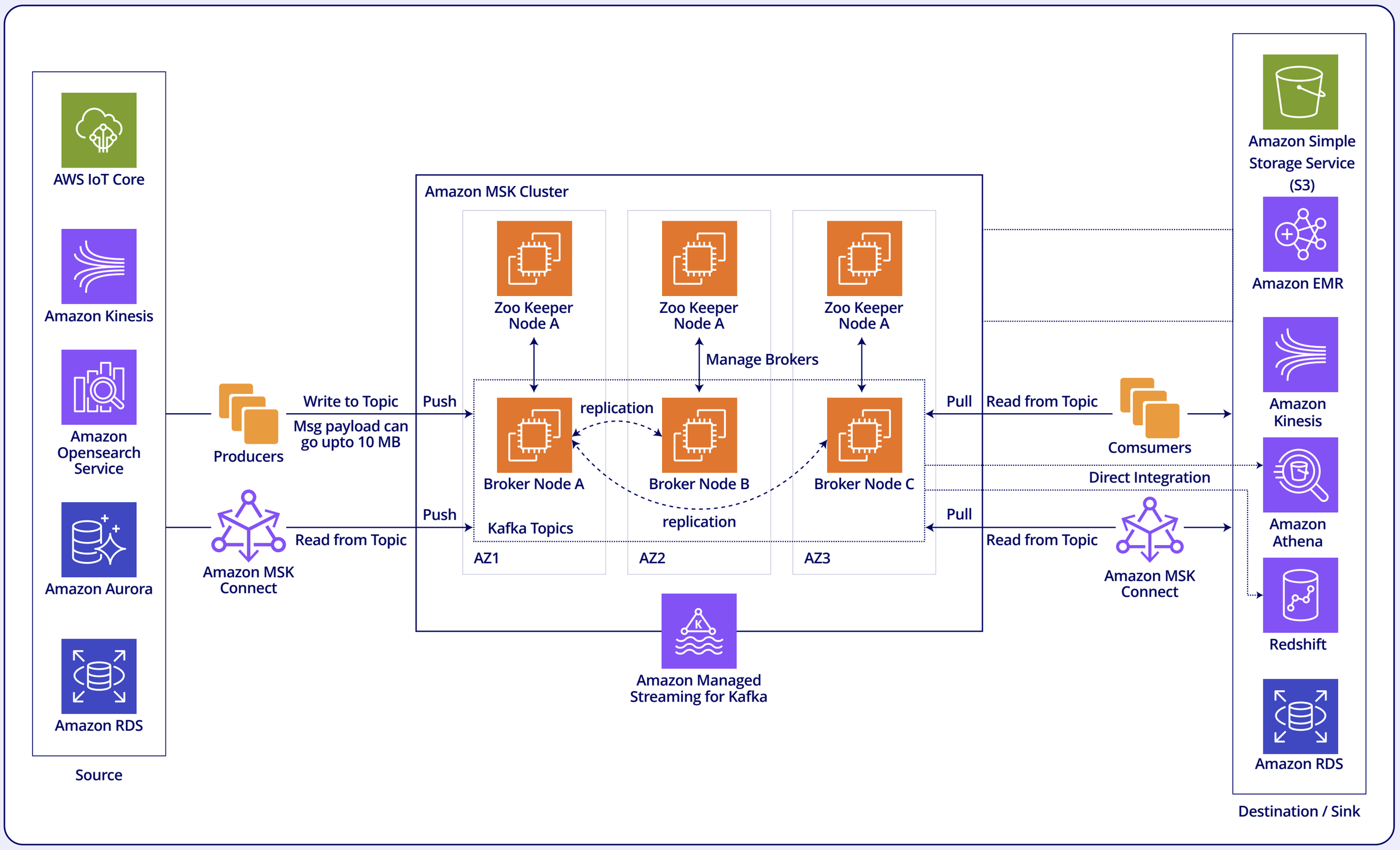


- **Kinesis Data Analytics**
  - Enables users to analyze and query real-time data.
  - Supports SQL queries on real-time kinesis data streams.
  - Supports streaming sources as well as static data.



