BIG Data

>Redshift Databases

- Amazon Redshift
    - Scalable data warehouse solution
        - Redshift is a data-warehousing service for storing and analyzing structured data from multiple sources, including relational databases and S3. Redshift can store much more data than RDS, up to 16 PB!
            - Handles exabyte-scale data
    - Data consolidation
        - When you need to consolidate multiple data sources for reporting
    - Relational Databases
        - When you want to run a database that doesn't require real-time transaction processing (insert, update, and delete)

Amazon Redshift uses SQL to analyze structured and semi-structured data across data warehouses, operational databases, and data lakes, using AWS-designed hardware and machine learning to deliver the best price performance at any scale.
https://www.youtube.com/watch?v=IWwFJV_9PoE
Gain up to 3x better price performance than other cloud data warehouses out of the box. Automatically optimize design to improve query speed for complex and critical workloads. Pay only for what you use.

The 3 V's of Big Data

- Volume
    - Ranges from terabytes of petabytes of data
- Variety
    - Includes data from a wide range of sources and formats
- Velocity
    - Business require speed. Data needs to be collected, stored, processed and analyzed within a short period of time.

>Processing Data with EMR (Elastic MapReduce)

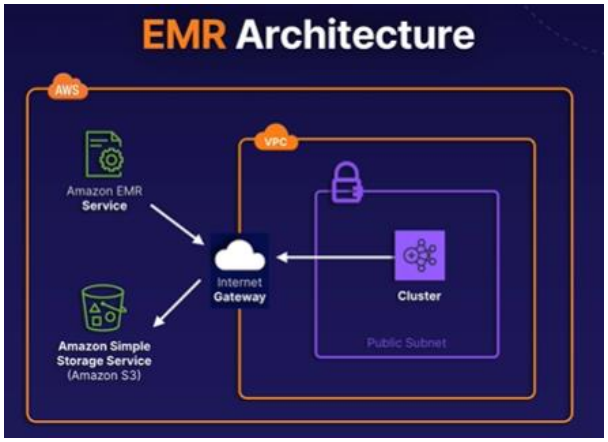
Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

Elastic MapReduce (EMR) lets you analyze enormous amounts of data stored in the cloud. EMR supports the Apache Hadoop, Apache Spark, HBase, Presto, and Flink big data platforms. For more information, visit

In EMR, there are 3 options how it works

1. Upload – Upload your data and processing application to S3
2. Create – Configure and create your cluster by specifying data inputs, outputs, cluster, size, security settings, etc.

3. Monitor – Monitor the health and progress of your cluster. Retrieve the output in S3



>Streaming Data with Kinesis

Kinesis allows you to ingest, process, and analyze real-time streaming data. Which there are 2 versions/types of Kinesis

Allows you to analyze data and video streams in real time

It's useful for analyzing large amounts of streaming data including access logs, video, audio, and telemetry. For more information, visit https://aws.amazon.com/kinesis/.

1. Data Streams
    a. Real-time streaming for ingesting data (Purpose)
    b. Real time (Speed)
    c. You're responsible for creating the consumer and scaling the stream (Difficulty)

Data streams – Real-time data capture > Ingest and store data streams from hundreds of thousands of data sources:

- Log and event data collection
- IoT device data capture
- Mobile data collection
- Gaming Data Feed

2. Data Firehose
    a. Data transfer tool to get information to S3, Redshift, Elasticsearch, or Splunk (Purpose)
    b. Near real time (60 secs) (Speed)
    c. Plug and play with AWS arch. (Difficulty)

