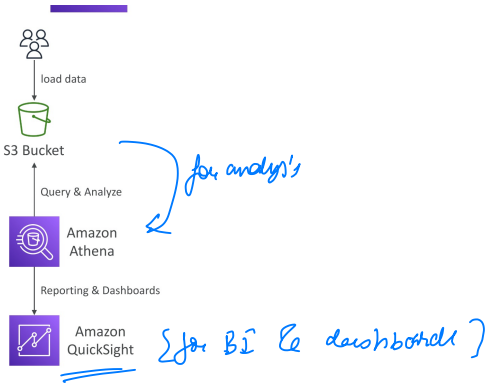
§ Data & Analytics

Amazon Athena

- Serverless query service to analyze data stored in Amazon S3
- Uses standard SQL language to query the files (built on Presto)
- Supports CSV, JSON, ORC, Avro, and Parquet
- Pricing: $5.00 per TB of data scanned
- Commonly used with Amazon Quicksight for reporting/dashboards

- Use cases: Business intelligence / analytics / reporting, analyze & query VPC Flow Logs, ELB Logs, CloudTrail trails, etc...
- Exam Tip: analyze data in S3 using serverless SQL, use Athena

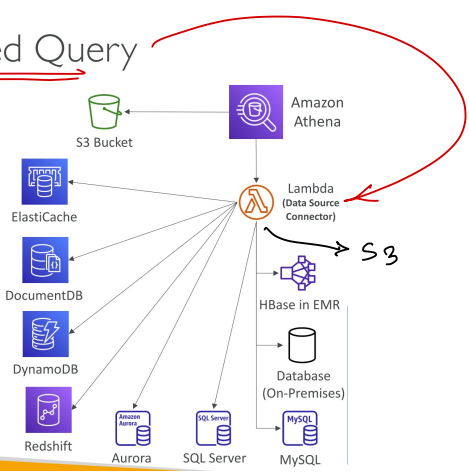


Performance improvements

- Use columnar data for cost-savings (less scan) } since we pay for data scanned }
  - Apache Parquet or ORC is recommended
  - Huge performance improvement
  - Use Glue to convert your data to Parquet or ORC } Glue for converting as an ETL job }
- Compress data for smaller retrievals (bzip2, gzip, lz4, snappy, zlip, zstd…)
- Partition datasets in S3 for easy querying on virtual columns
  - s3://yourBucket/pathToTable
    /<PARTITION_COLUMN_NAME>=<VALUE>
    /<PARTITION_COLUMN_NAME>=<VALUE>
    /<PARTITION_COLUMN_NAME>=<VALUE>
    /etc…
  - Example: s3://athena-examples/flight/parquet/year=1991/month=1/day=1/
- Use larger files (> 128 MB) to minimize overhead

Amazon Athena – Federated Query

- Allows you to run SQL queries across data stored in relational, non-relational, object, and custom data sources (AWS or on-premises)
- Uses Data Source Connectors that run on AWS Lambda to run Federated Queries (e.g., CloudWatch Logs, DynamoDB, RDS, …)
- Store the results back in Amazon S3



② Redshift

Analytics + data warehousing.

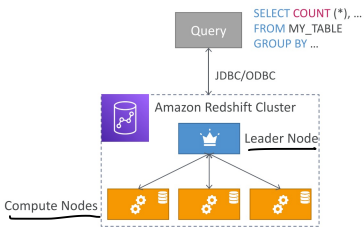
Redshift Overview

- Redshift is based on PostgreSQL, but it's not used for OLTP
- It's OLAP – online analytical processing (analytics and data warehousing)
- 10x better performance than other data warehouses, scale to PBs of data
- Columnar storage of data (instead of row based) & parallel query engine
- Pay as you go based on the instances provisioned
- Has a SQL interface for performing the queries
- BI tools such as Amazon Quicksight or Tableau integrate with it
- vs Athena: faster queries / joins / aggregations thanks to indexes

→ on top of EC2 instances.

→ column as data storage

Redshift Cluster



- Leader node: for query planning, results aggregation
- Compute node: for performing the queries, send results to leader
- You provision the node size in advance
- You can used Reserved Instances for cost savings

EC2 optimization.

