AWS BIG DATA SPECIALITY

By QuickTechie.com

QuickTechie

Contents

Introduction

AWS Certified BigData – Specialty (BDS-C00) examination is really one of the tough certification exam for sure. Even many of the learners are unable to complete this 3 hrs certification exam, because they are not well prepared and they have to re-appear for the exam. The learners which comes to HadoopExam, always complain that they have landed to wrong website for certification preparation and they wanted to learn in depth before appearing the certification exam, so that they can clear this certification exam with full confidence and can work on the same technology with the full confidence. HadoopExam assure you that, you would learn as well clear the certification exam with full confidence, we highly recommend you would go through entire material provide to you. Learners who are already working the Big Data and Data Analytics fields are facing various problems to clear this certification exam. Even people with 10+ yrs of experience in the Software Industry are facing various challenges with regards to this BigData Specialty Certification.

As you know HadoopExam is already in the fields of BigData, Analytics and Cloud Computing learning field since more than 7 years, and whatever experience we have gained during this time, we will also try to provide using this book. If you are already having a BigData background then it would certainly help for this certification, because many of the concepts remain same only the way solution should be provided is different because it depends which AWS service is being used for solving a particular problem. If you are already done some AWS Certification like AWS solution architect then it would make you're your learning curve faster, however if you have not done any AWS certification till now then also it is fine, because we would be trying to cover end to end solution in this book, we may be dividing this book in multiple part as well, if needed.

Yes, it is required that you understand the basics of various AWS services to clear BigData Specialty certification. There are 100's of AWS services and you don't need to learn all of these, we would cover whatever is required.

Concept you need to learn

- Service-to-service integration
- IOT concepts
- Each Service Security use case and Anti Use Cases
- Important White Paper for BigData Analytics :

BigData Analytics	https://d1.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf?did=wp_card&trk=wp_card

Kinesis	https://aws.amazon.com/kinesis/whitepaper/
EMR	https://aws.amazon.com/blogs/big-data/best-practices-for-securing-amazon-emr/
YouTube Video	https://www.youtube.com/playlist?list=PLhr1KZpdzukdeX8mQ2qO73bg6UKQHYsHb
AWS BigData Blog	https://aws.amazon.com/blogs/big-data/
Medium Blog	https://medium.com/@simonleewm/my-path-to-aws-big-data-speciality-certification-4baff3a8150
AWS BigData Case Studies	https://aws.amazon.com/solutions/case-studies/?customer-references-cards.sort-by=item.additionalFields.publishedDate&customer-references-cards.sort-order=desc&awsf.customer-references-location=location%23apac&awsf.customer-references-segment=customer-segment%23enterprise
Kinesis Firehose Video	https://www.youtube.com/watch?v=8L3lLSPPxpY&feature=youtu.be&t=2088

Scenario Based Questions

If you see in this exam all questions are MCQ (Multiple choice question & answer with single or multiple response). Almost all the questions are based on the various scenarios and you need to understand each specific word to answer the question, if you don't have enough knowledge then certainly you would get confused with the option given, sometime you would feel all the answers are given seems to be correct and it is very difficult to find the correct answer, however if you have practiced well with all the questions provided by the HadoopExam and all the trainings provided then this exam would be quite easy for you. We want you to go through the concepts in detail because even if you are having good experience working with these technologies, you may have not come across the situation which is being asked in the exam.

Exam Blueprint

If you look at the exam blue print, you may not be able to conclude, what exactly this exam would be testing they AWS given only overview and domain in which they would evaluating the candidate, they don't even mention which AWS service they would be asking in such a costly exam. Anyway, if you follow the material provided by HadoopExam then you would get to know all the required detail, because we are trying as much as possible to pass on the information needed. As of now currently below 6 domains have been given for this certification, which we would be discussing in detail later on through services.

- **Collecting (17%):** It focuses, how you collect the data using AWS webservices among the AWS service itself or from the external sources. For example, collection data from Mobile Application to your S3 bucket or collecting data from S3 bucket to one of the Redshift cluster or transferring data between the redshift cluster with the security. Collecting streaming data using Kinesis Firehose or Kinesis Streaming or using Apache Spark etc.
- **Storage (17%):** There are various possible storage, for example for structured data you can use AWS RDS, AWS DynamoDB, HBase etc. Similarly, for unstructured data we need to use S3 etc.

- **Processing (17%):** You need to identify that how you can securely process the data which is already stored or on the fly data processing while getting streaming data.
- **Analysis(17%):** You may be asked to use various Machine Learning Algorithm to analyze the data, we have not seen that they are asking you write proper queries to select the data etc. which is usually the role of the Data analysts, AWS is testing their services knowledge and not actually testing your analytical skills on which you would be working, if you see Cloudera Data Analytics (CCA159) certification, where they are testing not only their platform but also your data analytical skills for writing the queries etc.
- **Visualization (12%):** This is where your data analytics skills would be tested for selecting correct chart and how to utilize various available tool from AWS to generate analytics report. You need to understand various plots like bar chart, heat map, box plot etc.
- **Data Security(20%):** In many questions AWS would add the security aspects like Authentication, Authorization, Single Sign On, IAM Roles, Active Directory etc. And some regulatory requirement.

We can't say that for each individual domain there would be 12 questions or some other number, because in the question one or more domain are mixed to frame the scenario, which makes things more complicated and confusing. We should really appreciate the question designer; questions are really framed smartly.

Hands on Exercise

In this book we would not be considering the Hands-On exercises, rather we would cover the concepts because they are being tested in the real exam, because our learners have quite a good hand on with the AWS services then also they are not able to clear this certification exam.

Topics which are tested in the real exam

- Important Services which are mostly asked in the exam are
  - Kinesis
  - Redshift
  - EMR
  - S3
- AWS S3 (More than 80% questions would use S3 in the options or in the question, hence you should be well aware about the S3)
- Redshift Cluster is one of the important topics, like collecting, loading, transferring data securely in parallel for optimization.
- 15-20% questions would be talking about Kinesis Stream & Firehose
- Networking & Security: VPC, Availability Zones, STS (Security Token Service)
- Encryption: Encrypting data at rest as well as while in transit.
  - Client-Side Encryption
  - Server-Side Encryption
  - Hardware Security Module (HSM)
- Database based services like

  - DynamoDB & Accelerator
  - RDS (Relational Database Service)
  - Redshift Cluster
- Machine Learning
  - Understanding of Classification
  - Labeling data Manually as well as using another Machine Learning Algorithm
  - Understanding of NPL
  - Sage Maker
  - GPU Limits
- EMR Services & In-memory processing
  - Understand the Apache Hive
  - Understanding of Apache HBase
  - Apache Spark
  - Apache Spark Streaming solution
- Serverless Architecture
  - Lambdas
- ETL, ELT & Data Pipeline
  - Data pipeline
  - Data Migration Service
  - Direct Connect understanding
- Visualization
  - QuickSight
  - Various Charts
    - Bubble Chart
    - Heat Map
- Athena, Glue, Pipeline etc.

AWS Services

Kinesis Streams

- KPL, KCL, Kinesis Agents, Kinesis API, Connector Library
- Sharding, Retention Period, Autoscaling
- SQS vs Stream
- Batching, Aggregation, Collection
- KCL Checkpointing
- Monitoring & Exceptions

Kinesis Firehose

- Integrating Kinesis Firehose with S3/Redshift/ElasticSearch
- Kinesis Agents

- Monitoring & Exceptions

IOT

- Overview of IOT

Data pipeline

- Integration with AWS services

S3 & Glacier

- Policies
- Various storage standards

DynamoDB

- Integration with AWS services
- Choice of Partition/SortKey, LSI/GSI
- Understanding of Partition Size
- Throttling reads/writes, and mitigations
- DynamoDB streams

Lambda:

- Integration with AWS services

EMR

- Instance types, Storage, compression
- Consistent View
- S3Distcp
- Resizing and Autoscaling cluster
- Hadoop ecosystem with Hive, HBase, Presto, Spark
- Spark integration with Kinesis
- File formats Text/Parquet/ORC/Avro their usage and general knowledge.

Redshift

- Understanding of NodeSlice
- Distribution style
- Sort Key
- Data Types
- Compression
- Constraints
- Workload Management/Queues
- Data Loading techniques, encryption and compression
- Upsert
- Vaccum and Deepcopy
- Snapshots, Cross Region Snapshots, Restore from Snapshots

Elastic Search

Data Visualization

- QuickSight
- Zeppline, Jupyter, D3.js, Microstrategy

Athena

Glue

Machine Learning

Security

- Encryption Data At rest/in-transit
- SSE/CSE
- KMS
- Private Subnet /VPC endpoints
- Redshift Security
- EMR Security

Amazon Big Data Analytics Solution Introduction

There are various technologies which are available for solving BigData problem one of the popular one is Hadoop (Which changed entire Big Data world a decade before and still it is dominated) and similarly Spark computation engine where you can run where data processing these two are open source framework which AWS is providing using their EMR (Elastic Map Reduce) solution, AWS does not only incorporate these open source solution but they are also having their own solution to support the BigData like for storage they are using S3, and for structured data you can use DynamoDB, RDS (Relational Data Service), Redshift Cluster etc.

They are also providing solution to incorporate various Machine Learning Algorithms, and to visualize the data they have their own solution as well as open source solutions are also supported. Which makes AWS BigData solution quite vast as well as various options are available. Hence, based on your knowledge you can use respective solution. We will talk about each individual service in detail.

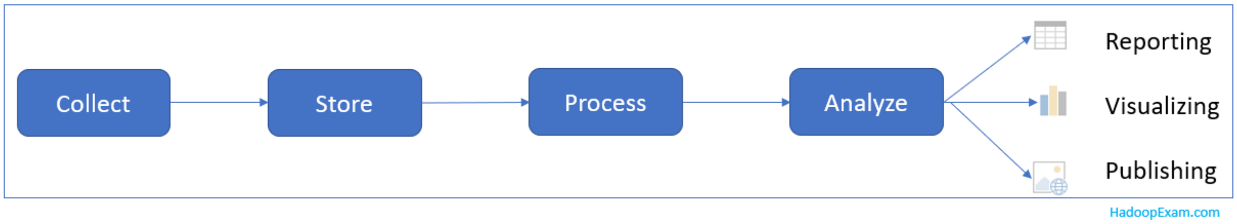
Analyzing the BigData

Once you have collected the data, generally your business intelligence team, data scientists, data analytics engineer would be analyzing this data using various available tool. Why do you want to analyze the data? There are few examples to answer this question

- Know your customer preferences in a set of products

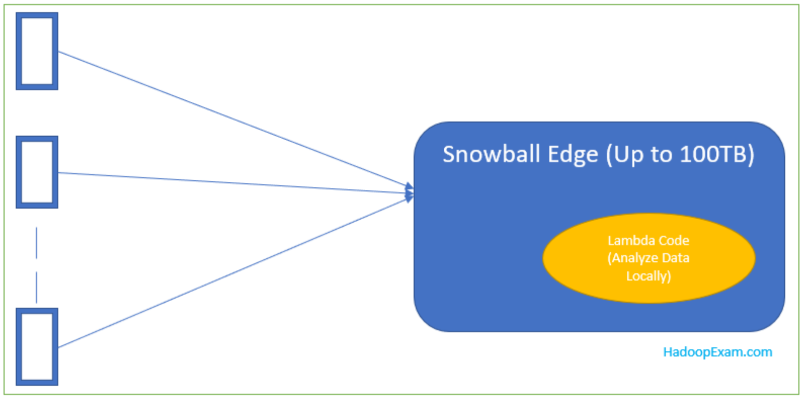
- How to compete in the market by providing required solution or products to the customer for his satisfaction.
- Finding frauds in the credit card industry or share market
- Understanding the performance of the business
- Finding security threats in the data