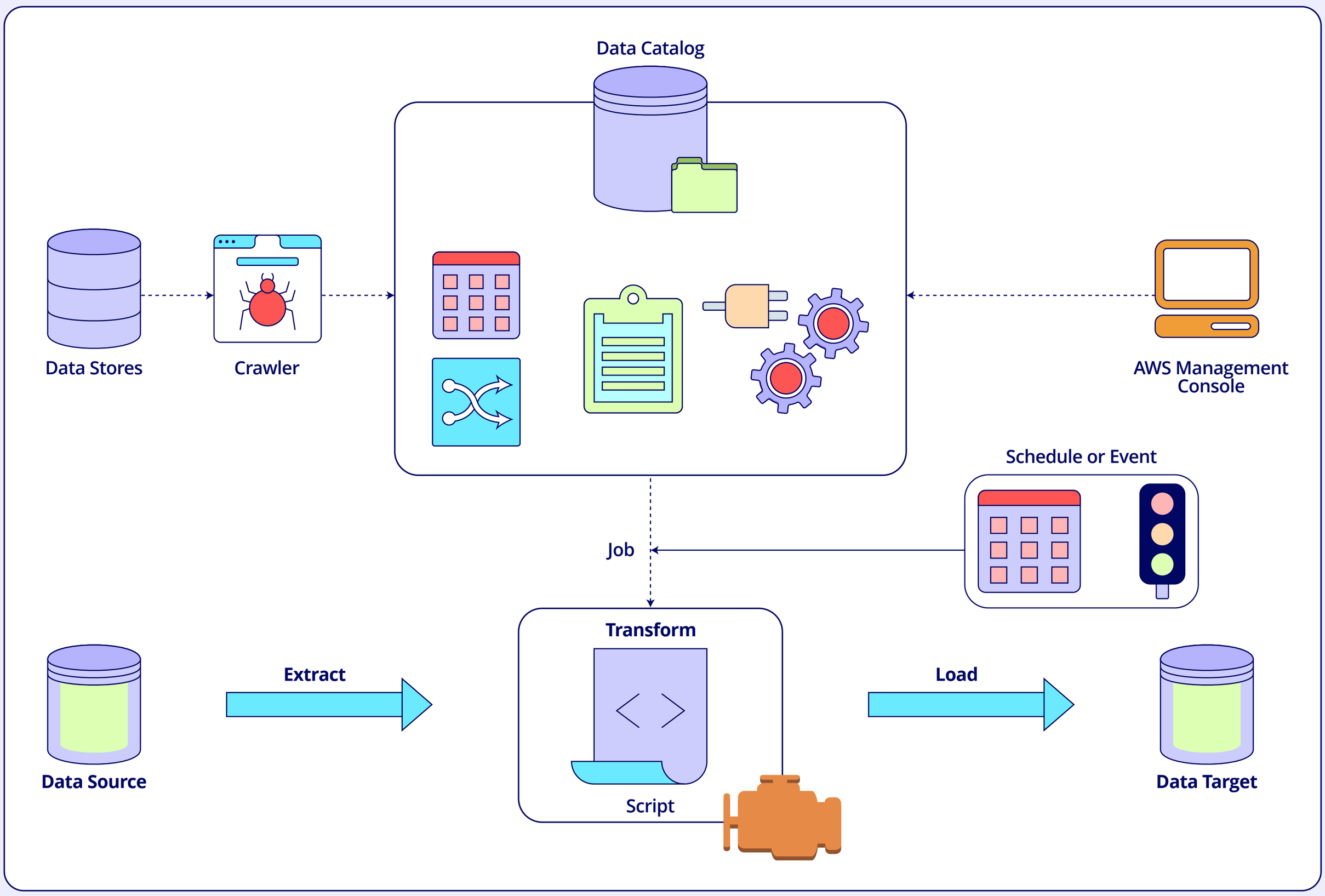
- **Amazon MSK**
  - Ingest data in real-time with a fully managed Apache Kafka service using Apache Kafka clusters.
  - MSK Connect copies data from streaming sources to MSK topics and from MSK topics to external data sinks.

