Data Lakes in AWS

Exabyte, the new megabyte

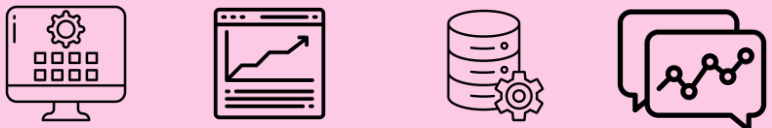
11,500 stores in 27 countries

700,000 associates server 100 million customers weekly

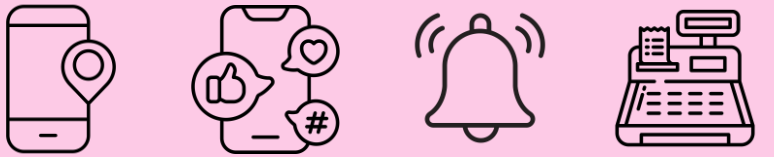
2.5 petabytes of data from 1 million customers every hour

Data Lake on AWS

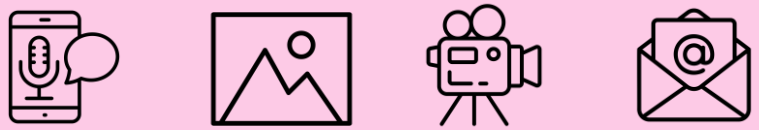
Structured Data
Data that are highly normalized with common schema and stored in relational databases, powering transactional line-of-business applications



Semi-structured data
Data that contain identifiers without conforming to a predefined schema



Unstructured data
Data that do not conform to a data model and are typically stored as individual files



Batch load
Extracts data from various data sources at periodic intervals and moves them to the data lake

Amazon EMR
AWS Glue

Streaming
Ingests data that are generated from multiple sources such as log files, telemetry, mobile apps, and social networks

Amazon MSK
Amazon Kinesis

Amazon S3 Data Lake
Cloud-scale centralized and scalable architecture that enables enterprise data science

Amazon S3

Analytics

Amazon Redshift
Amazon Athena
Amazon QuickSight

Machine Learning

Amazon EMR
Amazon SageMaker
AWS Deep Learning AMIs

Data warehouse

Vs.