Redshift – Snapshots & DR

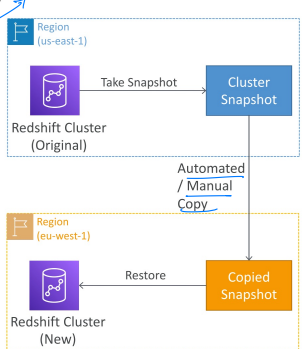
for DR

- Redshift has "Multi-AZ" mode for some clusters

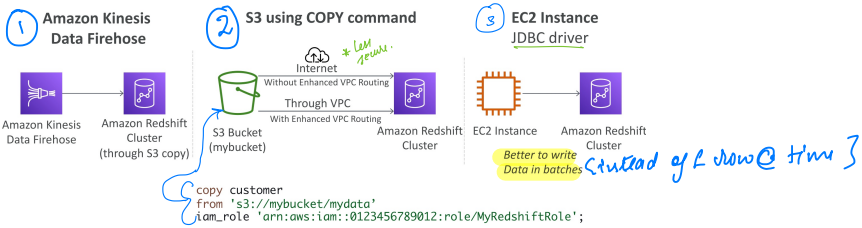
But it might not AZ

- Snapshots are point-in-time backups of a cluster, stored internally in S3.
- Snapshots are incremental (only what has changed is saved)
- You can restore a snapshot into a new cluster
1. Automated: every 8 hours, every 5 GB, or on a schedule. Set retention between 1 to 35 days
2. Manual: snapshot is retained until you delete it

- You can configure Amazon Redshift to automatically copy snapshots (automated or manual) of a cluster to another AWS Region

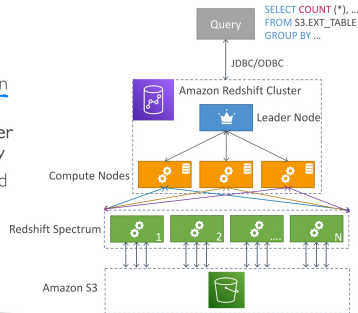


Loading data into Redshift:
Large inserts are MUCH better



Redshift Spectrum

- Query data that is already in S3 without loading it
- Must have a Redshift cluster available to start the query
- The query is then submitted to thousands of Redshift Spectrum nodes



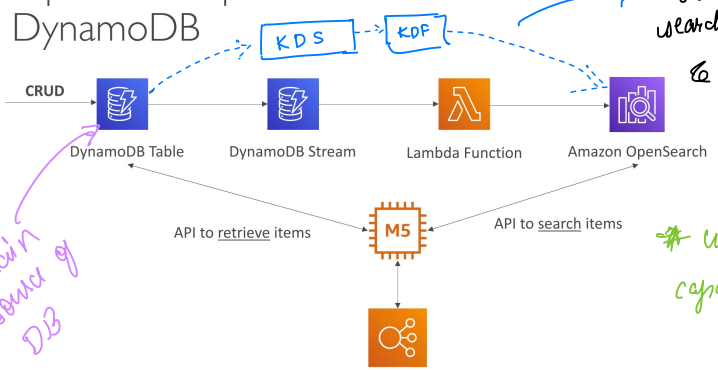
(3) Amazon Open search { elastic search }

Amazon OpenSearch Service

Analytic queries + search

- *Amazon OpenSearch is successor to Amazon ElasticSearch*
- In DynamoDB, queries only exist by primary key or indexes…
- With OpenSearch, you can search any field, even partially matches
- It's common to use OpenSearch as a complement to another database
- Two modes: (1) managed cluster or (2) serverless cluster
- Does *not* natively support SQL (can be enabled via a plugin) { has its own query language }
- Ingestion from Kinesis Data Firehose, AWS IoT, and CloudWatch Logs
- Security through Cognito & IAM, KMS encryption, TLS
- Comes with OpenSearch Dashboards (visualization)

OpenSearch patterns DynamoDB

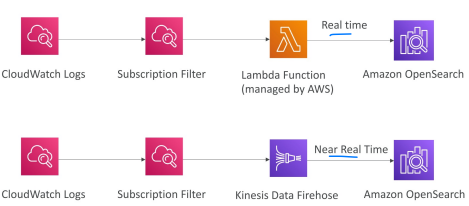


main source of DB

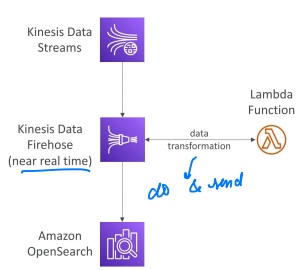
Application is now able to do a search with the item name to retrieve the item ID, & call dynamo DB to retrieve that item.