AWS Analytics Services



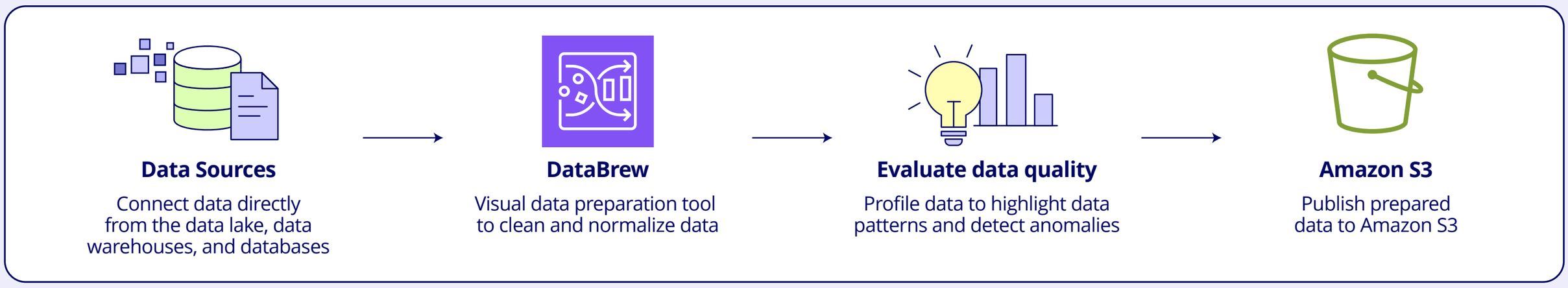
Data Processing

- AWS Glue



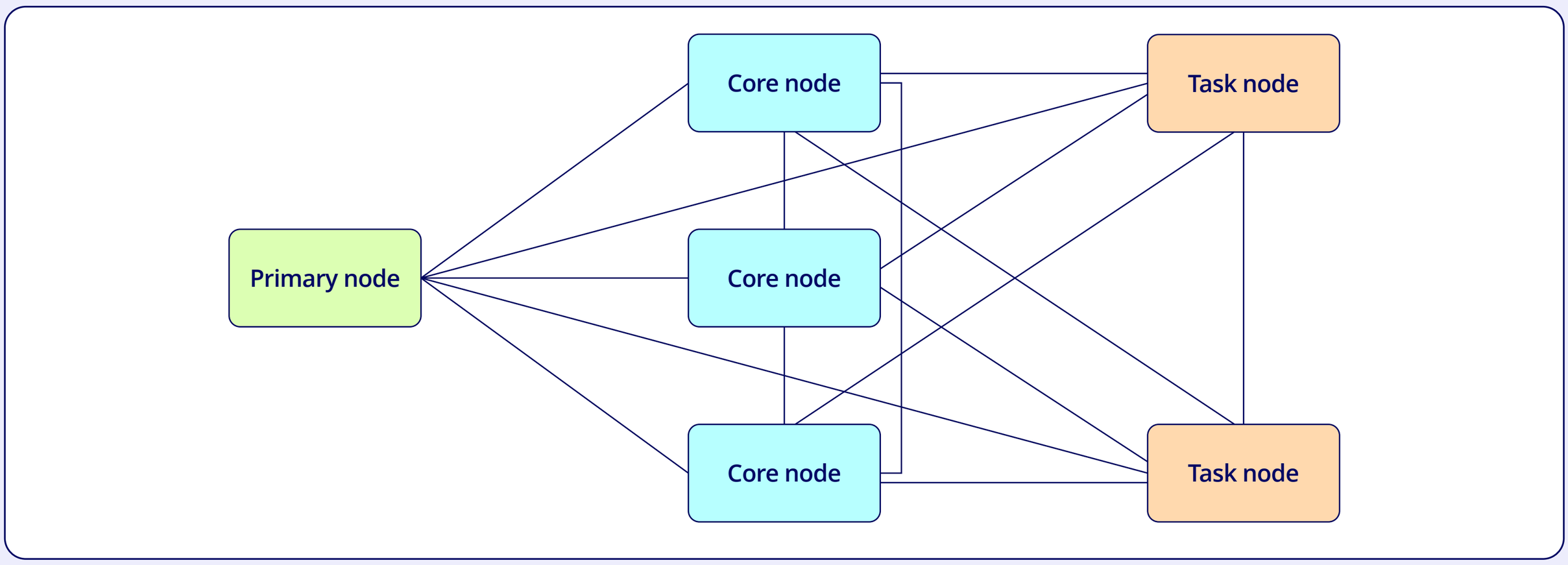
- It’s a serverless data integration service that allows accumulating, processing, cleaning, and moving data from multiple sources.
- Allows us to build complete ETL pipelines and schedule them based on events and demand.

- **Glue DataBrew**
  - Visually prepares data for analysis and training.
  - Cleans, normalizes and evaluates data without writing code.
  - Offers over 250 prebuilt transformations to automate filtering anomalies, converting data to standard format, and more.



