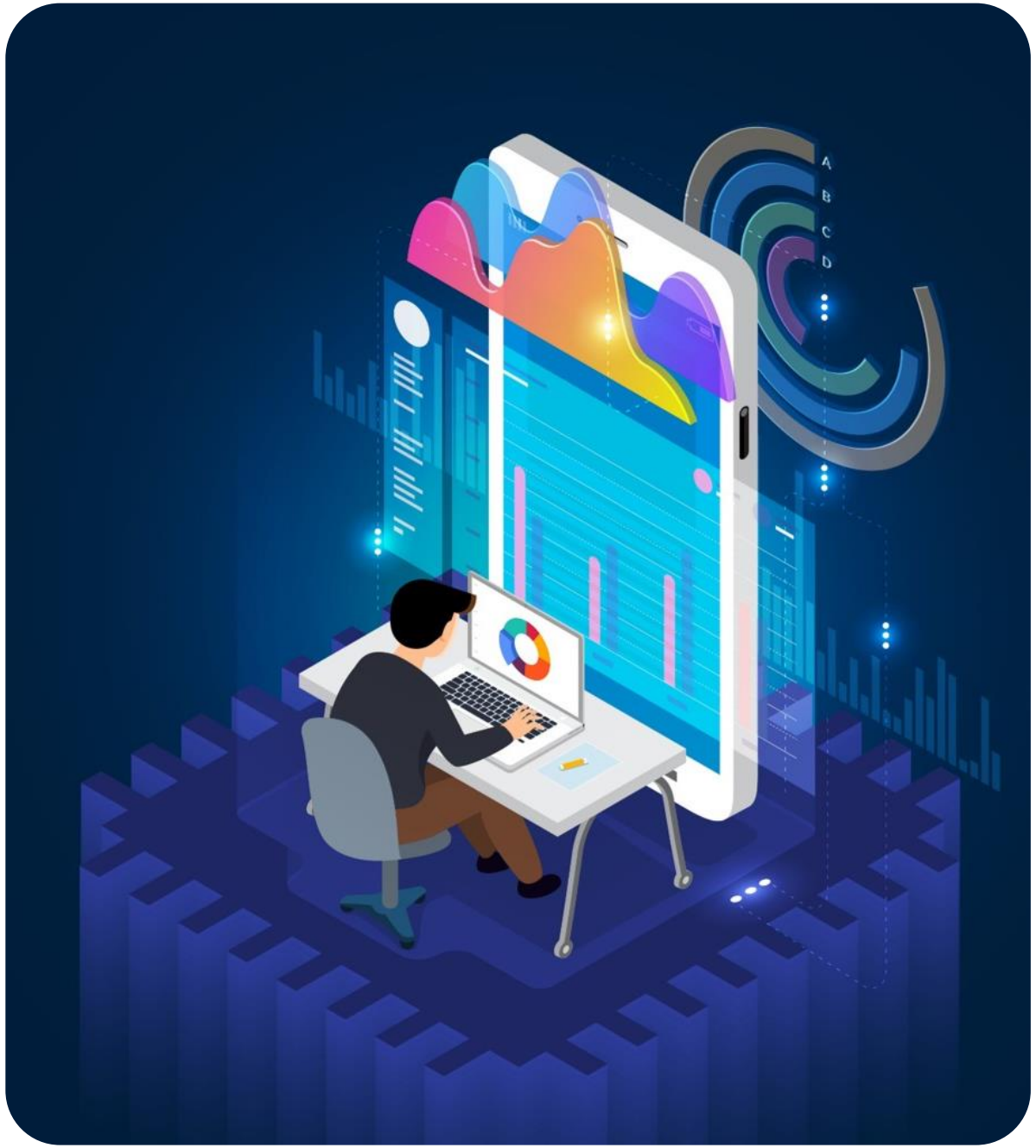
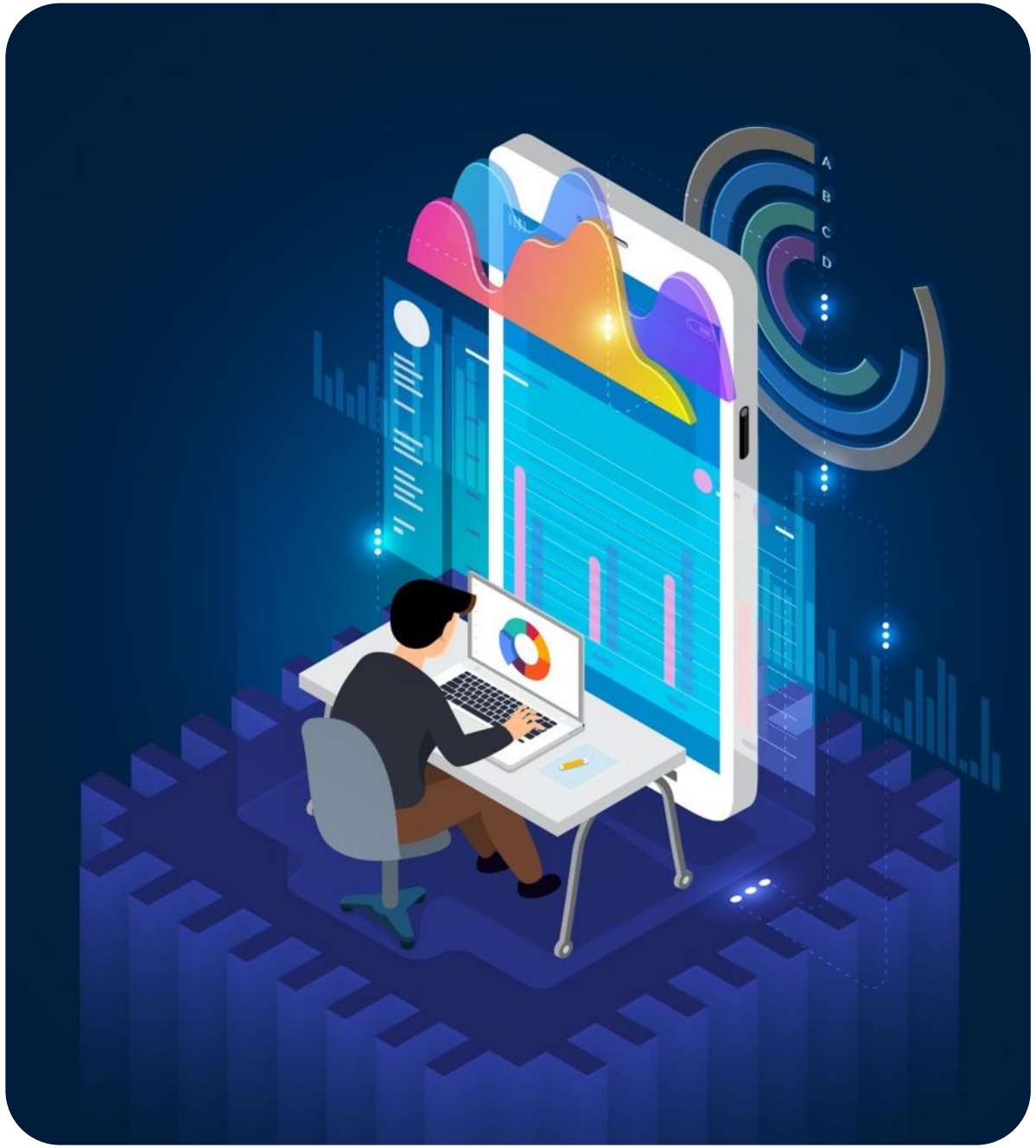
Data Analytics on AWS



Agenda

- What is Data Analytics?
- Basic Terminologies
- Data Analytics on AWS
  - AWS Glue
  - Amazon Athena
  - Amazon Redshift
  - AWS LakeFormation
  - Amazon EMR
  - Amazon Kinesis
  - Amazon QuickSight

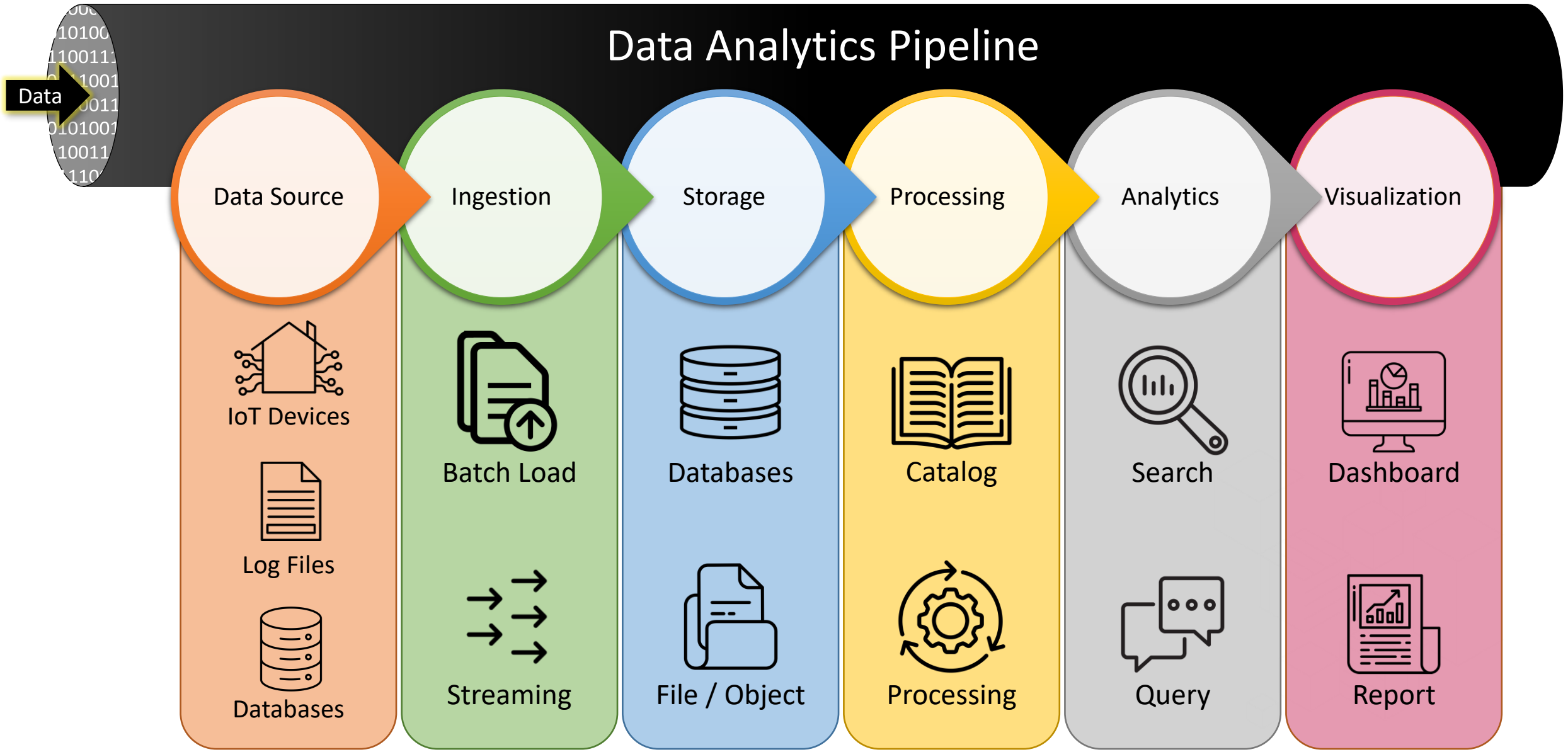


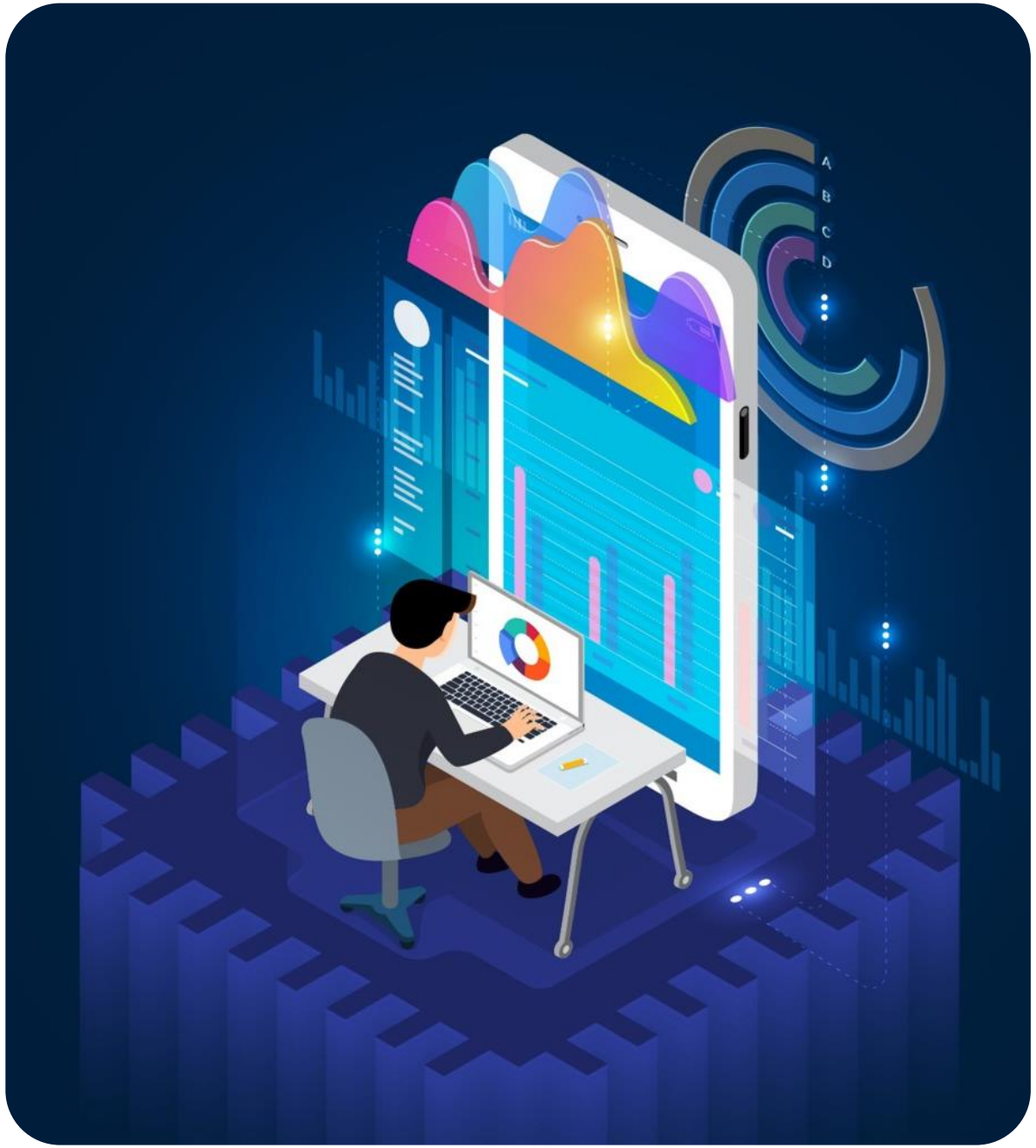
What is Data Analytics?

What is Data Analytics?

- Data analytics converts raw data into actionable insights.
- It includes a range of tools, technologies, and processes used to find trends and solve problems by using data.
- Data analytics helps companies gain more visibility and a deeper understanding of their processes and services.
- It gives them detailed insights into the customer experience and customer problems.

How does data analytics work?



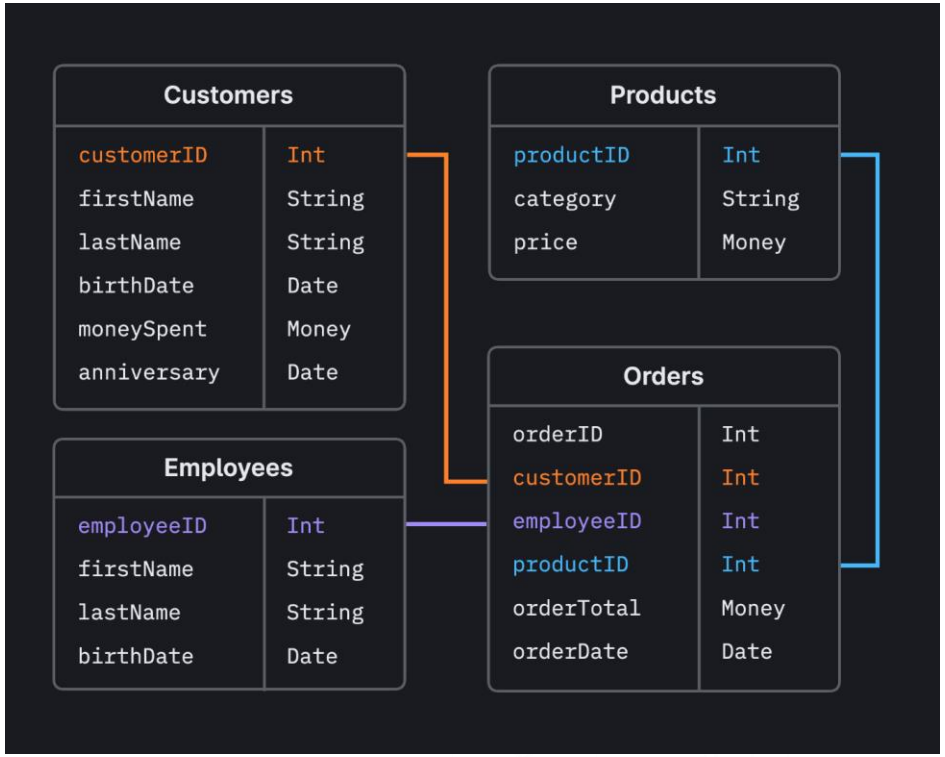


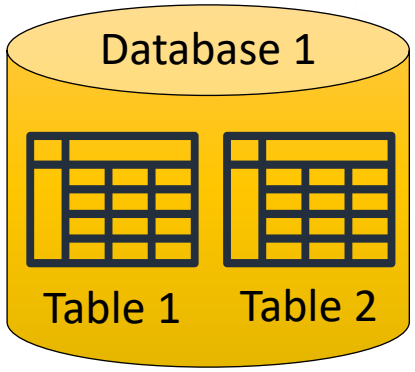
Basic Terminologies

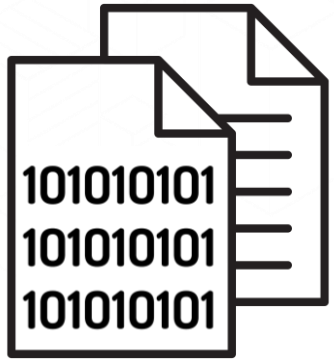
Database Schema

- Logical Representation of the Database
- It defines how the data is organized and how the relations among them are associated.
- It formulates all the constraints that are to be applied on the data.
- This process of database schema design is also known as data modeling.

