Relational Databases

aws training and certification

SQL Overview

Data Definition Language (DDL)

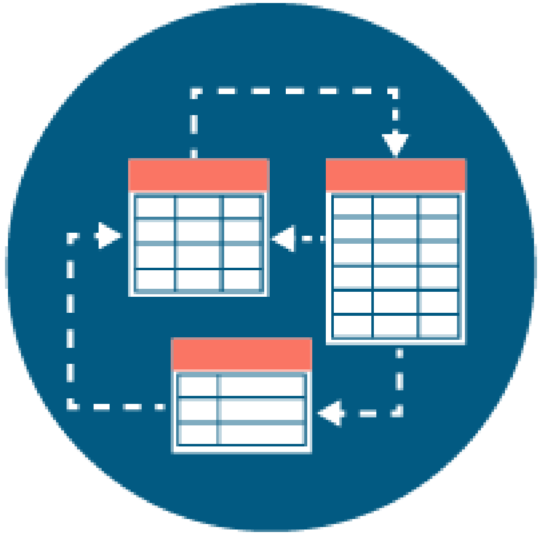
Data Control Language (DCL)

Data Manipulation Language (DML)

Structured Query Language (SQL)

Transactional Databases

- Relational Database
- Collects and manages transactional and operational data
- Examples include
  - Point-of-Sale (POS) systems
  - Registration systems
  - Auditing systems



Analytical Databases

- Relational Database
- Performs complex analytical queries
- Data is imported from other, often transactional, systems

Comparison of OLTP and OLAP Systems

aws training and certification

	Transactional	Analytical
Data source	Origin	Consumer
Purpose	Capture data	Analyze data
Workloads	INSERT, UPDATE, DELETE, short and fast queries	Batch jobs to import data, JOINs, complex queries
Database design	Highly normalized using many distinct tables to reduce duplication	Denormalized using fewer tables in star and snowflake schemas with some duplicated data
Database size	Depends on the amount of data but typically from MB to TB in size	Grows over time and typically ranges from TB to PB in size.

Data Warehousing Concepts

aws training and certification

Data and Insights

Data is not information.

Data sources

Data transformations

Analytical data

What is a Data Warehouse?

A data warehouse is:

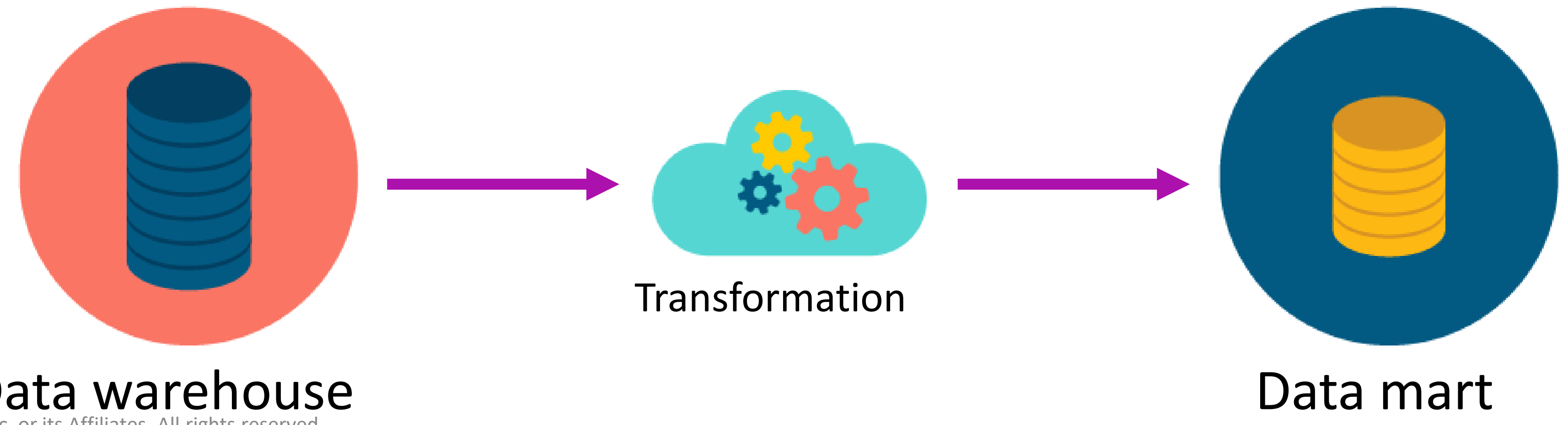
- A central repository of business data from disparate sources.
- A type of relational database that enables analysis of data.
- A collection of approved and trusted historical corporate data.

Data Warehousing Goals

- Provide access to corporate or organizational data.
- Ensure consistency and quality of the data.
- Enable analysis of data in different ways based on measurements that are defined by the business.
- Integrate with query, analysis, visualization, and reporting tools.

What Are Data Marts?

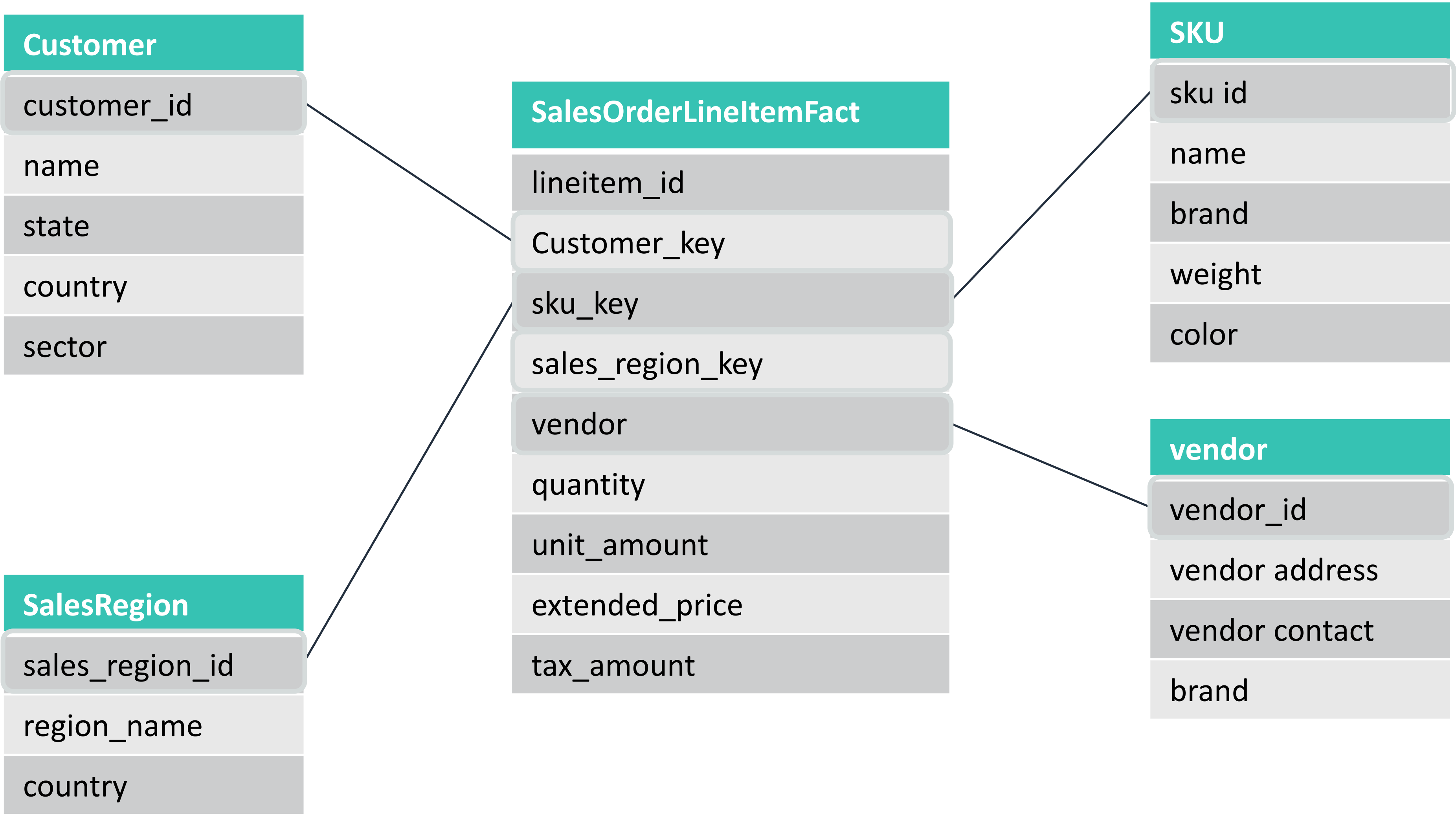
- A subset of a data warehouse
- Used to model a single subject or dimension, such as product or customer
- Often organized in a star or snowflake schema
- Fed from a data warehouse (top-down approach) or feeds a data warehouse (bottom-up approach)



Dimensional Modeling: Star Schema

- Developed by data warehouse architect, Ralph Kimball.
- The center is a fact table and the points are dimension tables.
- Many business intelligence (BI) applications support the star schema.