- **Glue Studio**
  - It’s a graphical interface to create and manage jobs to gather, transform, and load data.
  - Executes the job on an Apache Spark-based serverless ETL engine.
- **ETL Jobs**
  - Allows us to write ETL Job scripts in Apache PySpark and Scala to extract, transform, and load data.
  - Provides ETL visualizer to depict the data flow through source, transformers, and target.

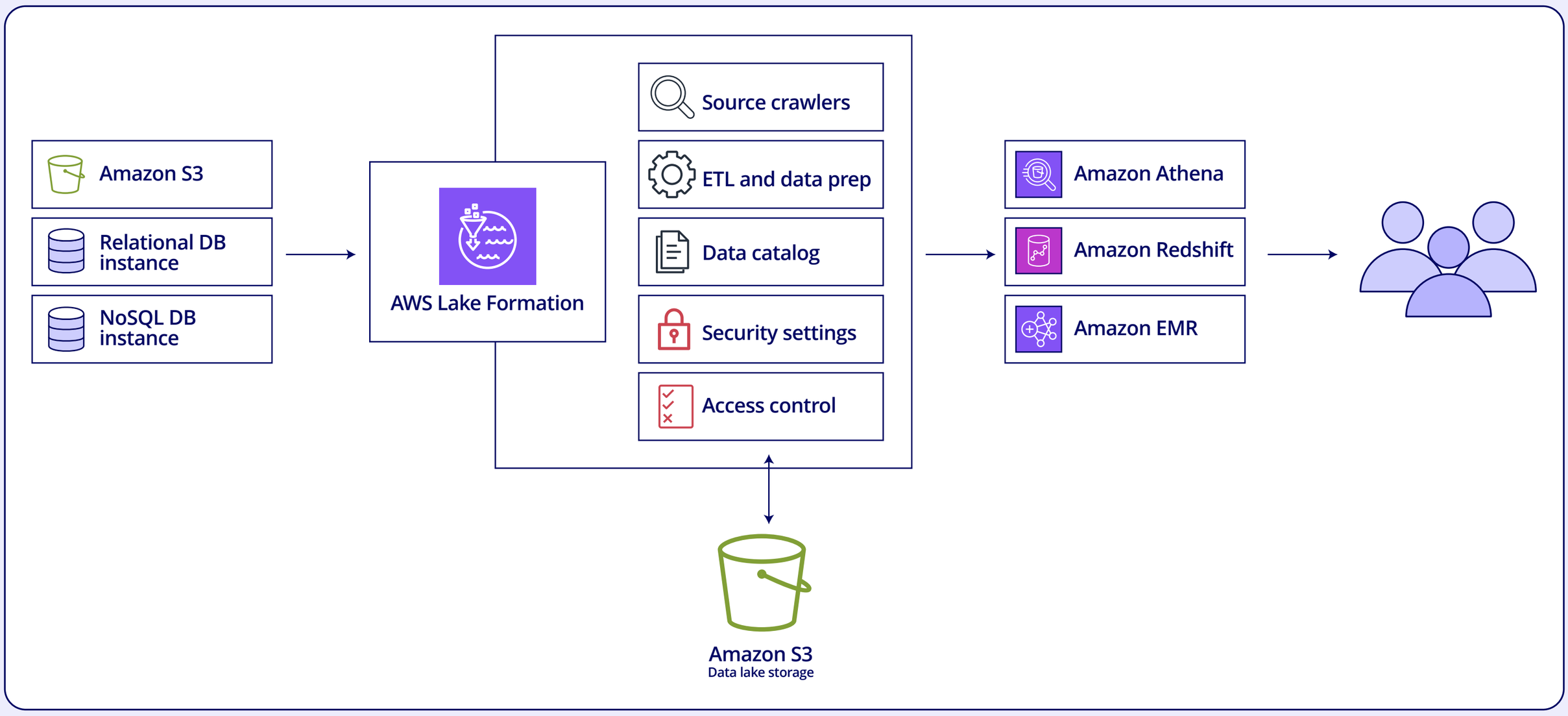
- **AWS EMR**
  - Runs clusters of EC2 instances to perform big data processing using open-source frameworks such as Apache Spark, Apache Hadoop, Apache Hive, and Presto.
  - Moves data into and out of the other AWS data stores and databases.