Data Lake

Data Warehouse vs. Data Lake

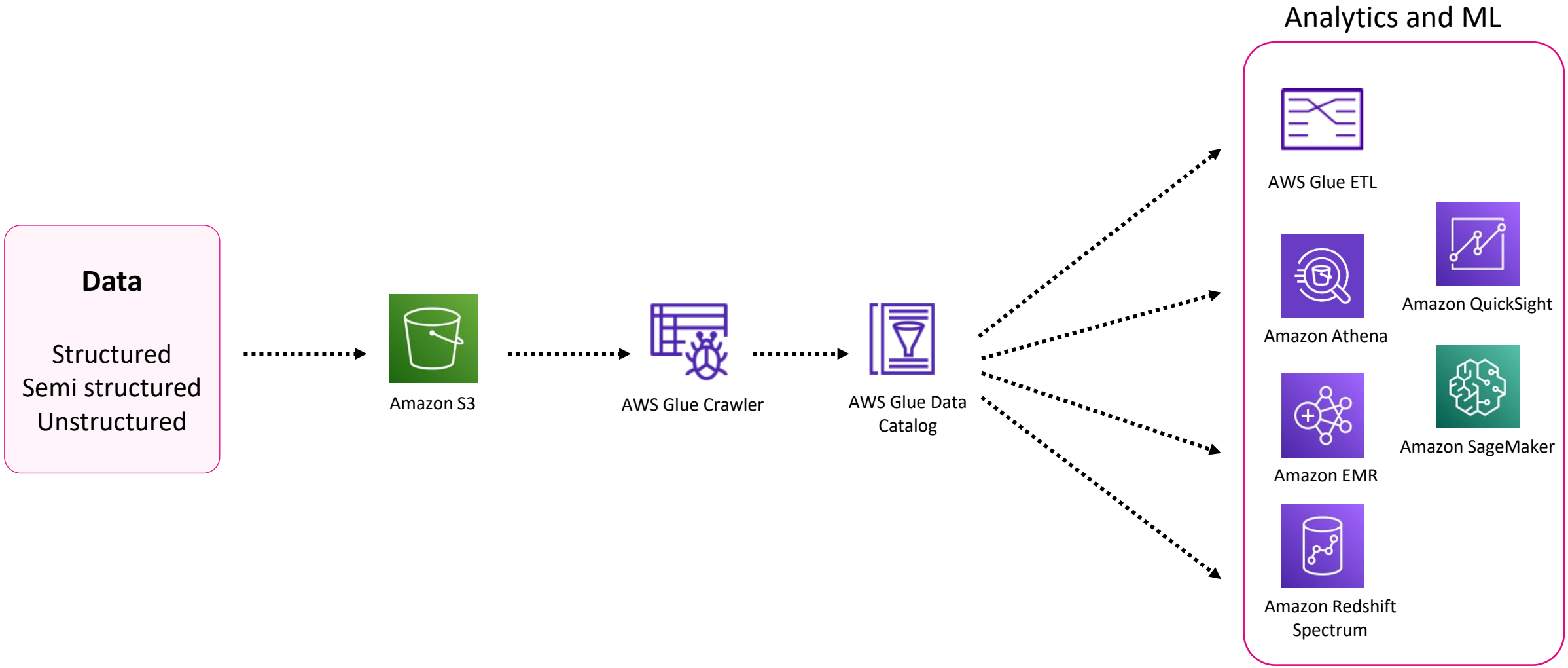
	Data Warehouse	Data Lake
Philosophy	Understand data first, load later	Load first, understand later
Data	Relational, structured data (databases)	Non-relational (object) and relational data
Schema	Schema-on-write	Schema-on-read
Data quality	Highly curated data	Raw data, unstructured data, many formats
Flexibility	Relatively difficult to change as the data is highly structured	Adapts to changes easily as requirements evolve
Users	Operational users - Business analysts	All kind of users – Data scientists, data analysts and business analysts
Performance	Faster query results: table structure	Less performant: Indexes and Catalog