Star Schema Example



Customer
customer_id
name
state
country
sector

SalesRegion
sales_region_id
region_name
country

SalesOrderLineItemFact
lineitem_id
Customer_key
sku_key
sales_region_key
vendor
quantity
unit_amount
extended_price
tax_amount

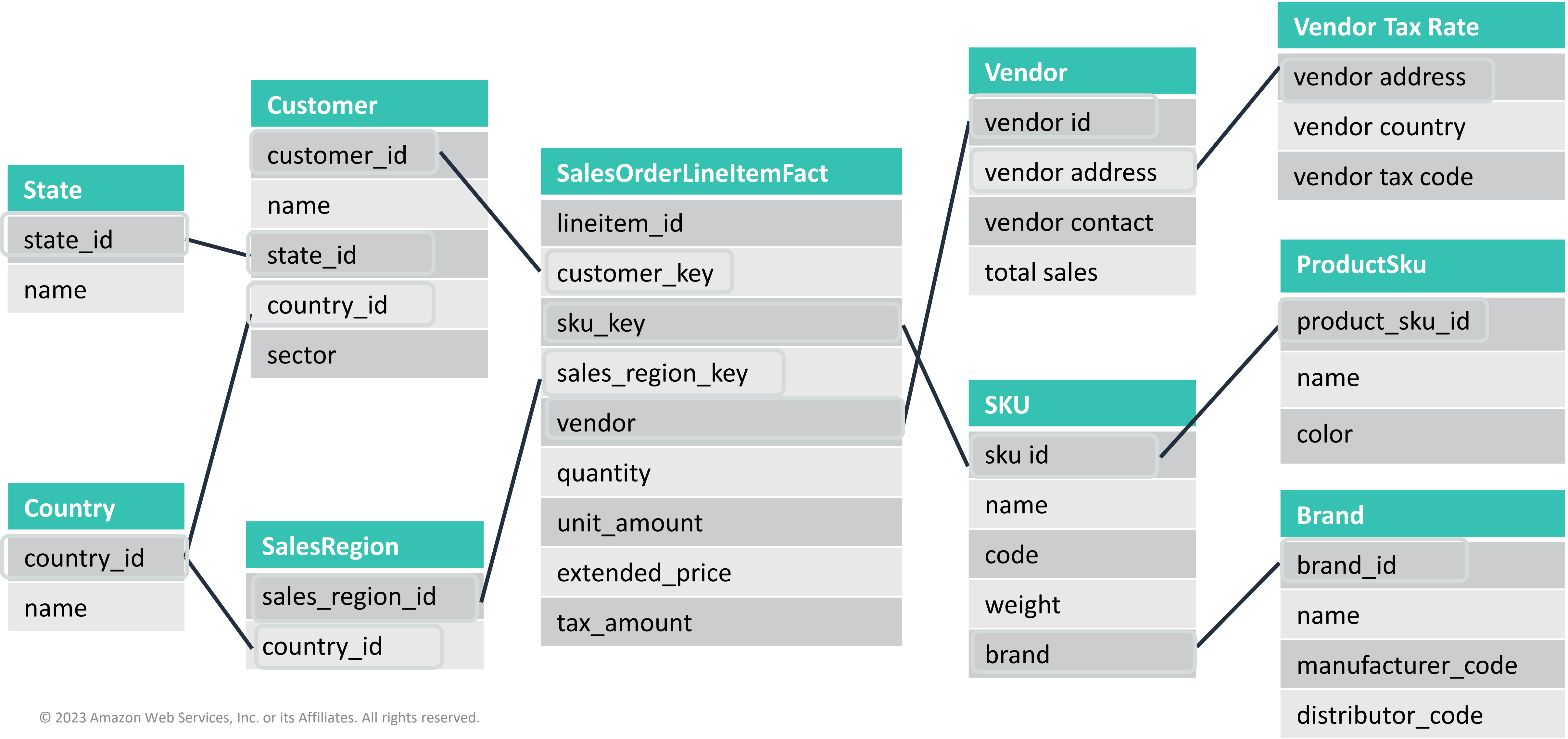
SKU
sku id
name
brand
weight
color

vendor
vendor_id
vendor address
vendor contact
brand

Normalized Modeling: Snowflake Schema

- Developed by “the inventor of the relation model,” Edgar F. Codd.
- Data is stored in related tables that are normalized to third normal form (3NF) to reduce data redundancy.

Snowflake Schema Example



State
state_id
name

Country
country_id
name

Customer
customer_id
name
state_id
country_id
sector

SalesRegion
sales_region_id
country_id

SalesOrderLineItemFact
lineitem_id
customer_key
sku_key
sales_region_key
vendor
quantity
unit_amount
extended_price
tax_amount

Vendor
vendor id
vendor address
vendor contact
total sales

SKU
sku id
name
code
weight
brand

Vendor Tax Rate
vendor address
vendor country
vendor tax code

ProductSku
product_sku_id
name
color

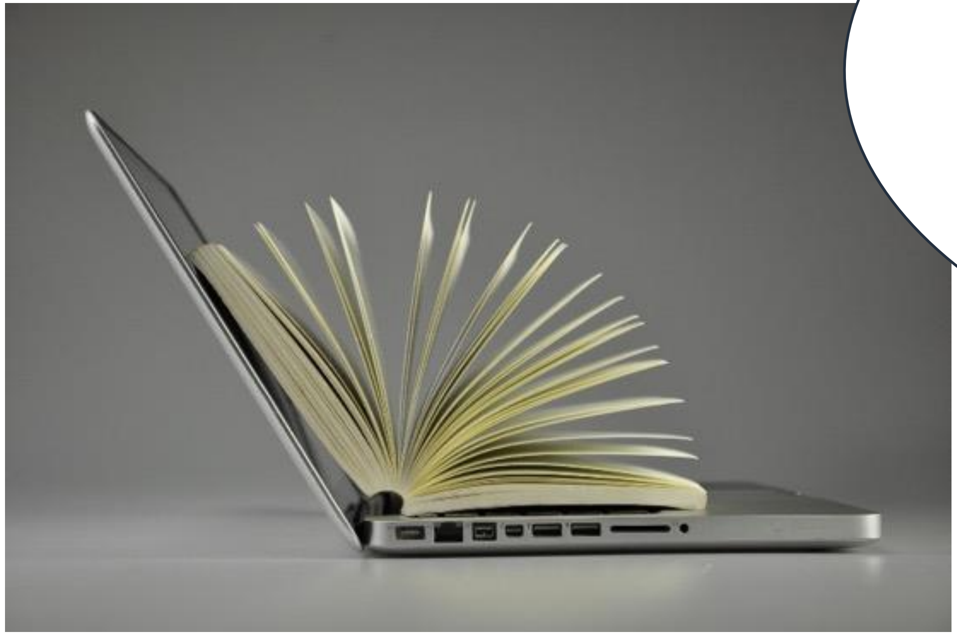
Brand
brand_id
name
manufacturer_code
distributor_code

Data Models and Redshift

- Data models were created to handle the complexity of organizing data from multiple data sources.
- Amazon Redshift already works with these common schemas and more. We recommend that you load your data in Amazon Redshift to try it out and adjust as needed.

The Intersection of Data Warehousing and Big Data

What Is Big Data?

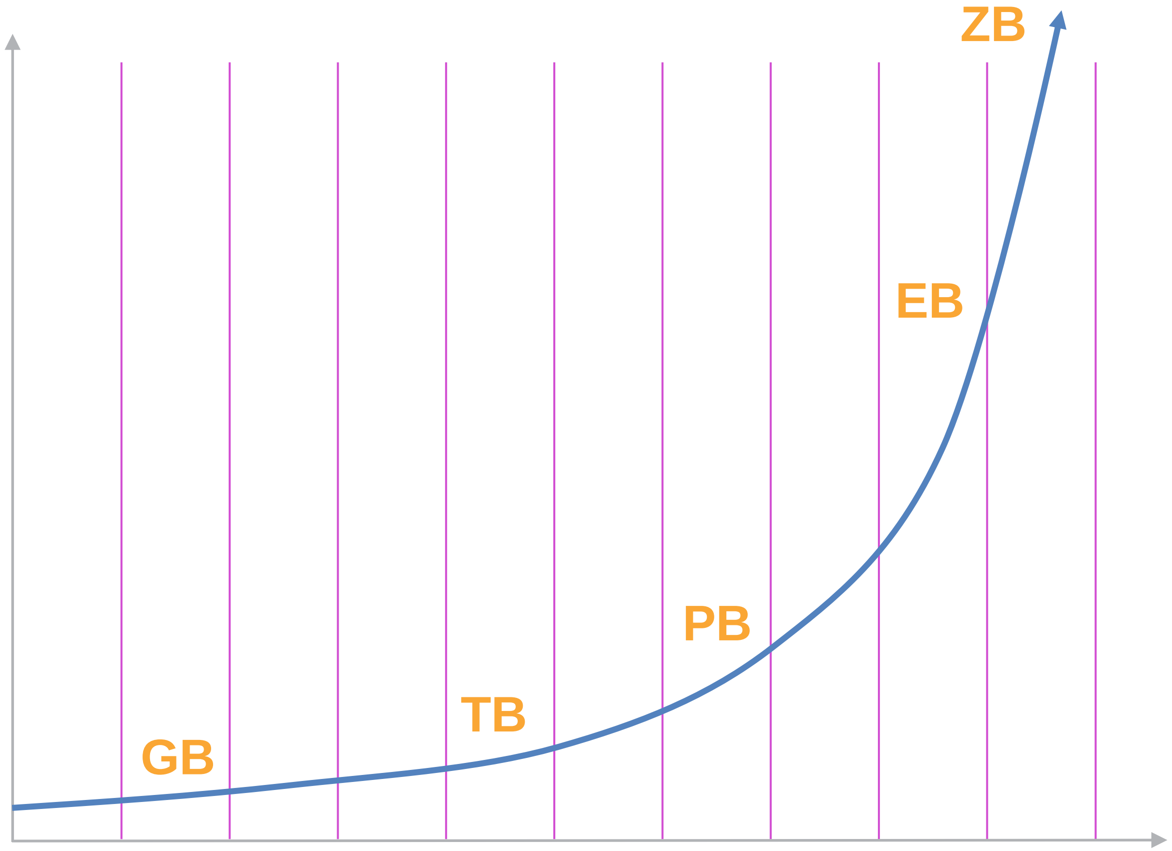


“Big data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.”

Gartner IT Glossary

Unconstrained Big Data Growth

- IT/Application server logs
  - IT Infrastructure logs, Metering, Audit logs, Change logs
- Websites/Mobile apps/Ads
  - Clickstream, User Engagement
- Sensor data/IoT
  - Weather, Smart Grids, Wearables
- Social media, user content
  - 450MM+ Tweets/day



Characteristics of Big Data

The 5 Vs

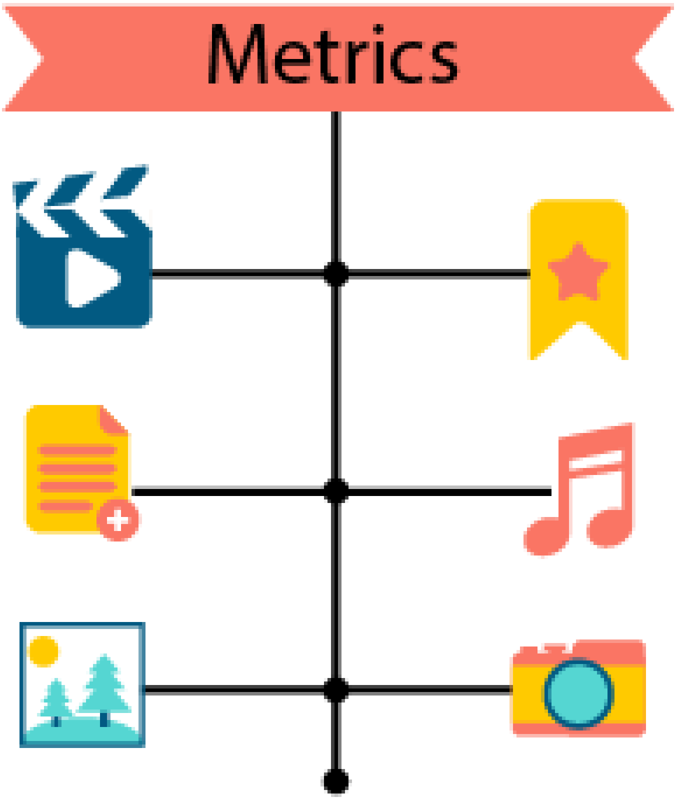
- **V**olume
- **V**ariety
- **V**elocity
- **V**ariability
- **V**eracity