* unlocking the search capability in dynamo DB

OpenSearch patterns CloudWatch Logs



OpenSearch patterns Kinesis Data Streams & Kinesis Data Firehose



4 Amazon EMR

Amazon EMR

- EMR stands for "Elastic MapReduce"
- EMR helps creating Hadoop clusters (Big Data) to analyze and process vast amount of data
- The clusters can be made of hundreds of EC2 instances
- EMR comes bundled with Apache Spark, HBase, Presto, Flink…
- EMR takes care of all the provisioning and configuration
- Auto-scaling and integrated with Spot instances

- Use cases: data processing, machine learning, web indexing, big data…

Amazon EMR – Node types & purchasing

- **Master Node:** Manage the cluster, coordinate, manage health – long running
- **Core Node:** Run tasks and store data – long running
- **Task Node (optional):** Just to run tasks – usually Spot (task spot instances)
- Purchasing options:
  - On-demand: reliable, predictable, won't be terminated
  - Reserved (min 1 year): cost savings (EMR will automatically use if available)
  - Spot Instances: cheaper, can be terminated, less reliable

- Can have long-running cluster, or transient (temporary) cluster

5 Amazon Quicksight

Amazon QuickSight

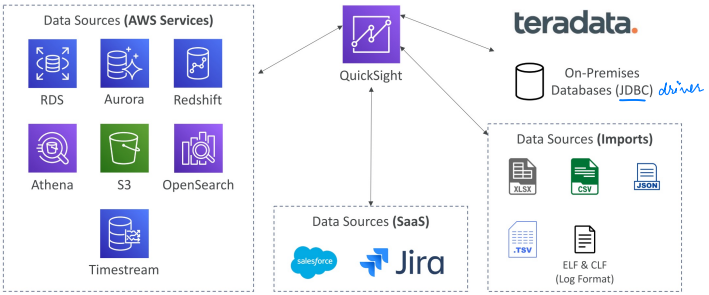
- Serverless machine learning-powered business intelligence service to create interactive dashboards
- Fast, automatically scalable, embeddable, with per-session pricing
- Use cases:
  - Business analytics
  - Building visualizations
  - Perform ad-hoc analysis
  - Get business insights using data
- Integrated with RDS, Aurora, Athena, Redshift, S3…
- In-memory computation using SPICE engine if data is imported into QuickSight
- Enterprise edition: Possibility to setup Column-Level security (CLS)

https://aws.amazon.com/quicksight/

only in enterprise to display user specific columns.

in-memory computation using spice engine for only data loaded in Quicksight not from the data which is connected to QS.

QuickSight Integrations



QuickSight – Dashboard & Analysis

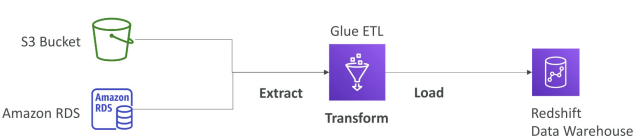
- Define Users (standard versions) and Groups (enterprise version)
  - These users & groups only exist within QuickSight, not IAM !!
- A *dashboard*…
  - is a read-only snapshot of an analysis that you can share
  - preserves the configuration of the analysis (filtering, parameters, controls, sort)

- You can share the analysis or the dashboard with Users or Groups
- To share a dashboard, you must first publish it
- Users who see the dashboard can also see the underlying data

AWS Glue

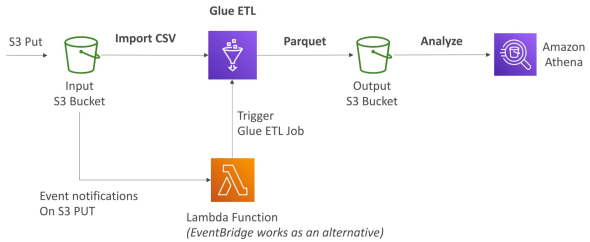
AWS Glue

- Managed extract, transform, and load (ETL) service
- Useful to prepare and transform data for analytics
- Fully serverless service



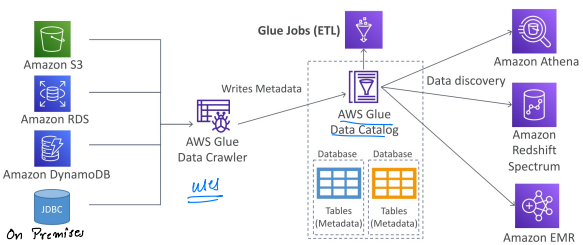
AWS Glue – Convert data into Parquet format

useful to convert to columnar data for cost saving in athena.



Glue Data Catalog: catalog of datasets

The metadata is used by other services to do data discovery or schema discovery.