Water Lake vs. Data Lake

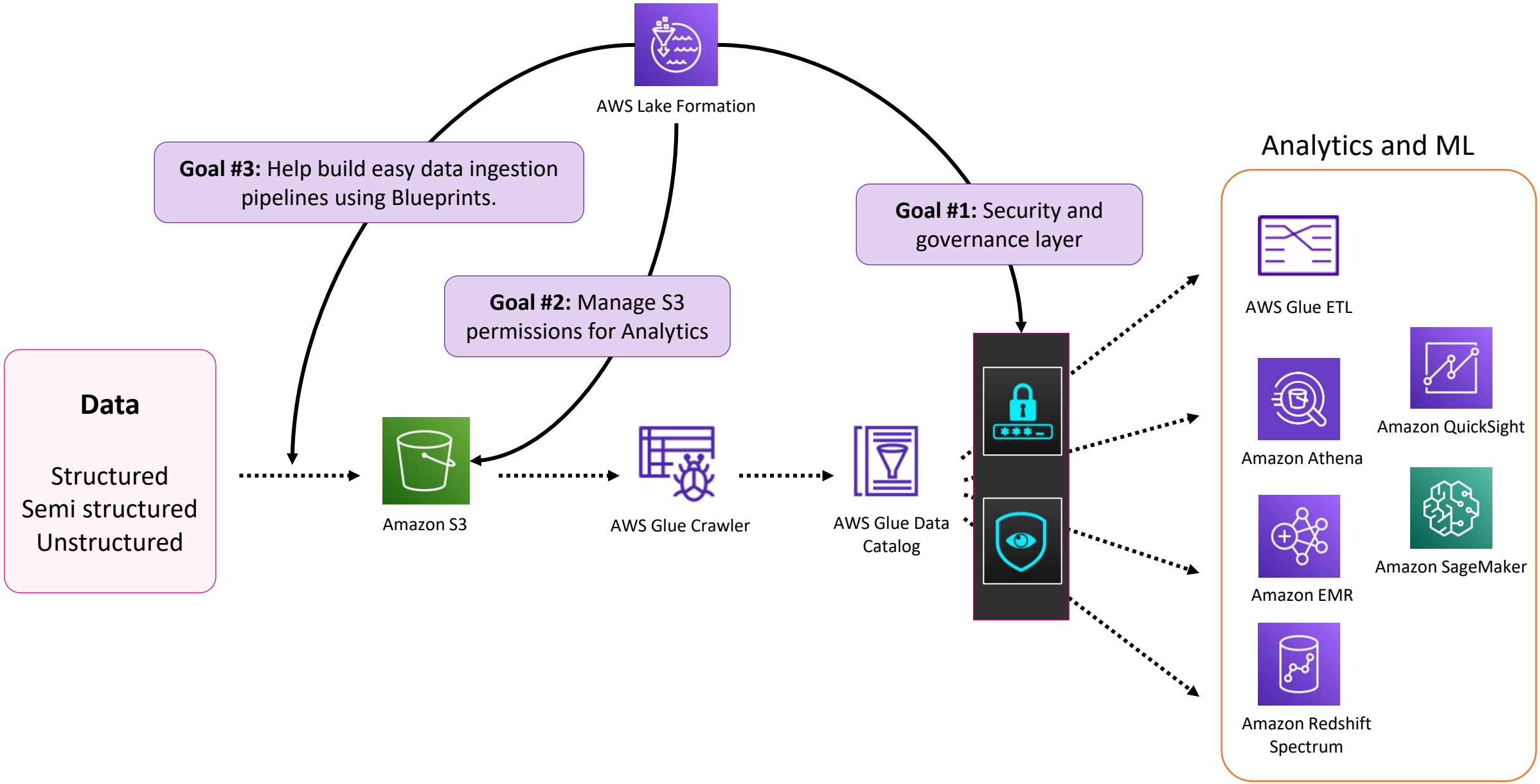
Water Lake	Data Lake
Holds water	Holds Data
Pick a location for the water lake	Pick a data storage Location (Amazon S3)
Dig a lake of certain size	Identify / create an S3 bucket
Identify water sources – creek, river, rain	Identify data sources – Database, Data Warehouse, IoT feed
Identify an overflow approach for water – another lake, reservoir/dam	Identify an overflow approach for data – S3 to Glacier (or other tiers)
Connect water sources to the lake – dig a trench, run a pipe	Connect data sources to the data lake – DMS, Direct Connect, Kinesis Firehose
Bring water to the lake from water sources using one or more connection methods	Bring data to the data lake from data sources using one ore more connection techniques
Process incoming water using standard or custom purification techniques	Process incoming data using built-in and/or custom functions
Fill the lake with processed water	Fill the data lake (S3 bucket) with transformed data
Run different tests on lake water – chemical composition, bacterial load	Run different queries on data in the data lake
Visualize test results	Visualize query results