These are the very few examples of the Big Data analytics. Why do you need different set of tools for analyzing the data? Because there are various pattern, format and frequency which require data to be processed in real-time, near-real-time, structured and unstructured data. Below is the flow you generally follow with your BigData, obviously one individual can not do all these activities but should have knowledge how all can be achieved using various AWS services



For example for storing the data you can use AWS S3, Redshift etc. , if you want to move, transfer, and transform the data, you can use AWS Glue to orchestrate jobs. If your small devices like refrigerators sensors or sensors in car breaks etc. need to send data then you can use AWS IOT (internet of things) services. Few of the other examples of some AWS solutions which you can use in AWS BigData Analytics Solutions are below

- **AWS Snowball**: If you have tera-byte or peta-byte data to be transferred you can use this service.
- **Snowball Edge**: This is an 100TB data transfer device with the storage as well as compute capabilities. Suppose you have been asked in the question that you want some temporary storage on remote location or locally which can be upto 100TB then you can think of AWS Snowball Edge as a solution. And once data is collected and needs to be analyzed locally (On Snowball Edge itself) on the same remote location then you can use AWS Lambda code on the AWS Snowball Edge to analyze the data locally as well as analyzing the stream data.


Snowball Edge (Up to 100TB)
Lambda Code (Analyze Data Locally)
HadoopExam.com

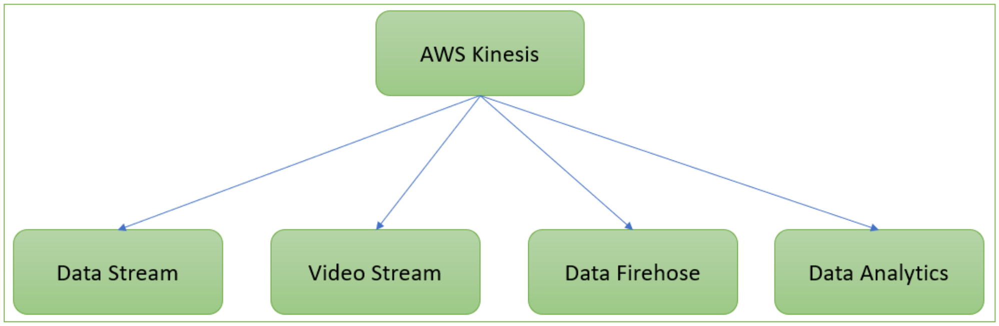
- **Kinesis Data Firehose**: If you are receiving data in stream continuously, they use this service for loading streaming data continuously.
- **AWS Mobile Hub**: Using this AWS Service we can collect and measure app usage and data, even we can transfer this data to another AWS service to store and to do custom analysis further.

Amazon Kinesis

AWS Kinesis platform is for the streaming data solution which can help you for the collecting, loading and analyzing streaming data in real-time or near-real-time. Below are the examples for the streaming data

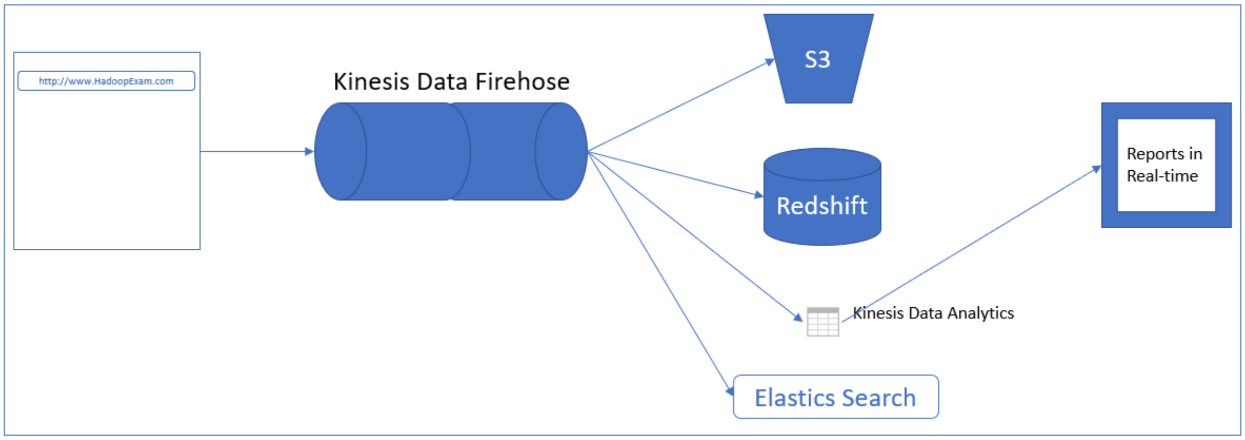
- Sensor data received from IOT devices continuously
- Application server or application logs
- Clicks on the webpages (Website Clickstream)

This all data can be stored in databases (e.g. RDS), Data warehouse (e.g. Redshift) or data lakes (e.g. S3 buckets). In Kinesis you can receive data in real-time as well as process it and analyze the same in near-real-time and respond accordingly. Amazon Kinesis has following four platform



HadoopExam.com

For the exam perspective most important service you need to learn is Amazon Kinesis Data Stream and Data Firehose. Using the Data Stream we can process and analyze the stream data very similar which you can do using Apache Kafka and Spark Structured streaming. Using the Data Firehose, we can collect the data in the AWS destination such as S3, Redshift, Kinesis Data Analytics, And Elastic Search Service.



HadoopExam.com

If you have been asked in the exam that which service to use to collect the stream data and destination could be either S3, Redshift, Kinesis Data Analytics or Elastic search then you can think Kinesis Data Firehose as the solutions. And also if you are asked to run the SQL queries on the collected data then you can say, you would be using the Kinesis Data Analytics because on S3 and Redshift you can not run the SQL queries. As I mentioned previously, if it talks about the destination then use the Kinesis Firehose and if it talks about the collecting data in real-time like data stream continuously then you can use Kinesis Data Stream for example capturing upto Terabytes of the data per hour from many sources like website clickstreams, equity market trading data, social media feeds, application server logs, geographical co-ordinates of moving vehicles etc.

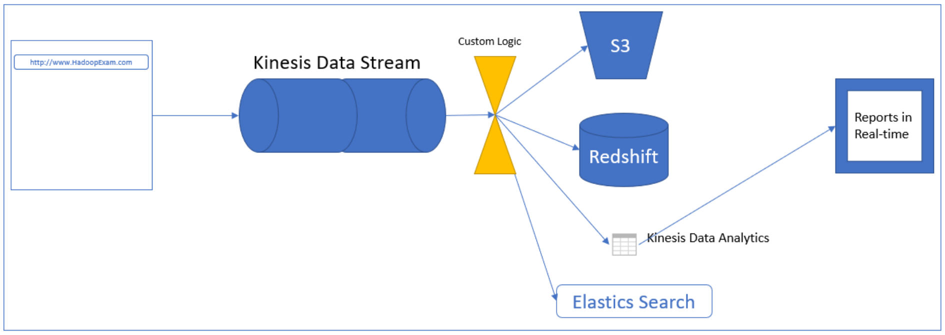
If it talks about capturing video stream data like capturing remote security camera's video data , video data from satellites etc. basically any data which is in video format then you have to use Kinesis video Stream solution.

Kinesis Client Library vs Kinesis Firehose

Both Kinesis Client library and Kinesis Firehose helps in ingesting data in S3, Redshift, Elastic Search, EMR, and AWS Lambda. Then what exactly is the difference in which scenario we should use which one. Few differences are below based on which you have to select correct answer

- Firehose is fully managed, scales automatically and stream needs to be manually managed
- In Kinesis Stream applications are build using the Kinesis Producer Library which put the data into a stream and then process it with application that uses the Kinesis Client Library and using the Kinesis Connector Library which send the processed data to S3, Redshift and DynamoDB etc.
- With the Kinesis Firehose it is simple, where we need to create the delivery stream and send the data to S3, Redshift etc. And you should have Kinesis Agent or API for that.
- Kinesis data stream can keep data for 7 days; hence it can be used as a storage as well. Which helps in custom processing before ingesting data to S3, Redshift or Elastic Search.
- Kinesis data stream is open-ended service at both the end on the producer side you will be configuring data producer to write the data in the Kinesis Stream, and this service will store your data in a continuous manner and able to replay as well, and order would be retained and on the other side, we would be configuring the data consumer to read the data out of the stream and

process it with the custom application. Kinesis data stream is a data storage system, it is more flexible and you can build your custom application as you want, even you have full control how to partition your data, how many shards you want to have for your particular stream.



HadoopExam.com

- **Kinesis Firehose**: It is an open ended only one side, you configure the you configure data producer to continuously push data into the Firehose Delivery Stream and on the other side you don't read the data from FireHose delivery stream and you don't write any application for that. Firehose automatically deliver the data to your destination like S3, Elastic Search, Redshift cluster etc. You can transform the data before data delivery it does not require manual administration.
- Even you can batch, compress and encrypt the data before loading it, which helps in minimizing the amount of storage used at the destination and to increase the security.

In Kinesis stream we need to provision input/output in a 1 MB/Sec block, and to adjust the size we don't need to restart the stream. Within one second data put in the stream would be available for analysis.

Tip: Hence, in the question if it is asked that producer needs to submit the data as soon as it is generated without batching it and even your producer e.g. Web Server fails then it should not lose the events/logs already generated. Consider the Kinesis data stream as a solution, even if in the question it is mentioned that report generations, extracting metrics, creating dashboard in real-time then select Kinesis data stream as an option.

Shards and Kinesis Data Stream

Kinesis data stream is made up from one or more shards, each shard gives 5 reads transaction per second (If 10 shards then 50 reads per second is possible, and also 2MB is the limit per second for each shard, hence with 10 shards 20MB is the limit for read in per second). And 1000 writes per second for each shard and 1MB is the maximum.

Availability Zones and Kinesis Data Stream

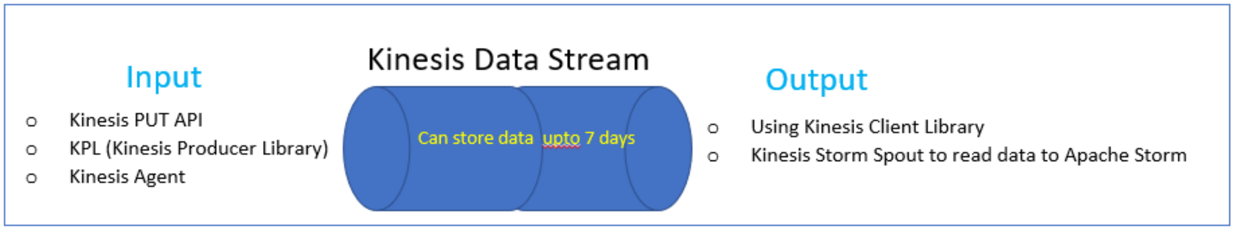
Amazon Kinesis Data Stream Synchronously replicates data across three availability zones in a single AWS region, (**Tip**: So, in a question if you see requirement with Multi-Az real time data then you can select Kinesis Data Stream).