Glue things

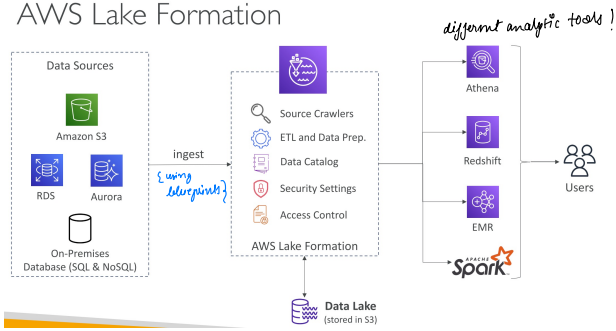
1. • **Glue Job Bookmarks:** prevent re-processing old data
2. • **Glue Elastic Views:**
   - Combine and replicate data across multiple data stores using SQL { create a view of RDS, Aurora or S3 }
   - No custom code, Glue monitors for changes in the source data, serverless*
   - Leverages a "virtual table" (materialized view) { across multiple data stores }
3. • **Glue DataBrew:** clean and normalize data using pre-built transformation
4. • **Glue Studio:** new GUI to create, run and monitor ETL jobs in Glue
5. • **Glue Streaming ETL** (built on Apache Spark Structured Streaming): compatible with Kinesis Data Streaming, Kafka, MSK (managed Kafka) { instead batch jobs run it as streaming jobs }

(7) AWS Lake formation

AWS Lake Formation

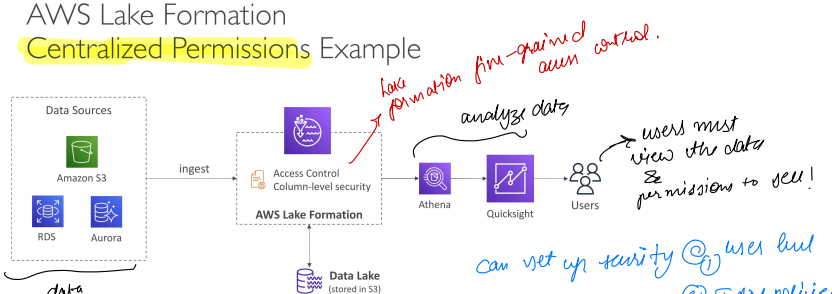
- Data lake = central place to have all your data for analytics purposes
- Fully managed service that makes it easy to setup a data lake in days
- Discover, cleanse, transform, and ingest data into your Data Lake
- It automates many complex manual steps (collecting, cleansing, moving, cataloging data, …) and de-duplicate (using ML Transforms)
- Combine structured and unstructured data in the data lake
- Out-of-the-box source blueprints: S3, RDS, Relational & NoSQL DB… { blueprints → to help migrate data from various data sources to data lake }
- Fine-grained Access Control for your applications (row and column-level)
- Built on top of AWS Glue

AWS Lake Formation



Q Why Lake formation ?

AWS Lake Formation
Centralized Permissions Example



can set up security @ (1) user level
(2) IAM policies
(3) Quicksight
(4) RDS or S3

But which one?

major advantage!

Therefore the data ingested into Lake will sit in a central S3 bucket but from within the lake formation all the access controls are managed. (for row & column level sec)

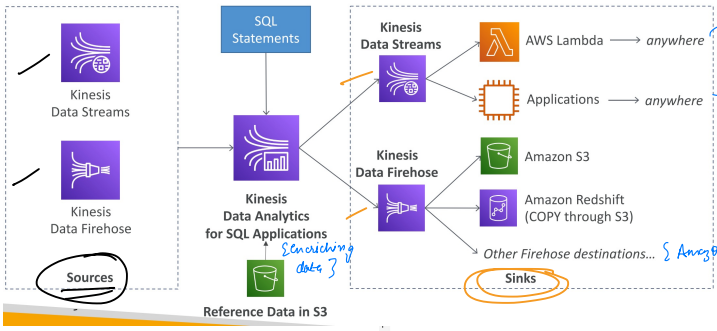
So any service connected to lake formation will have the permission to say view the data & here all the security can be controlled from one centralized location i.e. lake formation

② AWS Kinesis Data Analysis

① Kinesis Data Analytics for SQL applications {legacy}

creating a SQL application

② go to SQL analytics to do real time query. for the application



do real time processing

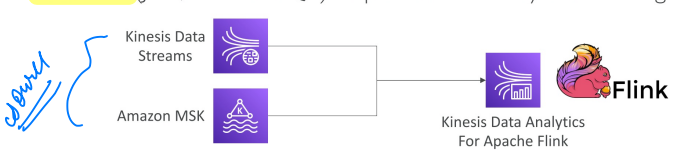
Kinesis Data Analytics (SQL application)