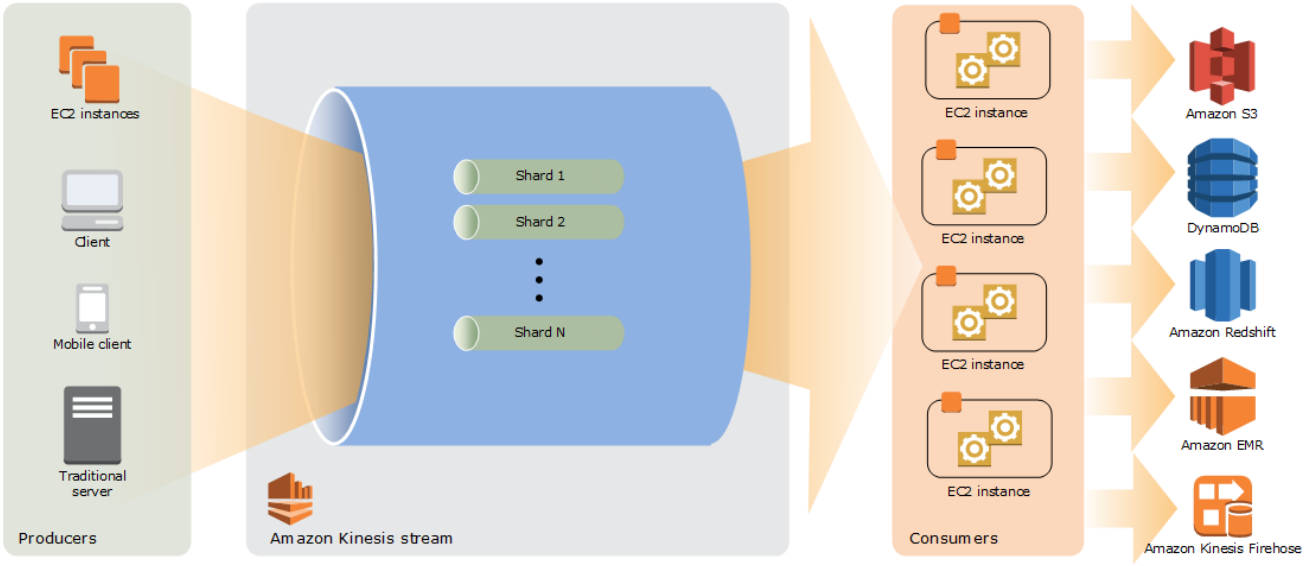
Firehose – Load real-time data > Load streaming data into data lakes, data stores, and analytics tools for..

Anything that talks about data streams that automatically scales… Think of Data Firehouse

- Log and event analytics
- IoT data analytics
- Clickstream analytics
- Security monitoring

How does Kinesis work?

- Collect and store data streams
  - Collecting GB of data per second
  - And storing the data streams are stored in shards in your stream temp
- Process and deliver data streams
  - Prepare and load real-time data streams into data stores and analytics tools
- Analyze Streaming Data
  - Get actionable insights from streaming data in real time



Shards can only handle a certain amount of data, so you need to scale how many you need to have. You also need a consumer which is something that is going to take the data in and process the content and put it in the endpoint you selected which can be anything.

AWS Kinesis stream was the first one, the 2nd one AWS made was Data Firehose

Firehose is much simpler. Data Firehose handles the scaling for you, building out that consumer so basically you don't have to write out that code.

- You can use an Amazon Kinesis Data Analytics application to process and analyze data in a Kinesis stream using SQL, Java, or Scala.

Case study…

When we're looking for a message broker, which do we pick?

Remember SQS is simpler, but doesn't offer real-time. Kinesis is a bit more complicated but it does have real-time.

>Amazon Athena and AWS Glue

- Athena
    - Query service for Amazon S3
    - Amazon Athena is not a database engine.
    - Athena lets you use SQL queries to find data stored in S3. If you have data stored in CSV, JSON, ORC, Avro, or Parquet format, simply upload it to S3 and use Athena to query it. Athena is serverless, so there's no need to provision your own database or import your data into it. For more information, visit https://aws.amazon.com/athena/.
    - Pay per query

Athena can take all the data that Glue has and structure it and run queries on it without having to load it into that database

- From there it would be very easy to use something like QuickSight, which at a really high level is effectively Amazon's version of Tableau. Which can help visualize this data

- Glue
    - Glue prepares your data for analytics
    - Data can live in a variety of places on AWS. AWS Glue can discover, clean, and bring this data together in one place for analysis using the Apache Spark big data framework. It can extract and analyze data from S3 objects and relational databases such as MySQL, Oracle, and Microsoft SQL Server. For more information, visit https://aws.amazon.com/glue/.
    - Better understand the data
    - Serverless data integration
    - Perform ETL workloads without managing underlying servers

For Amazon Redshift spectrum, you don't have to know this for the exam. But this allows you to use Redshift without having to load all that data into the redshift database itself

So, you structure with Glue, and query it with Athena and Quicksight to give you a dashboard with all of your needed insights

Scenario

If you're ever faced with a scenario that's looking for a serverless SQL solution. Think Athena

>Visualizing Data with QuickSight

QuickSight is a fully managed BI data visualization service. It creates dashboards and share them within your company.

Think about the stock market, that is just like QuickSight. We went from spreadsheets to something that looks better to read and makes more sense.

So above, Glue and Athena… Does the heavy lifting for us. And Quicksight gives us something to look at as it practically is a data visualization tool

Any questions about sharing your data, interpreting that data, or anything related to business intelligence look for any answer that includes QuickSight

- You see BI think QuickSight

>Analyzing Data with ElasticSearch

Turns out that ElasticSearch is not an AWS product, this is an analyzing tool/search engine.