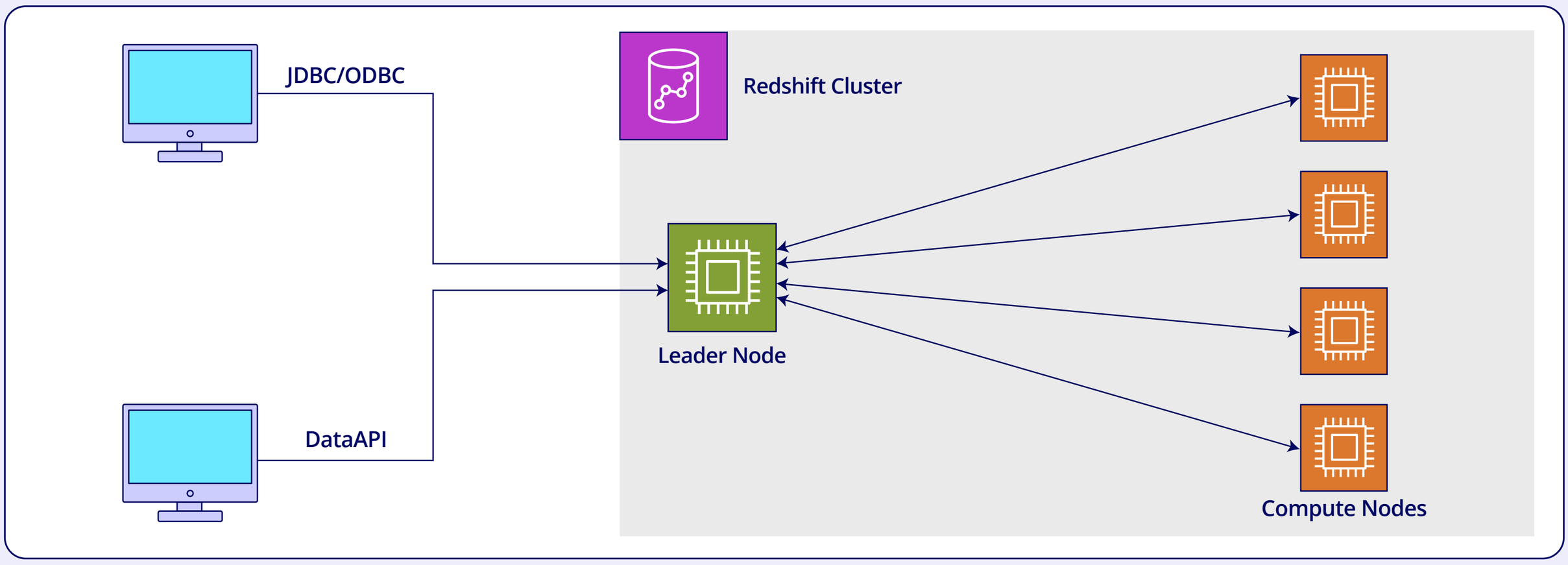
Primary node	Core node	Task node (optional)
Coordinates the distribution of tasks across core and task nodes, monitors the cluster’s health, and manages communication between nodes.	Stores and processes data. Typically runs the Hadoop Distributed File System (HDFS) and performs data replication for fault tolerance.	Computes resources for processing tasks in parallel. Often provisioned using spot instances to save cost.

Data Storage

- **LakeFormation**
  - Manages data lakes on S3, which stores and catalogs data from databases before transferring them to S3 lakes.
  - Effectively ingest, catalog, manage, secure, and share data.