Database Schema

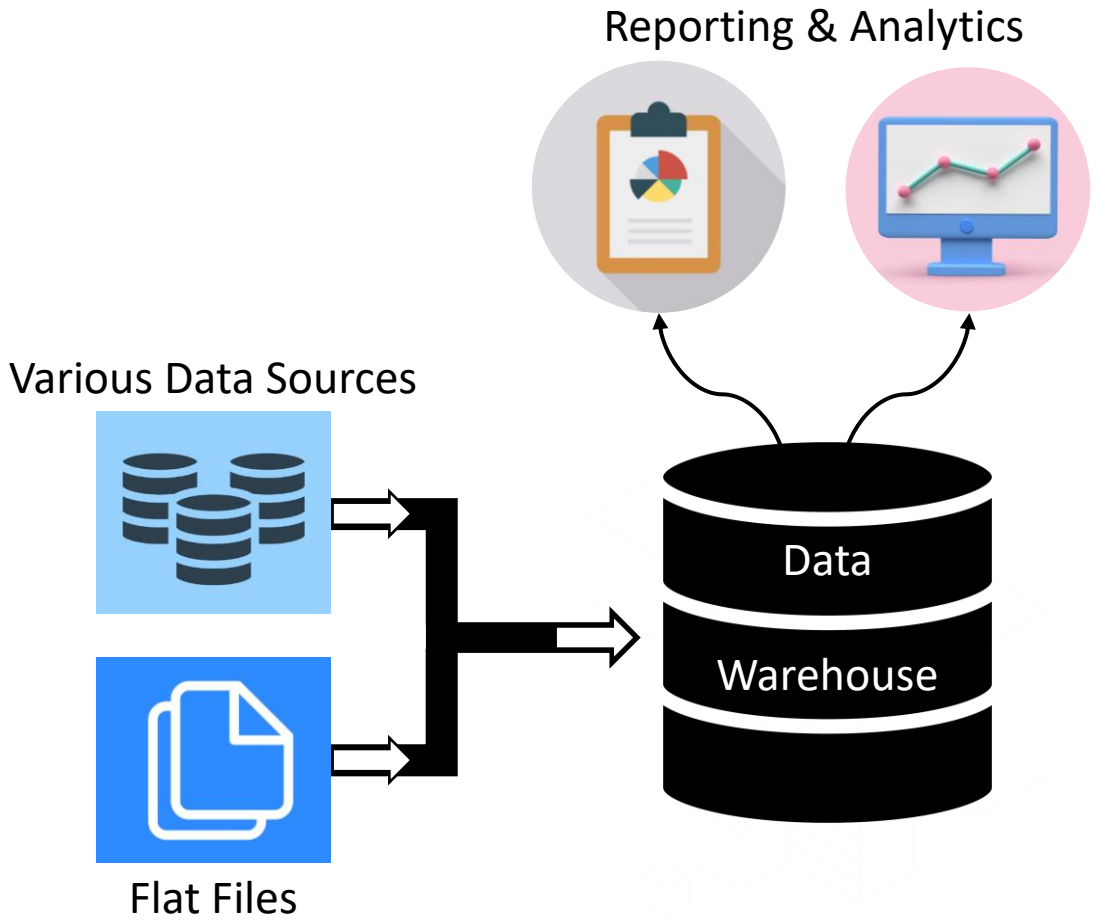






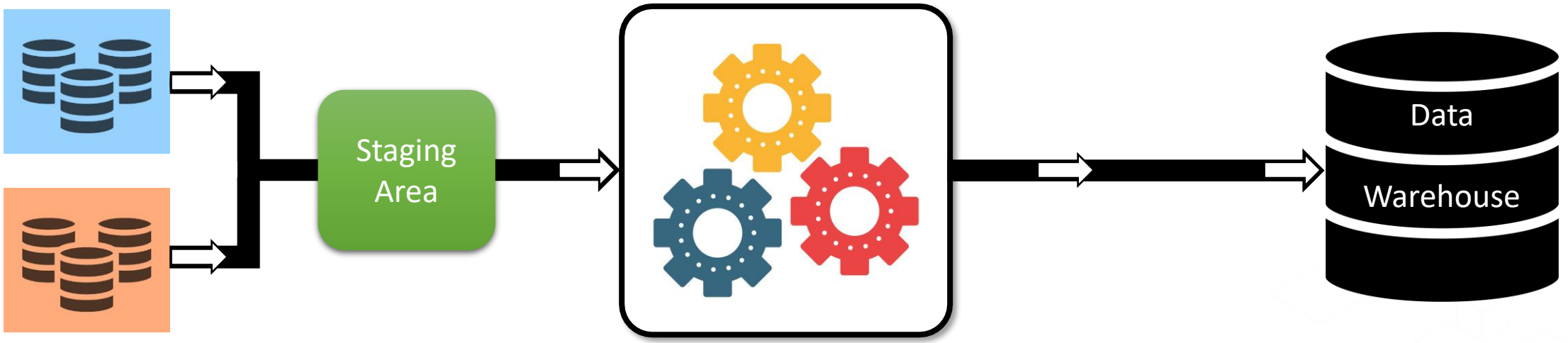
Datawarehouse

- A data warehouse is a central repository of information that can be analyzed to make more informed decisions.
- Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence.
- A database is designed to supply real-time information. A data warehouse is designed as an archive of historical information.



Extract, transform, and load (ETL)

- Process of combining data from multiple sources into a large, central repository



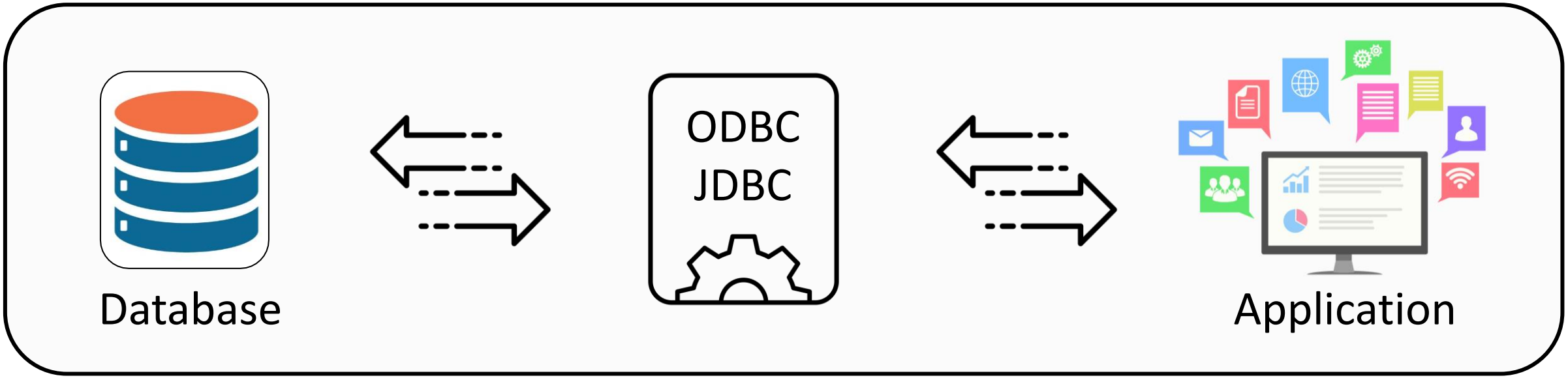
Extract
the relevant data from the source database

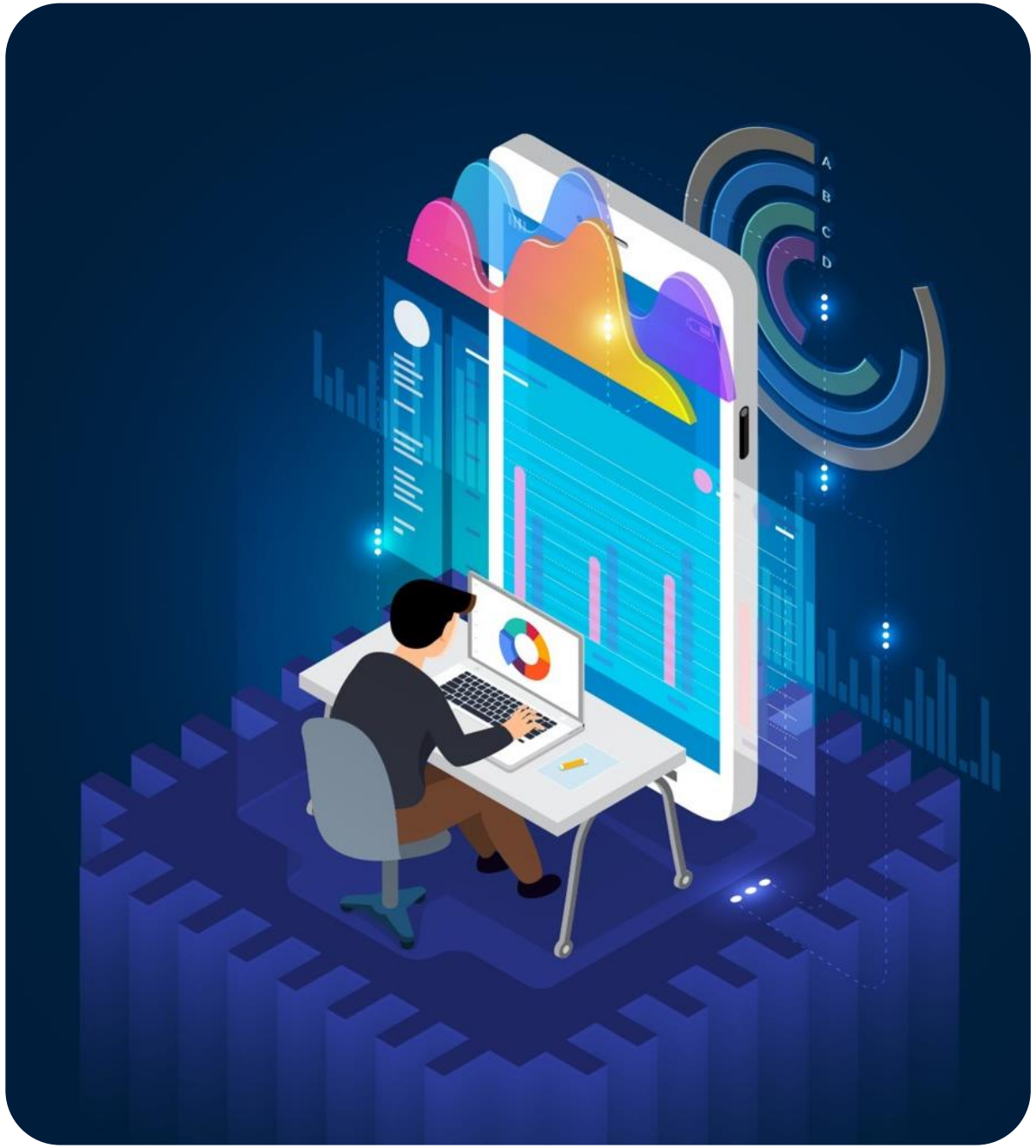
Transform
the data so that it is better suited for analytics

Load
the data into the target database

ODBC and JDBC

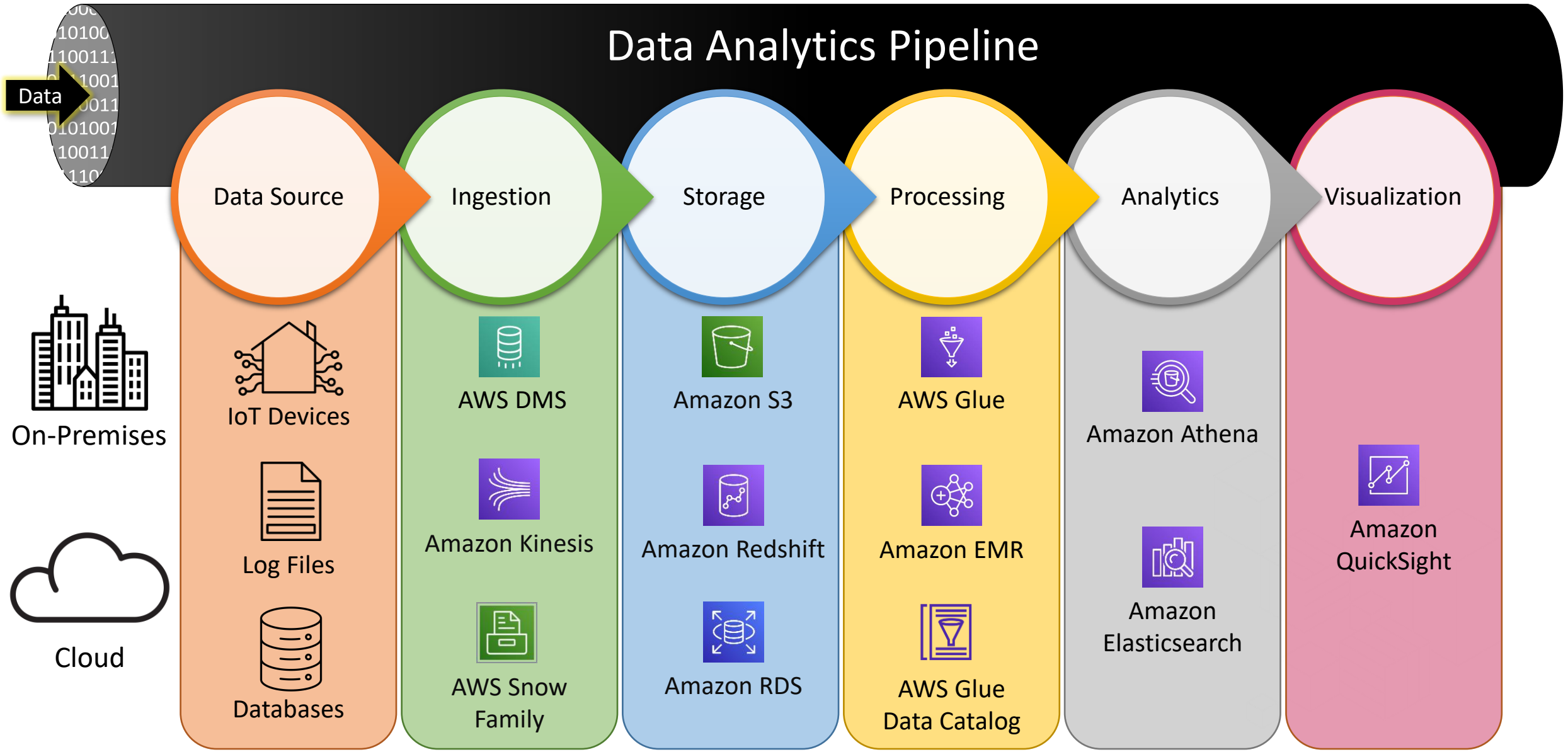
- Open Database Connectivity (ODBC) and JDBC (Java Database Connectivity), both are the API (Application Programming Interface) that help the applications on the client side to access the database on the server side.
- ODBC is created by Microsoft and JDBC created by Sun Microsystems.
- These APIs provide communications between an application residing on a client machine and a data source residing on the same client machine or on another server computer.

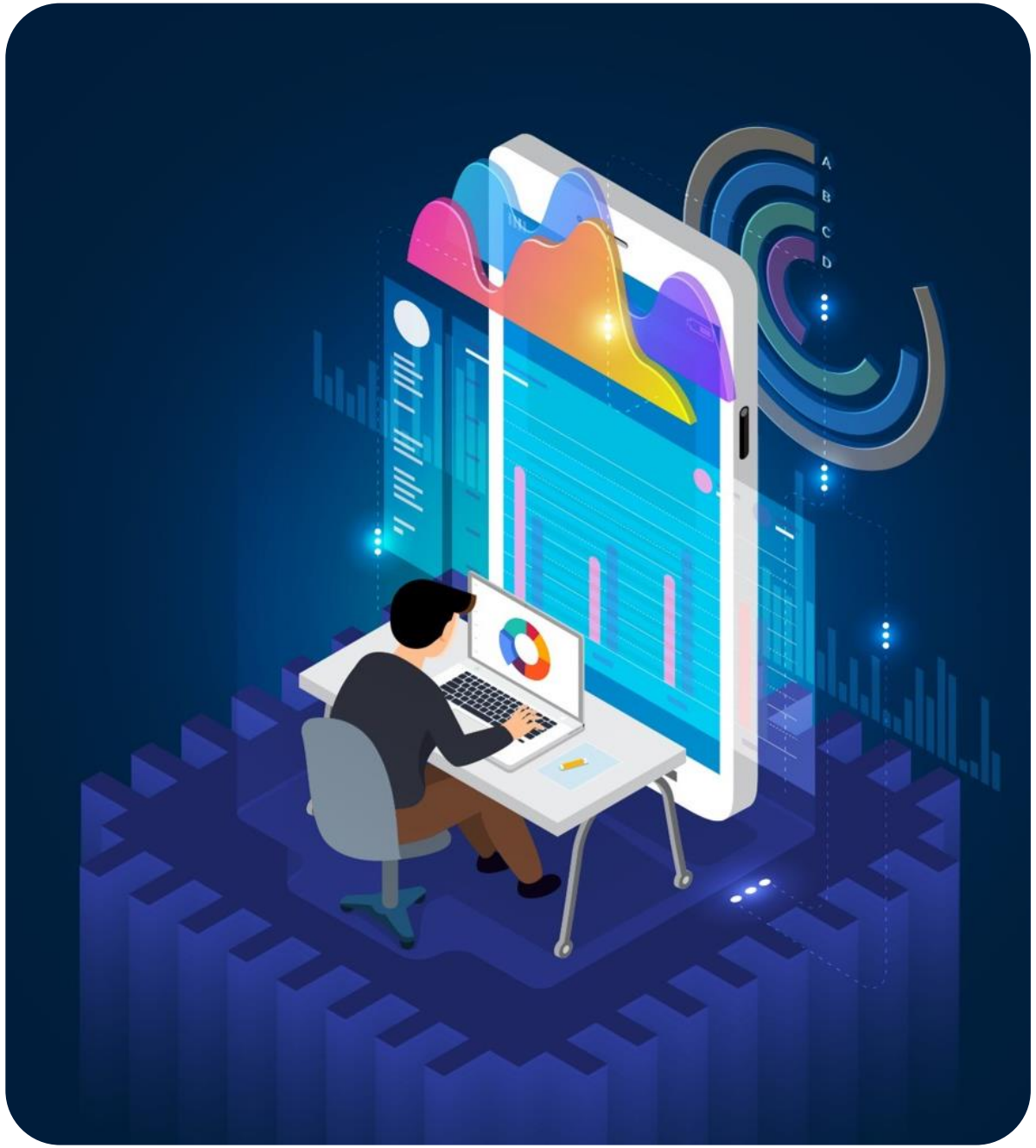




Data Analytics on AWS

Data Analytics on AWS





AWS Glue

AWS Glue

- AWS Glue is a serverless data integration service.
- It provides all the capabilities needed for data integration, so you can start analyzing your data and putting it to use in minutes instead of months.