AWS Lake Formation

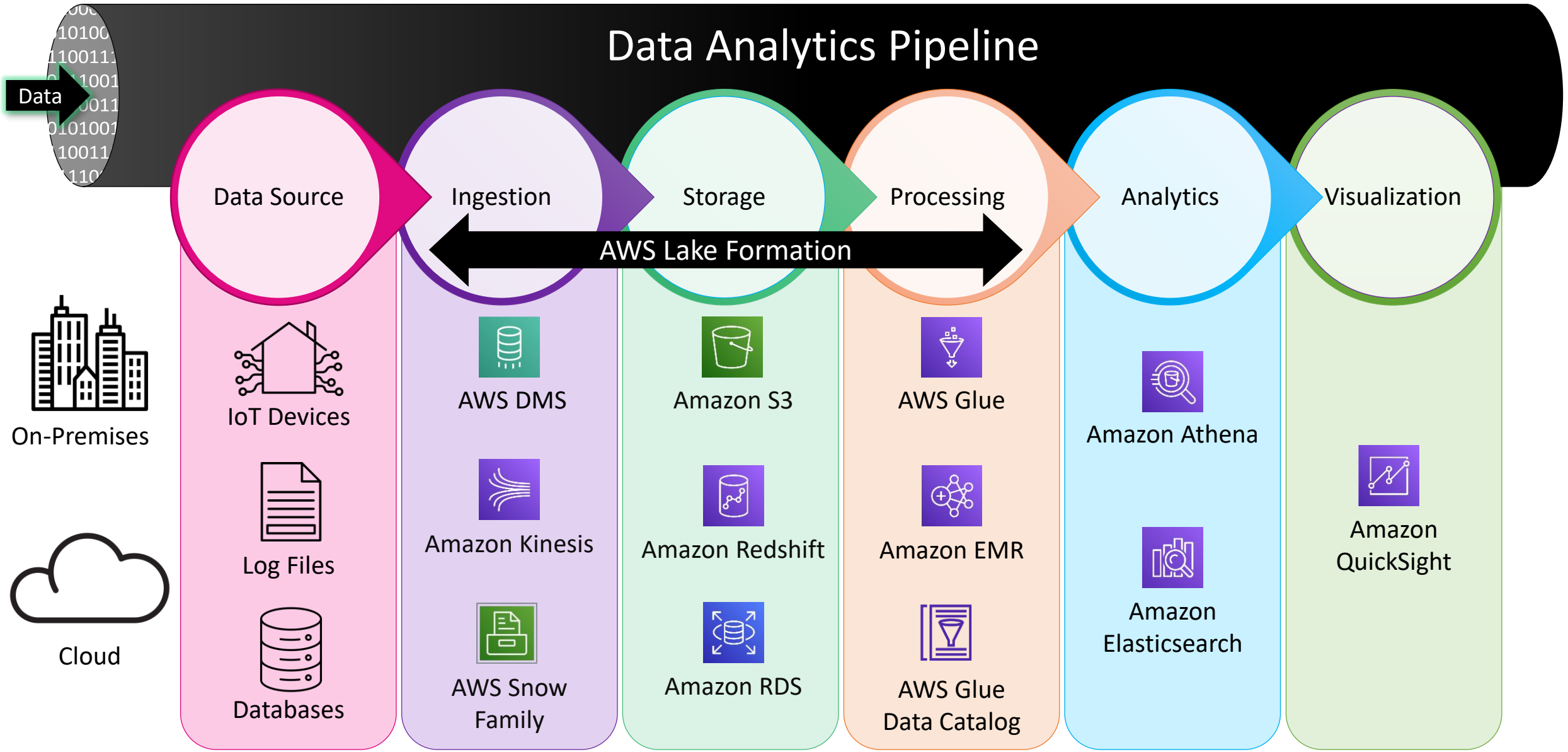
Data Lake on AWS without AWS Lake Formation