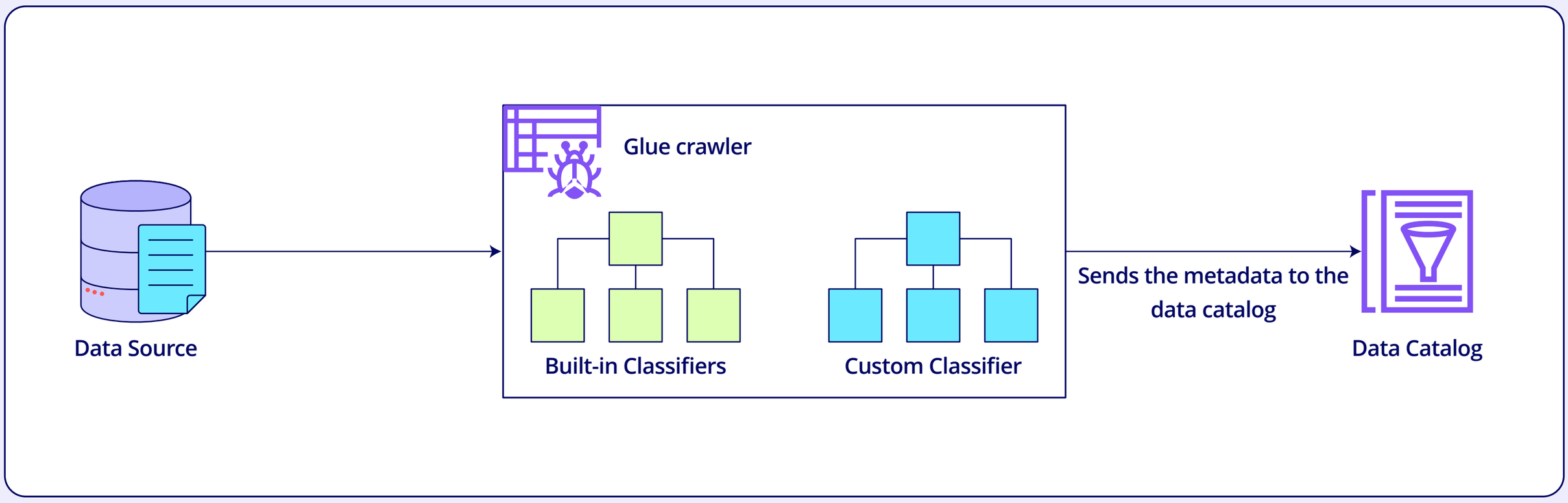


- **Redshift**
  - It’s a petabyte-scale data warehouse service with massively parallel architecture (MPP) to allow performing complex queries in less time.
  - Uses columnar storage, compression, and zone maps for efficient data retrieval, ideal for OLAP databases.