Typical Use Cases

- Customer segmentation
- Marketing spend optimization
- Financial modeling and forecasting
- Ad targeting and real-time bidding
- Clickstream analysis
- Fraud detection

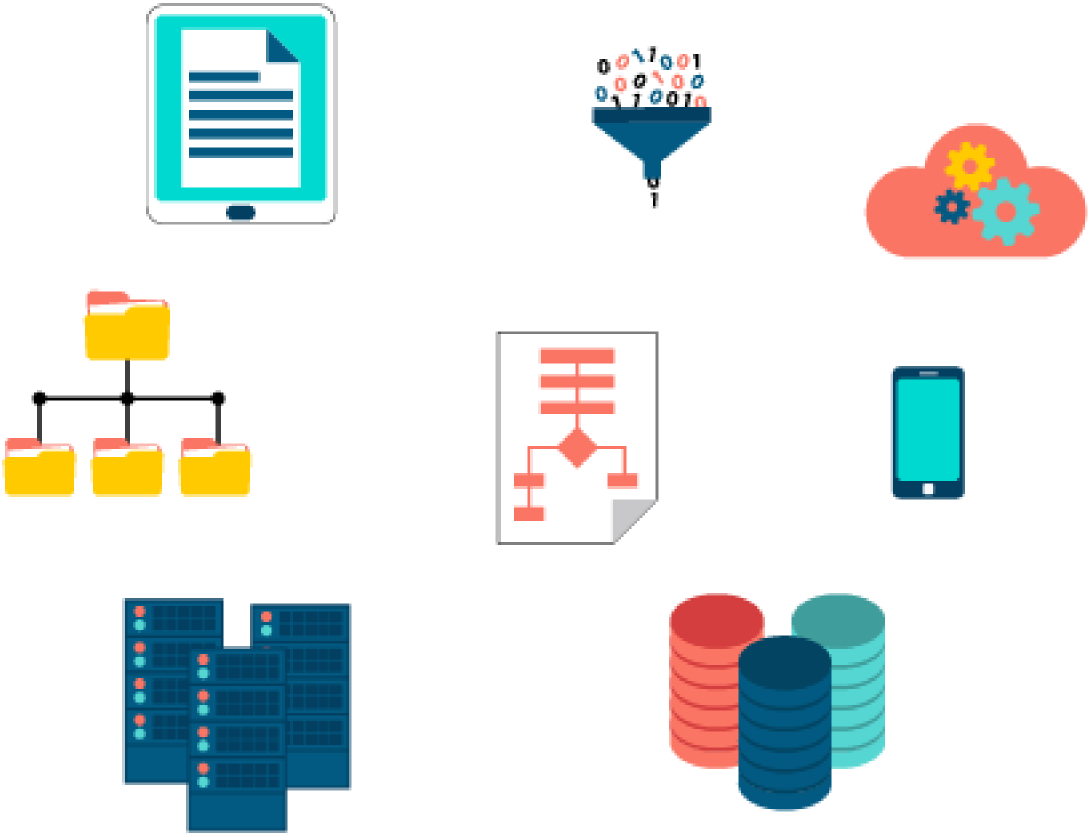
Metrics

- Visits, views, clicks, and purchases
- Sources, devices, locations, times
- Latency, throughput, uptime
- Likes, shares, friends, follows
- Prices, frequency



Data Sources

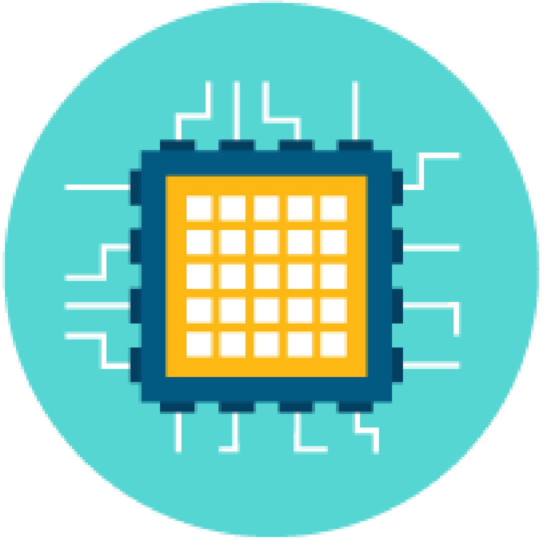
- Relational databases
- NoSQL databases
- Web servers
- Mobile phones
- Tablets
- Data feeds



Data Formats

- Structured, semi-structured, and unstructured
- Text
- Binary
- Streaming and near real-time
- Batched

Use the Right Tool for the Job



Compute-intensive

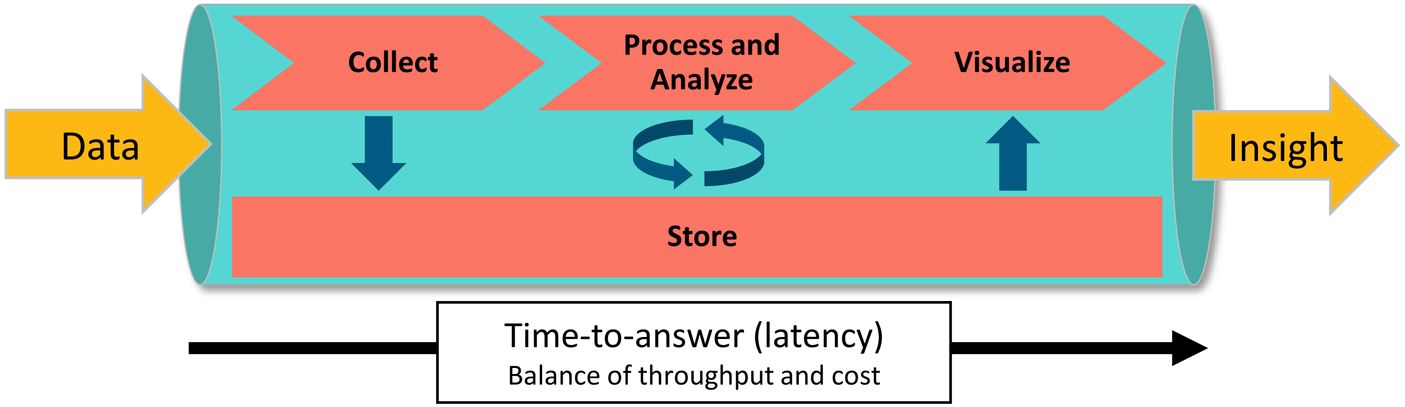
Streaming analytics

Data transformation

Multi-input operations

Big Data vs Data Warehouses

Big Data is a concept.



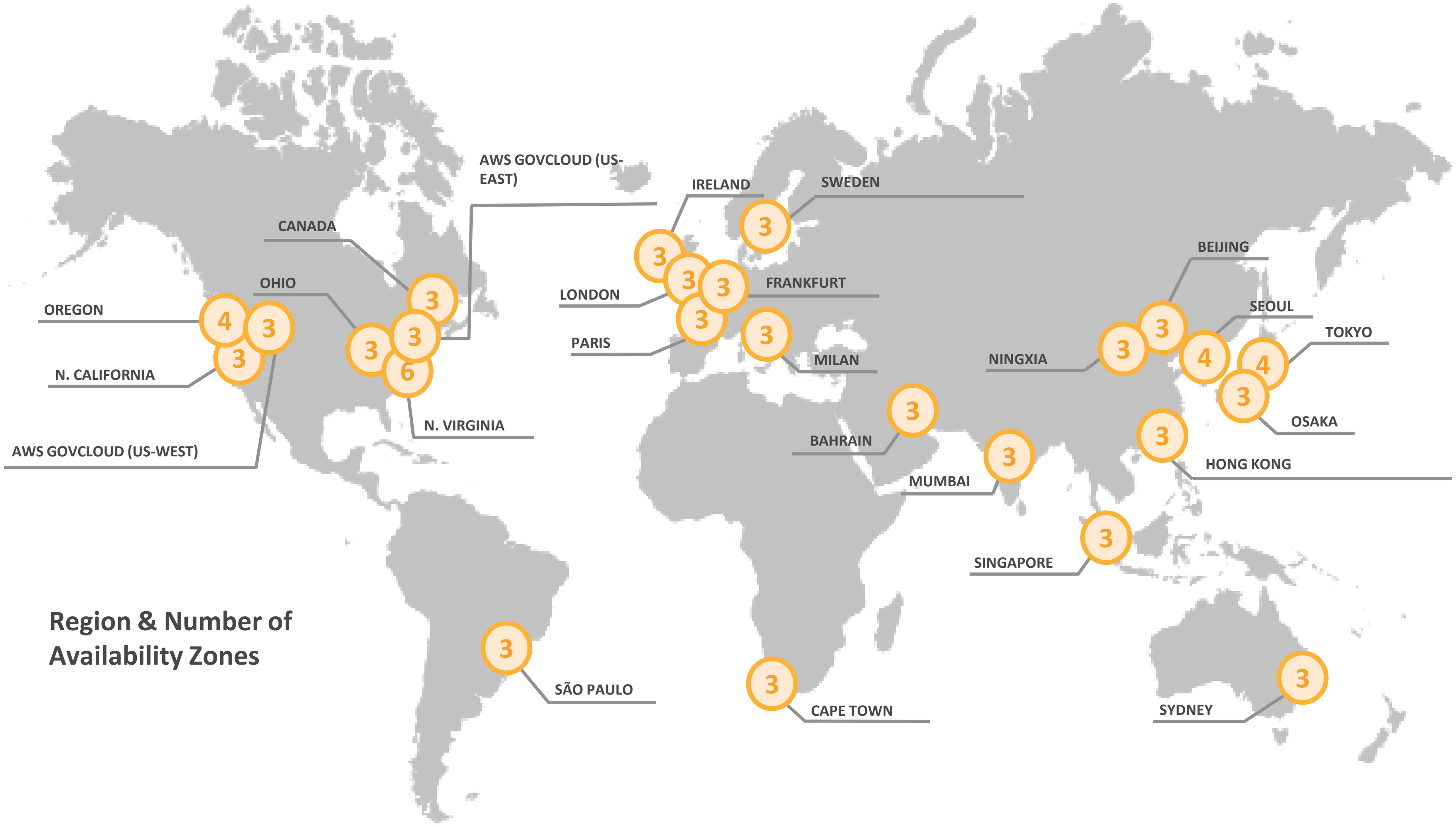
A data warehouse:

- Can be used with both small and large datasets
- Can be used in a Big Data system

Overview of Data Management in AWS

AWS Global Infrastructure: Current Regions

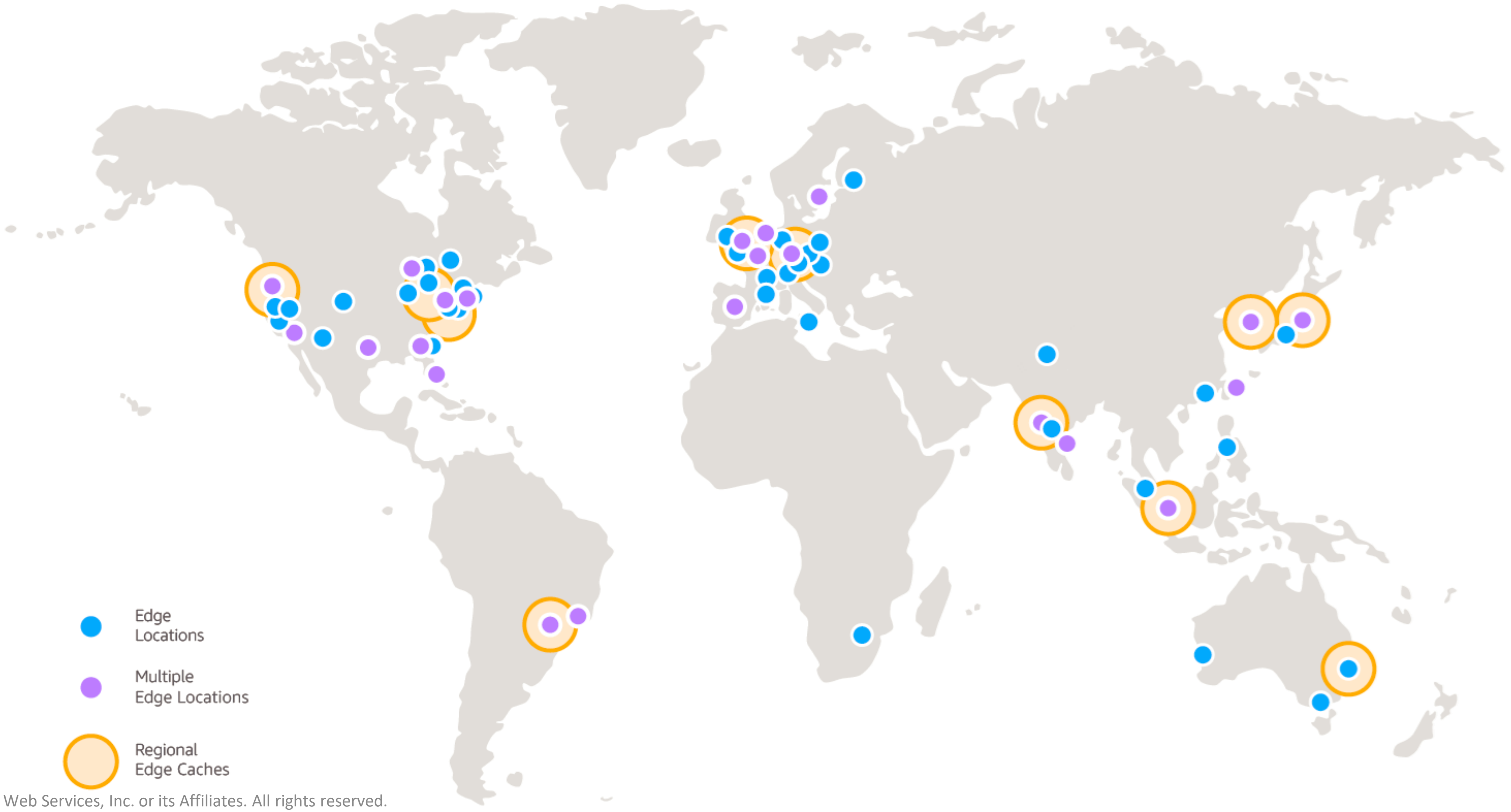
aws training and certification



Region & Number of Availability Zones

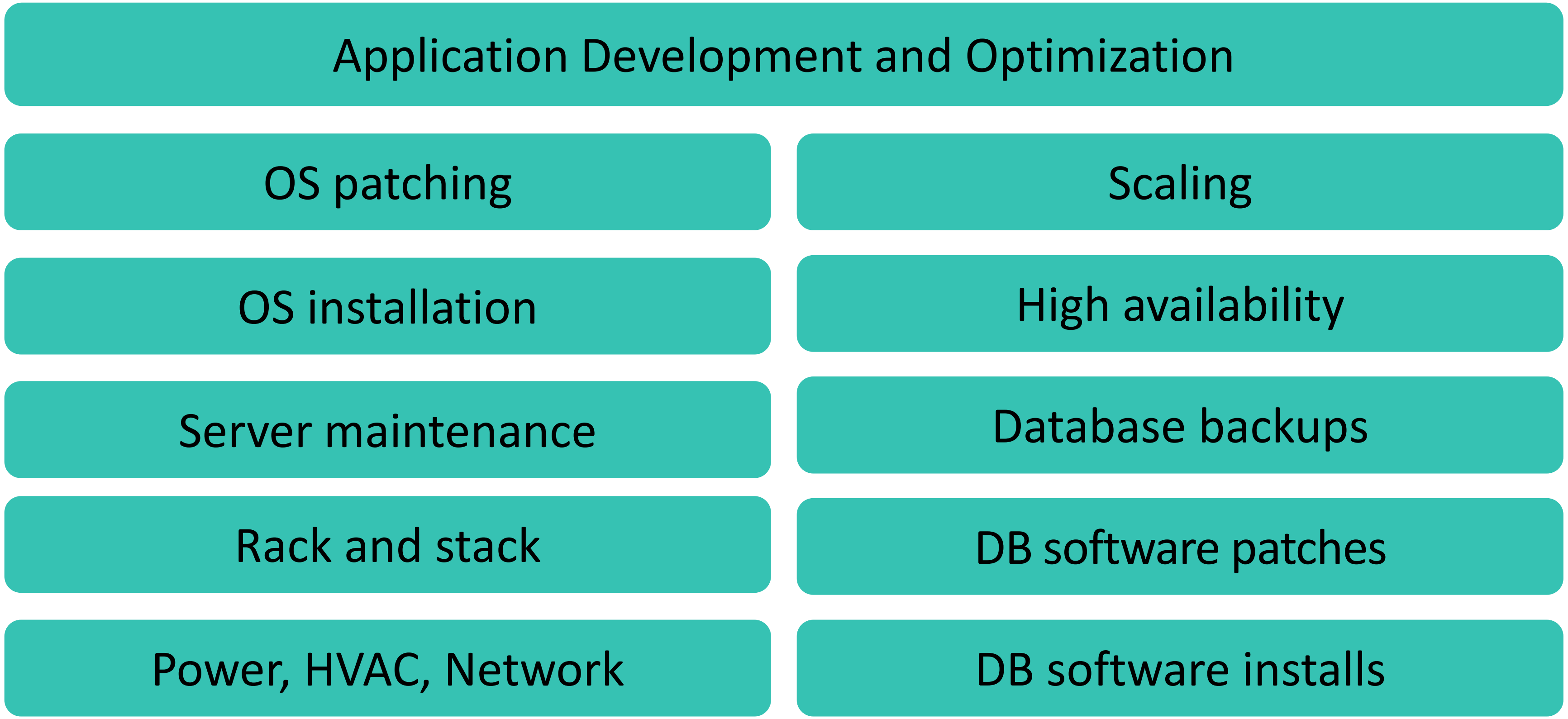
AWS Global Infrastructure: Edge Locations

aws training and certification



On-Premises Systems

In traditional systems, you are responsible for the following tasks:



Application Development and Optimization

OS patching

OS installation

Server maintenance

Rack and stack

Power, HVAC, Network

Scaling

High availability

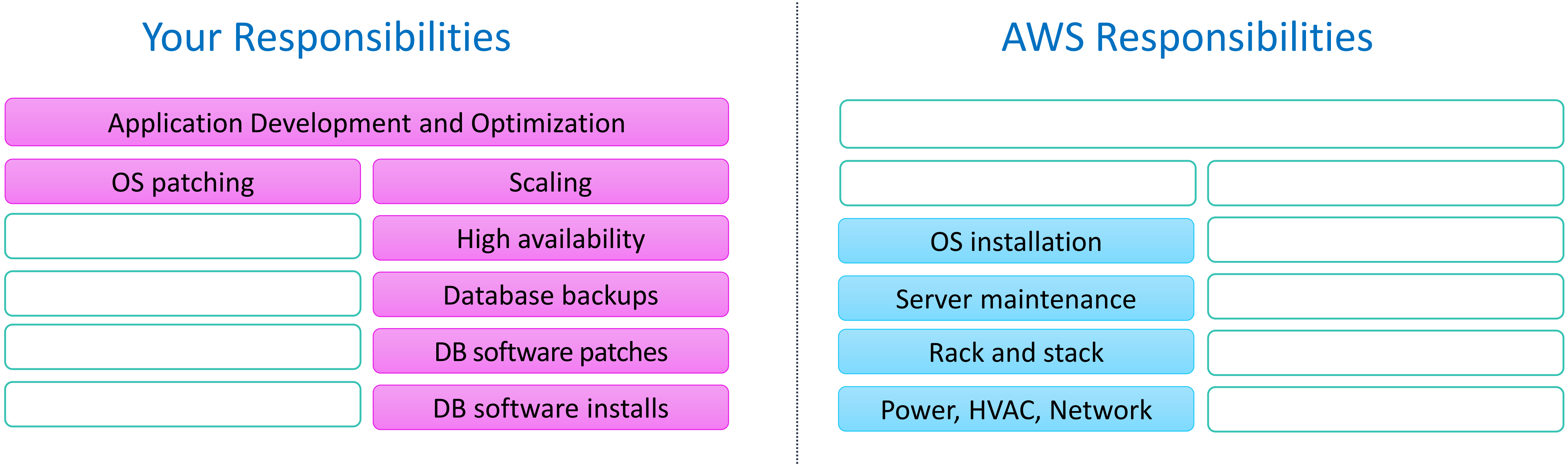
Database backups

DB software patches

DB software installs

Moving Systems to AWS

By redeploying your systems in Amazon EC2 instances, you unload some of the tasks to AWS.