Catalog

Crawler

AWS Glue
Data Catalog

Process

AWS Glue Studio

AWS Glue
Development Endpoints

Glue Workflow

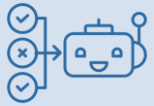
DevOps

Serverless Spark

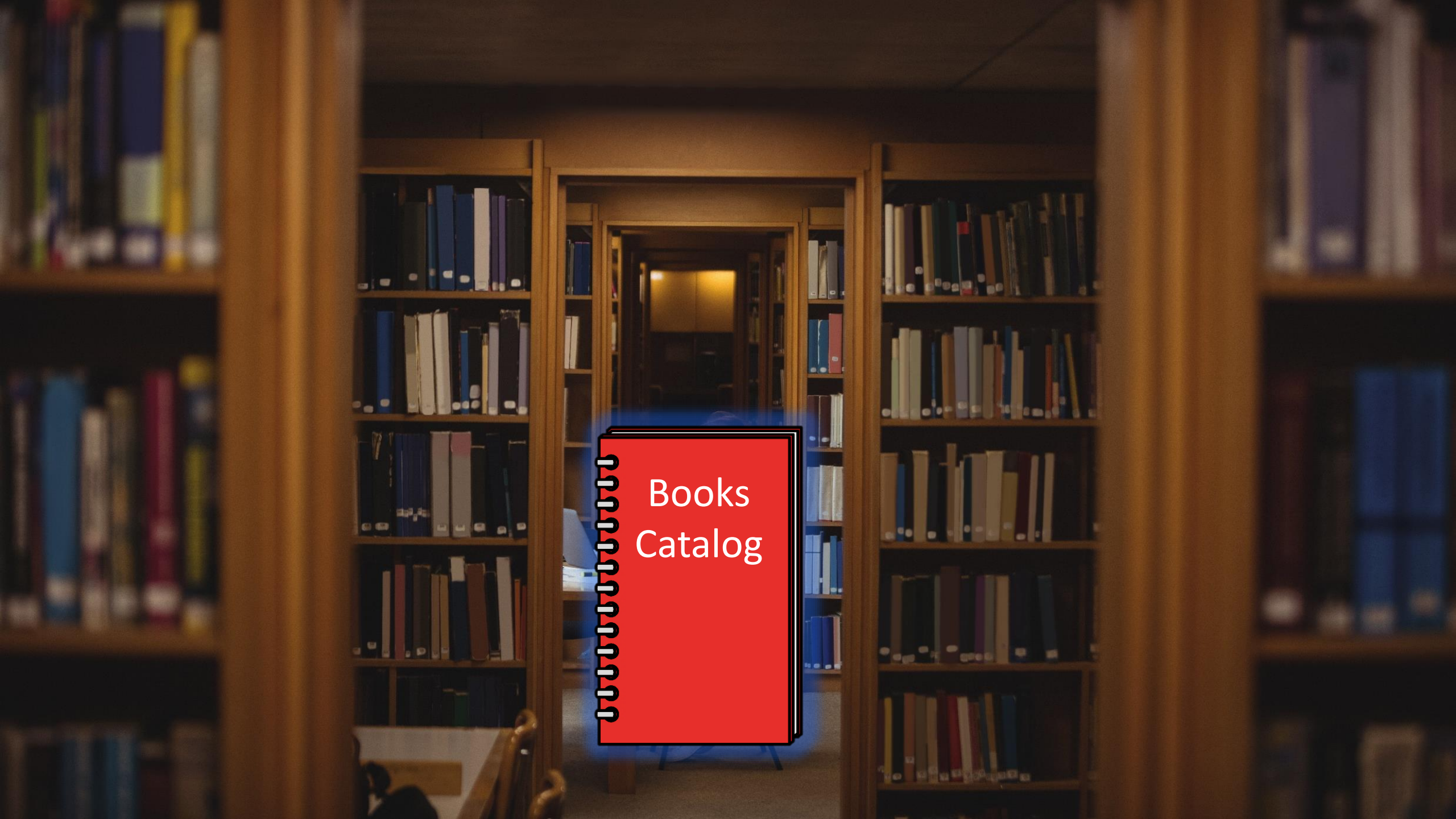
Dynamic Frames

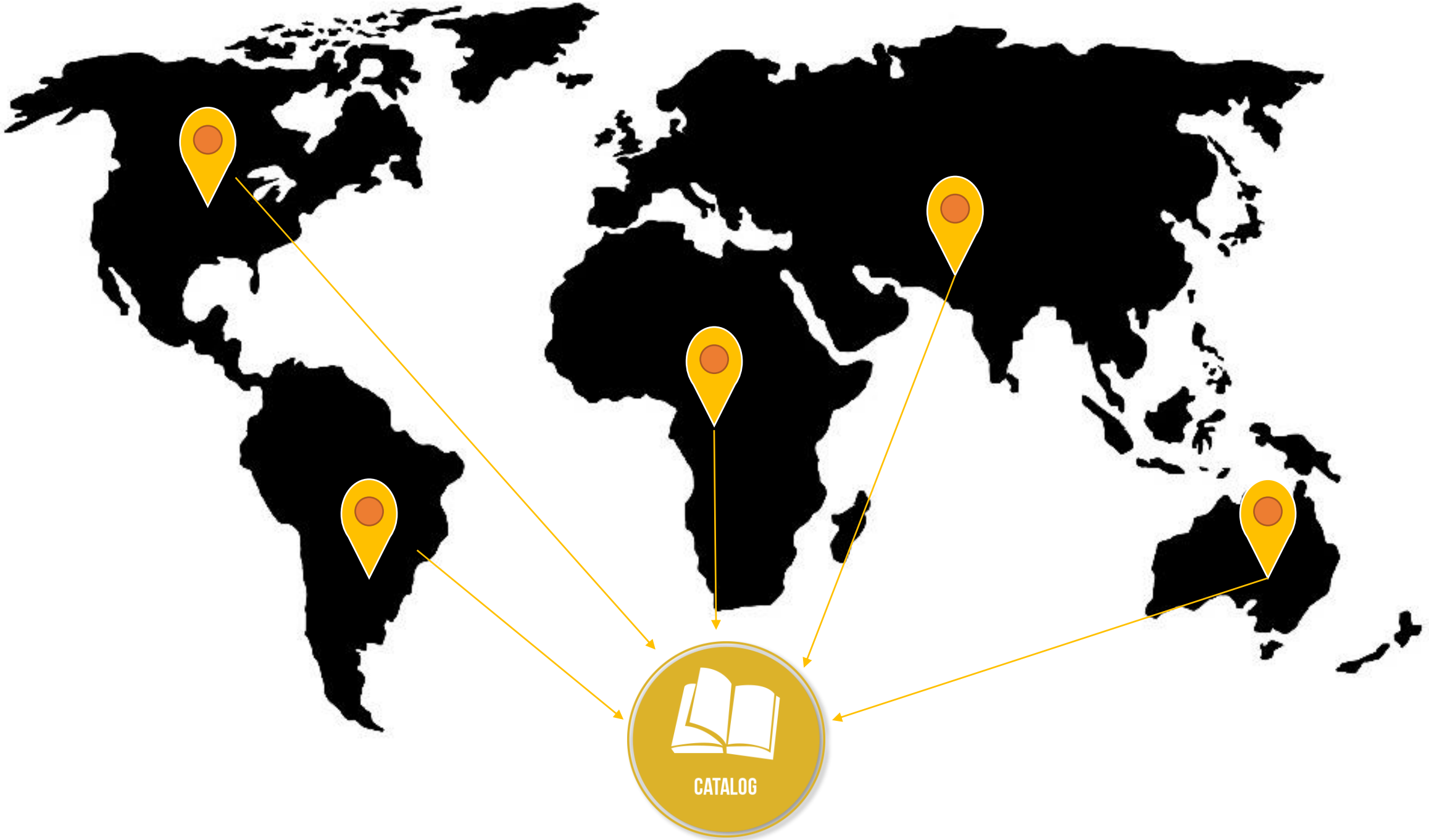
Glue Streaming

Quality



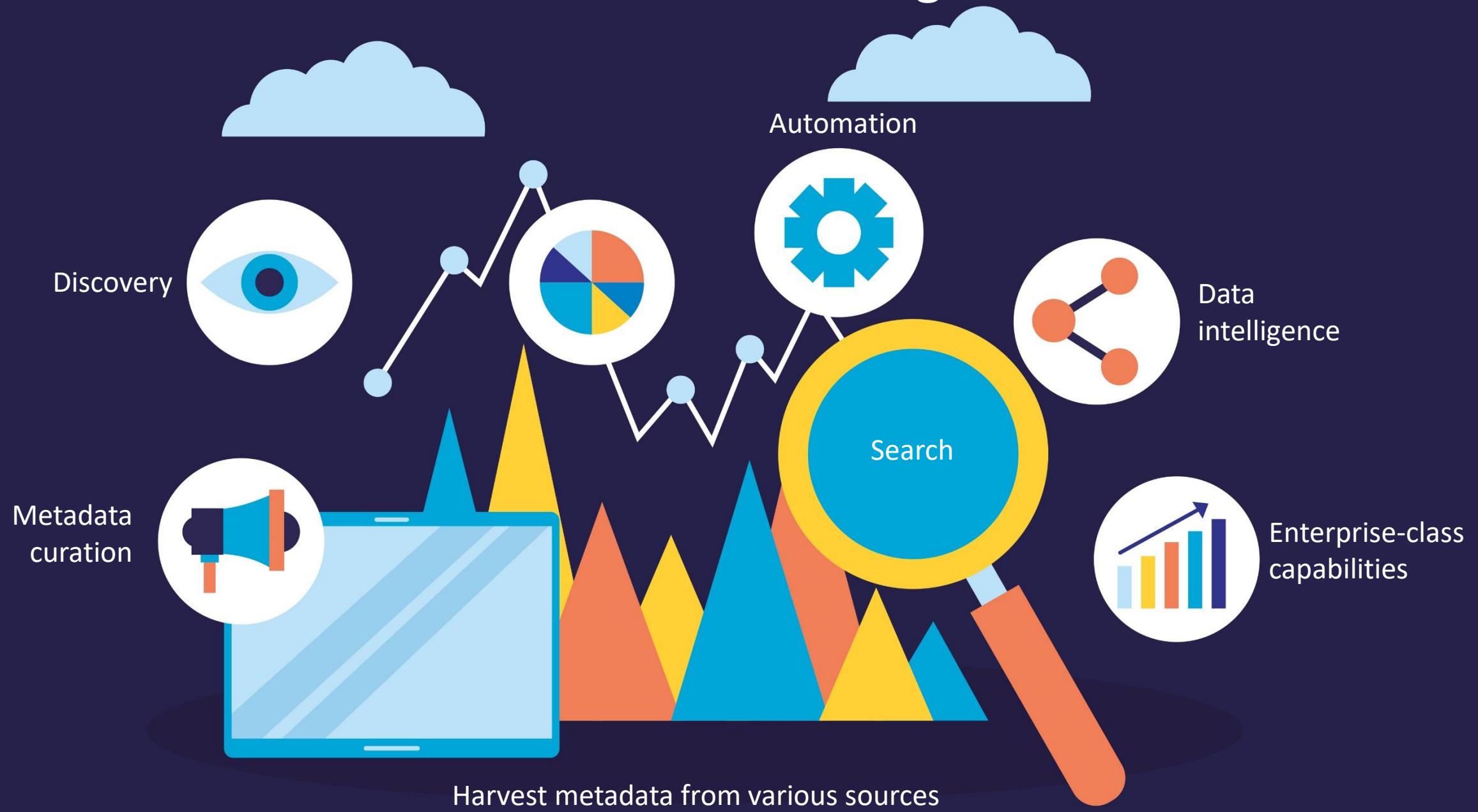
Machine Learning
powered data cleansing


Books
Catalog


CATALOG

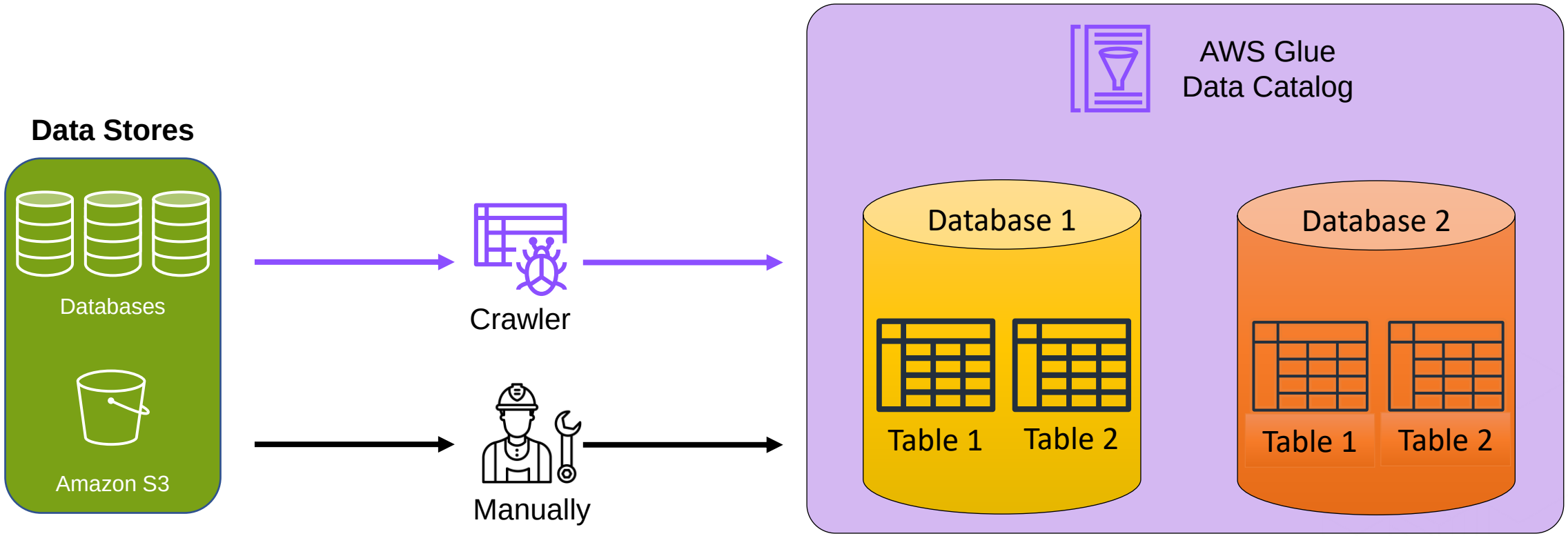
Data Catalog

What Should a Data Catalog Offer?



Glue Catalog

- Catalog - An index to the location, schema, and runtime metrics of your data.
- Databases - A set of associated table definitions, organized into a logical group.
- Tables - The metadata definition that represents your data, including its schema.



The AWS Glue Data Catalog is an Apache Hive metastore-compatible catalog.

Service Summary Cards (SSC)

Reference:

FAQs

Category:

Analytics

AWS Glue

More SSCs:

Click Here

Complete Book

Click Here

Created by:

Ashish Prajapati

What?

- AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

Why?

- AWS Glue provides all the capabilities needed for data integration, so you can start analyzing your data and putting it to use in minutes instead of months.

When?

- You should use AWS Glue to discover properties of the data you own, transform it, and prepare it for analytics.
- When you need unified catalog to find data across multiple data stores, explore data with self-service visual data preparation or create, run, and monitor ETL jobs without coding.

Where?

- AWS Glue is a Regional service.

Who?

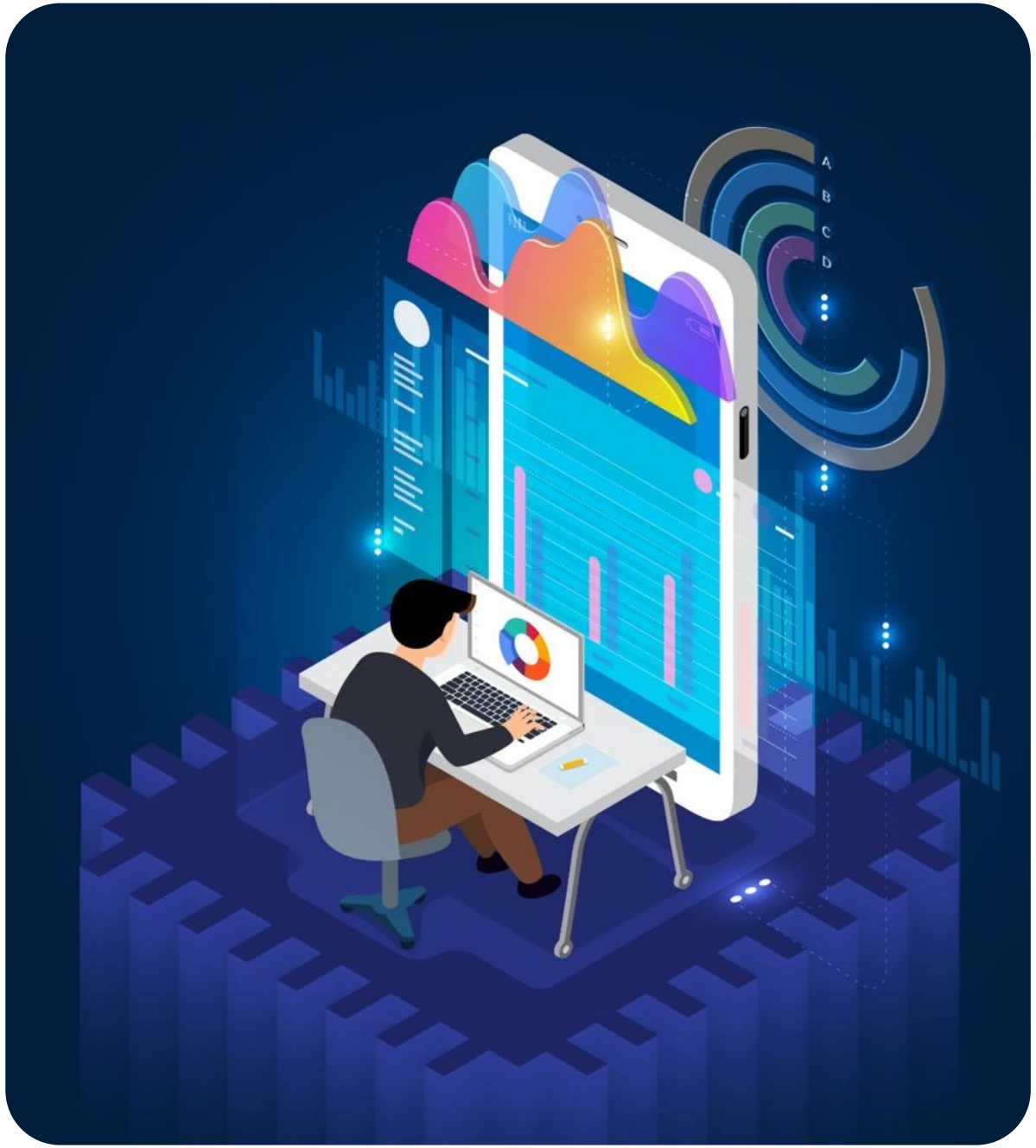
- AWS Glue provides a managed ETL service that runs on a serverless Apache Spark environment. This allows you to focus on your ETL job and not worry about configuring and managing the underlying compute resources.

How?

- You define jobs in AWS Glue to accomplish the work that's required to extract, transform, and load (ETL) data from a data source to a data target.
- AWS Glue can generate a script to transform your data. Or, you can provide the script in the AWS Glue console or API.

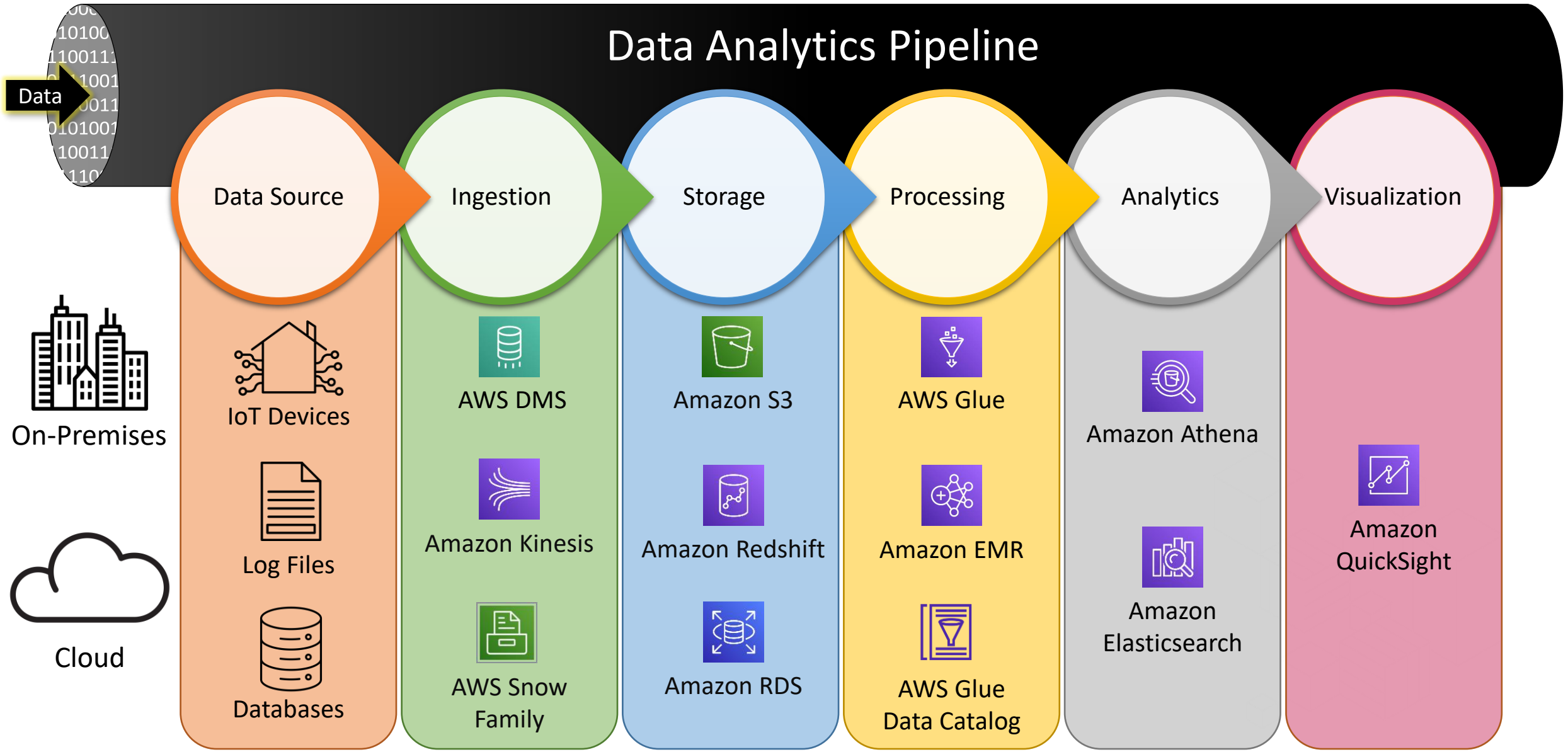
How much?