Elasticity of Kinesis Data Stream

For increasing or decreasing the capacity of the stream we don't have to stop the streaming solution, rather we can use the API or development tool to automate the scaling of Kinesis Data Stream based on the need dynamically.

Kinesis Data Stream and interfaces

- Input: Producer will push the data into Kinesis data stream using
    - Kinesis PUT API
    - KPL (Kinesis Producer Library)
    - Kinesis Agent
- Output: To read the data from the Kinesis Data Stream.
    - Using Kinesis Client Library
    - Kinesis Storm Spout to read data to Apache Storm


Input
Kinesis PUT API
KPL (Kinesis Producer Library)
Kinesis Agent
Kinesis Data Stream
Can store data upto 7 days
Output
Using Kinesis Client Library
Kinesis Storm Spout to read data to Apache Storm

HadoopExam.com

Tip: If in the exam, if you see the requirement to read data streaming with less than 200KB/Sec then try to check which other option is available because Kinesis Data Stream is good for larger data throughputs.

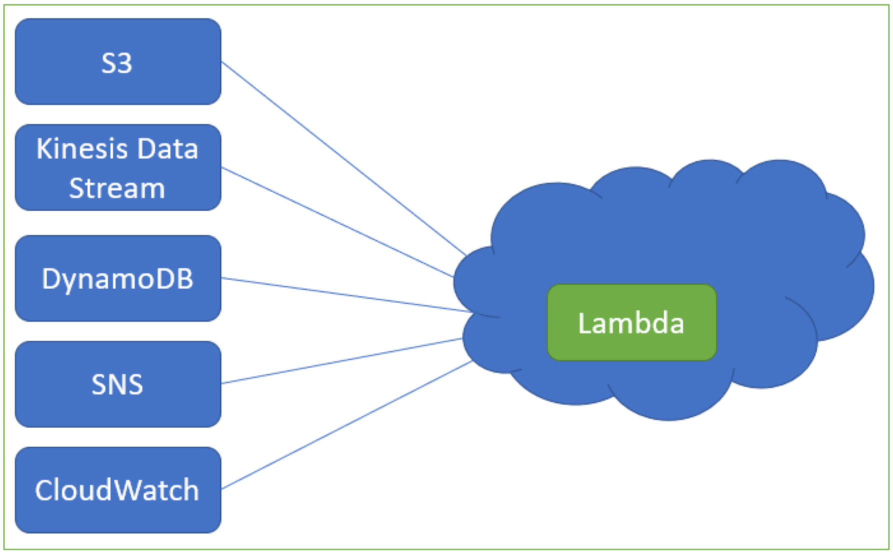
Tip: If you want to store stream data more than 7 days then, you have to move data from Kinesis Data stream out. Because it cannot retain data more than 7 days.

AWS Lambda

This service is used to run your code, without provisioning any AWS services, just upload your code and AWS Lambda will take care of running it whenever it is required. Even you can run as part of another AWS service or mobile. **Tip**: For example, with Kinesis Data Stream before sending data to destination e.g. S3 or Redshift cluster you wanted to apply custom logic or transformation on the data. Then use the AWS Lambda for that.

Lambda Usage

Sometime you wanted to execute the custom code based on the event or trigger like changes in data, changes in system state (e.g. Running to Stop), or once any action taken by a user. There are some services which can directly trigger the Lambda which are S3, DynamoDB, Kinesis Data Stream, SNS, CloudWatch, SES, Cognito etc.



HadoopExam.com

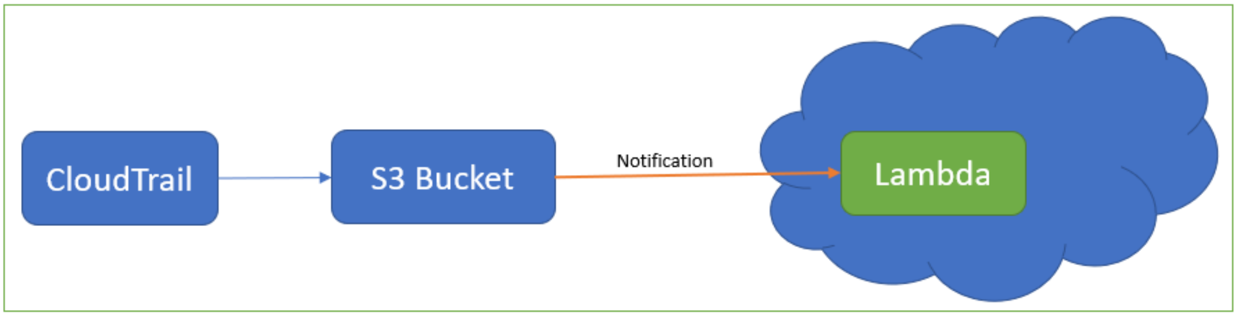
Tip: In question if it is asked that once file is uploaded in S3 bucket and need to apply some transformation logic on that file. Like Image file color change or converting the file format etc. this can be done using AWS Lambda.

Tip: If exam question if it is asked to transform the data which is received using Kinesis Data Stream from which can be either Website Click Stream, Application logs, IOT sensor messages etc. Then you can use AWS Lambda for that transformation in real time.

Tip: If you have been asked to extract data from one repository and load into another repository by applying some transformation then you can use the AWS Lambda for that transformation as well.

Tip: In question if it is being asked to implement scheduling solution and two options are given like running Cron (Schedular) on EC2 instance or AWS Lambda then give preference to the AWS Lambda because it is more available and cheaper option.

Tip: If even generated using CloudTrail require the Lambda Trigger then one of the possible way is CloudTrail will log the even in S3 bucket and you can use S3 bucket notification to trigger the Lambda function.



HadoopExam.com

Tip: Keep in mind in the questions if anywhere it is mentioned that your custom logic takes more than 15 mins or 900 seconds then you have to break your Lambda. Because each individual Lambda function can not run more than 900 seconds. Other alternative is using an EC2 instance and run your custom logic on that EC2 instance that is expensive.

Tip: AWS Lambda need to persist any state after or during the run then you should use S3, DynamoDB, RDS or any other storage. AWS Lambda itself should be stateless.

AWS EMR

If you have been already working with the BigData using Hadoop then AWS has managed solution for that which is available using the AWS EMR (Elastic Map Reducer). EMR provides cluster using EC2 instances on AWS and all the common BigData tools like Hadoop, Hive, Pig & Spark is available using EMR. Hence, creation of the cluster is not your headache, AWS would do the provisioning, managing and maintaining the infrastructure and software of a Hadoop Cluster.

If you have a GB/TB/PB scales data and want to processing in concurrent manner for 100's of instances then consider using the AWS EMR. Example of the things which you can do are

- Predictive Analytics
- Machine Learning
- Data Analytics (e.g. Click Stream, Log analytics, Customer Churn Analysis)
- You can Risk modeling, fraud analytics etc.

You can spin up the EMR cluster and once your job is done, you can shutdown the cluster, or you can keep it running if needed further then accordingly charges would be applied.

EMR performance is depend what EC2 instances you are selecting while creating the cluster, based on your memory, compute power, and storage requirement you have to select the EC2 instances.

EMR and EC2 Node failure behavior

In Hadoop cluster it is possible that nodes keep failing, whenever core node fails then EMR would provision a new node. However, keep in mind EMR is not going to replace the nodes if all the nodes in cluster fails. If you want to monitor the nodes in the cluster then use the CloudWatch.

HDFS

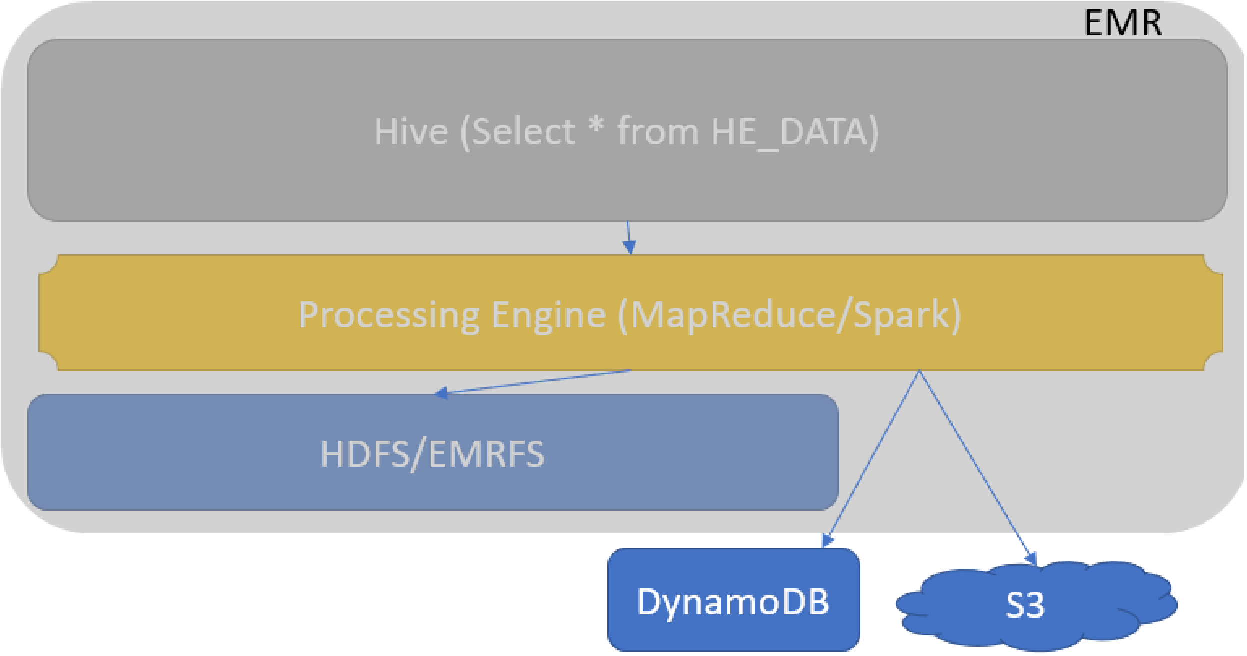
In EMR, core node holds the HDFS (Hadoop Distributed File System), hence if you wanted to increase the processing and storage capacity in EMR then you have to use additional EC2 instances. If you don't want to use EC2 instances as an storage then you can use the S3 natively, or EMRFS as an storage and you can have your EC2 instances for memory and compute power.

Compute Power in EMR

Compute power in EMR completely depend on the type and number of EC2 instances, if you want more compute capacity increase number of EC2 instances in EMR cluster or select the EC2 instances which has more CPU capacity.