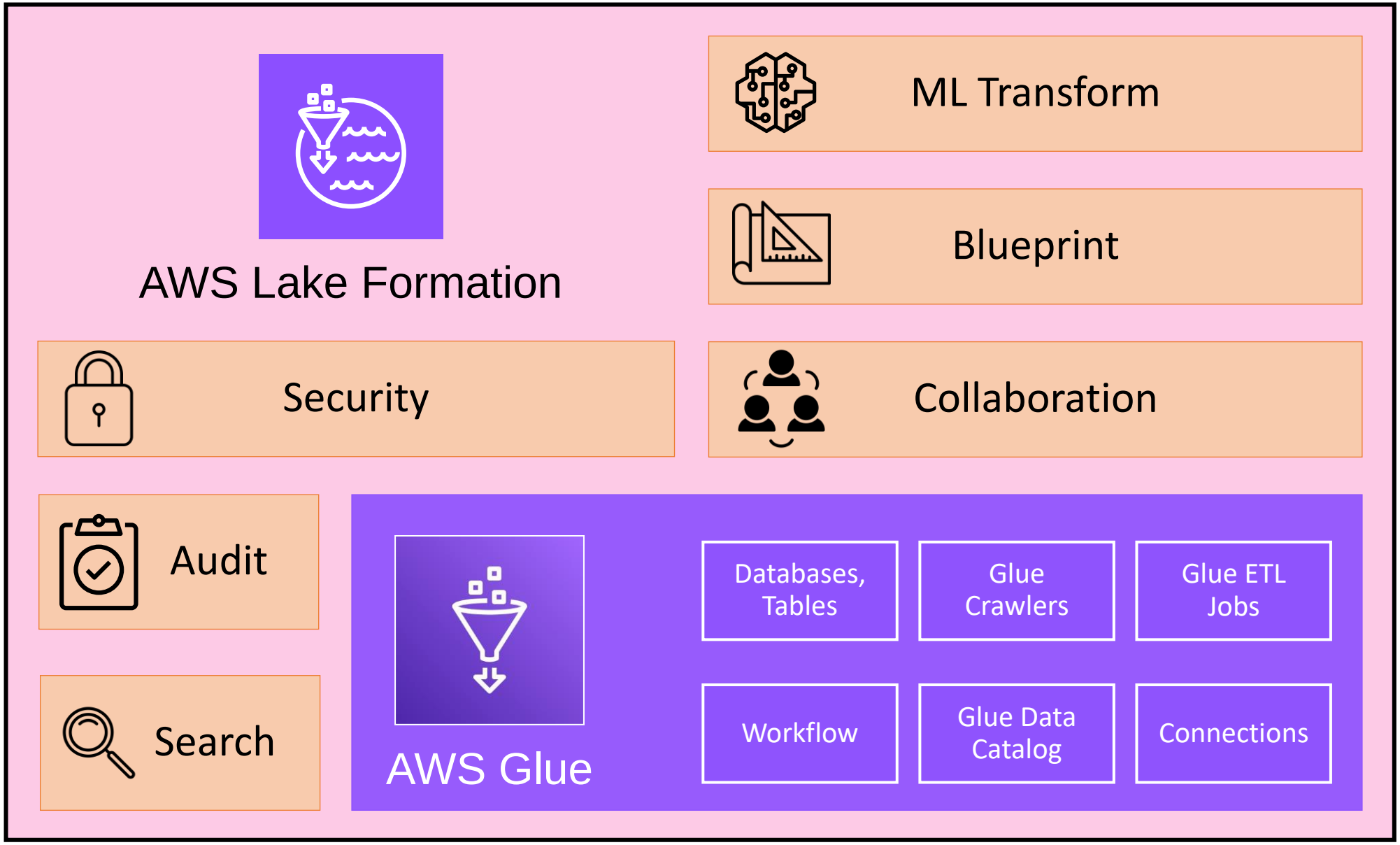
Data Lake on AWS with AWS Lake Formation



Data Analytics on AWS



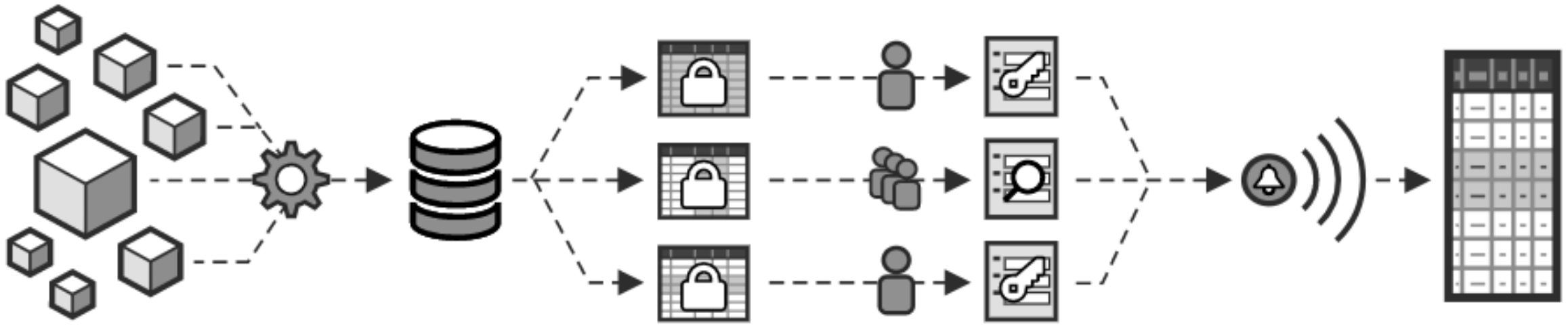
AWS Lake Formation and AWS Glue



How it works?

AWS Lake Formation

How it works



Ingest & Organize

Automatically ingest, clean, encrypt, and register existing Amazon S3 bucket content, including log data from CloudTrail, CloudFront, and Amazon ELB.

