Analytics & Data Science

Beyond applications & compute: anaylzing data

▶ Analytics & Data Science with Help of AWS

▶ Data Ingestion & Streams

▶ Transforming & Analyzing Data
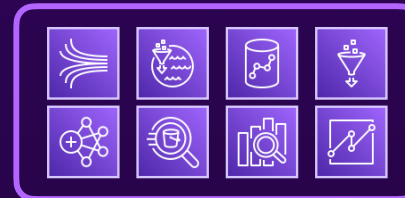
Data Analytics: What & Why?

Modern businesses generate huge amounts of data

Structuring & analyzing such data efficiently can be a competitive advantage

AWS offers various services that help with gathering, storing, extracting, transforming & analyzing data

Ingesting Data

There's a broad variety of data sources

Applications	Crawlers & Scheduled Tasks	Devices & Sensors	Manual Data Entry
e.g., orders, user data, website analytics, logs, …	e.g., crawled website data, weather data, …	e.g., temperatures, movement speeds, …	e.g., accounting data, documentation, …

Ingestion Frequency

Slow	Moderate	Fast
e.g., manual data entry, weather data	e.g., user orders	e.g., website logs, sensor data

Ingesting Data with AWS

There's a broad variety of data sources

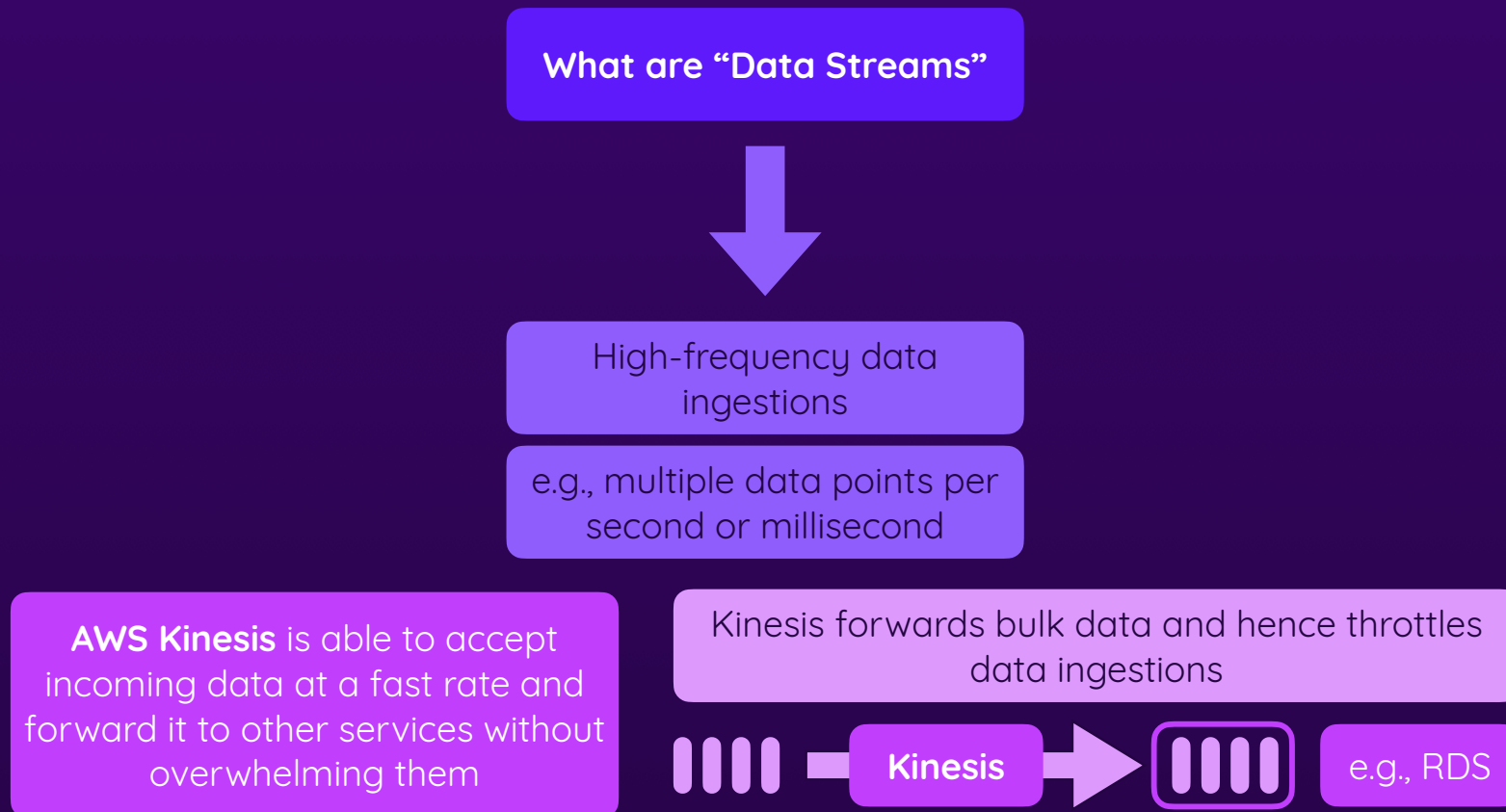| Applications	Crawlers & Scheduled Tasks	Devices & Sensors	Manual Data Entry
Application code can write data to S3, RDS etc.	(Batch) Tasks can store data to S3, RDS etc.	Kinesis helps with handling data streams	Backend can write data to S3, RDS etc.