- With AWS Glue, you pay an hourly rate, billed by the second, for crawlers and ETL jobs. For the AWS Glue Data Catalog, you pay a monthly fee for storing and accessing the metadata. For development endpoint, you pay an hourly rate, billed per second.
- For AWS Glue DataBrew, the interactive sessions are billed per session and the DataBrew jobs are billed per minute.



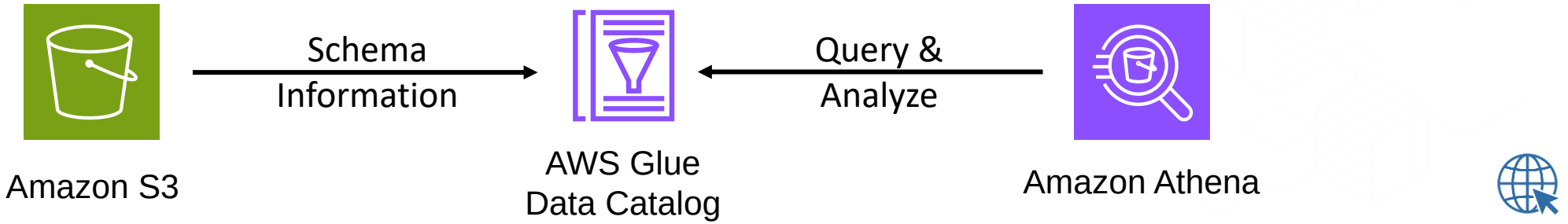
Amazon Athena

Data Analytics on AWS



Amazon Athena

- AWS Athena is an interactive query service to analyze data using standard SQL.
- Behind the scenes, Athena is built on top of the Presto engine and uses Amazon S3 as an underlying data store.
- With Athena, you can run interactive queries using SQL statements on multiple types of data stored on S3 – without being constrained by schema or having to load your data first.
- Popular formats
  - CSV, JSON, Parquet, Avro, ORC



Reference:

FAQs

Amazon Athena

What?

- Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.
- Amazon Athena can process unstructured, semi-structured, and structured data sets.
- Amazon Athena integrates with Amazon QuickSight for easy visualization.

Why?

- Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing data immediately.
- Athena is ideal for quick, ad-hoc querying but it can also handle complex analysis, including large joins, window functions, and arrays. It also supports Federated query to run SQL queries across variety of data sources.

When?

- You want to tap into your data without setting up complex processes to extract, transform, and load the data (ETL).
- You want to generate reports or to explore data with business intelligence tools or SQL clients, connected via an ODBC or JDBC driver.

Where?

- Amazon Athena is a regional service but it can access data in other Regions or other AWS accounts.
- You don't need to load your data into Athena, it works directly with data stored in S3.

Who?

- Amazon Athena allows you to control access to your data by using AWS IAM policies, Access Control Lists (ACLs), and Amazon S3 bucket policies.
- Athena is serverless, so there is no infrastructure to setup or manage.

How?

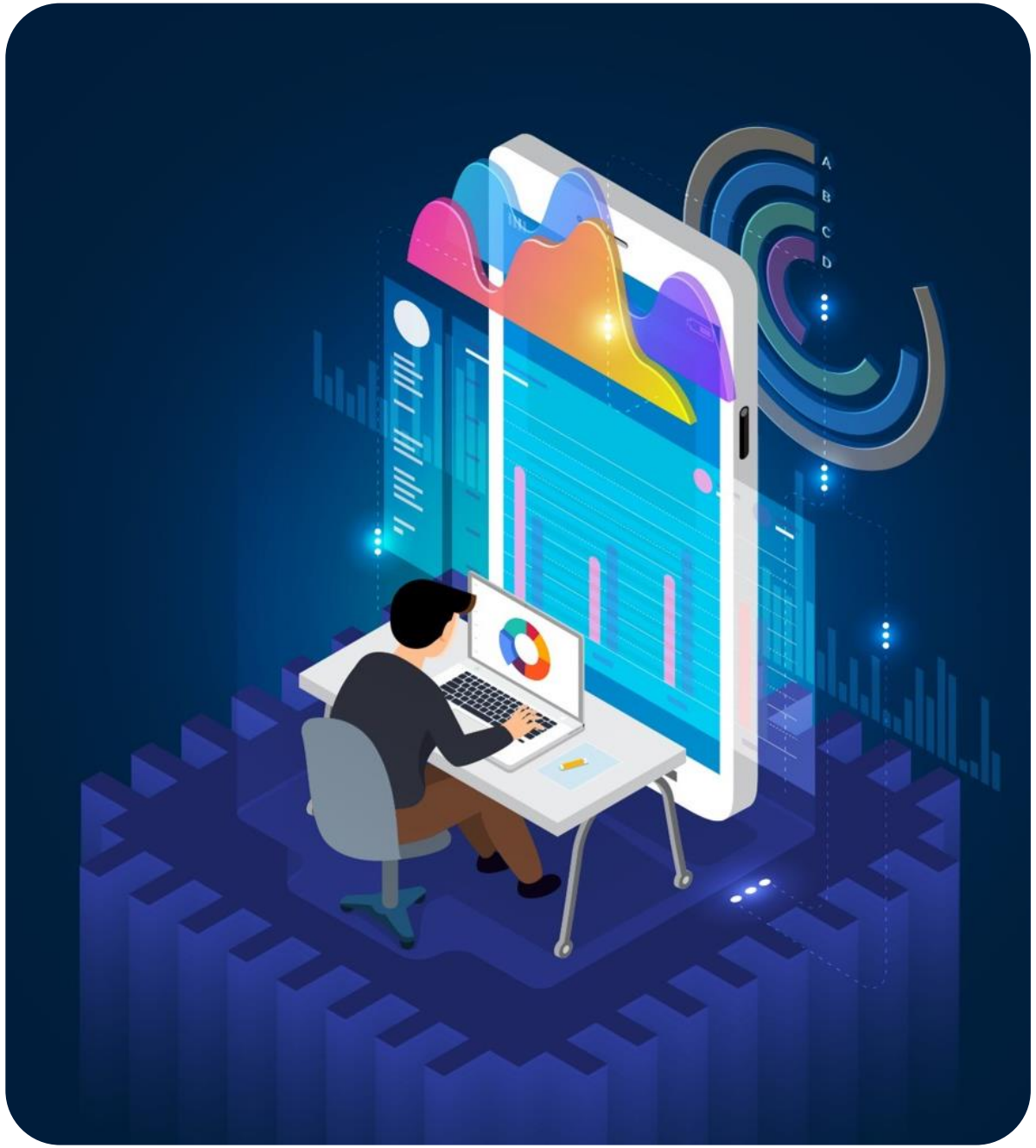
- To get started, just log into the Athena Management Console, define your schema, and start querying.
- Amazon Athena uses Presto with full standard SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet and Avro.

Category:

Analytics

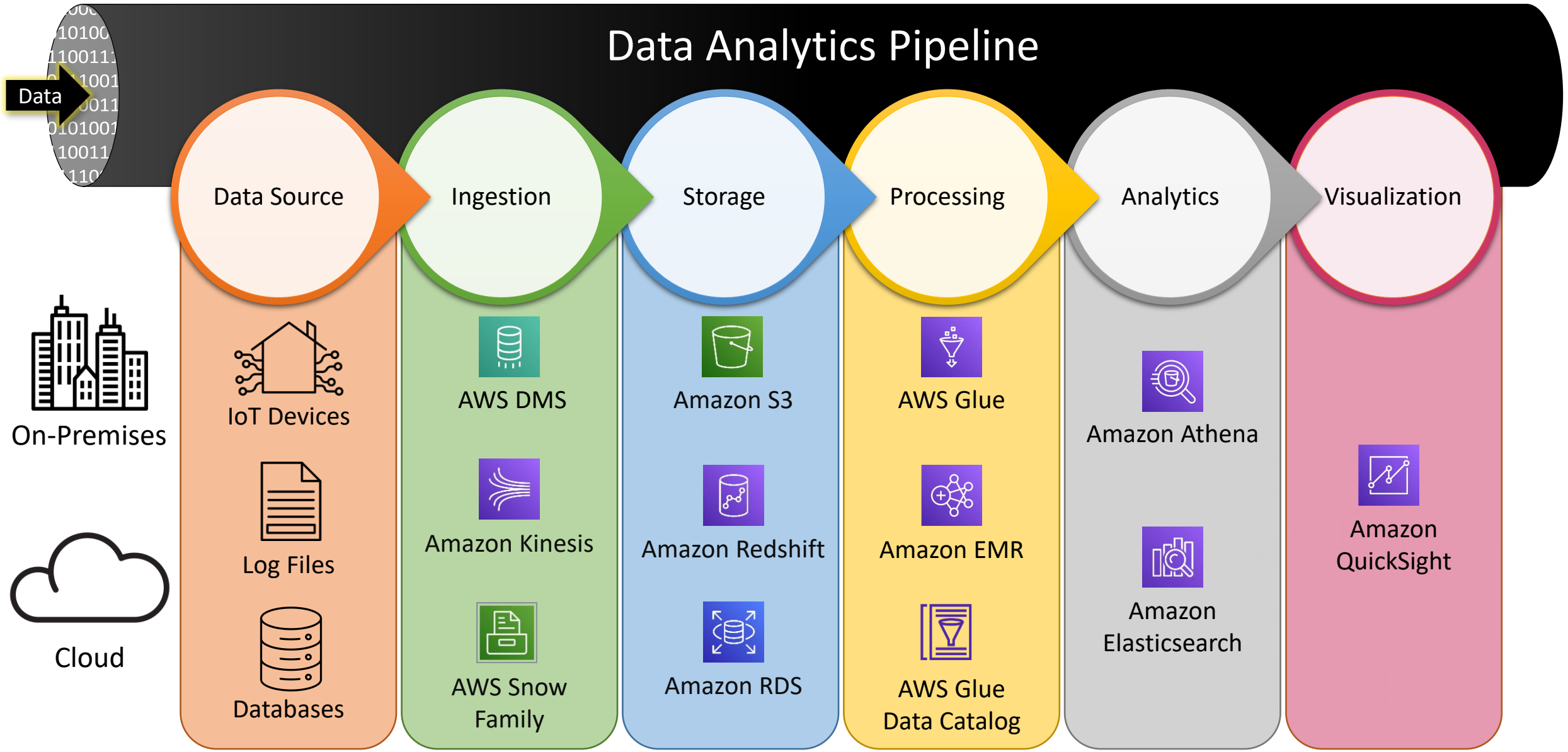
How much?

- Amazon Athena is priced per query and charges based on the amount of data scanned by the query. It queries data directly from Amazon S3, so your source data is billed at S3 rates.
- You are charged separately for using the AWS Glue Data Catalog.



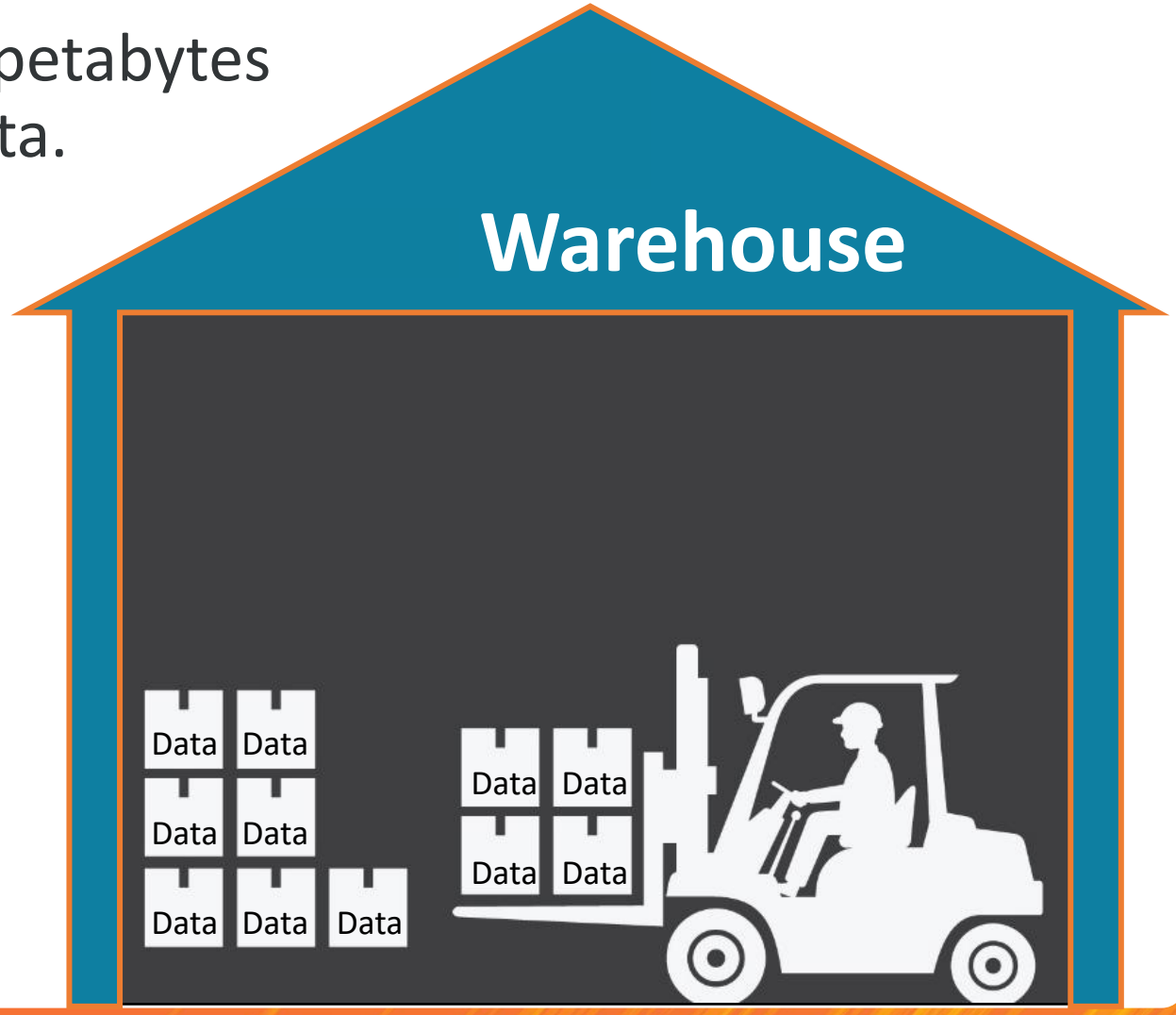
Amazon Redshift

Data Analytics on AWS



Amazon Redshift

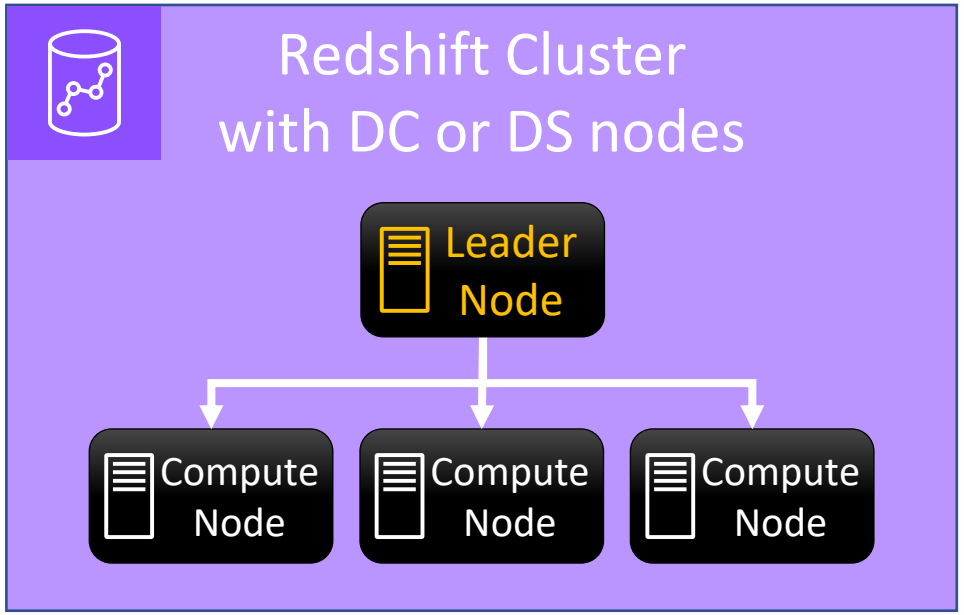
- Amazon Redshift is a purpose-built, fully managed data warehouse.
- It runs high-performance queries on petabytes of structured and semi-structured data.
- Allows you to analyze all the data using standard SQL and existing business intelligence (BI) tools.
- Fine tuned for analytics using features like columnar data storage, data compression, and massive parallel processing.