Secure & Control

Define access control that provides the right data to the right users, groups, and roles. Flexible database, table, and column permissions enable granular security.

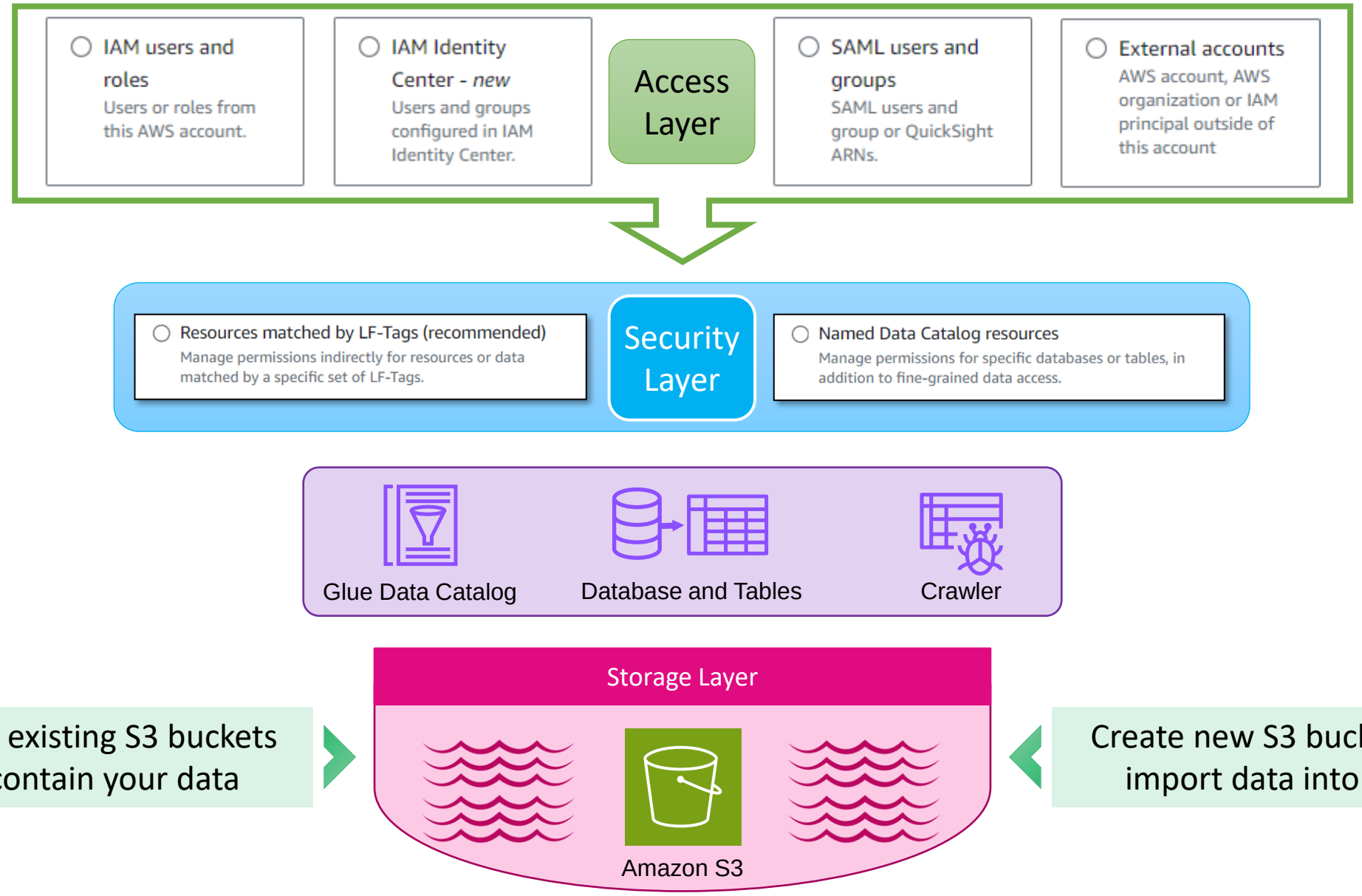
Collaborate & Use

Search and discover using catalog metadata. All access is checked against policy, so your data is protected even if tools change or new data arrives.