EMR & Data Warehouse Solution

One of the most popular Data warehouse solution for the Hadoop framework is Apache Hive, which can help you to query the data which is stored in the HDFS/EMRFS/S3, the data format can be simple CSV file, Avro file or ORC (Columnar data), Parquet data etc. To query the data Apache Hive uses either MapReduce computation engine or you can configure to use Apache Spark (in memory processing and relatively much faster than MapReduce). Apache Hive Support almost the same query as standard SQL.


EMR
Hive (Select * from HE_DATA)
Processing Engine (MapReduce/Spark)
HDFS/EMRFS
DynamoDB
S3

Tip: If in the question it is mentioned that some structured/unstructured data stored in the S3 bucket and you want to run SQL query on that data, and in option you find Hive/EMR etc. then you can consider that as one of the possible correct answer.

The queries used in the Hive are known as HiveQL which is even beyond the normal SQL. Hive can support custom data types and user defined function as well for implementing custom logic in your query. Your user defined function can be written using Java. AWS has done enhancements and now Hive can be integrated with the DynamoDB as well as S3. With the EMR we can load table partitions from S3, even for custom MapReduce job EMR can access the code/scripts stored in S3.

Apache Pig

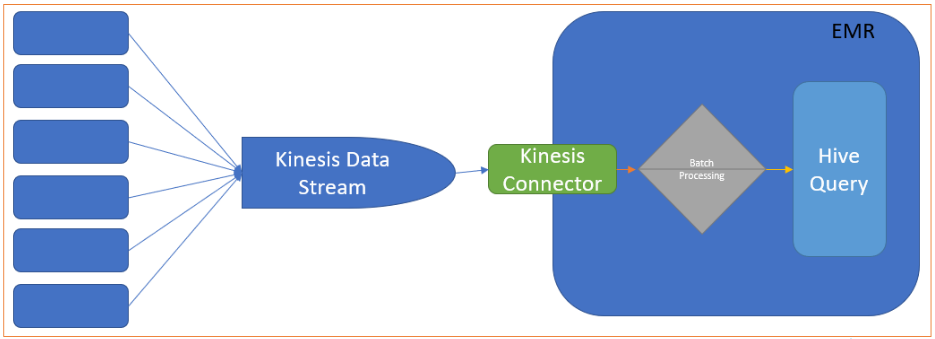
This is another interface to process the data in the EMR, this tool is also used for data analytics which uses different programming language which is known as PigLatin which is quite close to SQL, but exactly not an SQL language, it has some learning curve associated with it again you can write your own User Defined function using the Java, similarly you can do in the Hive. Using the PigLatin we can process structured/unstructured/complex data. Even you can load the additional require JARs from the S3 if needed.

Apache Spark

This is one of the most popular projects currently in the BigData world and its demand is growing very fast. Reason for its being so popular is its fastest compute engine and as well as user friendly API. HadoopExam has many learning resources for Apache Spark training and Certification and mostly subscribed package. It processes MapReduce kind of algorithm as well as many other computing algorithms by keeping the data in-memory. There are various API or Module available for different purposes like GraphFrame (for graph data processing), DataFrame (Data Processing), Machine Learning, Spark SQL, Structured streaming (Quite similar to AWS Kinesis Data Stream) etc.

Kinesis Connector

If you wanted to read data directly from the Kinesis Data Stream in the EMR and then apply any kind of processing on that data we can use the Kinesis Connector. And do the batch processing in the EMR using the tools like Hive, Pig, MapReduce etc.



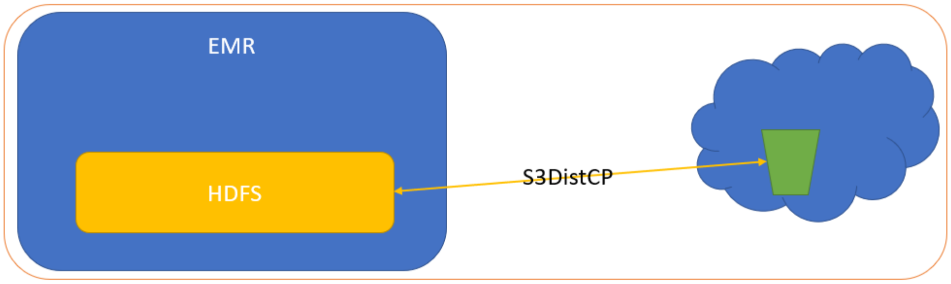
HadoopExam.com

Tip: In the certification exam if it is asked to you that you want to join the real time data stream with the data stored in either S3, HDFS, DynamoDB etc. Then you can use the Kinesis Connector to fetch the data from stream and then join with the already stored data.

Tip: Similarly, if you want to run the interactive queries on the real-time stream data then collect the data from the real-time stream using the Kinesis Data stream and then collect it in EMR using the Kinesis Connector. Once data is collected you can run the interactive query using Impala. Impala is an open source solution developed by Cloudera Inc. to support faster and interactive queries in real-time on the data.

S3DistCP

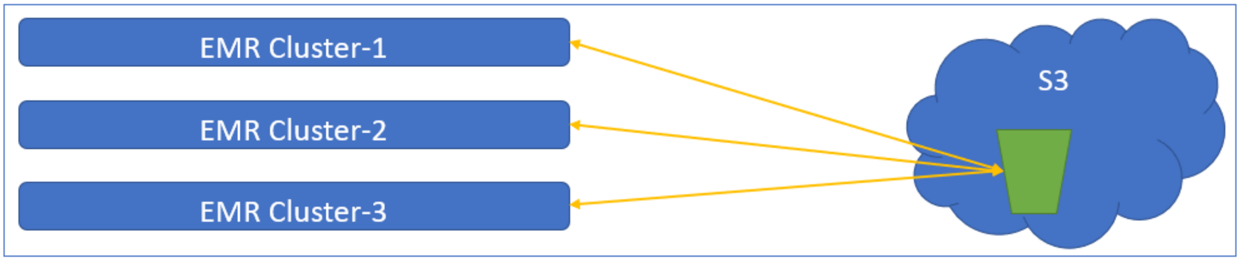
As we have discussed you can process the high volume of data using the EMR (Elastic Map Reduce :Hadoop based solution in AWS). Once data is processed you wanted to move the result back to S3 then you can use S3DistCP tool to copy the data from HDFS to S3 and vice versa.



HadoopExam.com

EMRFS

This is an alternative to the HDFS (Hadoop Distributed File System) on AWS, if you use EMRFS then AWS store your underline data in S3 and not on the native HDFS. And to provide the security you can enable the Sever-side and Client side encryption. So whatever metadata is required are stored in the DynamoDB. The advantage of doing that is more the one EMR cluster can access the data stored in S3, which is not easily (Without transferring) possible when you use the native HDFS as a storage.



HadoopExam.com

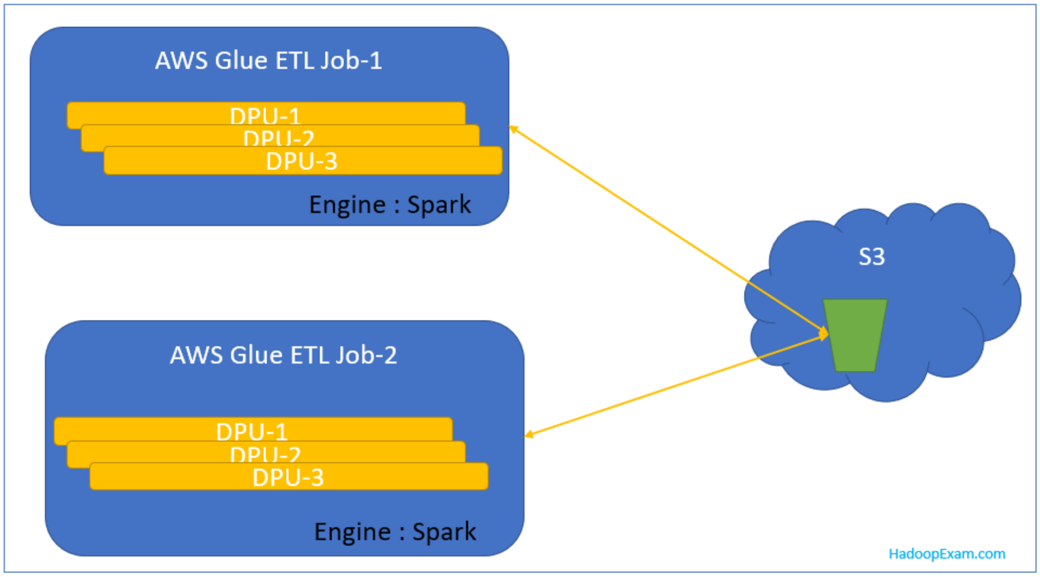
Tip: In the question if it is clearly mention that data volume is small enough to fit in a single machine, then never select EMR as a solution. Because EMR is good for the massive parallel computations where data size varies from 100’s of GB to PB.

Tip: If you see in the question it is mentioning that, they need operational data transaction features or OLTP (Online Transaction Processing) then also avoid considering EMR as a solution. Because EMR does not directly support ACID properties but rather you should see option related to AWS RDS (e.g. MySQL, Oracle etc.)

AWS Glue

Glue provides an ETL solution without any administration because it is serverless, and using AWS Glue, you can move data between data stores. AWS Glue can crawl the data and generate the code to execute data transformation and loading processes. Even you can integrate various AWS services like Athena, EMR and Redshift etc. And you can execute your code using Spark and Python. There are following kinds of the jobs you can implement with the AWS Glue

- ETL (Processing and Loading the data)
- Crawling (Data Discovery)



As you can see in the image, AWS Glue uses the Apache Spark for all ETL jobs processing in the Image two ETL jobs are shown. To process your ETL jobs you have to provide the DPU (Data Processing Unit) and minimum 2-DPUs are required. Even you can trigger the AWS Glue jobs from the external sources like AWS Lambda.

Tip: If you have been asked in the exam to have an ETL job for processing the data before loading into some RDS instance or DynamoDB etc. And needs to be triggered from external sources like AWS Lambda, S3 Bucket Notification etc. Then you can consider the AWS Glue as a solution. Undeline engine for the AWS Glue is Apache Spark.

Meta Data

As you already have data collected and wanted to have metadata generated or populated then AWS Glue can do this for you and done in the AWS Glue Data Catalog. Behind the seen if you permit then Glue would crawl the data and infer (derive) the schema of the data and even helps in partitioning information of the data. If you see that your new data would keep coming and structure could be different then better schedule the crawler so that on regular interval data catalog updated with the new meta data. Even if you already have metadata stored in the Hive Metastore then you can import that metadata from Hive to Glue as well. Because doing ETL processing, if metadata information is available then it makes ETL job faster.

Tip: If you see the requirement to process the data in real-time then avoid selecting AWS Glue as an answer. Because Glue is good fit for ETL processing which are batch oriented.