Ingestion Frequency

Slow	Moderate	Fast
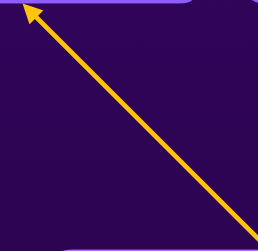
Ingesting Streaming Data with Kinesis

What are "Data Streams"

High-frequency data ingestions

e.g., multiple data points per second or millisecond

AWS Kinesis is able to accept incoming data at a fast rate and forward it to other services without overwhelming them

Kinesis forwards bulk data and hence throttles data ingestions

|||| ➜ **Kinesis** ➜ |||| e.g., RDS

A Closer Look At AWS Kinesis

Collection of features that simplify dealing with data streams (continuous high frequency data ingestion)

Kinesis Data Streams

Scalable service which captures incoming streaming data
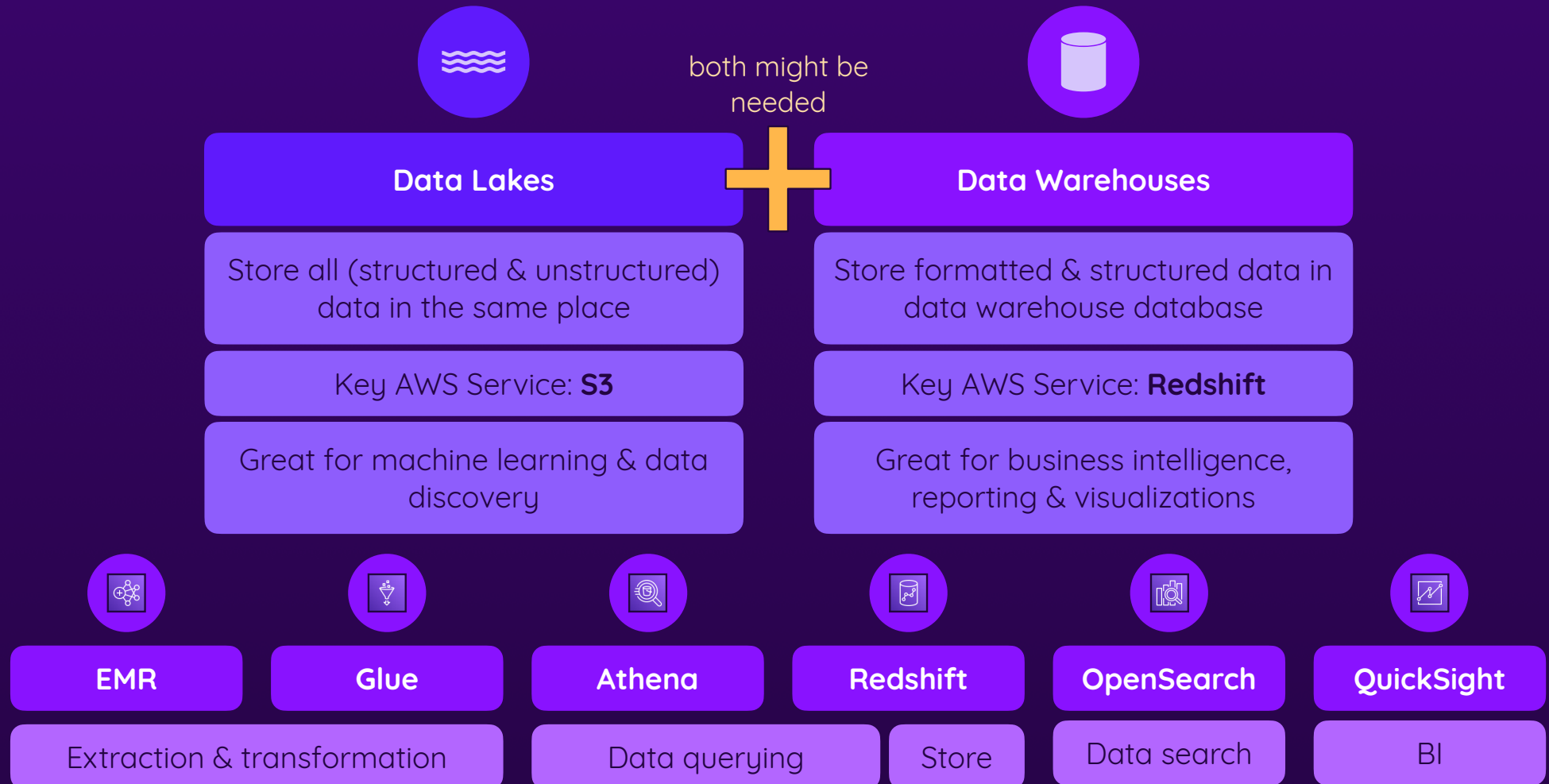
Kinesis Firehose