- Real-time analytics on Kinesis Data Streams & Firehose using SQL
- Add reference data from Amazon S3 to enrich streaming data
- Fully managed, no servers to provision
- Automatic scaling
- Pay for actual consumption rate
- Output:
  - Kinesis Data Streams: create streams out of the real-time analytics queries
  - Kinesis Data Firehose: send analytics query results to destinations
- Use cases:
  - Time-series analytics
  - Real-time dashboards
  - Real-time metrics

② Kinesis Data Analytics for Apache Flink

provides a way to create an apache flink application.

- Use Flink (Java, Scala or SQL) to process and analyze streaming data



Flink or special applications to be written as code & AWS allows us to run these Flink application on a cluster i.e. dedicated to it on KDA

- Run any Apache Flink application on a managed cluster on AWS
  - provisioning compute resources, parallel computation, automatic scaling
  - application backups (implemented as checkpoints and snapshots)
  - Use any Apache Flink programming features
  - Flink does not read from Firehose (use Kinesis Analytics for SQL instead)

→ Flink is more powerful than standard SQL {Adv. SQL queries.

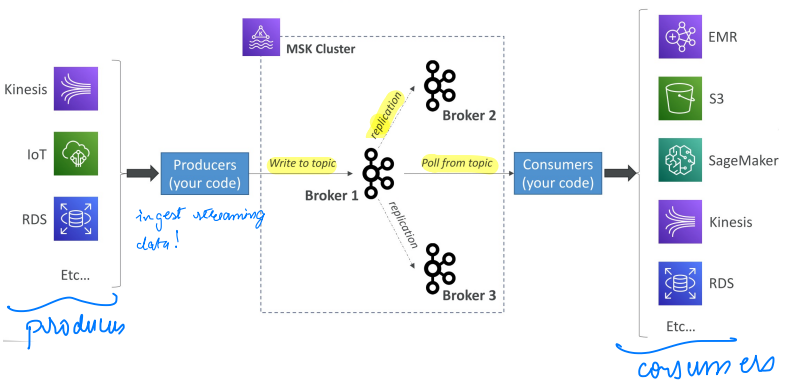
→ for just one of analysis rather than creating an application create a studio notebook {which runs on top of Flink}

(9) Amazon Managed Streaming for apache Kafka [MSK]

Amazon Managed Streaming for Apache Kafka (Amazon MSK)

- Alternative to Amazon Kinesis
- Fully managed Apache Kafka on AWS
  - Allow you to create, update, delete clusters
  - MSK creates & manages Kafka brokers nodes & Zookeeper nodes for you
  - Deploy the MSK cluster in your VPC, multi-AZ (up to 3 for HA)
  - Automatic recovery from common Apache Kafka failures
  - Data is stored on EBS volumes for as long as you want
- MSK Serverless
  - Run Apache Kafka on MSK without managing the capacity
  - MSK automatically provisions resources and scales compute & storage

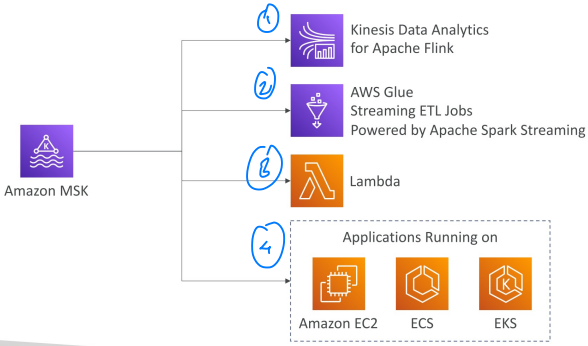
Apache Kafka at a high level



Kinesis Data Streams vs. Amazon MSK

	Kinesis Data Streams	Amazon MSK
(1) message size limit	• 1 MB message size limit	• 1MB default, configure for higher (ex: 10MB)
(2) method	• Data Streams with Shards	• Kafka Topics with Partitions
	• Shard Splitting & Merging	• Can only add partitions to a topic
(3) Encryption	• TLS In-flight encryption	• PLAINTEXT or TLS In-flight Encryption
	• KMS at-rest encryption	• KMS at-rest encryption
		• Keep data for as long as we want (pay for underlying EBS volumes too)

Amazon MSK Consumers



Big data ingestion Pipeline