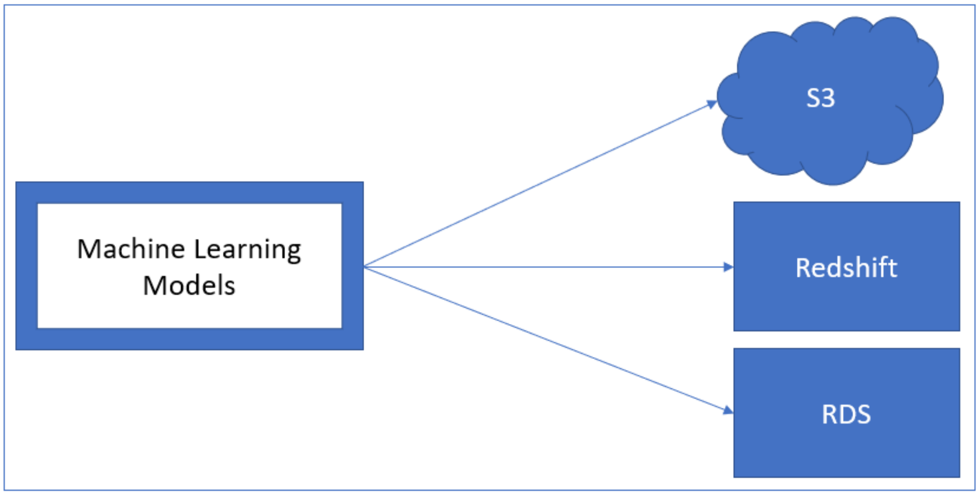
Tip: As in the Glue minimum batch interval is 5 mins; hence you cannot do the stream data processing. So, avoid Glue as an answer for stream data processing.

Tip: If you know that your ETL jobs required various technologies like Hive, Pig etc. Then don't use the Glue. Because Glue usage the only one framework that is PySpark. Again avoid using Glue and consider the EMR or AWS Data Pipeline as a better fit.

Tip: If your data is already stored in NoSQL databased like DynamoDB, HBase etc then AWS Glue is not supported. Reason NoSQL databases don't have strict schema in place, which is the basic requirement for Data source for the Glue.

Machine Learning

Machine Learning is one of the most popular technology since last 4-5 years and highly in demand with the higher pay packages. Using the ML you can do the predictions and for the AWS provides the visualization tools and wizards so that you can create various machine learning algorithm without knowing their underline implementations. Once you productionize the ML prediction model you can use it with your newly receiving data to do the various predictions. Machine Learning Models can be created using the data stored in S3, RDS & Redshift.



HadoopExam.com

As part of ML, you can do the following things in general using AWS Wizard

- Data exploration
- Training the models
- Evaluating the Model Quality
- Adjusting the Model Output to align with your business goals.

Once the ML Model is ready, we need to deploy it so that predictions can be made. There are two possibility to making the predictions either using the batches or in real-time.

Generally, identifying questions based on Machine Learning is easy to identify because we don't see they would be asking too complex questions based on the algorithm. You would be using Machine Learning to identify the pattern in the new data which is not yet seen based on the training data we have created our Machine Learning Model. If you see machine learning lifecycle works as below

- Collect the data set
- Divide this dataset in two parts (Training Dataset & Testing Dataset)
- Build/Select the more than one Machine Learning Model
- Train this Machine Learning Models based on the training data set
- Test the Machine Learning Model with the available testing dataset
- Select the Model which has performed the best among all
- Deploy the best selected model in production
- Use this deployed model in the new dataset and make the predictions

Example of the predictions you can do are

- Whether the new financial transaction made is fraud or not or done by some terrorist group etc.
- Based on the previous order can we recommend the new product to the customer or not
- Classify the data into various categories, best and simple example is whether the email received is spam or not-spam.
- Based on the social media data, can you predict the which political party is going to win etc.

AWS Machine Learning is a managed service and you don't have to provision the servers and become an administrator.

Tip: Suppose you have been given more than 100GB data and apply Machine Learning Model on it, is not supported. In this case don't select AWS Machine Learning solution as an answer. Rather go to the AWS EMR and use the Spark MLLib to build the model.

Tip: You must remember, what all are the Machine Learning algorithms are supported as of now by AWS ML and for all other uses case you have to consider AWS EMR with Spark MLLib. Below are the supported ML Models in AWS.

- Binary Classifications
- MultiLabel Classification
- Numeric Regression (Predicting a number)

Some of the example for unsupported Machine Learning Algorithms are

- Sequence Predictions
- Unsupervised Clustering

Tip: You must avoid using another Machine Learning Model output as an input for the Machine Learning model, if highest accuracy is required. Because Machine Learning model can have error and may not be able to predict with highest accuracy.

Machine Learning Concepts

- **Data sources**: As it has a confusing name that it can contain your input data to train ML algorithm. But this is not the case.
- ML Models:
- Evaluations:
- Batch Predictions:
- Real-time predictions:

Amazon DynamoDB

This is NoSQL solution from the AWS side, which is managed service and you don't have to administer. It supports high volume of traffic as well and provide the single digit millisecond latency. We have seen many people use it as a caching solution as well. However, we don't recommend that as a caching solution there are other services available for building the data cache. DynamoDB stores the structured data in a table with the Primary Key.

- DynamoDB is good for structured data with the Primary Key
- DynamoDB supports the document store as well using JSON, HTML and XML
- There are three datatypes in its number, string and binary
- There is no fixed schema for the table.
- **Tip**: If you need fixed schema in your database then don't use the DynamoDB
- Primary key can be single value or composite (Based on more than one attribute)
- DynamoDB provides the both local and global secondary indexes.
- **Tip**: So, you can write query based on the filter criteria which are not part of primary key, using the secondary indexes.
- DynamoDB is eventually consistent database.
- **Tip**: If you need strong consistency than don't use DynamoDB.
- DynamoDB supports individual item level transactions and not across the multiple attributes.
- **Tip**: If you need transactions support based on the ACID properties then consider using RDS.
- **Tip**: DynamoDB does not provide the SQL query interface to read the data.
- **Tip**: DynamoDB is a good for the key-value store.
- **Tip**: In your application whether mobile or something else with the low read and write latencies then use the DynamoDB.
- If you need specific throughput then you can provision it accordingly.

- **Replication**: DynamoDB by default replicate your data in three availability zones.
- **Across region replication**: If you want to replicate the data across the region then use the AWS Dynamo Streams.
- **Data selection**: To select the data from DynamoDB table you have to use REST API and you can not write SQL like queries.
- **Tip**: If in the questions it talks about the joining the data then ignore using the DynamoDB as a correct option and rather look for the RDS as a solution.
- **Tip**: If you want to store the metadata about the videos or images then its good choice. But storing the video itself or image is not a good choice. Best solution for that is S3.
- **Tip**: If in the question you find that your data would not be accessed very frequently then avoid using the DynamoDB as a solution.

Question-7: You are working in a web hosting company which manages the more than 10,000 webservers for supporting various websites. There is every hours various web server logs are generated which is managed to get stored in the S3, now you wanted to write some ETL to generate a partitioned data, and wanted to run regularly some SQL queries every 3 hours to find any hacking activities is being done on any of the webserver, if yes then they wanted to generate the report out of this. Which of the following is suitable solution for this because every hour around 300GB of the logs generated from all the servers?

A. You would be using Kinesis FireHose and Kinesis Data Analytics

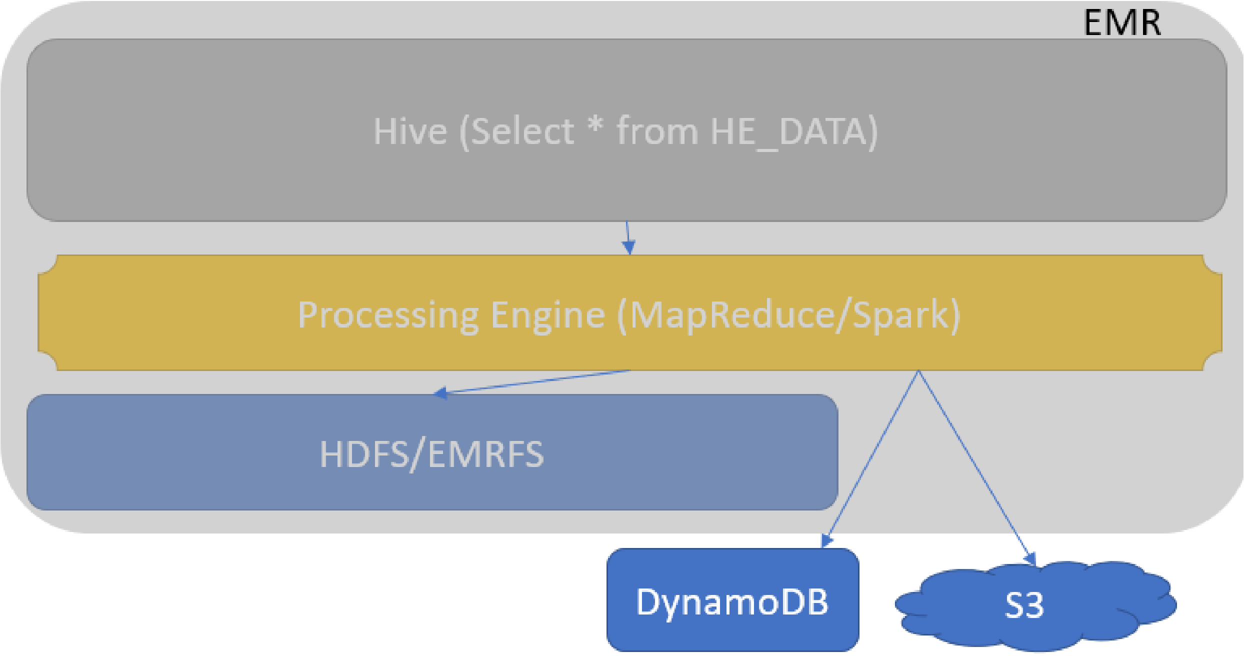
B. You would be using EMR and Hive Data warehouse

C. You would be using Kinesis Data Stream and Kinesis Data Analytics

D. You would be using Redshift cluster, Lambda and DynamoDB

Ans: B

Exp: As there is no need of having the real time data processing so we can safely ignore the options provided like Kinesis Data Stream and kinesis Data analytics. As most of this processing needs to be done in the Batch every 3 hours and looking at the data volume we can say EMR is the Good fit, even we can use the Apache Hive to query the data stored in AWS S3. Hence, we can use the EMR as compute engine and S3 as a data storage and using the Hive we can query the data stored in S3, even we need to do any transformation then we can use the EMR MapReduce job as well, if needed.


EMR
Hive (Select * from HE_DATA)
Processing Engine (MapReduce/Spark)
HDFS/EMRFS
DynamoDB
S3

Question-8: You are working in a Web Hosting company which has 10,000 web hosting servers and running more than 1 million websites, you want to collect the data regularly in real time from all the websites and webserver. And in every 30 mins you wanted to run the batch processing so that you can find the most top 50 errors happening on the various websites, so which of the following tools combination can help in this scenario?