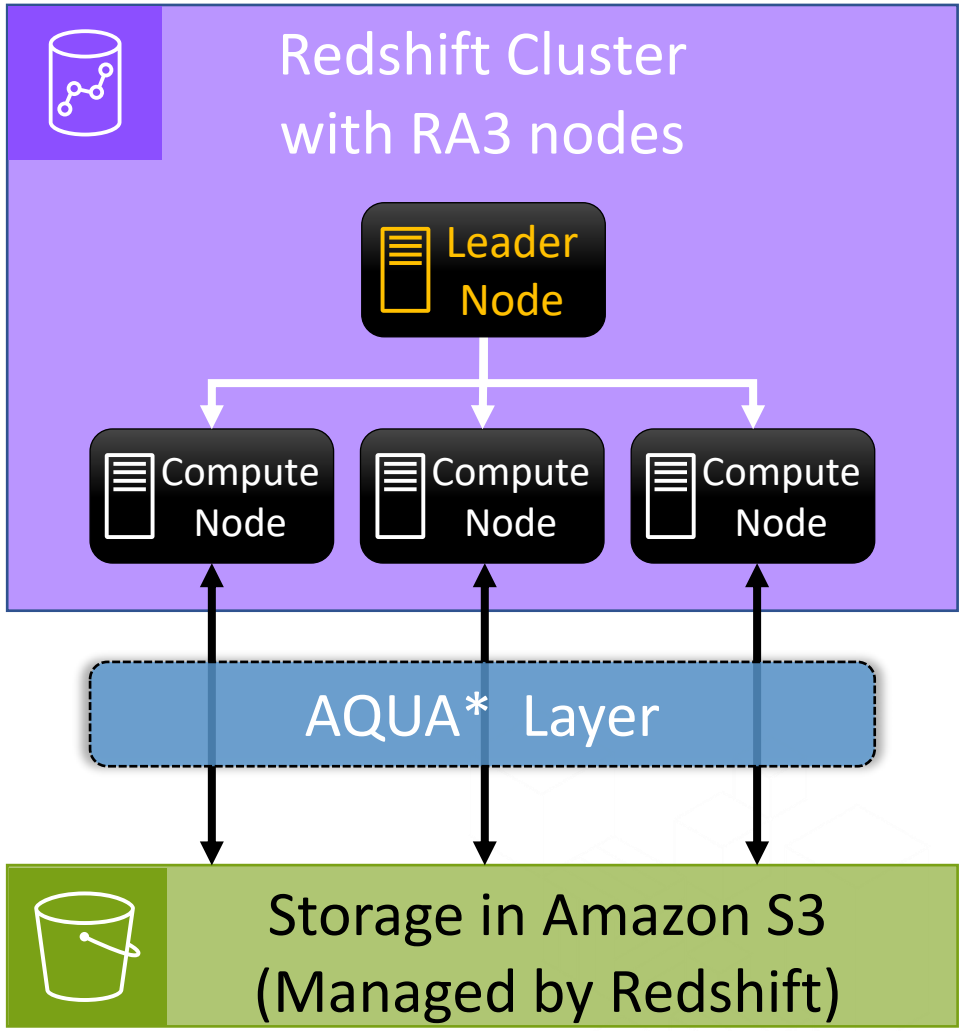
Storage in Kitchen vs. Warehouse

A typical warehouse

Amazon Redshift Architecture

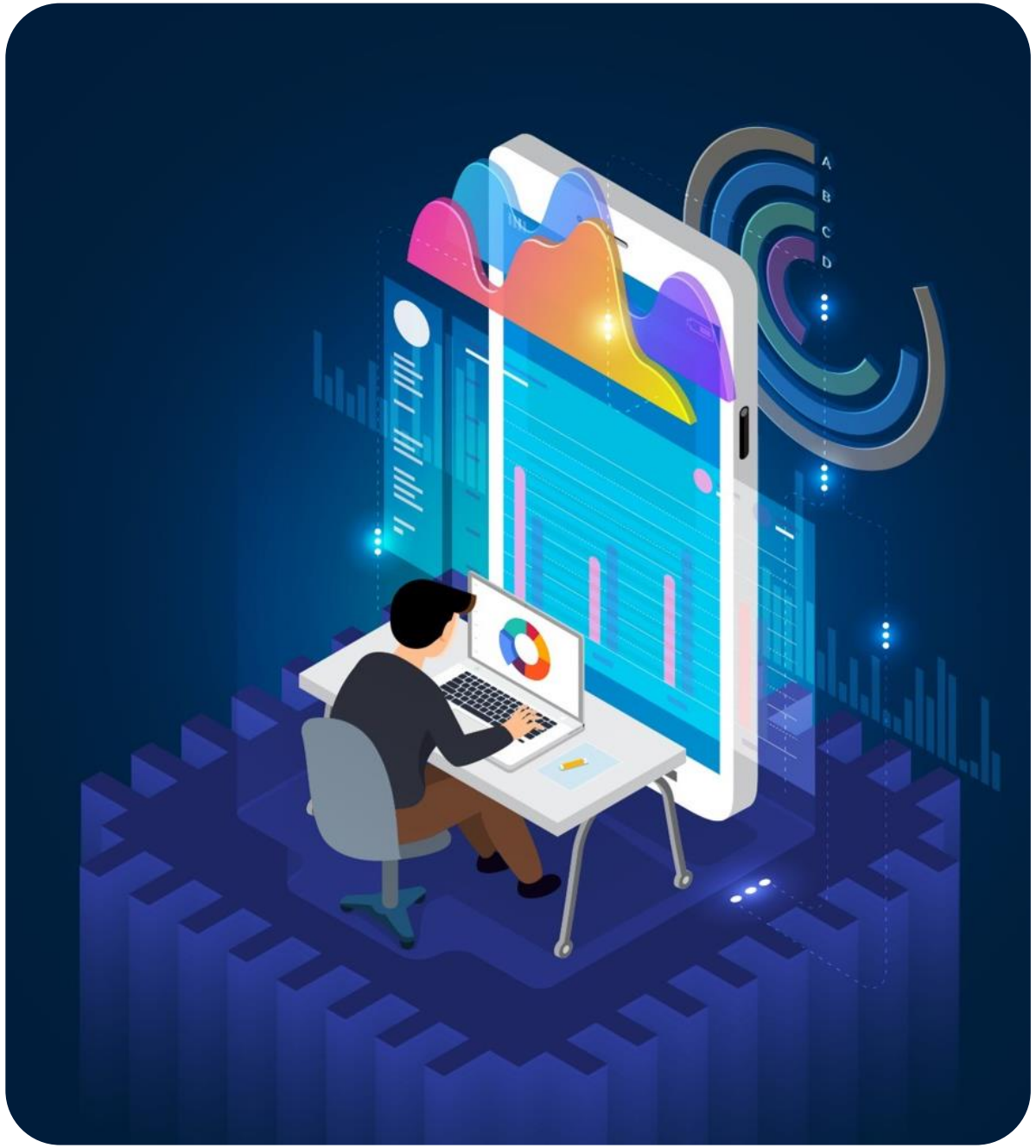


Scale compute and storage together



Scale compute and storage independently

*AQUA (Advanced Query Accelerator)
A Speed Boost for Your Amazon Redshift Queries



AWS Lake Formation

Exabyte, the new megabyte

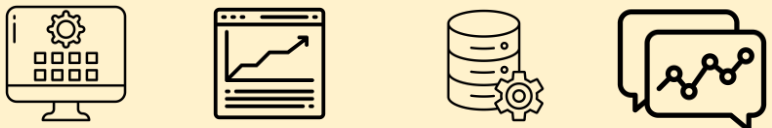
11,500 stores in 27 countries

700,000 associates server 100 million customers weekly

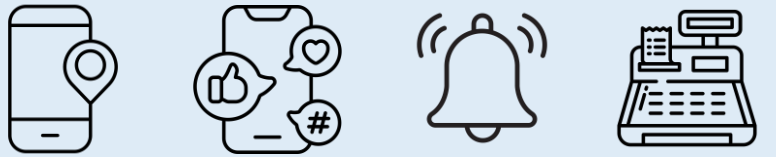
2.5 petabytes of data from 1 million customers every hour

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



Semistructured 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

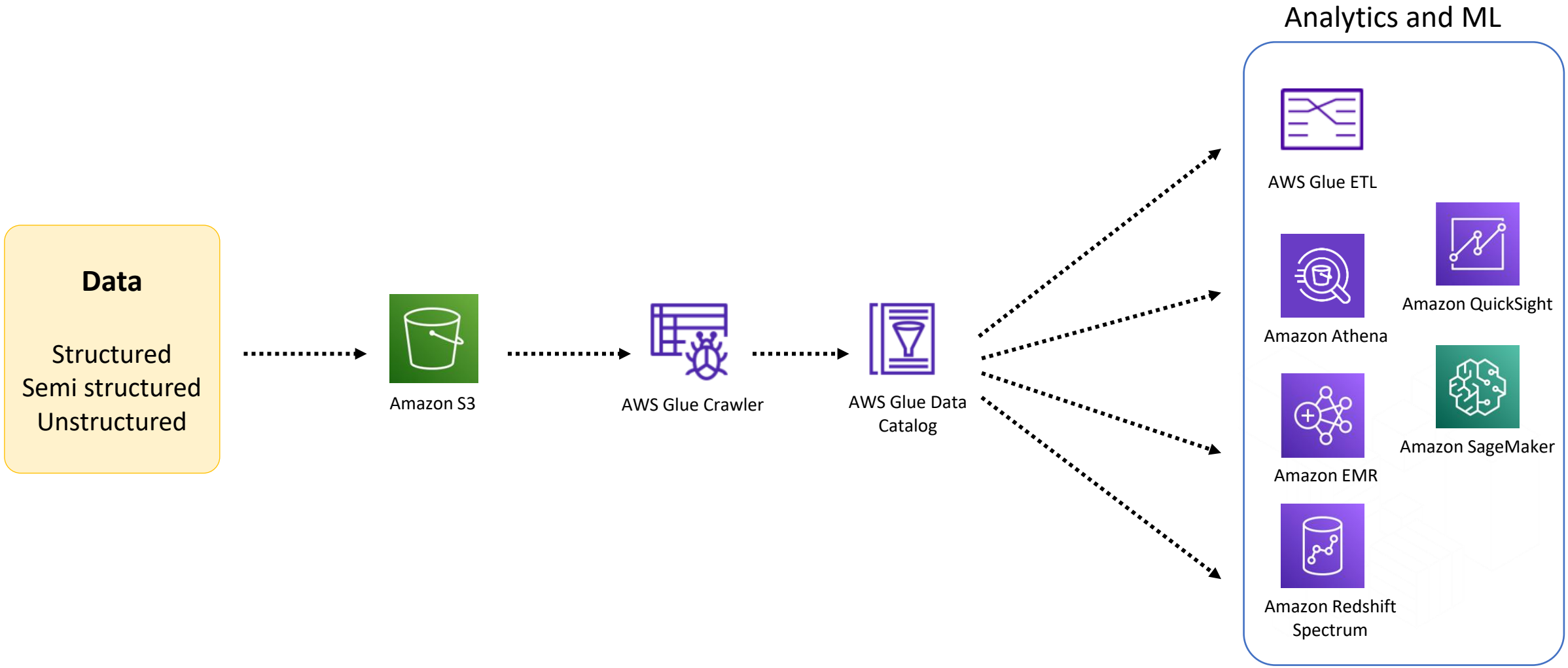
Data Warehouse vs. Data Lake

	Data Warehouse 	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

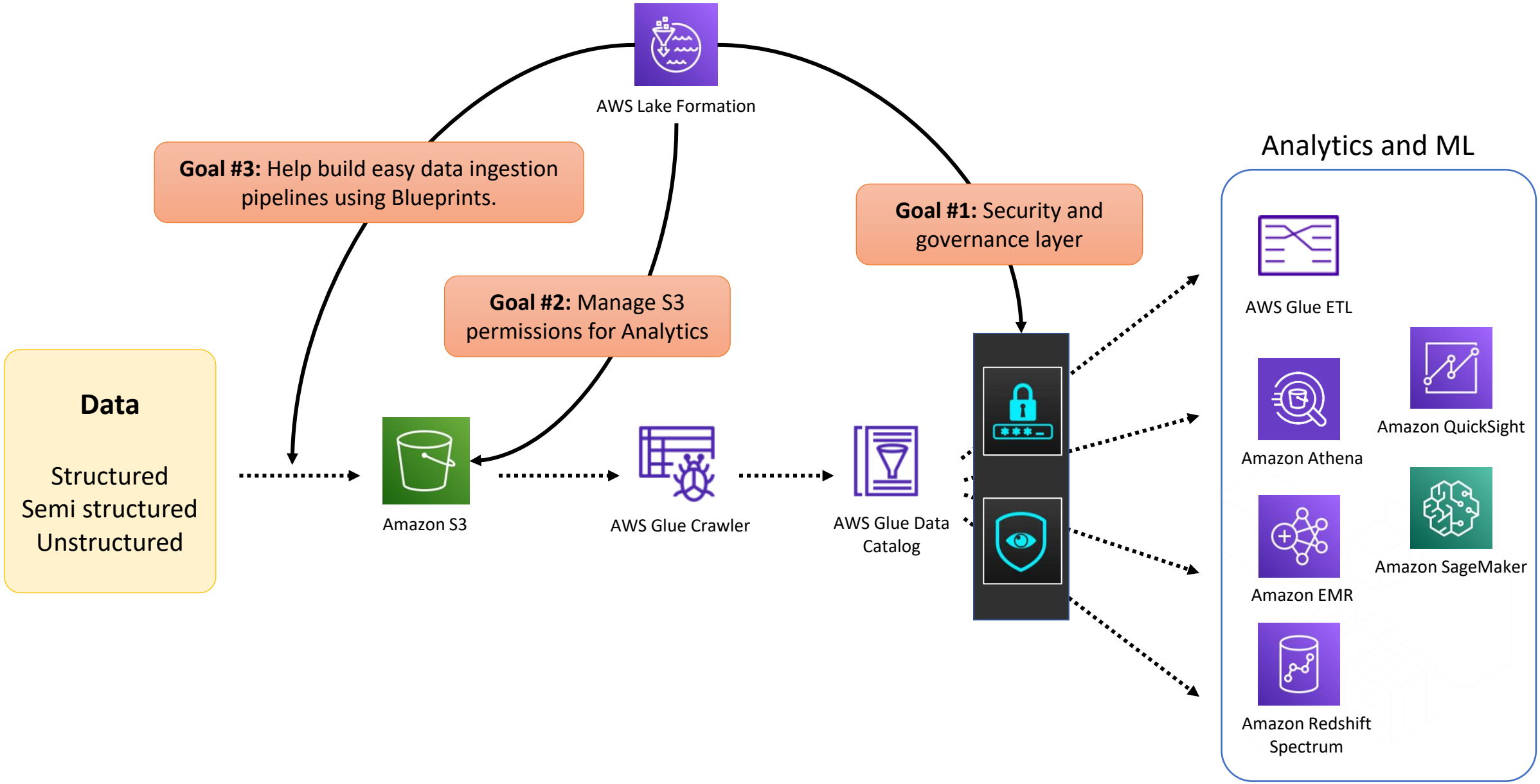
Water Lake vs. Data Lake

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

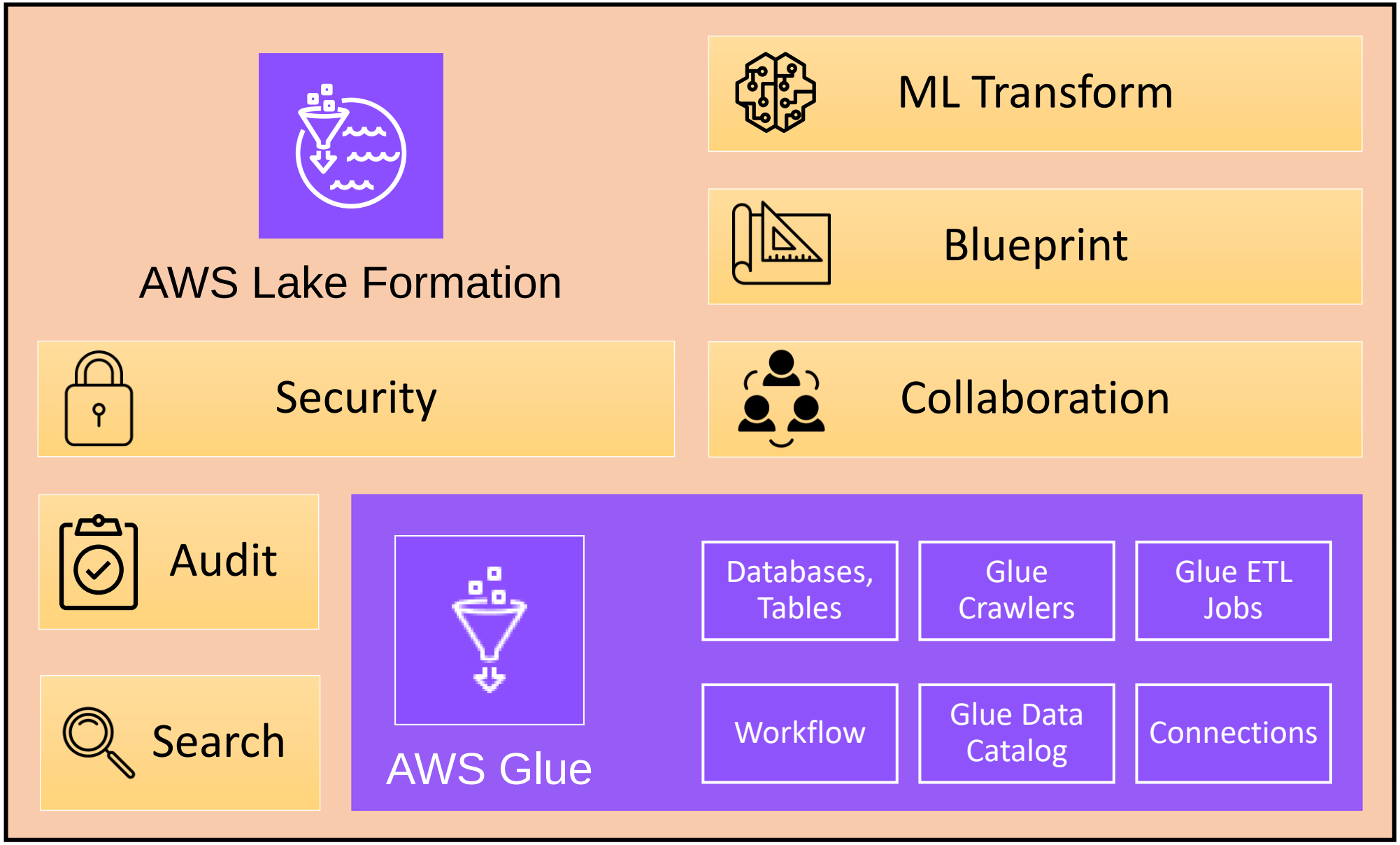
Data Lake on AWS without AWS Lake Formation



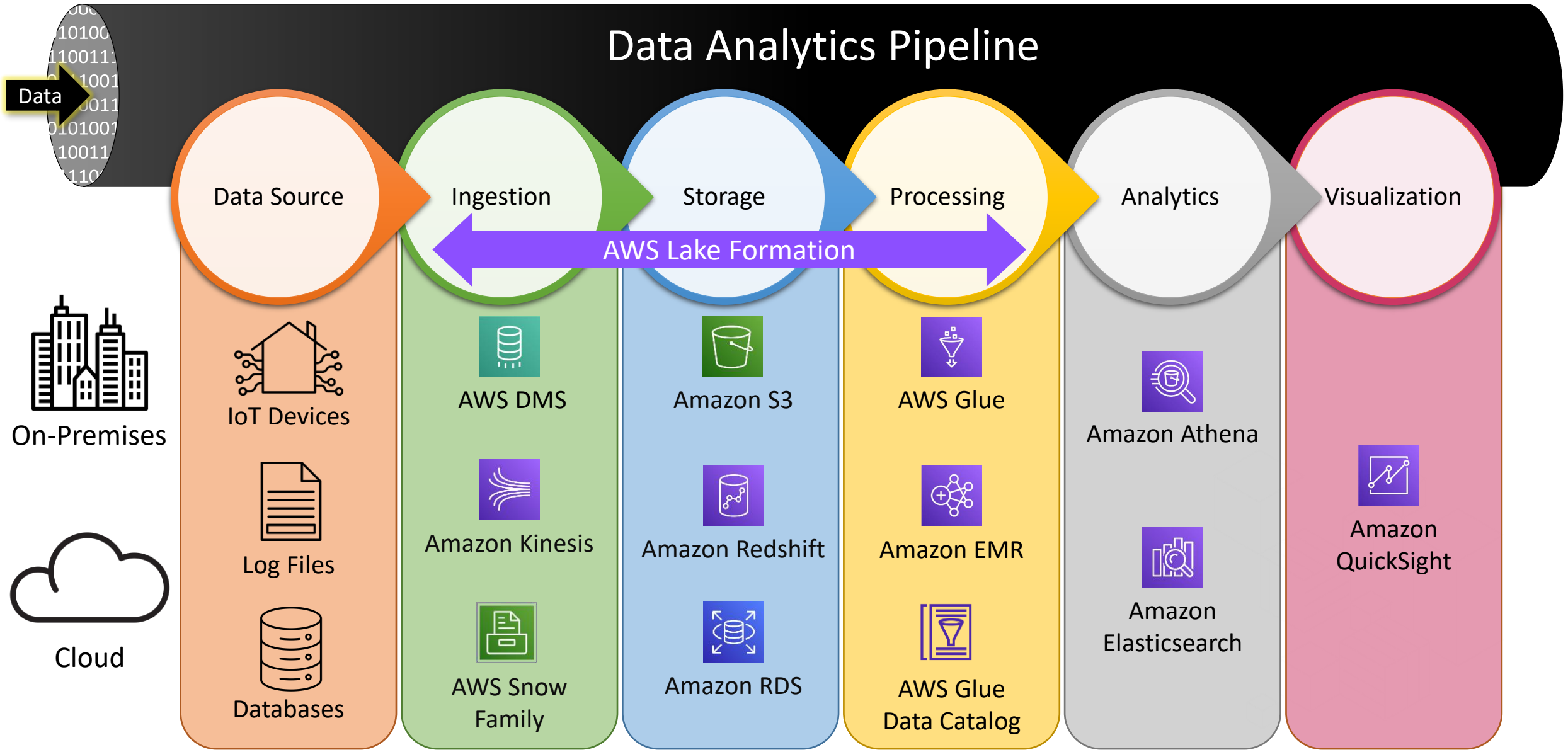
Data Lake on AWS with AWS Lake Formation



AWS Lake Formation and AWS Glue



Data Analytics on AWS



Reference:

FAQs

AWS Lake Formation

What?

- AWS Lake Formation is a fully managed service that makes it easy to build, secure, and manage data lakes.
- A data lake is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. Lake Formation provides its own permissions model that augments the IAM permissions model.

Why?

- It simplifies and automates many of the complex manual steps (collecting, cleansing, moving, and cataloging data, and securely making that data available for analytics and machine learning) that are usually required to create data lakes.
- It has built-in Machine Learning (ML) Transform to deduplicate and find matching records to increase data quality.

When?