Your Responsibilities

Application Development and Optimization

OS patching

Scaling

High availability

Database backups

DB software patches

DB software installs

AWS Responsibilities

OS installation

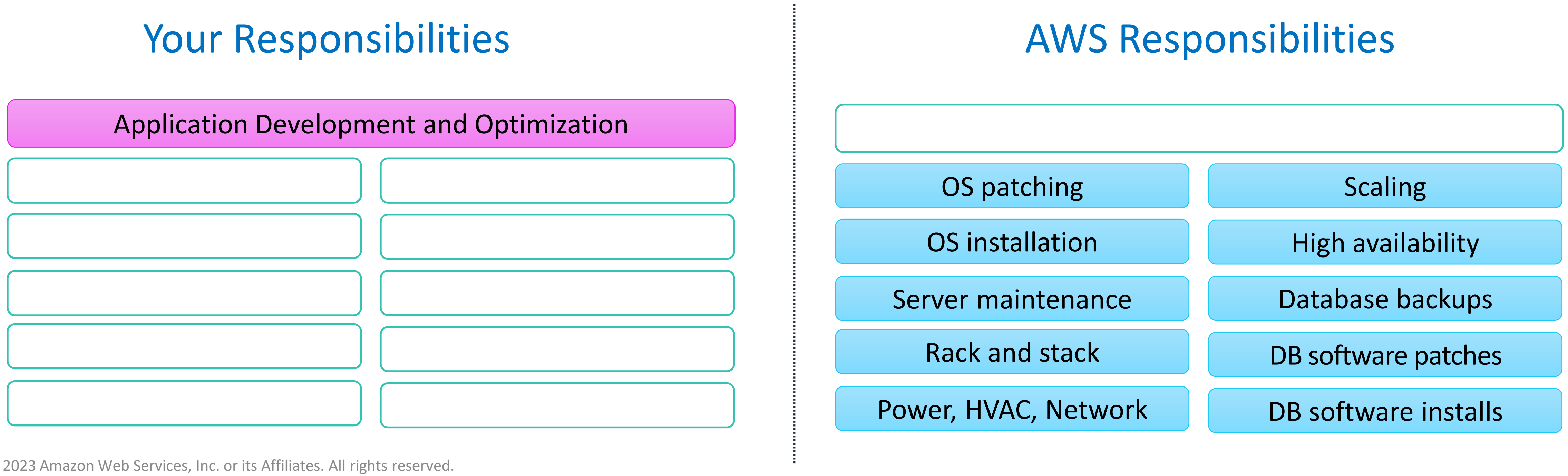
Server maintenance

Rack and stack

Power, HVAC, Network

AWS Managed Services

By transitioning your systems to AWS-managed services, you unload many tasks to AWS and can focus on your applications and optimizations for your business.



Your Responsibilities

Application Development and Optimization

AWS Responsibilities

OS patching

Scaling

OS installation

High availability

Server maintenance

Database backups

Rack and stack

DB software patches

Power, HVAC, Network

DB software installs

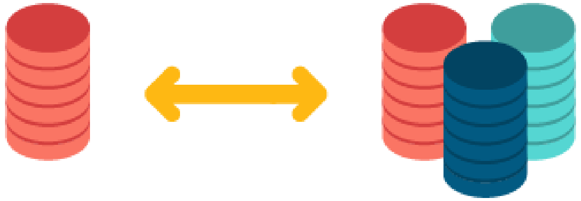
Benefits of AWS Database Services

Managed

AWS installs, patches, and manages the services.

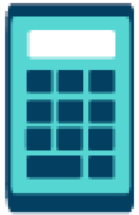
Scalable

You can grow or shrink resources as needed.



No up-front cost

You pay only for what you use.



Integrated

The database services are already integrated with other AWS services.