A. You can use the AWS Kinesis FireHose and Kinesis Data Analytics, with the DynamoDB

B. You would be using Redshift cluster and AWS RDS

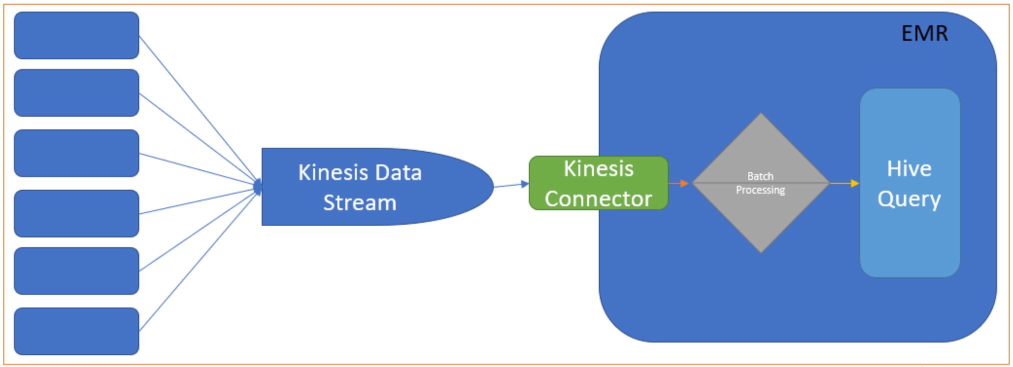
C. You would be using Redshift cluster, S3 and Apache Hive

D. You would be using Kinesis Connector, Kinesis Data Stream, EMR and Apache Hive

E. You would be using S3, Kinesis FireHose, Kinesis data analytics

Ans: D

Exp: As we are going to read data in real time, hence we can use the Kinesis Data Stream for that, now we need to read this data from Kinesis Data Stream to EMR and in the EMR we can do the batch processing and generate the report by running Hive queries in every 30 mins.



HadoopExam.com

Question-9: You are working with an eCommerce company which has more than 1 lakh different products which are getting sold online. You want to advertise different product in real time to the user who are currently browsing the website through eMail, recommendation or text messages. Which of the following AWS services you can combine to implement this requirement?

A. Kinesis Data Stream, HDFS, Kinesis Connector

B. Kinesis FireHose, HDFS, Kinesis Data Analytics

C. SQS, HDFS, Kinesis Data Analytics

D. EMR, S3 and DynamoDB

Ans: C

Exp: In the question requirement is showing advertisement in near-real time. Hence, we need to collect the data asap and the best solution for this is Kinesis Data Stream, and we can store all the pre-created advertisement in the DynamoDB, S3 or HDFS anywhere. Once we get the real-time stream data and join this data with the pre-created advertisement so that relevant ads can be selected and shown in the recommendations or send an email with the recommendations.

Question-10: You have data collected in an S3 bucket, and this data you want to load in one of the AWS RDS instances. But the data collected in S3 bucket is not well formatted and hence before inserting this data into RDS instance you want to convert in a particular format so that it can be easily loaded in the RDS instance as part of ETL job. There is a one custom function written using Lambda to check the size of the collected data, as soon as it reaches 10 GB then ETL job should be initiated to load the data in the RDS instance. How can you achieve this requirement?

A. You would be using EMR

B. You would be using AWS Lambda

C. You would be using AWS Glue and AWS Lambda

D. You would be using Apache Spark and EMR

Answer: C

Exp: As in the question it is given that you wanted to load the data in the RDS instance and before loading you want to clean up the data and pre-process the same to convert in required format so that it can be easily loaded in RDS instance. So this entire requirement is an ETL job, which can be easily implemented using the AWS Glue, which helps in creating ETL jobs and to trigger the ETL jobs from AWS Lambda function where we would be implementing the custom logic for checking the size of the data regularly and as soon as it reaches to 10 GB, we can trigger the AWS Glue job.

WL Paper-2 (Question Started from-1 only)

Question-11: You are working in retail bank as a business analyst, this bank offers the various retail products like Commercial Loan, Mortgage Loan, Credit Card, saving accounts, Current account etc. as regular retail bank does. Your bank has done collaboration with one of the eCommerce company to share their customer information with the bank as well as providing discounts if customer use the credit card issued by a bank. There is around 10GB of the data on daily basis in the evening you are receiving as well as in the real-time also you are receiving user information as soon as he does the transactions. You have been assigned with the task to identify below two things

- Whether customer can buy a new product offered by bank like loan, credit card tops up etc. and prediction should be Yes or no only.
- Similarly, at the real time the information you are receiving about the user transaction you need to find whether this is a fraud transaction or not.

Which of the following you would be using to solve the given analytical problem?

A. You would be using Apache Spark and EMR to run your Machine Learning algorithm in real time.

B. You would be using AWS ML batch predictions from the data which you are receiving in the evening every day.

C. You would be considering to use the Binary Classification Model

D. You would be considering Multiclass Machine Learning Model because of more than two products offered by bank.

E. You would be using AWS ML to make the predictions in real-time for fraud transactions.

F. You would be using batch predictions for asynchronously predicting the fraud transactions for the data you are continuously getting.

Ans: B, C, E

Exp: You need to understand the specific thing which is being asked in this question. Read carefully and summarize the things.

- Bank offering multiple products but they wanted to predict whether user/customer will subscribe or not the product like Yes or No
- To find whether the transaction is fraud or not, again need to predict either of the two values.

Hence, best fit for this requirement is a binary classification (where you need to predict either of the one value). Hence 3rd option is correct.

Next, we are received data in both batch as well as in real-time and AWS ML supports predictions on both kind of the data. If data volume is less than 100GB.

Hence, we can use Batch predictions in the evening for asynchronously predicting the values. And if we want to predict the values in real-time it means synchronous prediction is required. Hence, we would be using the AWS ML Real-time predictions. Hence, in this case option-2 and option-5 are also fit well.

Option-4 is given to confuse with the type of the products offered by the bank and that is wrong answer.

Option-1 is a possible solution, but why do you want to create such a complex solution even it is possible with the simple setup using AWS ML.

Question-12: You are working with a matrimonial website named HighMariage.com in which users can like or dislike the profiles and if users are having paid subscription then he can see the contact detail of other users. And can establish communications with each other using chat application, as well as can send email as well. Each user can create up to 5 albums which can have images in it and user can make it visible publicly or may not. Each user data is stored in a DynamoDB table named ProfileDataTable there are on daily basis 1000's of profiles are created however millions of views are happening. You need to implement Auto scaling with the DynamoDB table which all things you can do and also select other correct options.

A. You would be storing the images in the same ProfileDataTable so that it makes your website highly performant while viewing.

B. It would publish consumed capacity metrics to AWS CloudTrail

C. If the consumed capacity exceeds your target utilizations (or below the target) for a specific length of time, AWS send an alert using SQS.

D. If the consumed capacity exceeds your target utilizations (or below the target) for a specific length of time, AWS CloudWatch triggers an Alarm using SNS

E. If the consumed capacity exceeds your target utilizations (or below the target) for a specific length of time, AWS CloudWatch triggers an Alarm using SQS

F. The Alarm invokes Application Auto Scaling to evaluate your scaling policy and process request using ProfileDataTable

G. Publishes Consumed Capacity metrics at AWS CloudWatch

Ans : D, F,G

Exp : If you read the entire question it is more about asking the question regarding the working with specific table in DynamoDB, how being a managed service it can help you in reducing the cost. And you should get notified with various events.

Its about the dynamically accessing the managed resources from the DynamoDB table. Suppose during DayTime your application need more resources and in night time it require less resources and DynamoDB table should help you in getting such resources and cost you less.

AWS DynamoDB provides auto scaling which uses the AWS application Auto Scaling service to dynamically adjust the provisioned throughput capacity on your behalf. Here, on your behalf is important aspect, because DynamoDB is a managed service and you would not be doing administration tasks your own. And based on the traffic received DynamoDB table would do the performance. You can define auto scaling on both Table as well as any Global Secondary indexes which you have created.

What you have to do is create an auto scaling policy for the table or a global secondary index. And also you have to make sure to define whether you need auto scaling for the read or write or for both. As in the given question we need read scaling because that varies a lot.

In the scaling policy we would be defining the target utilization which defines the percentage of consumed provisioned throughput at a point in time and can set up upto between 20 and 90 percent. These are the steps you need to follow for auto scaling of the DynamoDB

1. You create an Application Auto Scaling policy for your DynamoDB table named ProfileDataTable.
2. DynamoDB publishes consumed capacity metrics to Amazon CloudWatch (Remember its CloudWatch). CloudTrail is more for logging perspective and for Alarm you should use Cloudwatch.
3. If the table's consumed capacity exceeds your target utilization (or falls below the target) for a specific length of time, Amazon CloudWatch triggers an alarm. You can view the alarm on the console and receive notifications using Amazon Simple Notification Service (Amazon SNS). Always you want to create a Notification then use the SNS and not the SQS.
4. The CloudWatch alarm invokes Application Auto Scaling to evaluate your scaling policy. Hence, whatever you defined in the scaling policy that would be considered during auto scalaing.
5. Application Auto Scaling issues an UpdateTable request to adjust your table's provisioned throughput.
6. DynamoDB processes the UpdateTable request, dynamically increasing (or decreasing) the table's provisioned throughput capacity so that it approaches your target utilization.

Question-13: You are working with a retail company as BigData expert, name of the company is BigRetail Inc. which has more than 3000 stores in the country. As so many stores and under the store there are around 30 sub-stores are created. Being an Big Data architect you decided to use the IOT solutions for all the stores like 3000X30=90,000 with each one having almost 5 IOT devices. You are using this IOT devices for the monitoring of theft, improving overall performance for fetching continuous data from the IOT device as well as increasing the cross sell be recommending the products to customer and taking feedback from the customer to increase their buying experience. You would be implementing this solution using the AWS IOT Core, as part of supporting so many IOT devices you need to follow the standard approach for provisioning and registering the IOT devices. Select the correct answer for provisioning the IOT devices and setting it up, assuming all the IOT devices already having X.509 certificates

A. You would be registering X.509 certificate with AWS IOT

B. You would be creating an IOT policy for the IOT device operations and attach it to the existing X.509 certificate

C. You would be creating an IOT policy for the IOT device operations and attach it to IOT thing

D. You would be creating a device shadow service for getting and setting the state of the device over MQTT and HTTP

E. You would be defining an IOT Rule

Answer: A, B, D

Exp: From the question summarize what it wants

- 90,000X5 IOT devices needs to be setup across all the store.
- IOT devices already having the certificate

Based on our knowledge When you provision a device with AWS IoT, you must create resources so your devices and AWS IoT can communicate securely. Other resources can be created to help you manage your device fleet. The following resources can be created during the provisioning process:

- **An IoT thing:** IoT things are entries in the AWS IoT device registry. Each thing has a unique name and set of attributes, and is associated with a physical device. Things can be defined using a thing type or grouped into thing groups.
- **An X.509 certificate**: Devices use X.509 certificates to perform mutual authentication with AWS IoT. You can register an existing certificate or have AWS IoT generate and register a new certificate for you. You associate a certificate with a device by attaching it to the thing that represents the device.
- **An IoT policy**: IoT policies define the operations a device can perform in AWS IoT. IoT policies are attached to device certificates. When a device presents the certificate to AWS IoT, it is granted the permissions specified in the policy.
- **Device shadow**: Device's shadow is a JSON document that is used to store and retrieve current state information for a device. The Device Shadow service maintains a shadow for each device you connect to AWS IoT. You can use the shadow to get and set the state of a device over MQTT or HTTP, regardless of whether the device is connected to the Internet.