Loads data into other AWS services (e.g., S3)

Kinesis Data Analytics

Perform real-time data stream analytics

Storing Data: Data Lakes & Warehouses

both might be needed

Data Lakes

Store all (structured & unstructured) data in the same place

Key AWS Service: **S3**

Great for machine learning & data discovery

Data Warehouses

Store formatted & structured data in data warehouse database

Key AWS Service: **Redshift**

Great for business intelligence, reporting & visualizations

EMR	Glue	Athena	Redshift	OpenSearch	QuickSight
Extraction & transformation		Data querying	Store	Data search	BI

A Data Warehouse Solution: Redshift

A Scalable & Flexible Data Warehouse
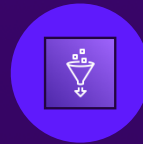
A data warehouse database → SQL-based database, optimized for analytics usage

Store data & run analytics queries against data

Redshift can also be used to query data stored in other data sources (e.g., S3)

RDS: Transactional data / daily operations
Redshift: Analytics data / data warehouse

Extracting & Transforming Data with Glue

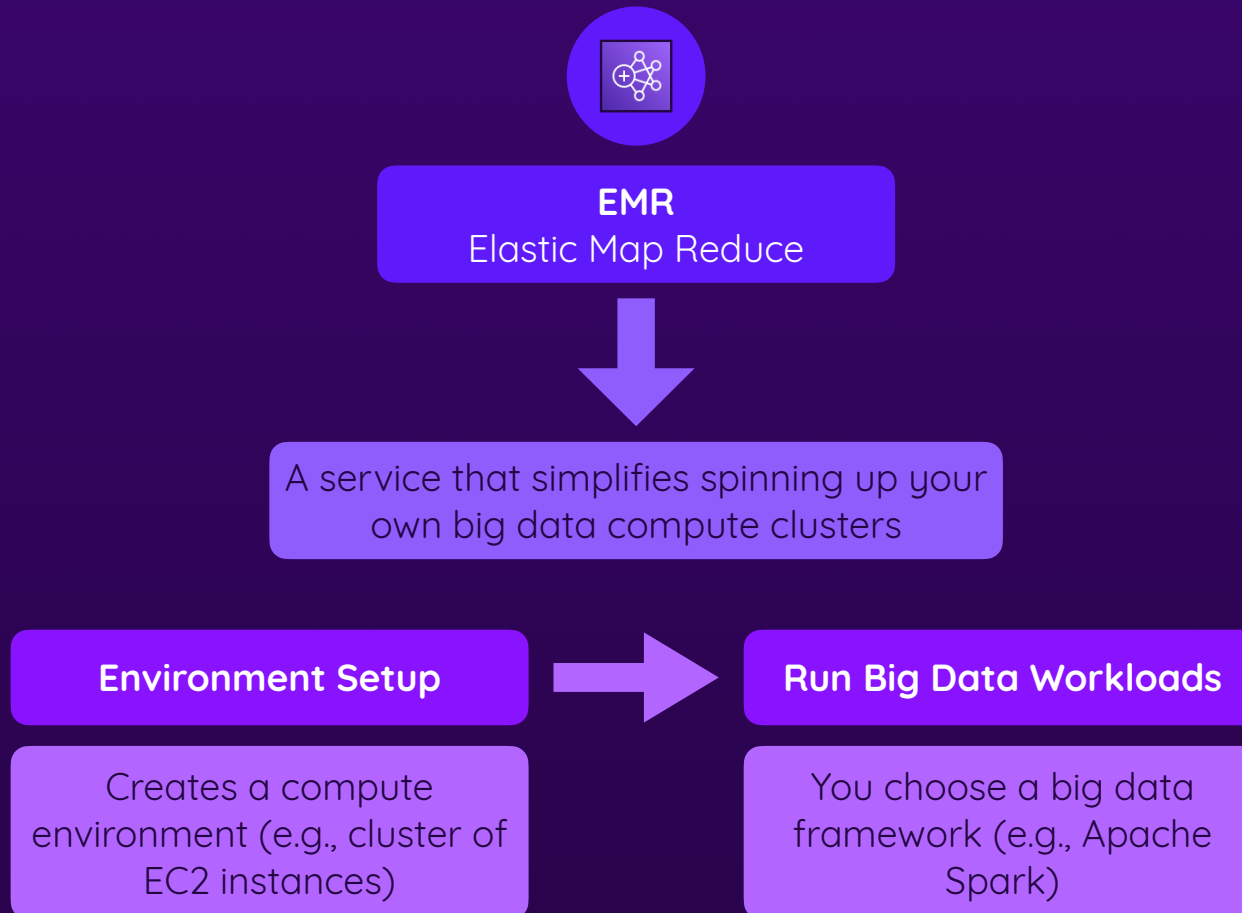**Serverless, managed ETL service**

ETL: Extract, Transform, Load

Simplifies the process of crawling, parsing & transforming raw data

Main features: Data schema creation, data transformation jobs, visual job editor

Self-managed Big Data Computation: EMR

EMR
Elastic Map Reduce

A service that simplifies spinning up your own big data compute clusters

Environment Setup

Creates a compute environment (e.g., cluster of EC2 instances)

Run Big Data Workloads

You choose a big data framework (e.g., Apache Spark)