Amazon ElasticSearch is a full managed version of that open-source tool. This tool allows us to easily search over our data. This will be primarily used in an ELK stack (**E**lasticsearch, **L**ogstash, **K**ibana)

We are collecting information (Logs, message, metrics, etc) > Bringing it in and Elasticsearch analyzes this > Using this tool you can look for problems, which look for real-time problems as they happen. Very similar to CloudWatch logs

If given a scenario on the exam it talks about creating a 3rd party logging solution… think of any below

Elasticsearch = ELK = Logs

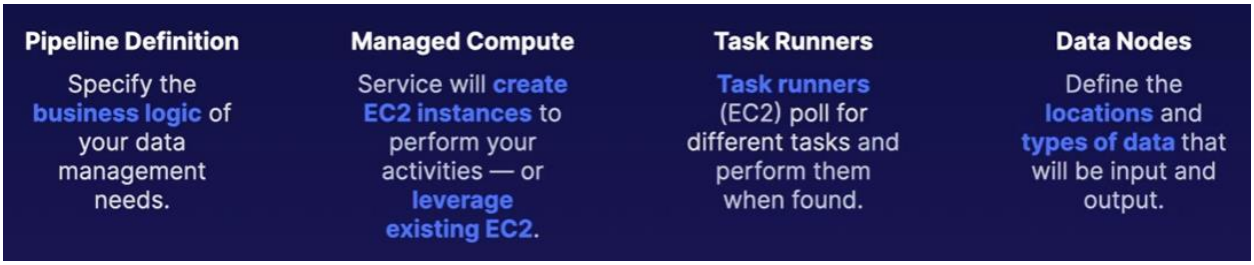
>Moving Transformed Data Using AWS Data Pipeline

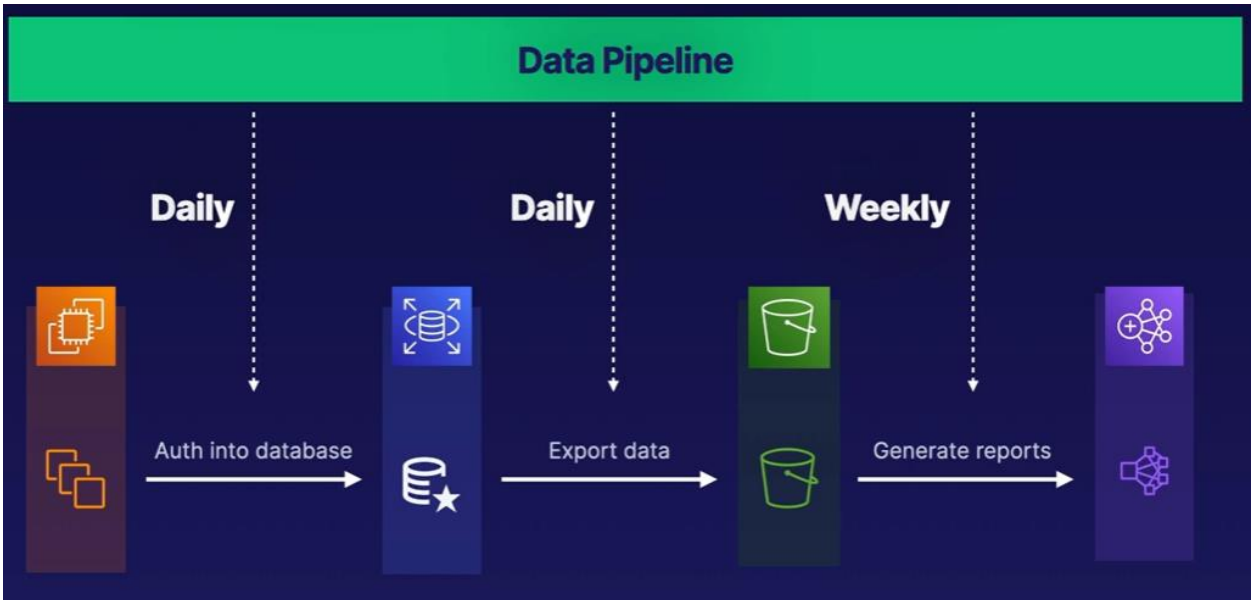
AWS Data Pipeline

- Data Pipeline is a managed extract, transform, load (ETL) service for automating movement and transformation of your data

Overview

- Data-driven workflows.
- Define your parameters for data transformations. Which this enforces your chosen logic.
- High available and distributed infrastructure. Also fault tolerant
- Automatically retires failed activities. Configure notifications via Amazon SNS
- Integrates easily with Amazon DynomoDB, RDS, Redshift, S3





>Implementing Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Amazon MSK is a streaming service for apache kafka. Fully managed service for running data streaming applications that leverage Apache Kafka. Leverage Kafka data-plane operations for producing and consuming streaming data.