Question-14: You are working with a company which offers online training courses which are developed by the university professors. This university professor records various training courses (Known as master videos) with different kind of recorder and the format based on the video recorder they have. However, your website can be opened by the subscribed students in the Browser, Mobile, Tablets, TVs etc. Hence, you need to transcode all the master videos which are created by the University professor. You are running your website on the EC2 instances in auto-scaling mode, you need to develop faster and efficient solution to process the master videos in the various required format. However, please note that this master videos may or may not required in future. Which of the following options help in achieving this requirement?

A. You would be initially storing master content in S3 and once transcoding is done move them to AWS Glacier.

B. You would be always storing the master contents in the S3 bucket, because it is required while learner access the transcoded format.

C. You will be using Elastic transcoder from AWS to convert the video in different format

D. You would be storing the transcoded content in the AWS S3 bucket.

E. You should be using AWS CloudFront to distribute the content on your website

F. You would be using Kinesis Data Stream for streaming the videos

G. You can monitor the health of the transcoding using the CloudWatch

Answer: A,C,D,E,G

Exp: As you can see you need to develop end to end solution for video Hosting website and video should be played efficiently on any device. Hence, it is required you convert the video format and the process is known as transcoding. Amazon provide the Elastic Transcoding service based on the per minute transcoding of the video (which is bit expensive from our view) for 10 mins transcoding to a particular format it charges $0.30

Once the original (master) videos once converted need to be shifted to AWS Glacier (Low cost storage) because you don't want to access them on regular basis. So, option-1 is correct. As a source video location initially, you would be adding videos in S3 bucket as well as after transcoding you need to put in S3 bucket as an output/destination. Option-2 is wrong because we are not going to keep the videos in the master bucket always.

Elastic transcoder is the obvious choice for this requirement, hence option-3 is also correct. CloudFront is a content distribution service from AWS, hence this is also a good choice for the given requirement, which makes option-4 & 5 as correct choice.

Kinesis data stream is not required here, because there is no requirement processing data in real-time for which this service is required.

In most of the cases you want to do monitoring of a particular AWS service and again AWS CloudWatch is an obvious choice.

Question-15: You are working with a company which has an online clothing store (ClothEra.com) and already having more than a million customer who regularly buy cloths from them and being an BigData solution architect you have been assigned a task to improve the overall customer experience by collecting the data in real time from the website clickstream. However, data generated is in quite high volume and various category. Once you receive the data it should be available in actual format at-least for 24 hours and then transformed data can be stored some persistence store for historical analysis. Having a more than 5000's customer on average being online generates almost 100 events per second, however each even data size is less than 1KB. the Select the correct option from below for the given requirement

A. You would be creating a single shared in Kinesis Data stream to collect the data

B. You would eb creating 20 or more shards in the Kinesis Data stream to collect the data

C. You would be using Kinesis Firehose for collecting all the clickstream data and transforming them in real-time

D. You can have more than one stream created in a single shard.

E. A data record is the unit of data stored in a Kinesis data stream. Data records are composed of a sequence number, a partition key, and a data blob, which is an immutable sequence of bytes.

F. You would be setting Data stream retention period as 24 hours once it is created.

G. You would be storing data in the DynamoDB

H. You would be using Elastic Cache to cache the data for 24 hours.

Answer: B, E, F, G

Exp: In the given question the requirement is to collect more than 100 event logs per second and obvious choice for this is Kinesis Data Stream. Let's understand the shard

A shard is a uniquely identified sequence of data records in a stream. A stream is composed of one or more shards, each of which provides a fixed unit of capacity. Each shard can support up to 5 transactions per second for reads, up to a maximum total data read rate of 2 MB per second and up to 1,000 records per second for writes, up to a maximum total data write rate of 1 MB per second (including partition keys). The data capacity of your stream is a function of the number of shards that you specify for the stream. The total capacity of the stream is the sum of the capacities of its shards.

If your data rate increases, you can increase or decrease the number of shards allocated to your stream.

Hence, we need to have at-least 20 shards (20X5=100) to support the 100 rad transactions per second. As Kinesis Data stream, data record is the unit of data stored in a Kinesis data stream. Data records are composed of a sequence number, a partition key, and a data blob, which is an immutable sequence of bytes. Which makes option-2 & 5 as correct.

Data retention: The retention period is the length of time that data records are accessible after they are added to the stream. A stream's retention period is set to a default of 24 hours after creation. You can increase the retention period up to 168 hours (7 days). Hence, you don't need different caching solution. Which make option-6 as correct and option-8 as wrong.

And for the persistence we can use DynamoDB to store data permanently.

Question16: You are working with a company which is already using the AWS as a solution and collecting the data from their websites and in the backend they are currently using AWS RDS for the structured data and S3 for storing unstructured data. Being an AWS BigData Architect you have been asked to create a Data Ware house solution so that every day analytical queries can be run on the tables. Hence, you decided to use the Redshift and come with the following tables in the Redshift cluster

- **FCT_ONLINE_ORDER**: This tables contains all the order placed by the customer.
- FCT_ONLINE_SALES: This table contains all sales transactions even failed transaction. Hence, on everyday in the evening you can generate the report for failed, half and completed transactions.
- DIM_CUSTOMER_INFO: This table contains the information about the customer.
- DIM_PRODUCT: This table contains the information regarding all the products ordered online.
- DIM_DATE: All the dates for the transactions are stored in this table, this is relatively small table.

As you need to join these two tables FCT_ONLINE_SALES with the DIM_DATE table to generate daily, weekly and monthly reports. Which distribution style you would be considering for these two tables?

A. FCT_ONLINE_SALES with the KEY Distribution strategy on the primary key, and DIM_DATE should be with the ALL distribution strategy on the Primary Key

B.FCT_ONLINE_SALES with the EVEN Distribution strategy on the primary key, and DIM_DATE should be with the EVEN distribution strategy on the Primary Key

C. FCT_ONLINE_SALES with the ALL Distribution strategy on the primary key, and DIM_DATE should be with the ALL distribution strategy on the Primary Key

D. FCT_ONLINE_SALES with the KEY Distribution strategy on the primary key, and DIM_DATE should be with the KEY distribution strategy on the Primary Key

E. FCT_ONLINE_SALES with the EVEN Distribution strategy on the primary key, and DIM_DATE should be with the EVEN distribution strategy on the Primary Key

Ans : A

Exp: This question is simple to answer of you understand the following stuff at the first place.

Redshift is a distributed database where your data can be stored on one or more than one node in the distributed system. And how the data should be distributed on various node is decided by the strategy you would be using and this is very critical factor not only for the storing the data, but also on the query pattern you would be applying on the database. When you run the query Redshift will send the data to the Compute node for doing the joins and aggregations. And while doing that redistribution can be done based on specific rows (Good choice for large table and queried based on some filter criteria) or your entire table (Good choice for small table). Based on the requirement if

- For Joining: Collocate the rows from the Joining tables.
- Even load : Distribute the data evenly across all the nodes in the cluster.

Lets understand the Distribution styles in Redshift cluster tables.

KEY distribution: The rows are distributed according to the values in one column. The leader node will attempt to place matching values on the same node slice. If you distribute a pair of tables on the joining keys, the leader node collocates the rows on the slices according to the values in the joining columns so that matching values from the common columns are physically stored together.

ALL distribution: A copy of the entire table is distributed to every node. Where EVEN distribution or KEY distribution place only a portion of a table's rows on each node, ALL distribution ensures that every row is collocated for every join that the table participates in.

EVEN distribution: The rows are distributed across the slices in a round-robin fashion, regardless of the values in any particular column. EVEN distribution is appropriate when a table does not participate in joins or when there is not a clear choice between KEY distribution and ALL distribution. EVEN distribution is the default distribution style.

Now in the question we need to understand that we are joining one large table (FCT_ONLINE_SALES) with the small table (DIM_DATE).

Hence, we can distribute the large table using the KEY DISTRIBUTION style, while small table can be replicated evenly so that it is collocated with the FCT_ONLINE_SALES data table and join query can be efficiently executed.

Question-17: You are working as a Big Data solution architect in a retail bank named American Bank of Retail, which is already having their various solution built on the AWS. And recently they partnered with one of online e-commerce retail giant. And while making the purchase user want to use the American Bank of retail credit/debit/internet banking solution and tax can be calculated in runtime and can be used for annual tax return filing for the customer as well as for the e-commerce website. To receive the data you decided to use Kinesis Data stream for real-time data accessing, you have already built the solution using Java Programming to consume the data from the stream in real-time, which of the following statements are correct with regards to writing the consuming the data?

A. You would be using Kinesis Data Stream API for connecting the stream and enumerating the shards.

B. You would be using KCL (Kinesis Client Library) for connecting the stream and enumerating the shards.

C. You would be using Hybrid solution both the API and KCL for reading data in real time

D. You would be using the design patterns and code for the consumer applications provided by the KCL.

E. You would be using KCL for checkpointing the processed record.

Ans: B, D, E

Exp: You need to basically understand the difference between KCL and Kinesis Data stream API to correctly selecting the options. You can develop a consumer application for Amazon Kinesis Data

Streams using the Kinesis Client Library (KCL). Although you can use the Kinesis Data Streams API to get data from a Kinesis data stream, we recommend that you use the design patterns and code for consumer applications provided by the KCL.

The Kinesis Client Library (KCL) helps you consume and process data from a Kinesis data stream. This type of application is also referred to as a consumer. The KCL takes care of many of the complex tasks associated with distributed computing, such as load balancing across multiple instances, responding to instance failures, checkpointing processed records, and reacting to resharding. The KCL enables you to focus on writing record-processing logic.

The KCL is different from the Kinesis Data Streams API that is available in the AWS SDKs. The Kinesis Data Streams API helps you manage many aspects of Kinesis Data Streams (including creating streams, resharding, and putting and getting records). The KCL provides a layer of abstraction specifically for processing data in a consumer role.

The KCL is a Java library; support for languages other than Java is provided using a multi-language interface called the MultiLangDaemon. This daemon is Java-based and runs in the background when you are using a KCL language other than Java.

At runtime, a KCL application instantiates a worker with configuration information, and then uses a record processor to process the data received from a Kinesis data stream. You can run a KCL application on any number of instances. Multiple instances of the same application coordinate on failures and load balance dynamically. You can also have multiple KCL applications working on the same stream, subject to throughput limits.

The KCL acts as an intermediary between your record processing logic and Kinesis Data Streams.

When you start a KCL application, it calls the KCL to instantiate a *worker*. This call provides the KCL with configuration information for the application, such as the stream name and AWS credentials.

Question-18: You are working as a BigData solution architect in a Global matrimony website, which provides the both Men and Women profiles on their application in the 20 different countries in different continent like United States, Europe, Australia and more. As you need high read capacity because at a time many thousands of prospects are online and can visit the free and paid profile. As the customers are global which of the best design for the DynamoDB table.