Analyzing Data with Athena & QuickSight

Athena

Query data in S3 (via SQL queries)

No data movement to databases required

Standard SQL commands supported

Other data sources are also supported (e.g., CloudWatch logs, DynamoDB)

QuickSight

A business intelligence service

Build charts, reports & dashboards

Perform various analyses (average, sums etc.)

Searching & Visualizing Data

OpenSearch

Managed search service for searching & analyzing data

Running OpenSearch server ("domain") can be used to connect & search or analyze data

Grafana

Managed Grafana service (helps with visualizing data)

Simplifies the creation of live, interactive data visualizations

CloudSearch

Not primarily focused on data analysis

Instead: Managed service that simplifies the creation & management of website or application search solutions

Summary

Utilizing Data: A Complex Problem

Data must be ingested, transformed, stored & analyzed

Data ingestion can be tricky because of frequency / size

Transformation & extraction tasks require efficient compute

Different analytics tasks need different tools (search vs visuals)

Data Ingestion & Storage

Kineses helps with ingesting high-frequency streaming data

Data is buffered and (typically) forwarded to other services

Data is often stored on **S3**, following a "Data Lake" approach

Data warehouses can be built with **Redshift**

Transformation & Analytics

Manual big data workloads can be executed with **EMR**

Glue is a managed, serverless ETL solution

Query raw data (e.g., in S3) with SQL & **Athena**

Perform BI with **QuickSight**, search & visual with **OpenSearch** & **Grafana**