Monitor & Audit

Be alerted of access requests and policy exceptions. Review activity history with detailed change logs and data lineage.

Data Lakes using AWS Lake Formation



AWS Lake Formation Personas

IAM Administrator ≠ Data Lake Administrator

IAM administrator
(Required)

User who can create IAM users and roles. Has the AdministratorAccess AWS managed policy.

Data Lake Admin
(Required)

User who can register Amazon S3 locations, access the Data Catalog, create databases, create and run workflows, grant Lake Formation permissions to other users.

Workflow Role
(Required)

Role that runs a workflow on behalf of a user.

Data Engineer
(Optional)

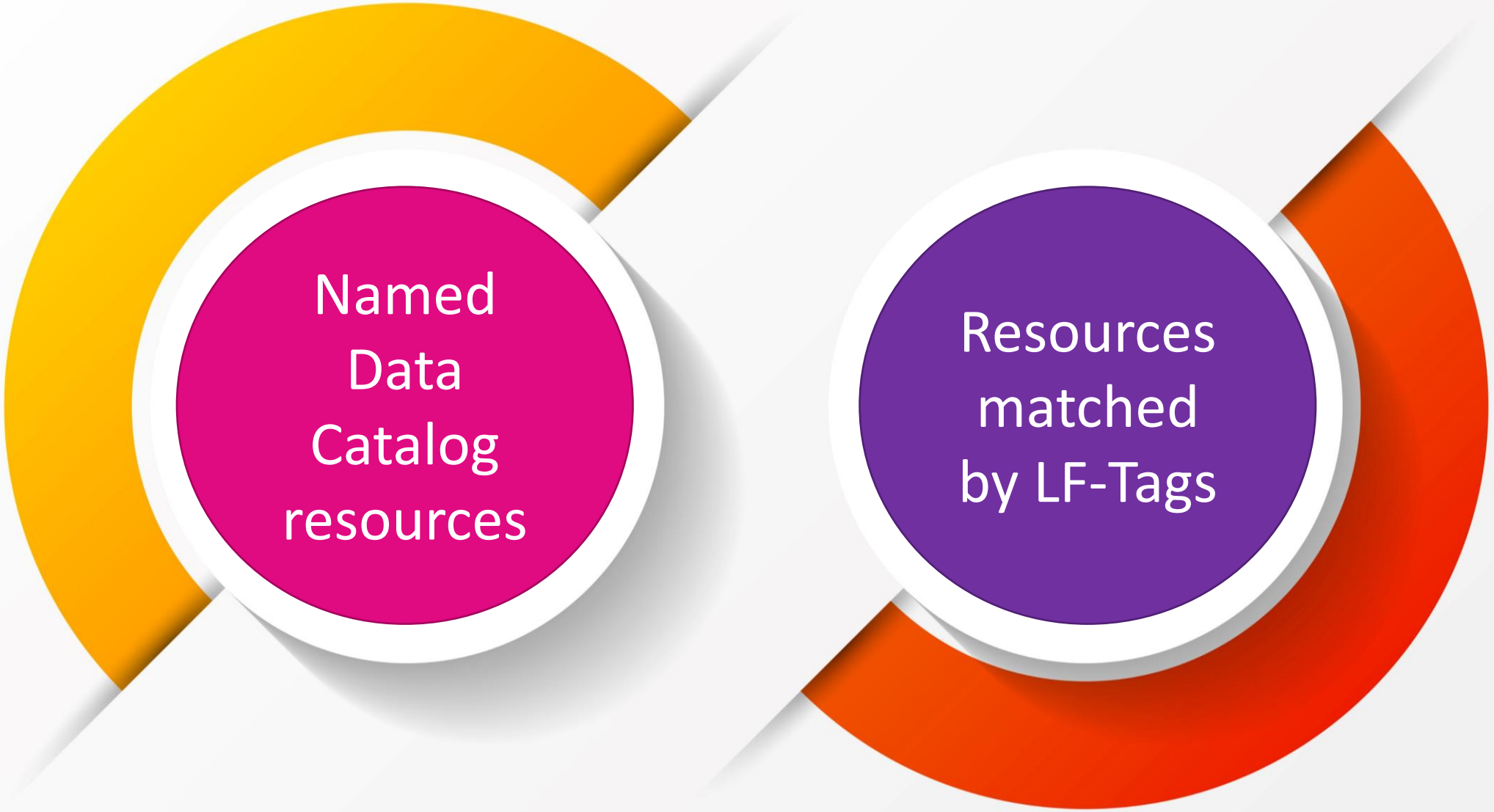
User who can create and run workflows and grant Lake Formation permissions on the Data Catalog tables that the workflows create.