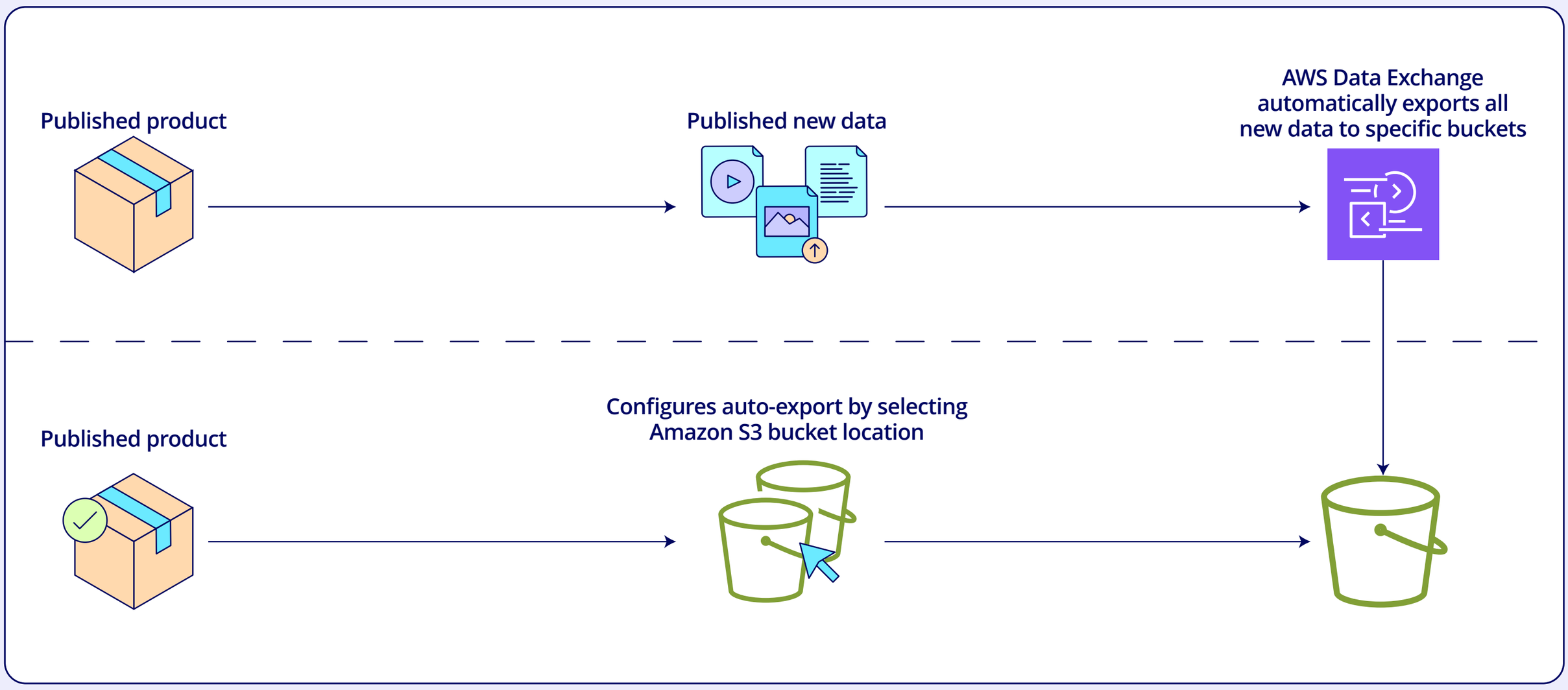


  - The reader nodes receive queries and plan execution.
  - The compute nodes execute these queries and return results to the leader nodes.
  - The leader nodes aggregate the results and return them to the client.

- **Glue Data Catalog**
  - It’s a centralized repository for storing metadata of the data sources.
  - It’s populated through Glue data Crawler, which can crawl file-based and table-based data stores.
  - The crawler determines the schema of the data.

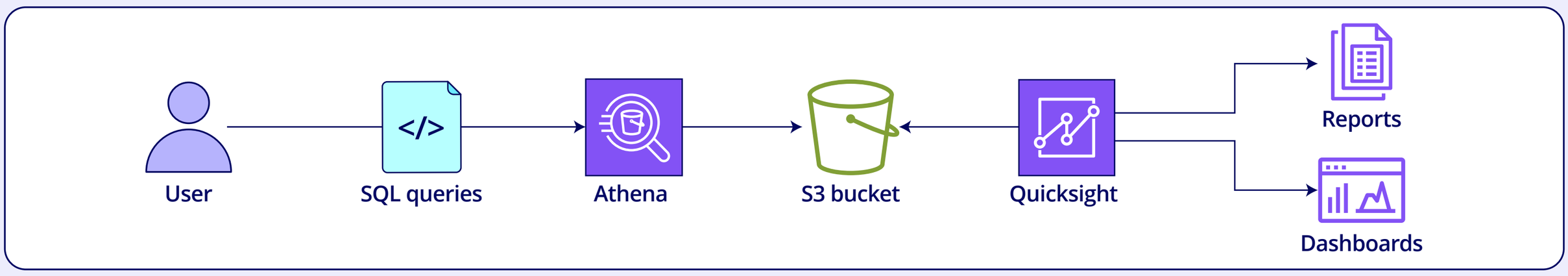


- **Data Exchange**
  - Allows us to publish, subscribe, and use third-party data through AWS Marketplace.