- You want to build data lakes quickly, provide self-service access to data and simplify security management.
- You need a single place to define and enforce access controls through a simple grant or revoke mechanism that operate at the table, column, row, and cell-level for all the users and services that access your data.

Where?

- AWS Lake Formation is a Regional services.
- Lake Formation uses the AWS Glue Data Catalog to store metadata about data lakes, data sources, transforms, and targets. Each AWS account has one Data Catalog (your persistent metadata store) per AWS Region.

Who?

- A data lake administrator can grant any principal any permission on any Data Catalog resource or data location.
- The following AWS services integrate with AWS Lake Formation and honor Lake Formation permissions - AWS Glue, Amazon Athena, Amazon Redshift Spectrum, Amazon QuickSight Enterprise Edition, Amazon EMR.

How?

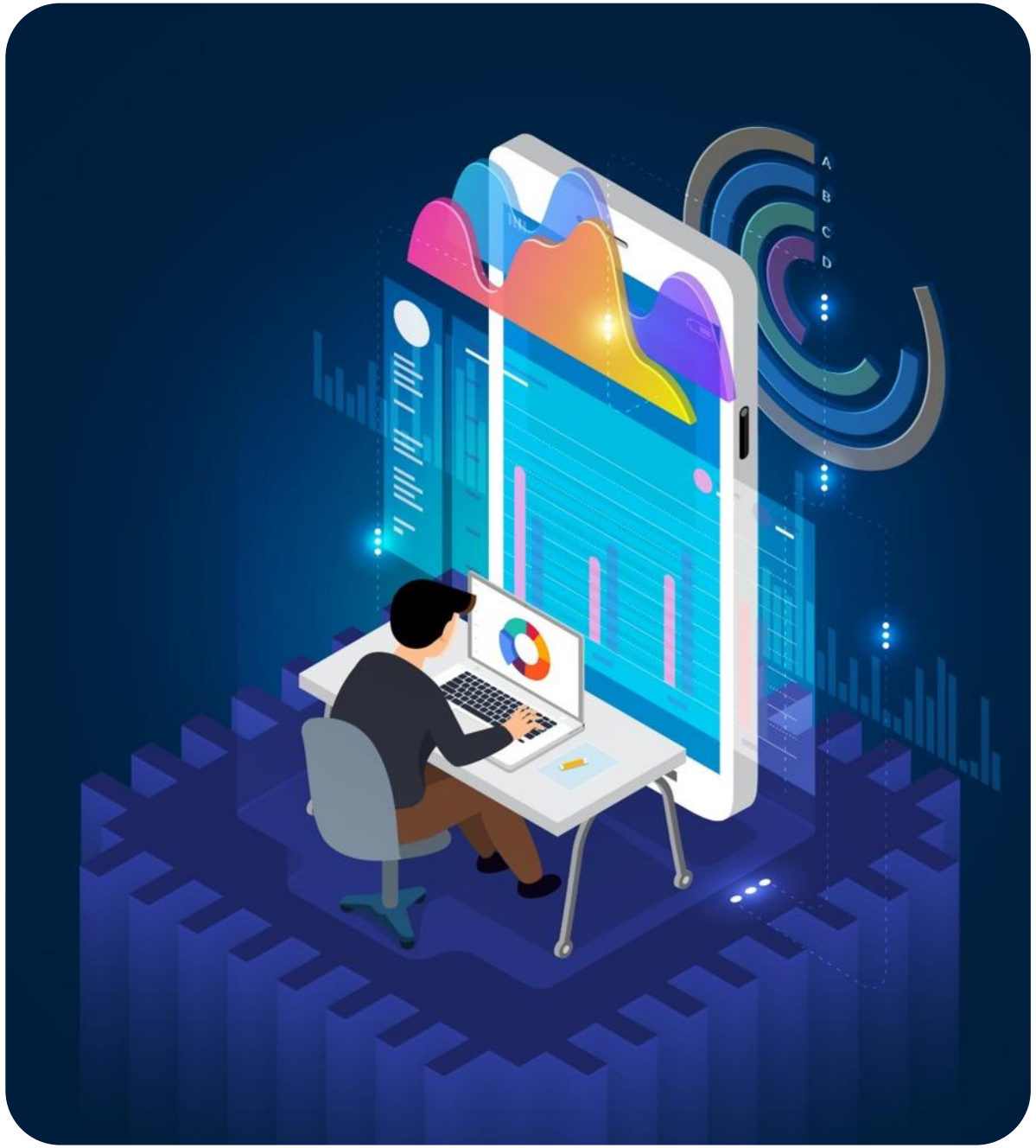
- First, identify existing data stores in S3 or relational and NoSQL databases, and move the data into your data lake. Then crawl, catalog, and prepare the data for analytics. Next, provide your users with secure self-service access to the data through their choice of analytics services. It uses AWS Glue to orchestrate jobs and crawlers.

Category:

Analytics

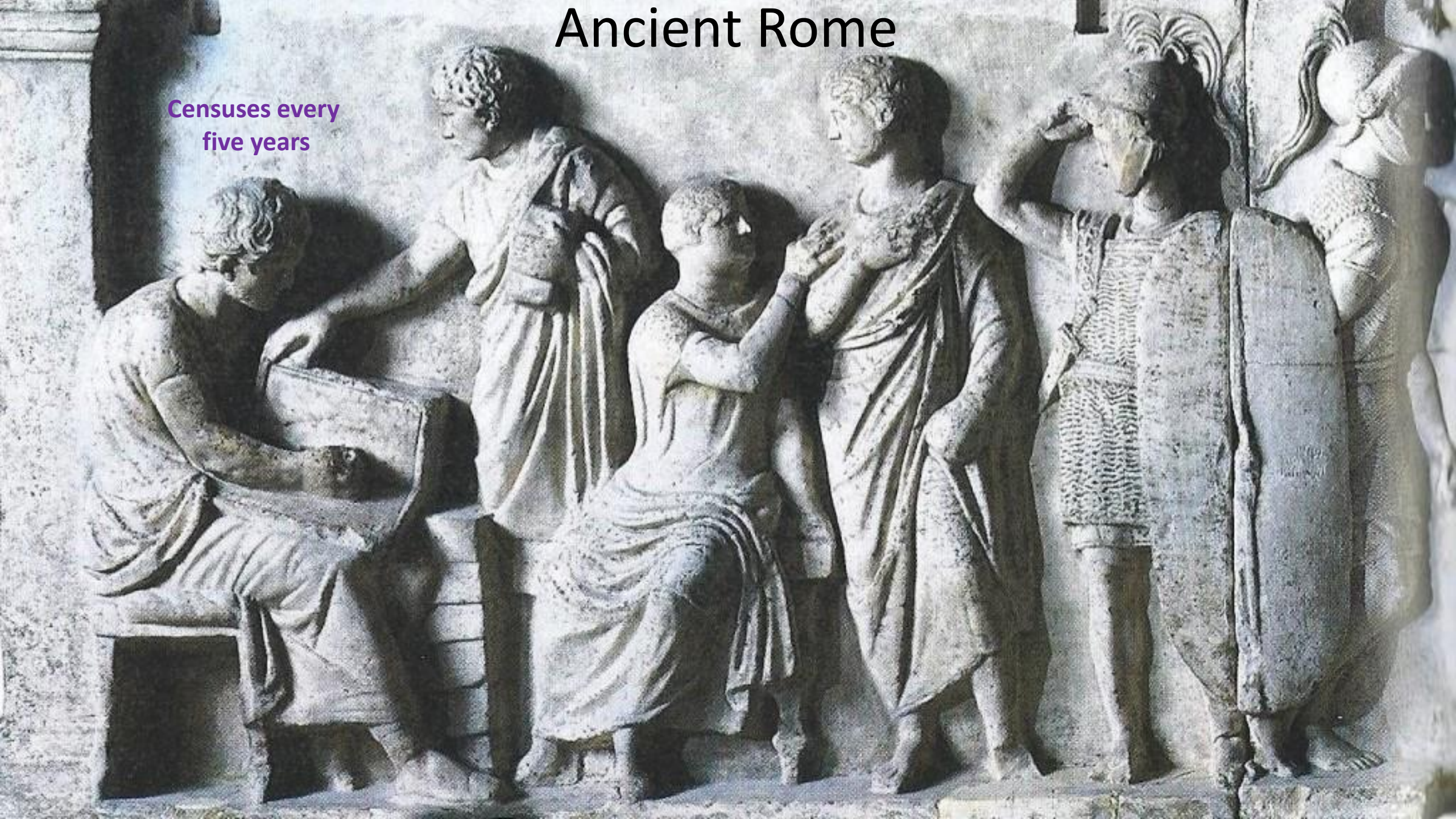
How much?

- AWS Lake Formation provides database, table, column and tag-based access controls, and cross-account sharing at no charge.
- Lake Formation charges a fee for data filtering (per TB of data scanned), data processed by the storage optimizer, metadata storage and the API requests.



Amazon EMR

Ancient Rome



Censuses every five years


Rome
1
Ephesus
2
Carthage
4
Antioch
3
Ancient Rome
Alexandria
7
Ephesus
6


Census Bureau
Rome
1
Carthage
4
Ephesus
2
Antioch
3
Ancient Rome
Alexandria
7
Ephesus
6

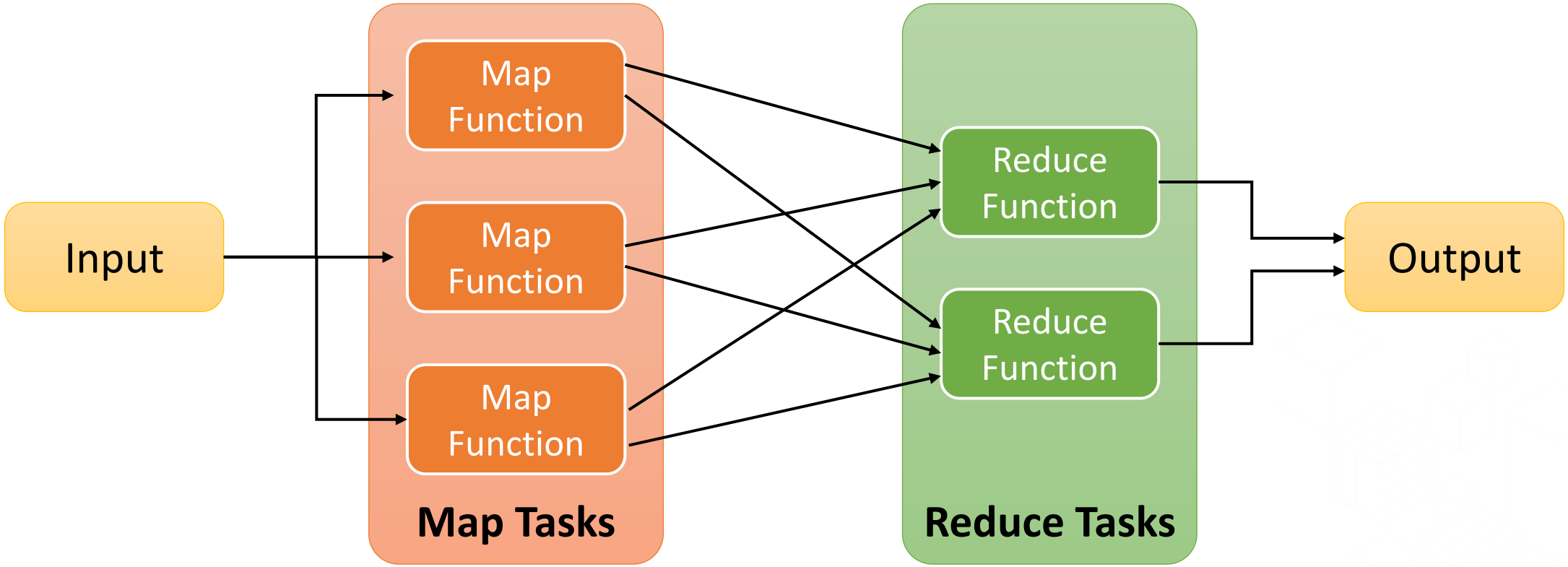

Census Bureau
Rome
1
Carthage
4
Ephesus
2
Antioch
3
Alexandria
7
Ephesus
6
Ancient Rome


Census Bureau
Rome
1
Carthage
4
Ephesus
2
Antioch
3
Alexandria
7
Ephesus
6
Ancient Rome

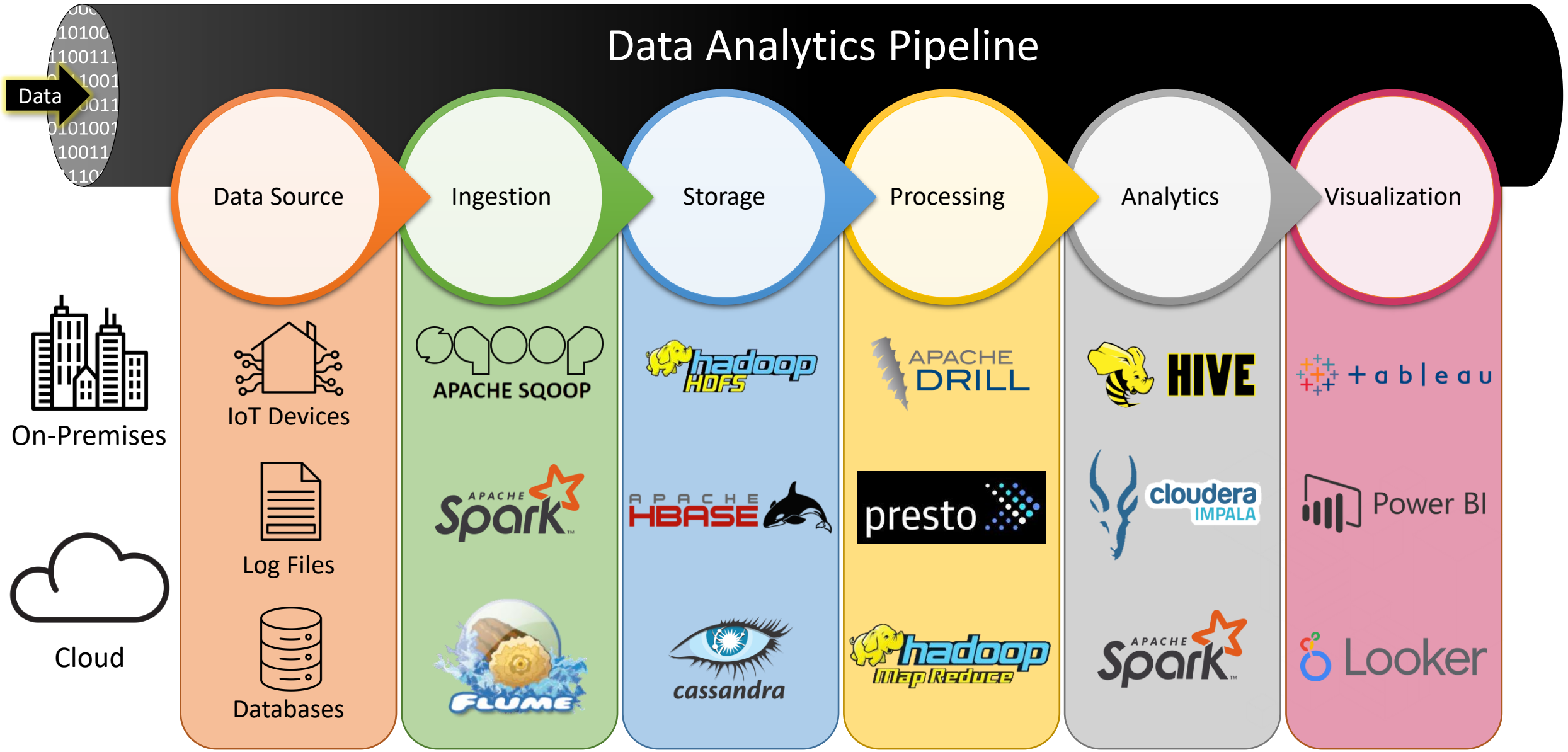

Census Bureau
Rome
1
Ephesus
2
Carthage
4
Antioch
3
Ancient Rome
Alexandria
7
Ephesus
6

What is MapReduce?

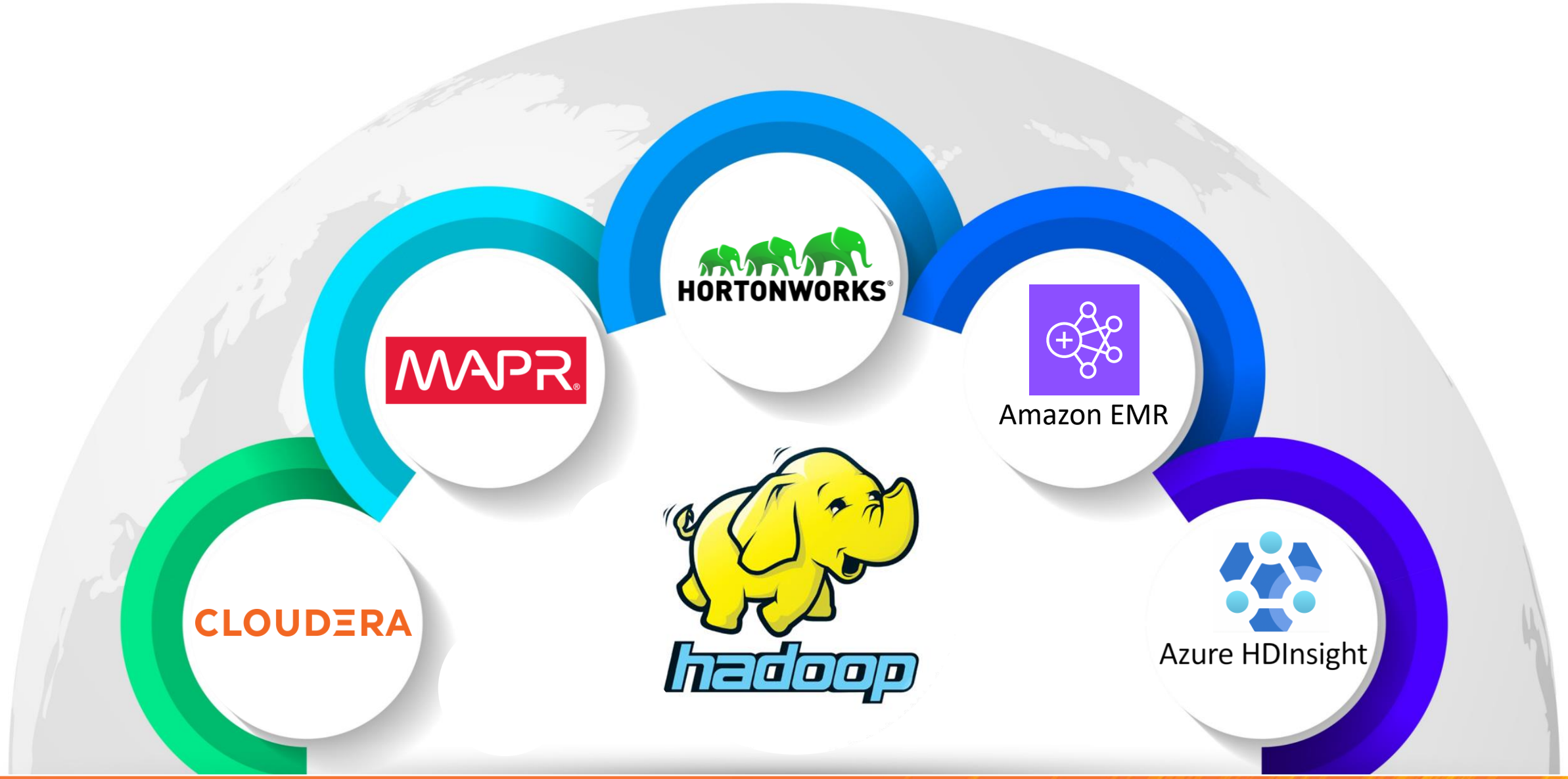
- MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment.