Data Analyst
(Optional)

User who can run queries against the data lake using, for example, Amazon Athena.

Permissions on Data Lake

Granting Permission



Named Data Catalog resources

- Manage permissions for specific databases or tables, in addition to fine-grained data access.

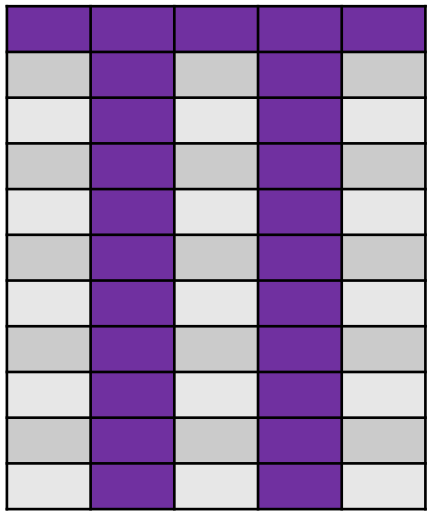
Database and Tables

Column name	Data type
marketplace	string
customer_id	bigint
review_id	string
product_id	string
product_parent	bigint
product_title	string
star_rating	string
helpful_votes	bigint
total_votes	bigint
vine	string
verified_purchase	string
review_headline	string
review_body	string
review_date	string
product_category	string

User 2

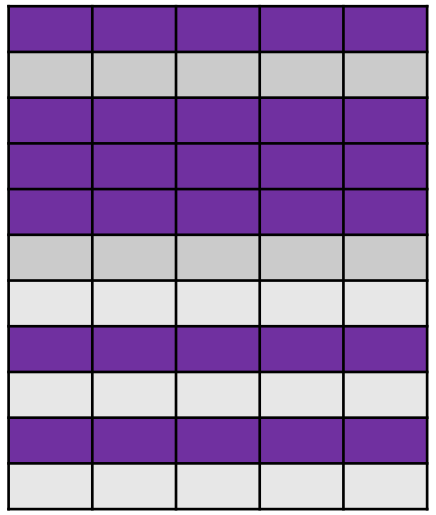
User 1

Columns



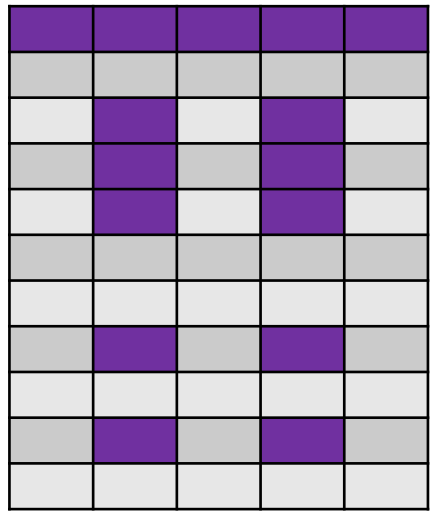
Specify include or exclude list

Rows



Specify row filter with PartiQL

Cells



Combine column and row filters

Resources matched by LF-Tags (Recommended)

- Manage permissions indirectly for resources or data matched by a specific set of LF-Tags.

Define LF-Tag creators and LF-Tags

Create an ontology of attributes or LF-Tags, and decide who can create/manage LF-Tags.

Assign LF-Tags to catalog

Associate combinations of LF-Tags (key & value) to specific databases, tables and columns.

Grant LF-Tag based access

Define scalable permissions that grant access to catalog resources via specific LF-Tag combinations.

Retire old resource access

Revoke direct resource access that are superseded by LF-Tag based permissions.

Open Table Format

Support for open table formats

Linux Foundation
Delta Lake

Apache
Iceberg

Apache
Hudi