Amazon KMS

Amazon S3

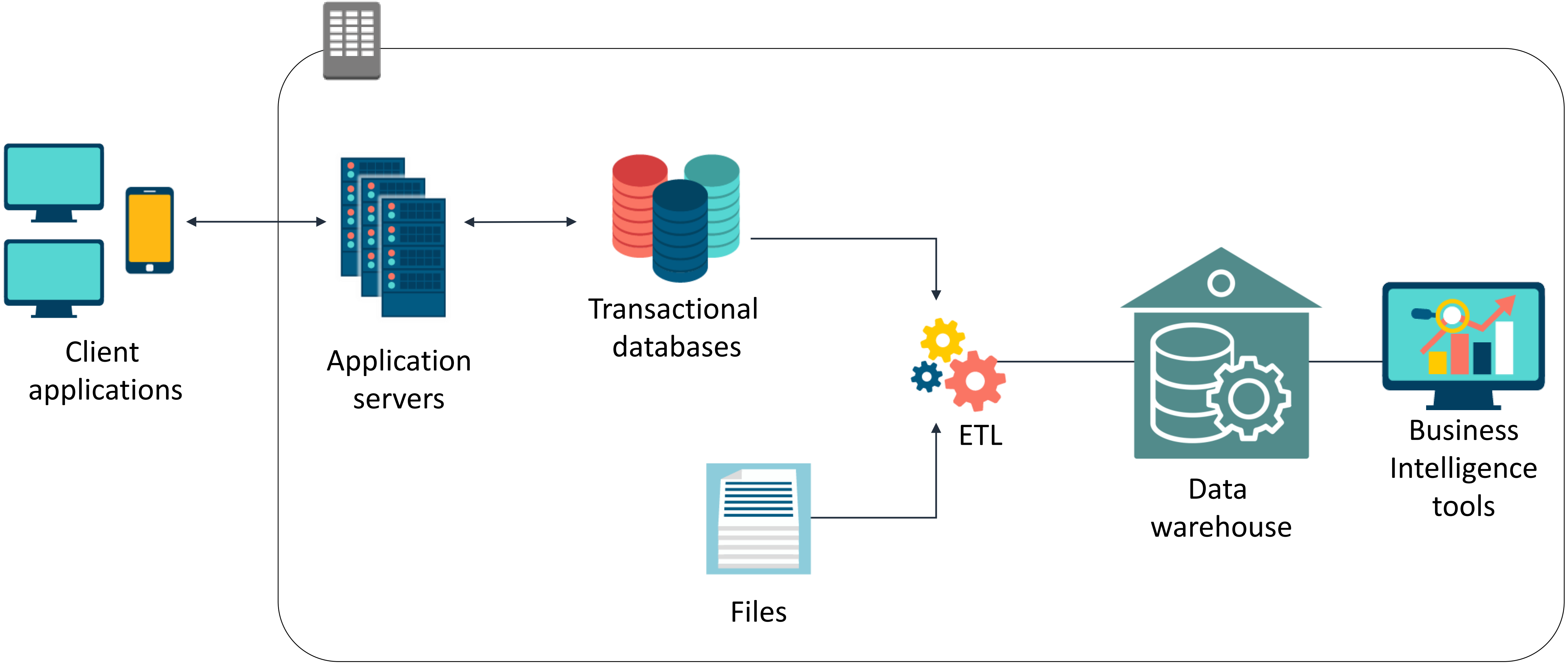
Amazon EC2

Amazon VPC

Amazon SNS

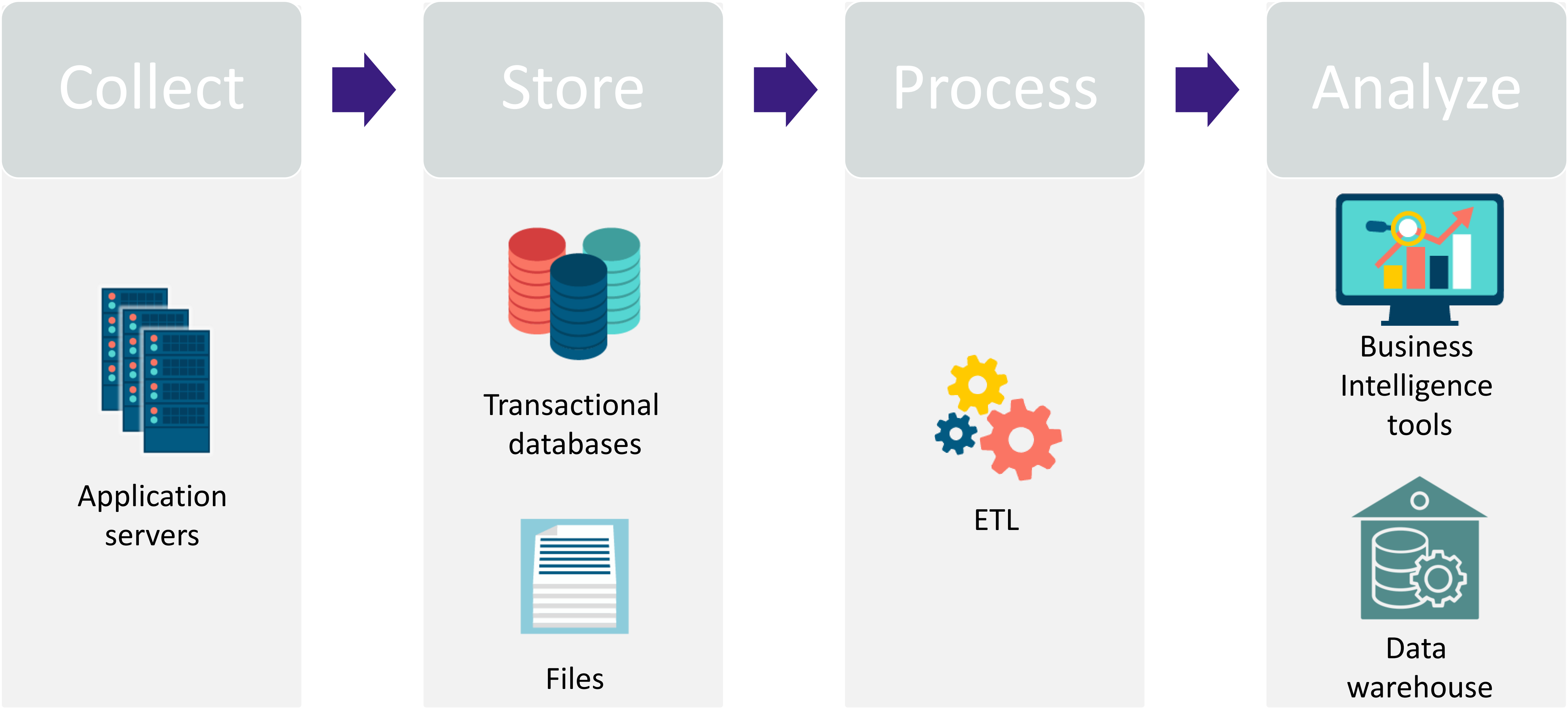
Amazon CloudWatch

Traditional Data Flow Overview



Data Flow With Traditional Systems

aws training and certification



Collect

Application servers

Store

Transactional databases

Files

Process

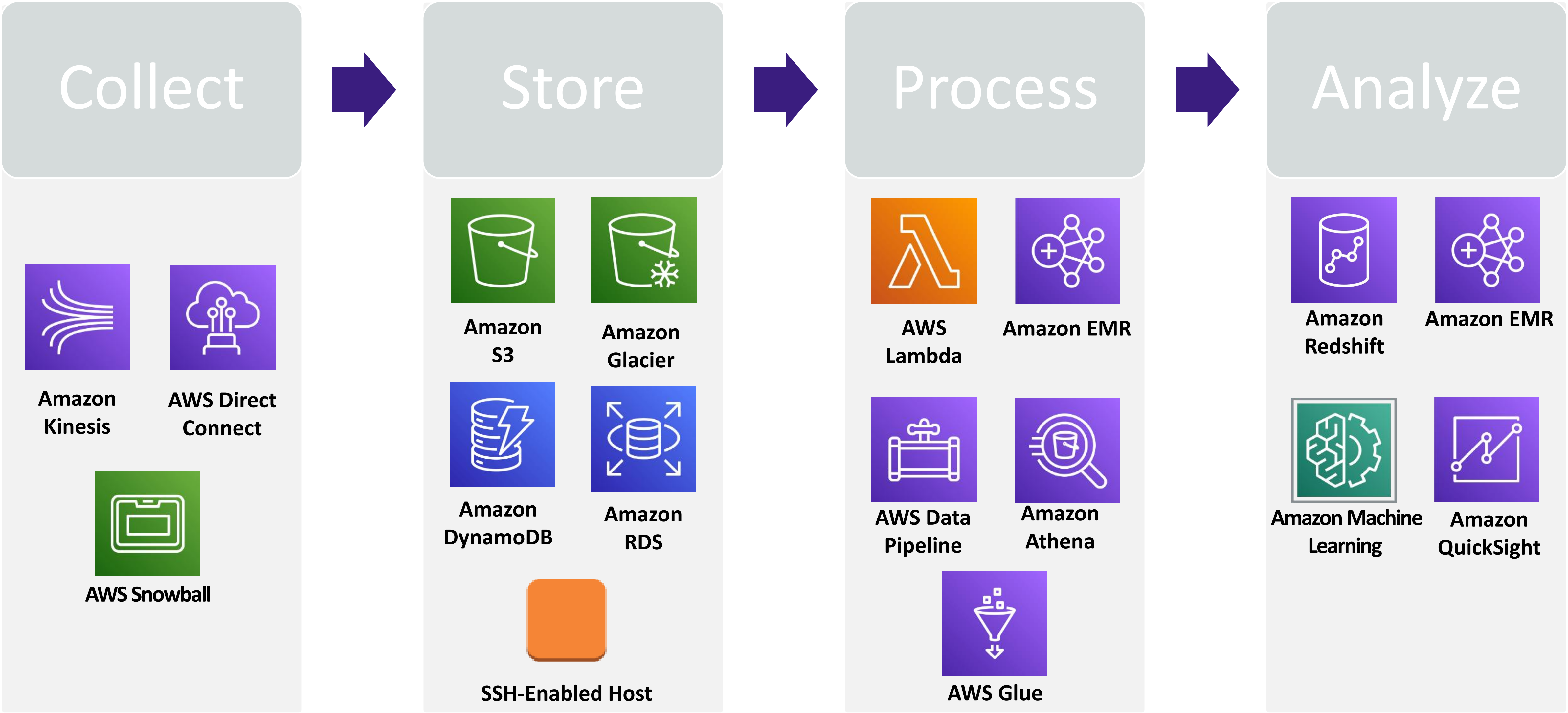
ETL

Analyze

Business Intelligence tools

Data warehouse

Data Flow With AWS Services



Collect → Store → Process → Analyze

Collect

- Amazon Kinesis
- AWS Direct Connect
- AWS Snowball

Store

- Amazon S3
- Amazon Glacier
- Amazon DynamoDB
- Amazon RDS
- SSH-Enabled Host

Process

- AWS Lambda
- Amazon EMR
- AWS Data Pipeline
- Amazon Athena
- AWS Glue

Analyze

- Amazon Redshift
- Amazon EMR
- Amazon Machine Learning
- Amazon QuickSight

RedShift

Amazon Redshift Is a…

fast,

simple,

cost-effective,

fully managed,

petabyte-scale,

enterprise-grade,

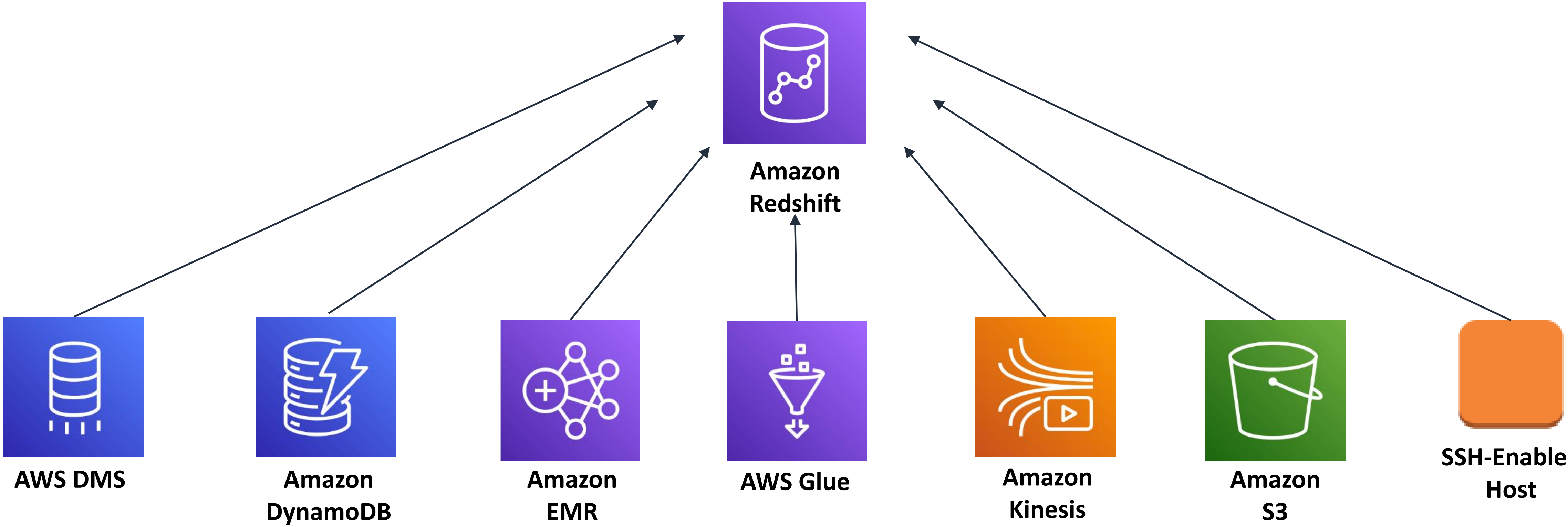
relational

…data warehousing service on AWS.

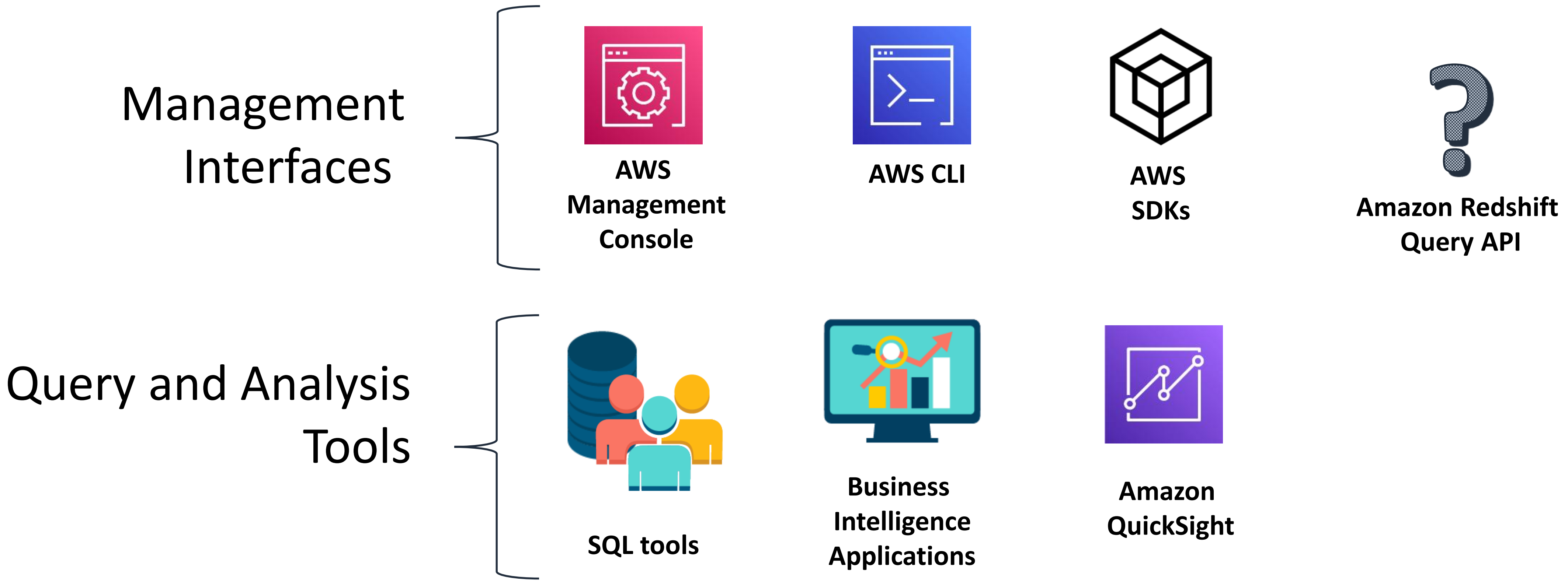
Amazon Redshift

Data Sources Integrated with Redshift

These services are integrated data sources loading data in parallel into Amazon Redshift. An SSH-enabled host can be either an Amazon EC2 instance or an on-premises computer.