Big Data Eco System

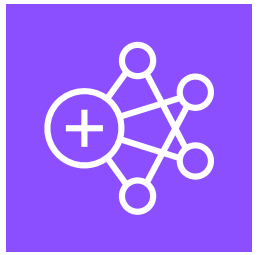


Various Hadoop Distribution / Vendors



Amazon EMR

- A managed big data platform in AWS Cloud
- Multiple ways to run your big data based workload
  - Amazon EMR running on Amazon EC2
  - Amazon EMR Serverless
  - Amazon EMR on EKS
  - Amazon EMR Studio (Manage Jupyter notebooks)
- Run Hadoop, Spark, Presto and other applications
- Resize a running cluster
- Use HDFS and S3 file systems



Amazon EMR

Reference:

FAQs

Amazon EMR

What?

- Amazon EMR is cloud big data platform for data processing, interactive analysis, and machine learning using open source frameworks such as Apache Spark, Apache Hive, and Presto.

Why?

- Amazon EMR simplifies building and operating big data environments and applications.
- Amazon EMR enables you to quickly and easily provision as much capacity as you need, and automatically or manually add and remove capacity. This is very useful if you have variable or unpredictable processing requirements.

When?

- When you want to focus on transforming and analyzing your data without having to worry about infrastructure provisioning, cluster setup, configuration, open-source applications or tuning.

Where?

- Amazon EMR launches all nodes (Master, Core, Optional Task nodes) for a given cluster in the same Availability Zone of a Region.

Who?

- After an EMR cluster is launched, customer can monitor and manage it. Amazon EMR provides several tools you can use to connect to and control your cluster.

How?

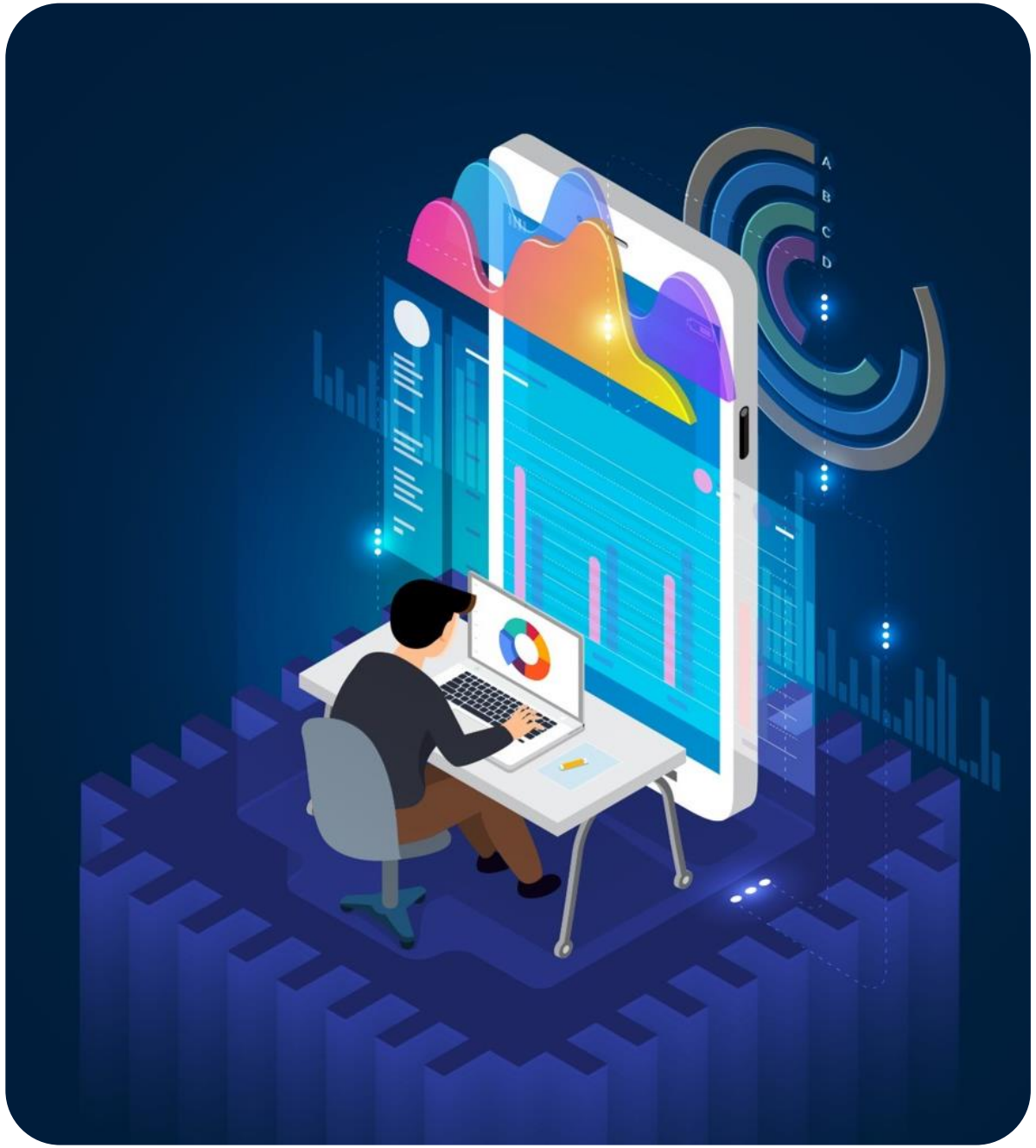
- You can launch a cluster by specifying the name of your cluster, the location in Amazon S3 of your input data, your processing application, your desired data output location, and the number and type of Amazon EC2 instances you'd like to use.

Category:

Analytics

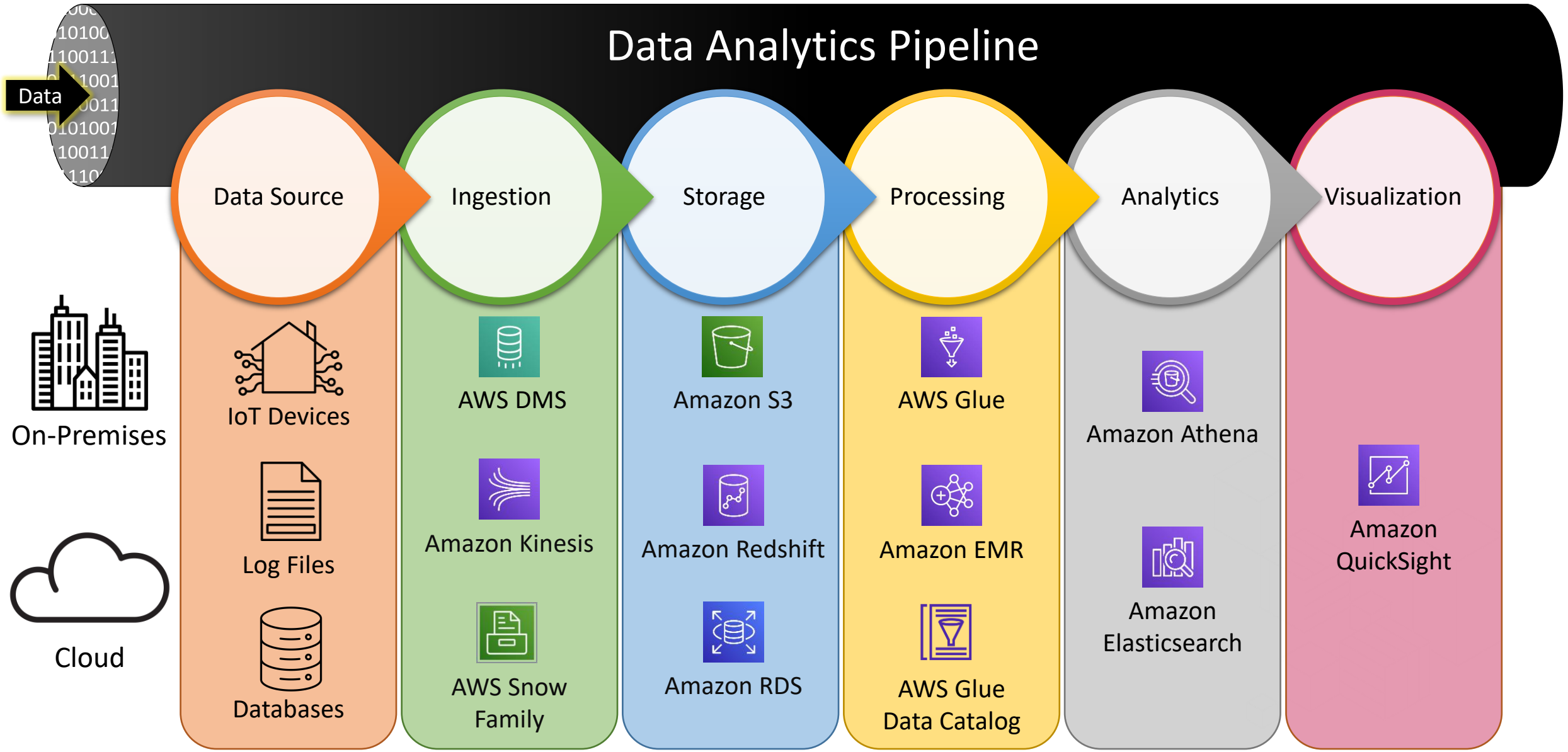
How much?

- You pay a per-second rate for every second you use, with a one-minute minimum. Customers pay for Amazon EMR plus backend compute price (Amazon EC2, Amazon EKS, AWS Outposts, Amazon EMR Serverless).



Amazon Kinesis

Data Analytics on AWS



What is streaming data?

- Streaming data is data that is emitted at high volume in a continuous, incremental manner with the goal of low-latency processing.
- It may be generated continuously through thousands of data sources typically in smaller sizes (order of Kilobytes).
- Examples:
  - Log files generated by customers using your mobile or web applications
  - E-commerce purchases
  - In-game player activity
  - Information from social networks
  - Stock market
  - Geospatial services
  - Telemetry from connected devices or instrumentation in data centers

Batch Processing vs Streaming

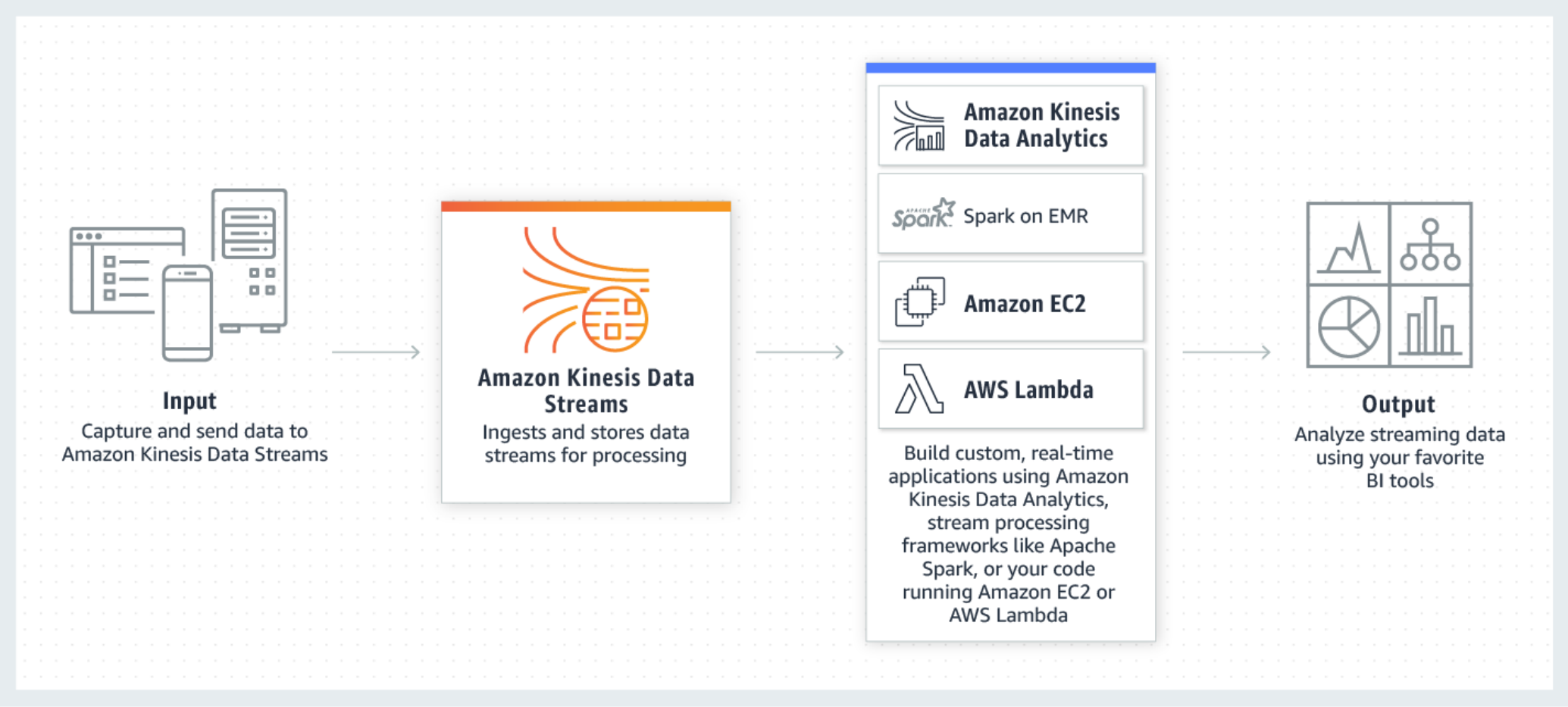
	Batch Processing 	Streaming 
Data scope	Queries or processing over all or most of the data in the dataset.	Queries or processing over data within a rolling time window, or on just the most recent data record.
Data size	Large batches of data.	Individual records or micro batches consisting of a few records.
Performance	Latencies in minutes to hours.	Requires latency in the order of seconds or milliseconds.
Analysis	Complex analytics.	Simple response functions, aggregates, and rolling metrics.

Amazon Kinesis

- Kinesis is a platform for streaming data on AWS, to load and analyze streaming data.

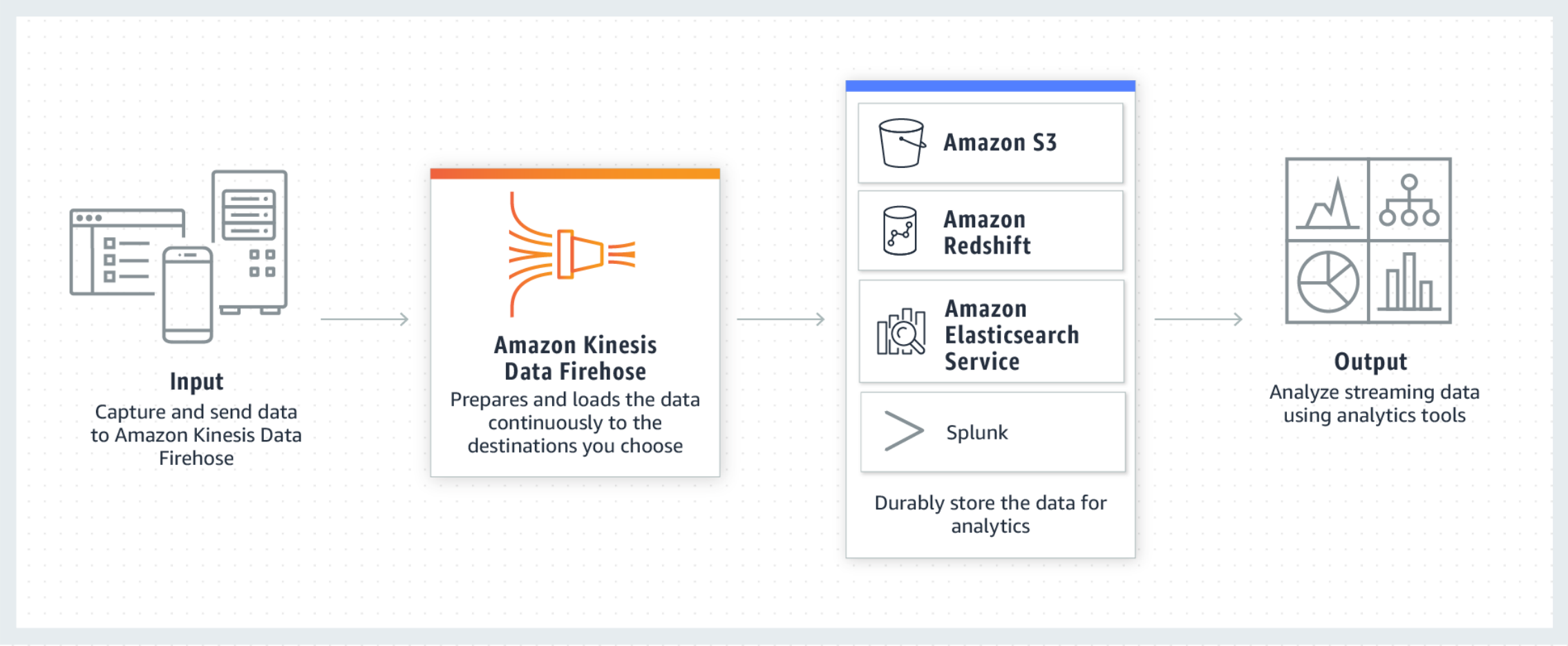
Amazon Kinesis Data Streams

- Collect and stream data for ordered, real-time processing.
- Sub-1 second processing latency.



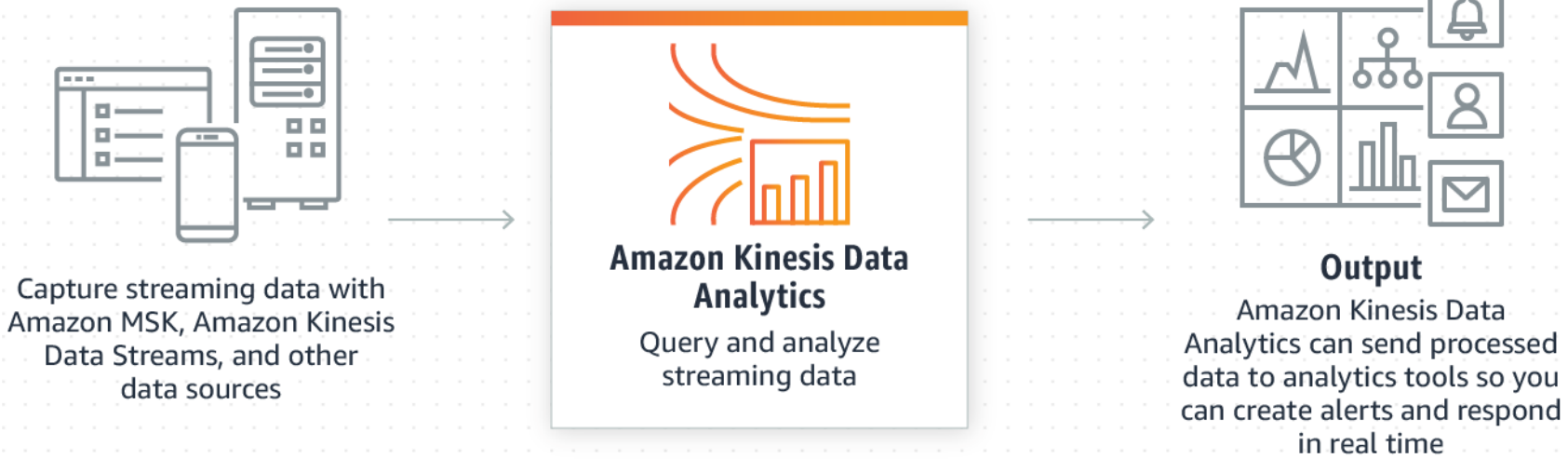
Amazon Kinesis Firehose

- Firehouse is about data movement from Point A to Point B and transformation. Useful when you do data processing at the destination.
- Processing latency of 60 seconds or higher.