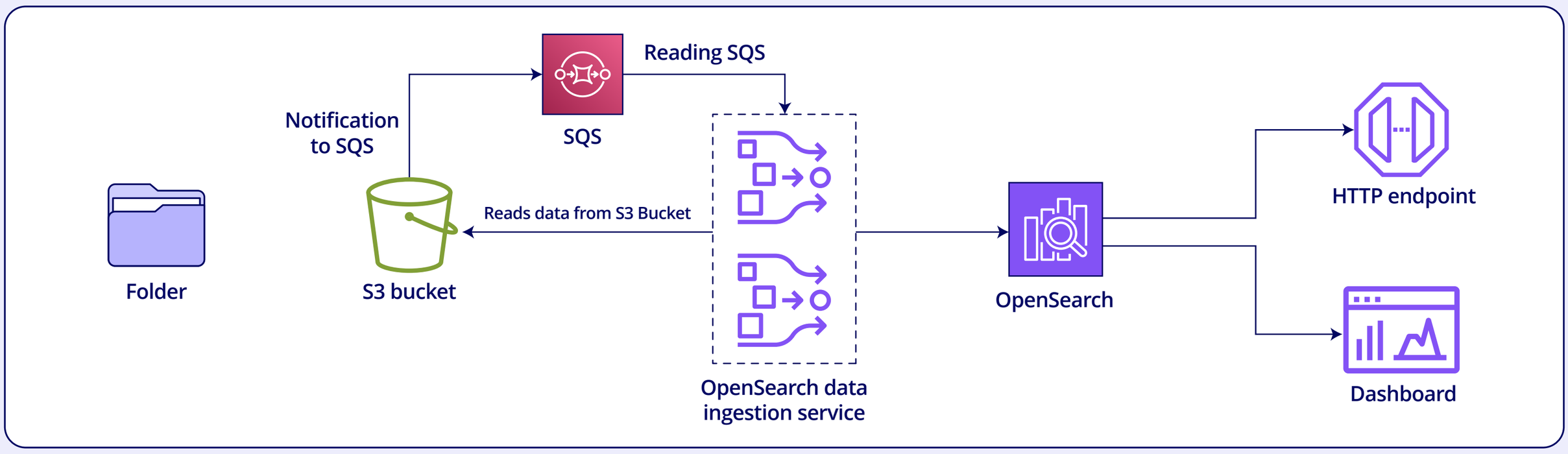
Data Analysis

- **Athena**
  - It’s a serverless analytical service to analyze and query data at its source like S3 using SQL.
  - Supports various formats such as JSON, CSV, and Parquet.
  - Integrates well with QuickSight for data visualization.
- **QuickSight**
  - It’s a business analytics service to visualize and perform ad hoc analysis on data.
  - Offers a library of visualizations, charts, tables, and filters to build interactive dashboards, email reports, and more.



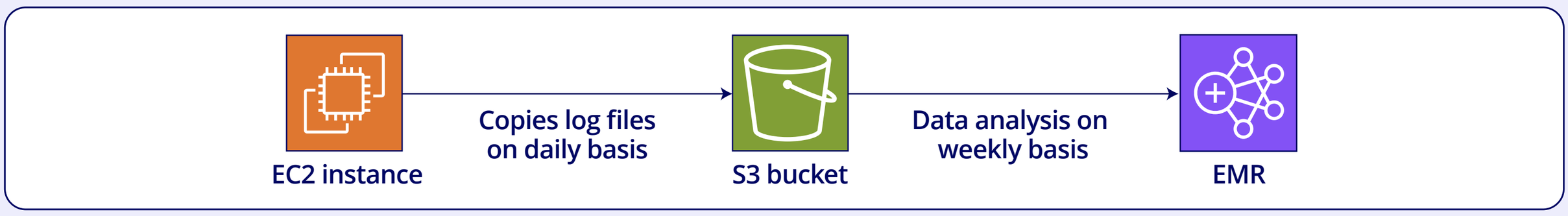
AWS Analytics Services

- **OpenSearch**
  - Allows us to create OpenSearch domains which are managed ElasticSearch clusters.
  - ElasticSearch is a real-time distributed analytics engine.
  - Offers open-source ElasticSearch APIs, Kibana dashboards, and integration with LogStash



Data Pipeline

- Defines the business logic for data movement and transformation.
- Automates data movement by specifying schedules.
- Defines workflows.



Pricing in Analytical Services

Service	Pricing	Service	Pricing
Kinesis Data Streams	**On-demand mode:** Pay per GB of data written and read. **Provisioned mode:** Pay per shard at an hourly rate.	Data Pipeline	Pay based on how often your activities and preconditions are scheduled to run
Kinesis Video Streams	Pay is based on the volume of data ingested, stored, and consumed.	Data Exchange	Pay for the subscribed data. The publisher determines the price of the data
Kinesis Firehose	Pay for the volume of the data ingested and for data format conversion.	LakeFormation	Pay for the integrated services, including S3 and Glue Data Catalog
Kinesis Data Analytics	Pay for the kinesis processing units (KPU) used in an hour.	EMR	Pay for the EC2 instances running in an EMR cluster
MSK	Hourly rate of Apache Kafka broker instance.	Athena	Pay for the data processed by each query
Glue Data Catalog	It can store up to a million objects for free. Charges for objects more than 1 million on a monthly fee basis.	Redshift	Pay per second for the compute instances and the number of bytes scanned by ResShift Spectrum
Glue ETL	Pay for the time the ETL job takes to run an hourly rate.	QuickSight	Allows users to choose between per-user pricing or capacity pricing
Glue Data Brew	Billed for every session of 30 minutes.	OpenSearch	We pay for the EC2 instances and the EBS volumes used within a cluster