Interacting with Amazon Redshift



Management Interfaces

- AWS Management Console
- AWS CLI
- AWS SDKs
- Amazon Redshift Query API

Query and Analysis Tools

- SQL tools
- Business Intelligence Applications
- Amazon QuickSight

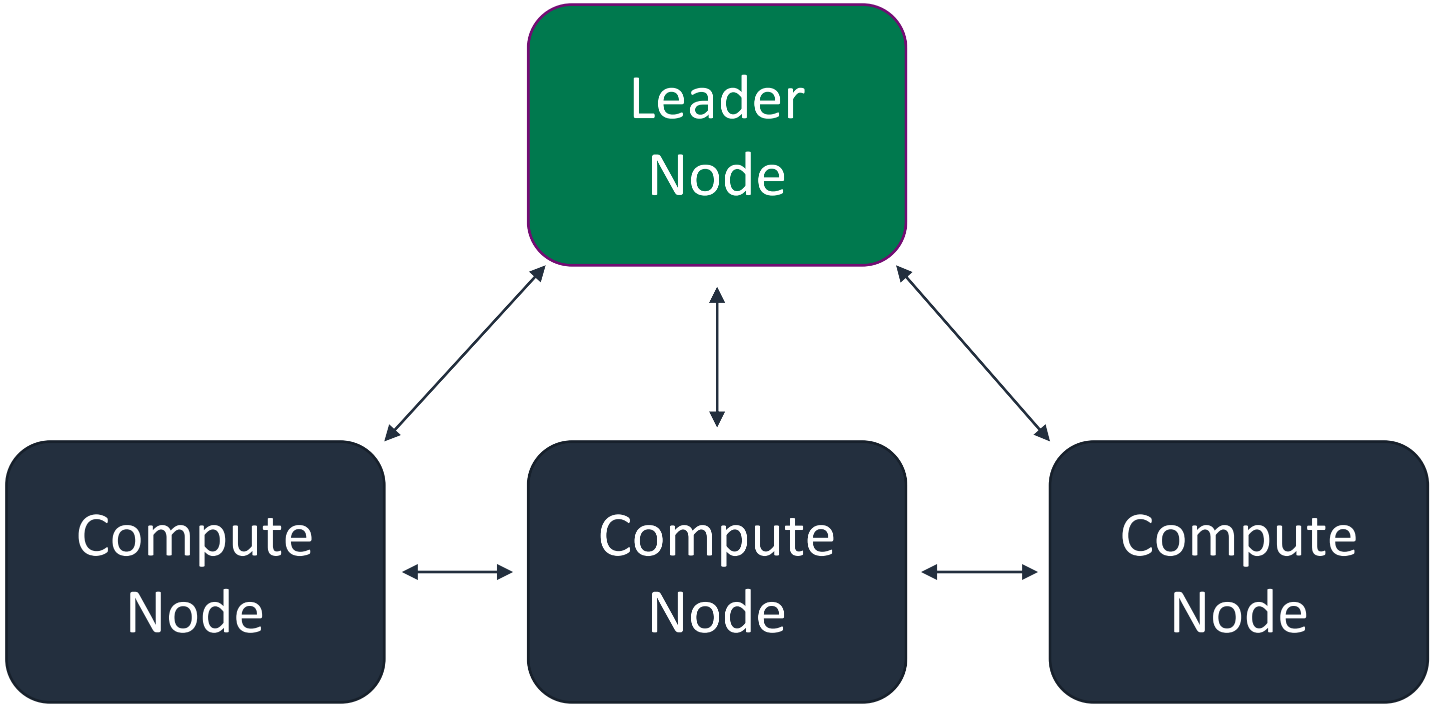
Massively Parallel Processing (MPP)

Amazon Redshift is built on MPP architecture, which:

- Is a distributed, shared-nothing architecture that is easily scalable by adding or removing nodes
- Is optimized for analytic workloads
- Provides the ability to distribute, scan, and process queries in parallel across nodes in the cluster
- Improves query performance significantly for complex analytical queries against massive data sets

Cluster Overview

An Amazon Redshift cluster comprises one leader node, and one or more compute nodes.

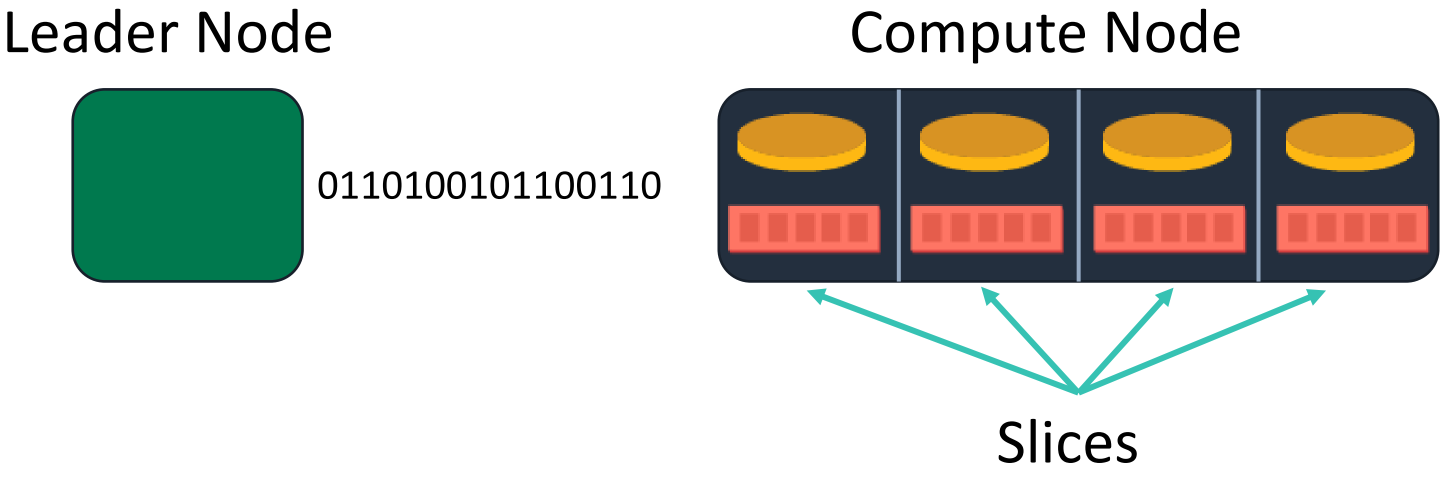


Node Slices

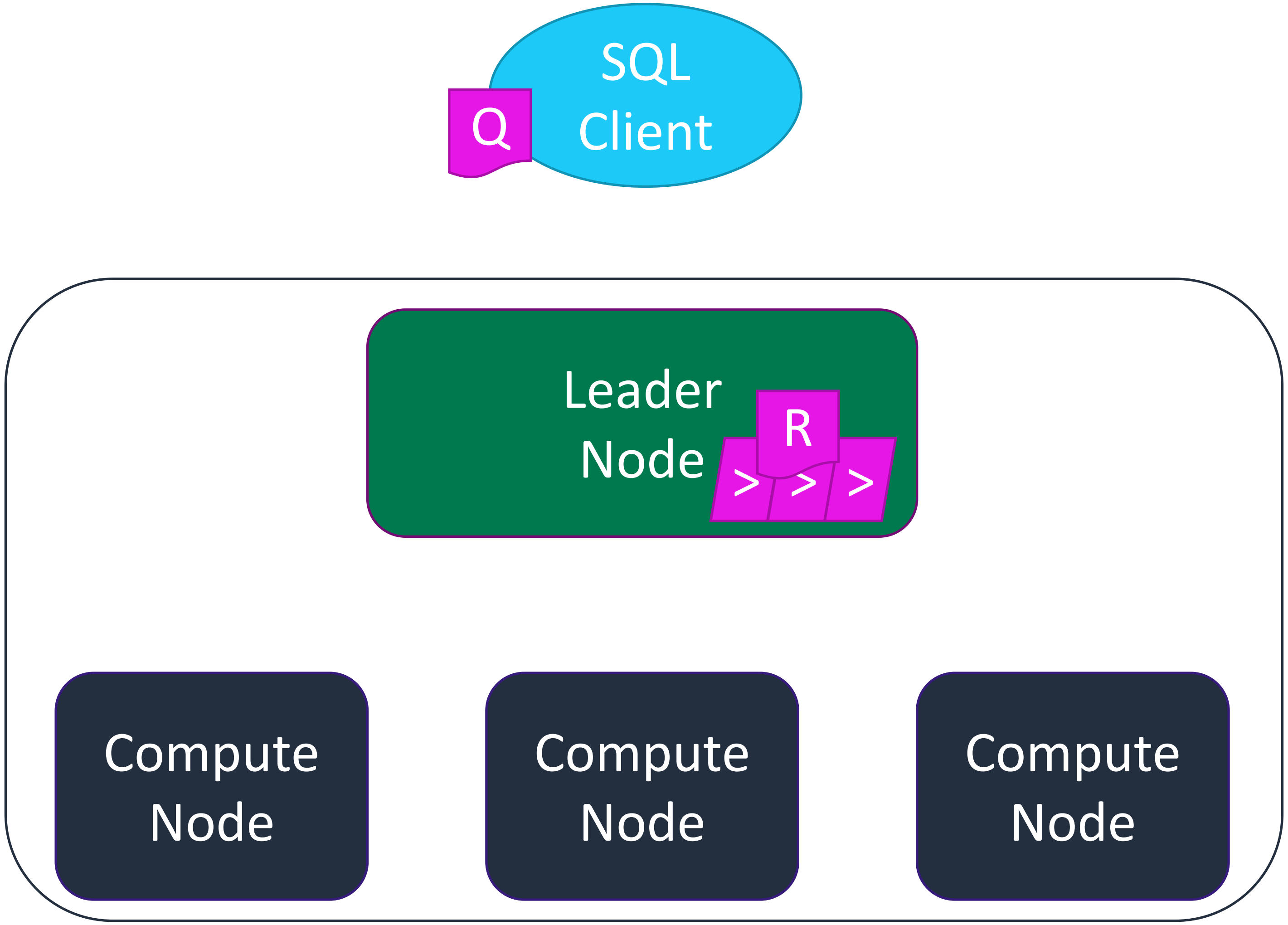
A compute node is partitioned into slices.

Each slice is allocated a portion of the cluster's memory and disk space.

The Leader node distributes data to the slices.



Query Processing



1. The SQL client submits a query.
2. The leader node parses the query and develops an execution plan.
3. The leader node distributes work among compute nodes.
4. The compute nodes run the query according to the plan, and return interim results to the leader node.
5. The leader node aggregates and returns results to the SQL client.