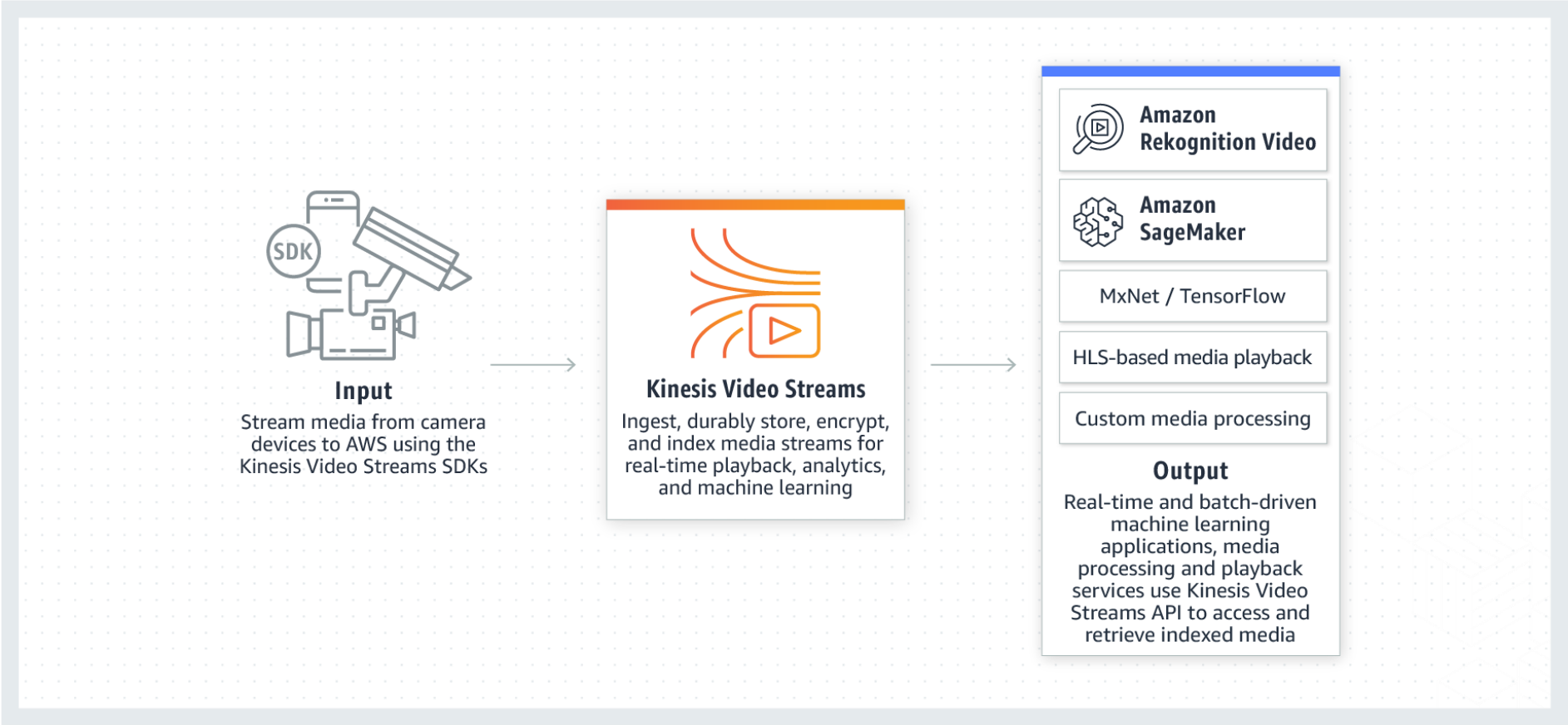
Amazon Kinesis Data Analytics

- Interact with streaming data in real-time by using SQL or integrated Java applications
- Build fully managed and elastic stream processing applications

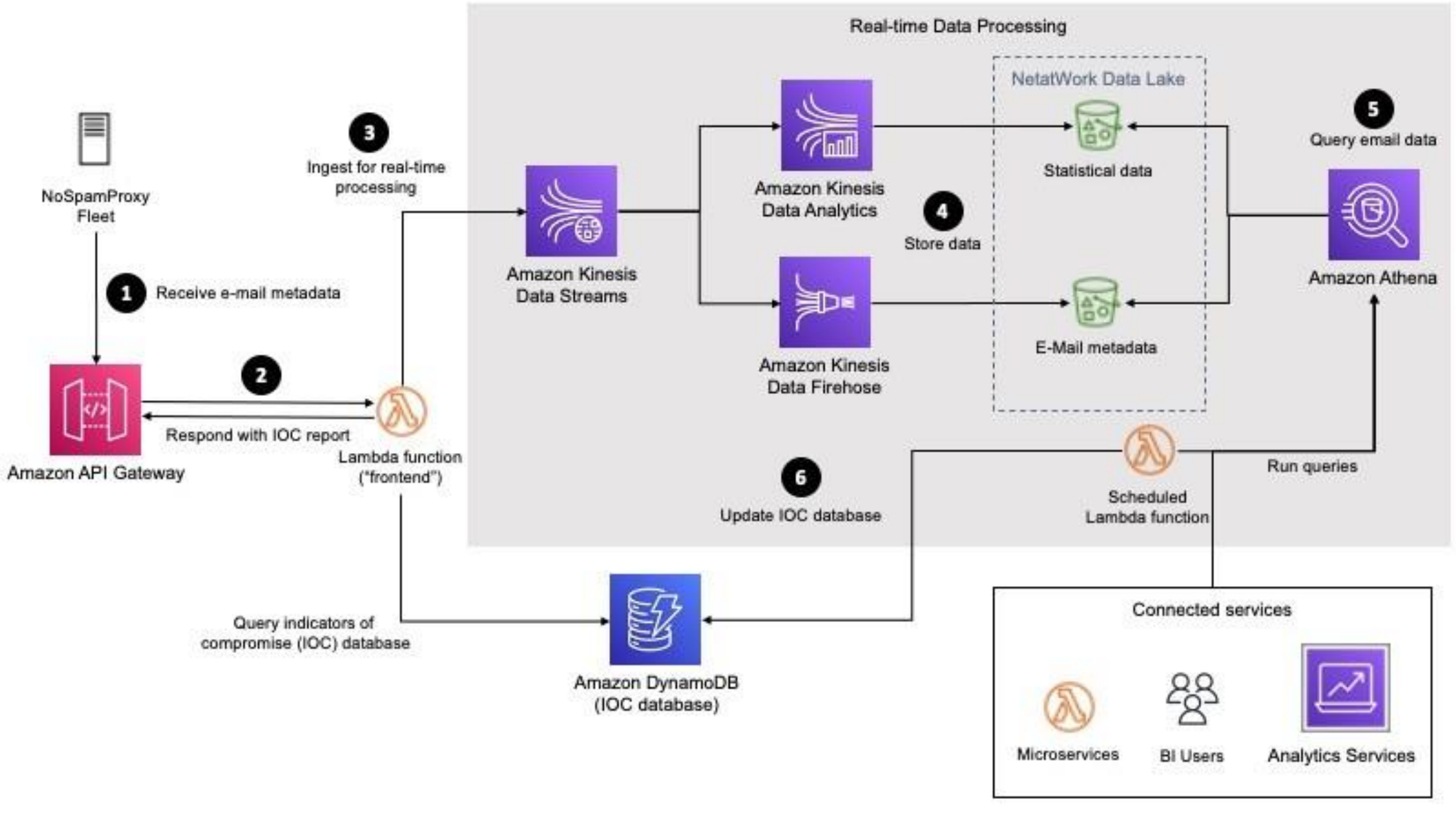


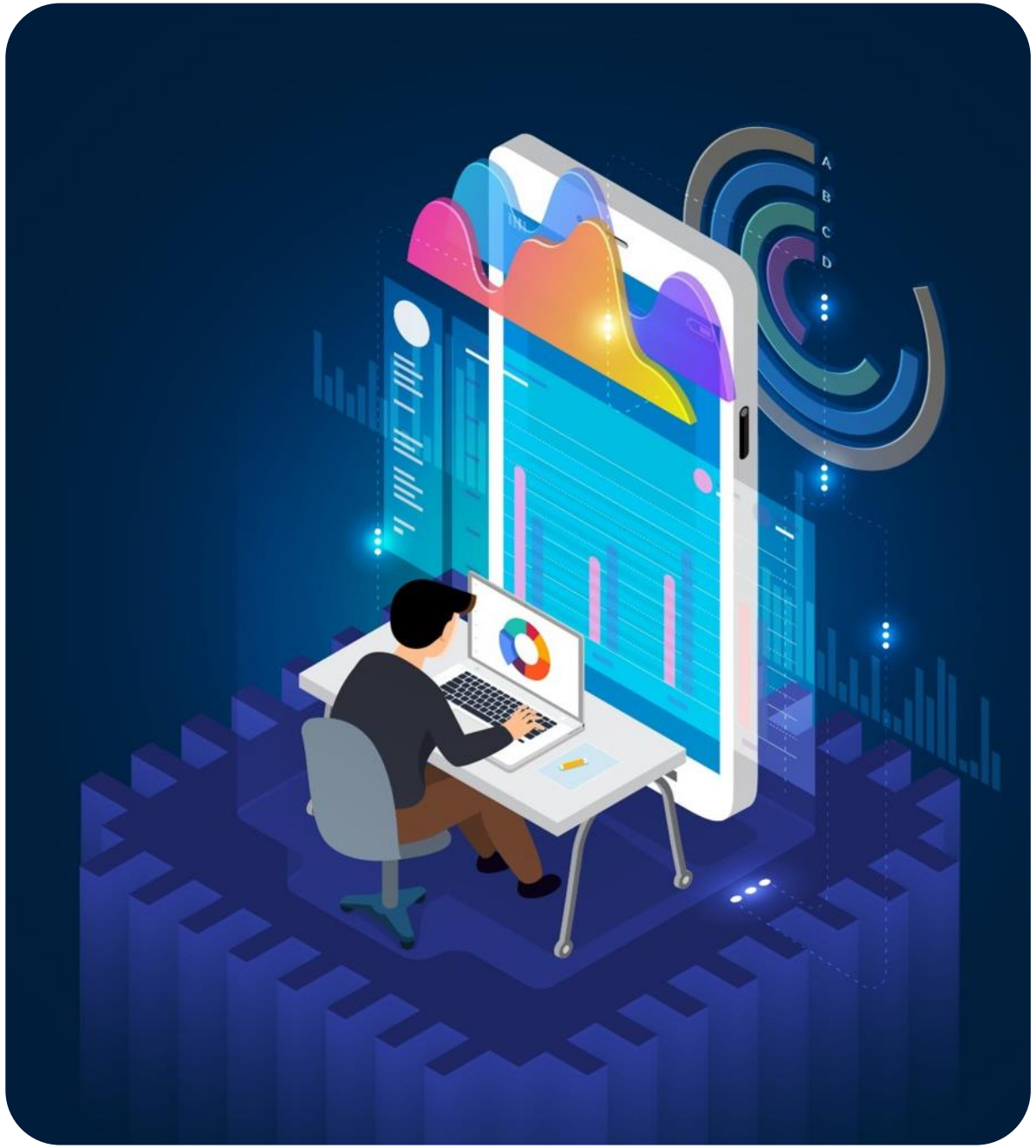
Amazon Kinesis Video Streams

- Capture, process, and store media streams for playback, analytics, and machine learning



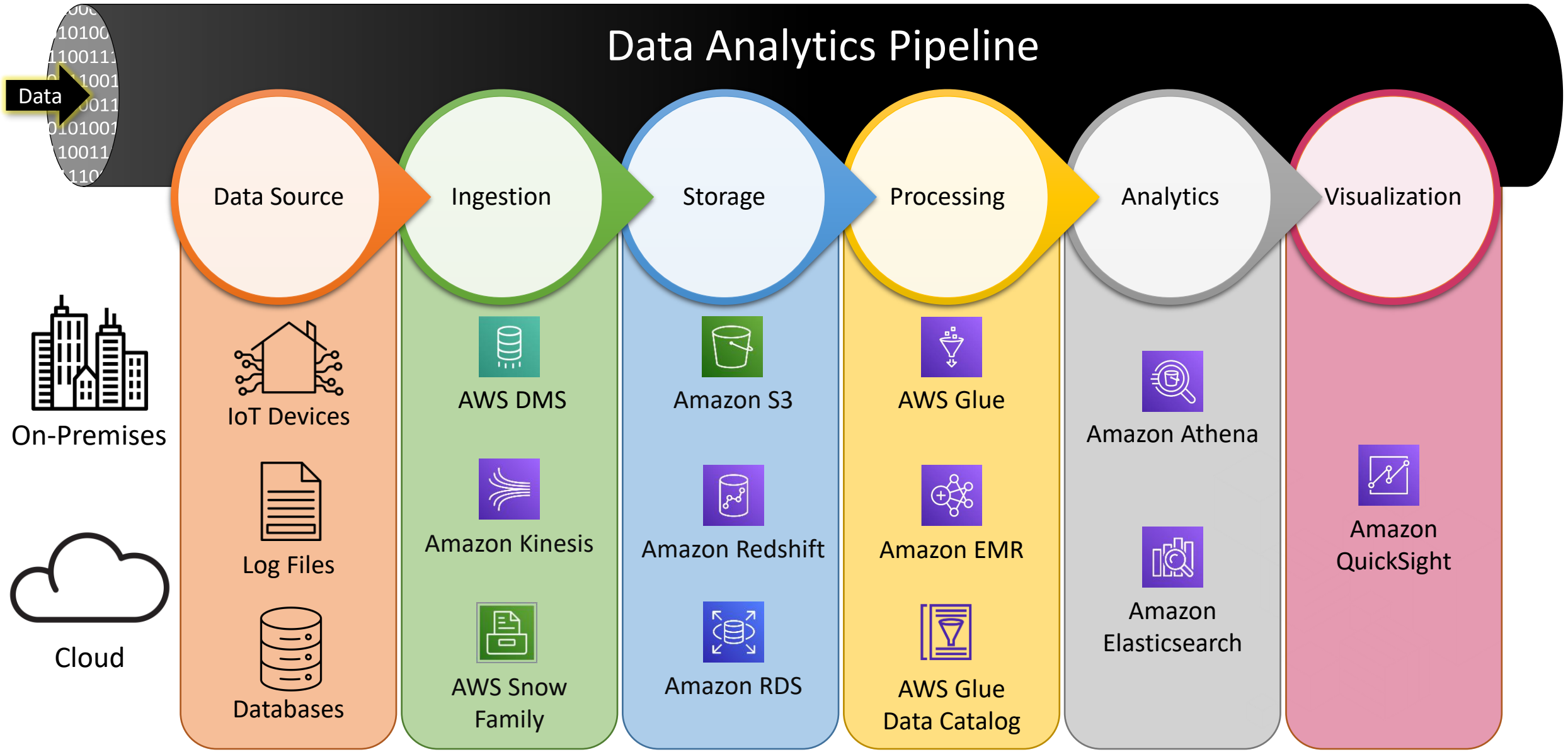
Reference Architecture





Amazon QuickSight

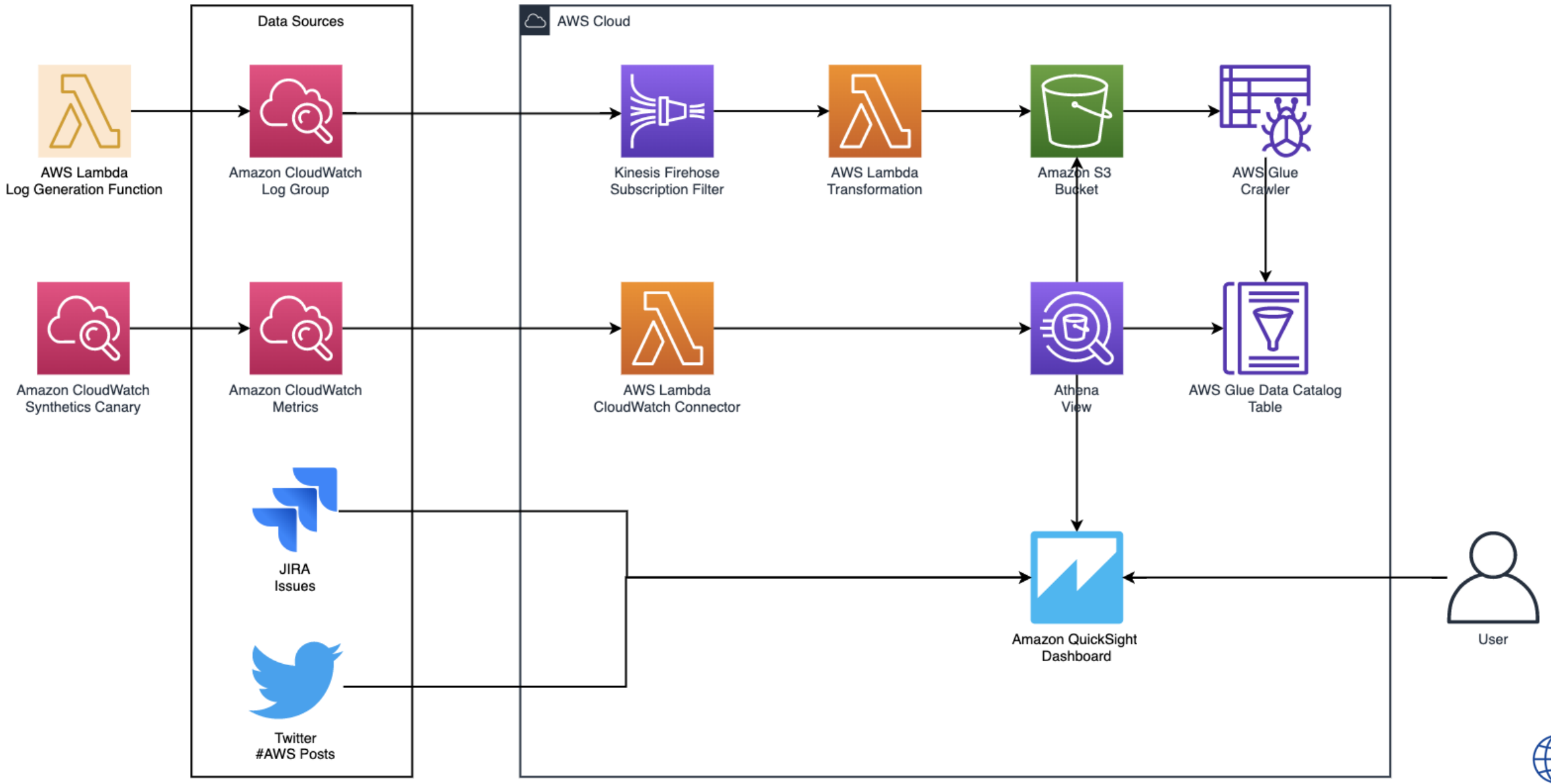
Data Analytics on AWS



Amazon QuickSight

- Amazon QuickSight is a cloud-powered business analytics service.
- Allows you to build visualizations, perform ad-hoc analysis, and quickly get business insights.
- Can visualize data by connecting to SaaS applications like Salesforce; access on-premises databases like SQL Server, MySQL, and PostgreSQL and AWS data sources such as Amazon Redshift, Amazon RDS, Amazon Aurora, Amazon Athena, and Amazon S3.
- Sample Dashboards:
  - https://aws.amazon.com/quicksight/gallery/
  - https://d1s0yx3p3y3rah.cloudfront.net/anonymous-embed?dashboard=cudos

Reference architecture



Service Summary Cards (SSC)

Reference:

FAQs

Category:

Analytics

Amazon QuickSight

More SSCs:

Click Here

Complete Book

Click Here

Created by:

Ashish Prajapati

What?

- Amazon QuickSight is cloud-powered business analytics service that makes it easy to build visualizations, perform ad-hoc analysis, and quickly get business insights from data.

Why?

- QuickSight enables organizations to scale their business analytics capabilities to hundreds of thousands of users, and delivers fast and responsive query performance by using a robust in-memory engine (SPICE).

When?

- When you don’t want to use traditional BI solutions that often require teams of data engineers to spend months building complex data models before generating a report.

Where?

- Amazon QuickSight is a regional service but you can also explicitly connect to other AWS data sources that are not in your account or in a different region by providing connection details for those sources.

Who?

- You can seamlessly grow your data from a few hundred megabytes to many terabytes of data without managing any infrastructure.

How?

- Amazon QuickSight discovers your data sources in AWS services such as Amazon Redshift, Amazon RDS, Amazon Athena, and Amazon Simple Storage Service (Amazon S3). You can connect to any of the data sources discovered by Amazon QuickSight and get insights from this data in minutes.

How much?

- Month-to-month pricing for Authors and Readers.
- Alerts are priced based on metrics evaluated.
- Accounts enabled with Q are charged a $250/month Q base fee plus Q questions capacity pricing.