Important components and concepts

- Resiliency in Amazon MSK
    - Automatic Recovery
        - Automatic detection and recovery from common failure scenarios.
    - Detection
        - Detected broker failures result in migration or replacement of unhealthy nodes
    - Reduce Data
        - Tries to reuse storage from older brokers during failures to reduce data needing replication
    - After recovery
        - After successful recovery, producers and consumers apps continue to communicate with the same broker IP as before.

Good to know

MSK Serverless

- Cluster type within Amazon MSK offering serverless cluster management. Automatically provisions and scales.

Fully compatible

- MSK Serverless is fully compatible with Apache Kafka.

MSK Connect

- Allows developers to easily stream data to and from Apache Kafka clusters

Security and Logging

1. Integration with Amazon KMS for SSE requirements
2. Encryption at rest by default
3. TLS 1.2 for encryption in transit between brokers in clusters

4. Deliver broker logs to Amazon CloudWatch, Amazon S3, and Amazon Kinesis, and Amazon Kinesis Data Firehose
5. By default, Amazon MSK will collect metrics and sent to CloudWatch, and if desired set up some type of alarm
6.

Exam Tips

1. Redshift
    a. 16 PB of data per cluster (PB is larger than a TB)
        i. 1 PB = 1,024 TB = 1,048,576 GB
    b. Great for BI applications, so only use Redshift when it comes to BI apps
    c. Not standard, don't use redshift in place of RDS
    d. Relational database
    e. Does not support mult-AZ deployments
    f. Backups are kept for 1 day by default, but this can be raised up to 35 days at most
2. EMR
    a. Managed fleet of EC2
    b. Open source cluster
    c. Rules apply
        i. Use RIs (Reserve instances) and Spot Instances to reduce your cost
    d. High Level
        i. EMR is used to process and move data
    e. VPC
        i. The architecture lives in VPC
    f. Valid case for this would be extract, transform, and load (ETL) jobs
3. Kinesis
    a. If you see anything "real-time" or processing or moving data
        i. Think Kinesis
    b. If you see anything about "near real-time"
        i. Think Data Firehose
    c. If you see anything real-time
        i. Think Data Streams
    d. Any questions based on streaming any sort of data, that is a dead giveaway to be looking for anything some forms of Kinesis
    e. Data analytics is the easiest way to process data going through kinesis using SQL
    f. Anything that talks about data streams that automatically scales… Think of Data Firehouse
    g. SQS and Kinesis can both act as queues. SQS is easier and simpler, and Kinesis is faster and can store data up to a year.
4. Athena and Glue
    a. Serverless
        i. Both solutions are fully managed serverless services
    b. Better Together
        i. While Athena can work by itself, Glue can design a schema for your data
    c. Knowing the 3,000-foot view of these services is good enough for this exam
    d. Athena lets you query data stored in S3
    e. You're not responsible for scaling Glue as AWS handles everything for you.
5. QuickSight
    a. Visualize
        i. Need to look at your data? Use QuickSight

   b. Knowing what services is good
   c. BI
      i. If this phrase comes up starting looking for QuickSight
6. Elasticsearch (OpenSearch Service now)
   a. Open source
      i. Fully managed service
   b. Know these services is good
   c. Elasticsearch will most likely be a distractor on the exam
      i. Remember anything with 3rd party logging think either ELK
7. AWS Data Pipeline
   a. Managed ETL workflows that automates movements and transformations of your data
   b. Data Driven workflows to create dependencies between task and activities
   c. Storage integrations such as DynamoDB, RDS, Redshift, S3
   d. Compute integrations with EC2 and EMR for managed compute needs
   e. Anything related to managedETL services, and automatic retries for data-driven workflows… Think of Data Pipeline
8. Streaming for Apache Kafka (Amazon MSK)
   a. Apache Kafka
      i. Fully managed AWS service for running and building Apache Kafka data streaming applications
   b. Control Plane
      i. Creation, updating, and deletion of clusters
   c. Data Plane
      i. Leverage the same Apache Kafka data-plane operations for producing and consuming data
   d. Automatic recoveries
      i. Service detects and automatically mitigates most of the common failures
   e. Logging
      i. Push broker logs to CloudWatch, S3, or Kinesis Data Firehose. API calls are logged to CloudTrail
   f.