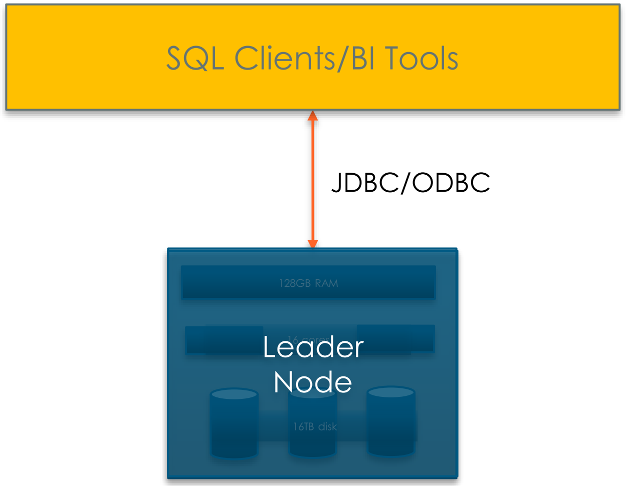
Internal Architecture

Network

- SQL endpoint
- No direct client application communication
- Client communication must be JDBC/ODBC
- Can be encrypted

Leader Node

- Execution engine
- Stores metadata
- Coordinates query execution



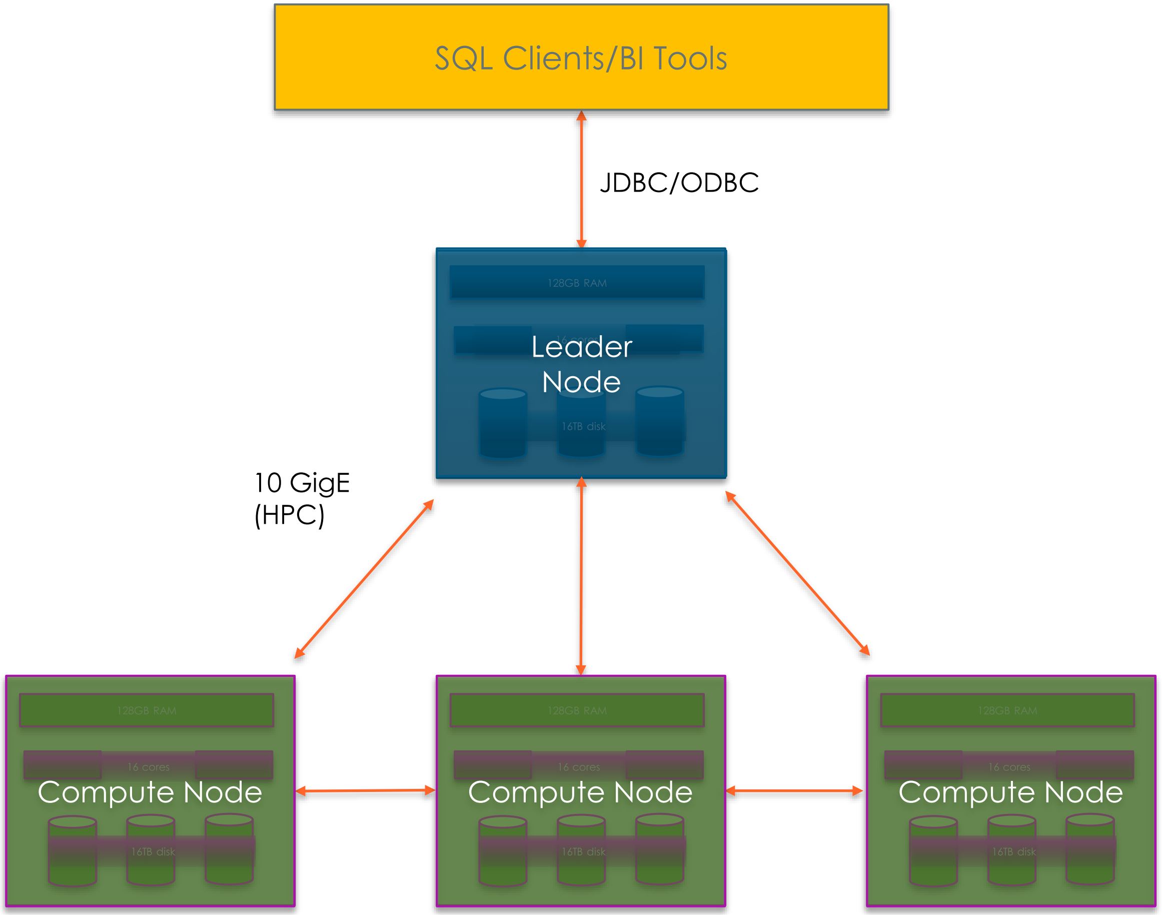
Internal Architecture

Compute nodes

- Local, columnar storage
- Run queries in parallel
- Load, backup, restore via Amazon S3
- Load from Amazon DynamoDB or SSH
- Fault tolerant

Three hardware platforms

- Optimized for data processing
- DS2: HDD; scale from 2 TB to 2 PB
- DC2: SSD; scale from 160 GB to 326 TB
- RA3: SSD; scale from 64 TB to 16,384 TB (16.38 PB)



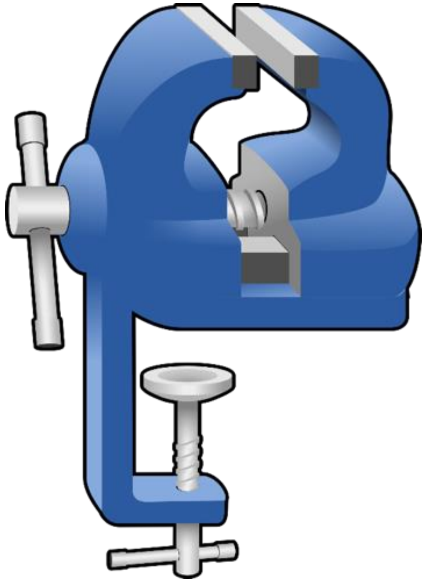
Columnar Storage

- Is optimized for scanning large data sets and complex analytic queries
- Enables a data block to store and compress significantly more values for a column compared to row-based storage
- Eliminates the need to read redundant data by reading only the columns that you include in your query
- Offers overall performance benefits that can help eliminate the need to aggregate data into cubes as in some other OLAP systems

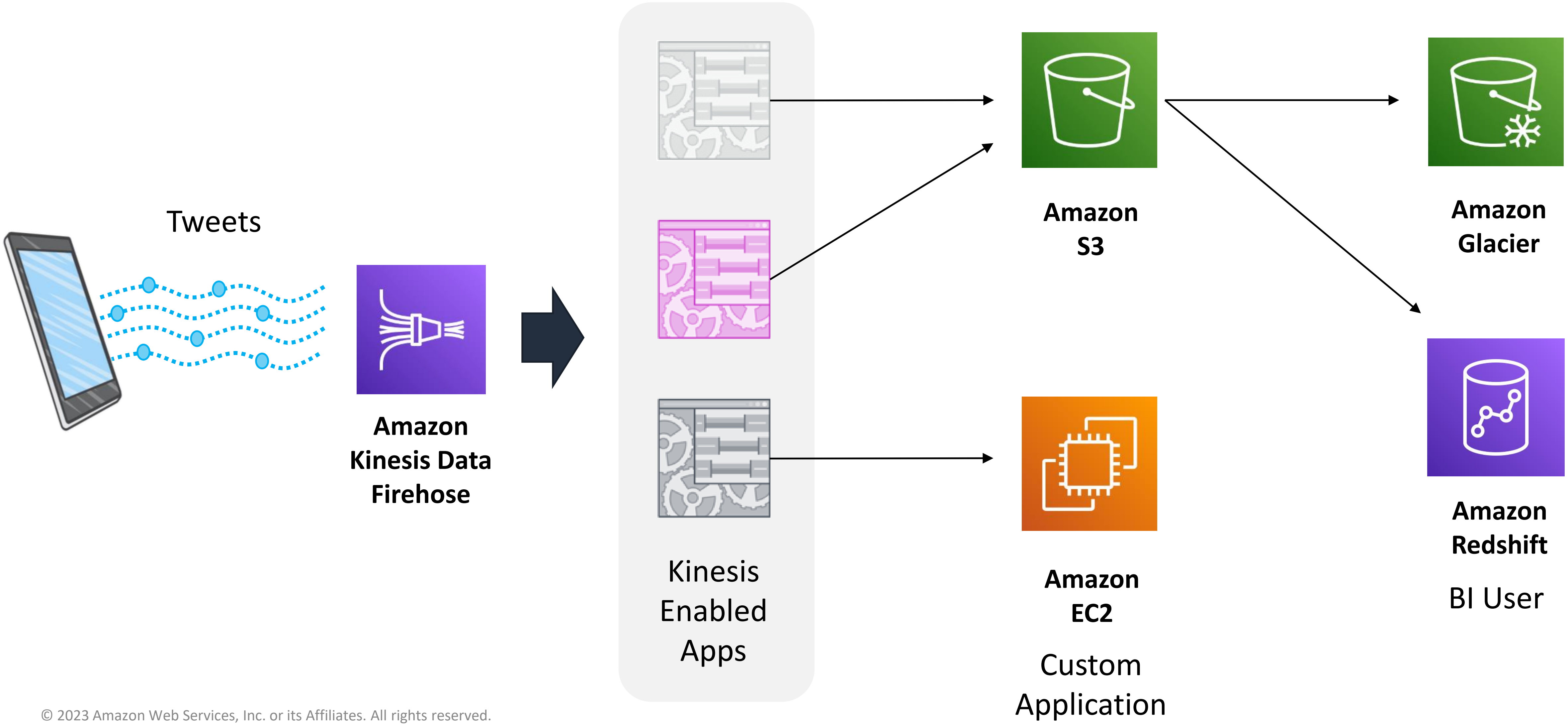
Columnar Compression

Amazon Redshift supports the following compression encodings:

- Raw
- Byte-Dictionary
- Delta
- LZO
- Mostly
- Runlength
- Text
- Zstandard
- AZ64 (A new compression algorithm to save disk space)



Analyzing Near Real–Time Tweets



Analyzing Near Real–Time Tweets from Twitter Firehose

A look at the numbers:

- 500 million tweets per day = ~5,800 tweets per second
- 2000 tweets = ~12 MB per second (~1 TB per day)
- $0.015 per hour per shard
- $0.028 per million PUT requests
- $0.765 per hour for Amazon Kinesis
- $0.850 per hour for a 2 TB node in Amazon Redshift
- $1.28 per hour for uncompressed data in Amazon S3

Total: $2.895 per hour