A. You would be creating a separate DynamoDB table for each country or at least region to provide better performance.

B. You would be creating a Global table in the DynamoDB and enable the replication.

C. For creating a separate table in each region you must have same partition key across all the tables.

D. For using the Global table each replica must have the same partition keys as all of the other replicas.

E. Across all the replica table must have the same write capacity and not necessary the same read capacity.

F. Across all the table for each region you must have the same write capacity and not necessary the same read capacity.

Answer: B, D, E

Exp: This question is easy to answer if you know the basic concepts of the Global table in DynamoDB. There is no point of creating a separate table for each country. Because customer can be across the region or country and can access the profile of different country as well. And having separate table for each country is highly in-efficient solution. And you should go for the Global Table in the DynamoDB, however you should keep the following things in mind while using Global DynamoDB table with the replication. Using Amazon DynamoDB global tables, you can replicate your table data across AWS Regions. It is important that the replica tables and secondary indexes in your global table have identical write capacity settings to ensure proper replication of data. If you want to add a new replica table to a global table, each of the following conditions must be true:

- The table must have the same partition key as all of the other replicas.
- The table must have the same write capacity management settings specified.
- The table must have the same name as all of the other replicas.
- The table must have DynamoDB Streams enabled, with the stream containing both the new and the old images of the item.
- None of the new or existing replica tables in the global table can contain any data.

If global secondary indexes are specified, the following conditions must also be met:

- The global secondary indexes must have the same name.
- The global secondary indexes must have the same partition key and sort key (if present). Write capacity settings should be set consistently across all of your global tables' replica tables and matching secondary indexes.

We strongly recommend that you enable auto scaling to manage provisioned write capacity settings. If you prefer to manage write capacity settings manually, you should provision equal replicated write capacity units to all of your replica tables. Also provision equal replicated write capacity units to matching secondary indexes across your global table.

Question-19: You are working in a company which provides the promotional sales for specific product based on the customer preference. For example, if you open mobile application with notification enabled every hour you would be notified new products available with the lower price than available in the market. And this promotion would be available only for 24 hours on lower price. As a BigData solution architect you decided to use DynamoDB table to store the promotional data and all the images would be stored in the DynamoDB table. You don't want to lost the data and should be available across two regions in United states for better performance. So you have one local copy of the data and two additional copy in different region. As there are around 20,000 product provider which are offering almost 0.5 millions of different products to the customer and you are expecting every second almost 3 new promotions are added. With this requirement what is the write capacity you need to provision if you do it manually and don't use the DynamoDB auto-scaling.

A. 36 WCU's

B. 9 WCU's

C. 18 WCU's

D. 20 WCU's

E. 6 WCU's

Answer: C

Exp: You need to understand the Global Replication of the DynamoDB table at the first place. When you use the DynamoDB global tables, you can replicate your table across various AWS pre-defined regions. As in the given question we need to do the same in two regions.

As we are not using AWS DynamoDB auto-scaling, we need to provision read/write capacity our own. The provisioned replicated write capacity units (rWCUs) on every replica table should be set to the total number of rWCUs needed for application writes across all Regions multiplied by two. This accommodates application writes that occur in the local Region and replicated application writes coming from other Regions. For example, suppose that you expect 3 writes per second to your replica table. In this case, you should provision 18 WCUs to each replica table (3 + 3 = 6; 5 x 2 = 12).

12 rWCU for two additional region. But we also have local table as well. If you enable on-demand mode on a global table, your consumption of replicated write request units (rWCUs) will be consistent with how rWCUs are provisioned. For example, if you perform 3 writes to a local table that is replicated in two additional Regions, you will consume 18 write request units (3 + 3 + 3 = 9; 9 x 2 = 18).

Question-20: You are working with an online market place where customer (shopkeeper, industries, retail customer etc. can look for the various products). You help them connect both supplier and customer and you also have premium subscription on the website which helps your subscriber to get more detailed analysis for the various requirement by the customer. Hence, this premium feature is more from the perspective of the self-data discovery platform which uses the AWS QuickSight. Using the QuickSight can help your premium subscriber to generate multiple analysis from the data even in various segments. Few of your premium subscriber is looking for a chart which show a measure for the intersection of two dimensions and also help them to color-code so that values can be easily differentiated based on the defined range. Which of the following can help with this requirement?

A. Bar Chart

B. Box plot

C. Tree Maps

D. Line Chart

E. Tabular reports

F. Heat Maps

Ans : F

Exp: This question is more regards with the Data Analytics knowledge and which visual chart best fit for the given requirement. Let’s understand each one detail

Tabular Reports: Table visuals can help you in finding the customized views. In the table visuals you can create a table and need to choose at least ne field of any data type, you can add as many columns as you want, even calculated columns can be part of the table.

Heat Map: Use heat maps to show a measure for the intersection of two dimensions, with color-coding to easily differentiate where values fall in the range. Heat maps can also be used to show the count of values for the intersection of the two dimensions.

Each rectangle on a heat map represents the value for the specified measure for the intersection of the selected dimensions. Rectangle color represents where the value falls in the range for the measure, with darker colors indicating higher values and lighter colors indicating lower ones.

Heat maps and pivot tables display data in a similar tabular fashion. Use a heat map if you want to identify trends and outliers, because the use of color makes these easier to spot. Use a pivot table if you want to further analyze data on the visual, for example by changing column sort order or applying aggregate functions across rows or columns.

To create a heat map, choose at least two fields of any data type. Amazon QuickSight populates the rectangle values with the count of the X axis value for the intersecting Y axis value. Typically, you choose a measure and two dimensions.

Line Chart: You should be using line charts to compare changes to measure values over period of time, for example few of the scenarios are below:

- Sales growth of a particular product over the time
- Gross sales based on time period like every month.
- One measure for a dimension over a period of time, for example number of product delivery delays per day by courier company.

Line charts: show the individual values of a set of measures or dimensions against the range displayed by the Y axis. Area line charts differ from regular line charts in that each value is represented by a colored area of the chart instead of just a line, to make it easier to evaluate item values relative to each other.

Tree Maps: To visualize one or two measures for a dimension, use tree maps.

Each rectangle on the tree map represents one item in the dimension. Rectangle size represents the proportion of the value for the selected measure that the item represents compared to the whole for the dimension. You can optionally use rectangle color to represent another measure for the item. Rectangle color represents where the value for the item falls in the range for the measure, with darker colors indicating higher values and lighter colors indicating lower ones.

Question-21: You are working in a HighValue Retail Bank which offers various products to the customer including retail load, credit cards, certificate of deposits, current account, credit account etc. Which has their establishment in more than 20 countries and having 36 million of customer globally. To acquire more customer HighValue bank signed an agreement with an eCommerce giant which has presence in more than 5 countries and over 50 million of customer purchasing products annually.

Your technical team already written many Java based applications which are running on EC2 hosts in the AWS public cloud to serve the various business functionality to the customer. Bank is also looking for the Click Stream behavior of the customer on the e-commerce website and collect the data and while collecting the data using AWS Web Service solutions, they wanted to make sure

- ClickStream records must be accumulated so that stream throughput can be increased.
- In existing Web application there should not be so much needs to be applied.
- On the consumer side you should be able to de-aggregate the records which are already aggregated at producer side.
- And once data retrieved that should be stored permanently in the DynamoDB for historical analysis.

Which of the following Kinesis Streaming solution can help for the given requirement?

A. On the existing application to at the producer side use the KPL (Kinesis Producer Library) to make necessary changes, so that data can be aggregated using the batching and at the client side using KCL De-aggregate the records.

B. Use the Kinesis Data Stream API and by looking at the record contents aggregate the records based on specific properties. And de-aggregates the records using the Kinesis Client Library.

C. You should use the Kinesis Agent to pre-process the data, aggregate them in a batch and using the Kinesis Data Stream de-aggregate the collected records.

D. You should use the Kinesis Client Library to store the data in DynamoDB.

E. You should use the Kinesis Connector library to de-aggregate and write the data in DynamoDB

Answer: A, E

Exp:

Option-1: Yes, we can use Kinesis Producer Library to create batch and collector for the streaming data.

Option-2: Kinesis Data stream API cannot look into the data content.

Option-3: Agent can not be used to pre-process the data.

Option-4: Kinesis Client Library is not a correct solution for storing data in DynamoDB

Option-5: Yes, Kinesis Connector is the write solution for de-aggregate and write data in the DynamoDB.

Please go through below for reading each component use.

Kinesis Agent: is a stand-alone Java Software application using this you can collect the data and send it to Kinesis Data Streams. Agent continuously monitors the set of files (for example log files of your application) and can send the data to your stream. It is the responsibility of the agent to handle file rotations, checkpointing, and when there is a failure it can re-try. Agent can also emit the CloudWatch metric to help and better monitor and troubleshoot the streaming process.

Kinesis Connector: Using the Kinesis Connector Library you can integrate Amazon Kinesis with the other AWS and NON-AWS services. As of now this connector provides the help for Amazon DynamoDB, Amazon Redshift, Amazon S3, Elasticsearch etc.

Consumer with the Kinesis Data stream API and SDK: Again using the Java. You can get the data from the Kinesis Data Stream.

Producer using Kinesis Data Stream API: You can develop producers using the Amazon Kinesis Data Streams API with the AWS SDK for Java. However, for most use cases, you should prefer the Kinesis Data Streams KPL library. Once a stream is created, you can add data to it in the form of records. A record is a data structure that contains the data to be processed in the form of a data blob. After you store the data in the record, Kinesis Data Streams does not inspect, interpret, or change the data in any way. Each record also has an associated sequence number and partition key. There are two different operations in the Kinesis Data Streams API that add data to a stream, PutRecords and PutRecord. The PutRecords operation sends multiple records to your stream per HTTP request, and the singular PutRecord operation sends records to your stream one at a time (a separate HTTP request is required for each record). You should prefer using PutRecords for most applications because it will achieve higher throughput per data producer. For more information about each of these operations.

Kinesis Client Library: You can develop a consumer application for Amazon Kinesis Data Streams using the Kinesis Client Library (KCL). Although you can use the Kinesis Data Streams API to get data from a Kinesis data stream, we recommend that you use the design patterns and code for consumer applications provided by the KCL.

The Kinesis Client Library (KCL) helps you consume and process data from a Kinesis data stream. This type of application is also referred to as a consumer. The KCL takes care of many of the complex tasks associated with distributed computing, such as load balancing across multiple instances, responding to instance failures, checkpointing processed records, and reacting to resharding. The KCL enables you to focus on writing record-processing logic.

Kinesis Producer Library:

Batching refers to performing a single action on multiple items instead of repeatedly performing the action on each individual item.

In this context, the "item" is a record, and the action is sending it to Kinesis Data Streams. In a non-batching situation, you would place each record in a separate Kinesis Data Streams record and make one HTTP request to send it to Kinesis Data Streams. With batching, each HTTP request can carry multiple records instead of just one.

The KPL supports two types of batching:

- Aggregation – Storing multiple records within a single Kinesis Data Streams record.
- Collection – Using the API operation PutRecords to send multiple Kinesis Data Streams records to one or more shards in your Kinesis data stream.

The two types of KPL batching are designed to coexist and can be turned on or off independently of one another. By default, both are turned on.

Question-22: You are working with a company which does the consulting for CSR (Corporate Social responsibility) and various NGS's and government organization. Which is having data from last 15 years in various formats like CSV, XML, JSON, eMail, modern formats like Parquet, Avro etc. You have been hired as BigData solution architect for collecting all this data in a central repository and build a service so that user can make queries on your data. And visualize the data in using various charting solution. And you can charge to customer for each query they fire. What would you use and why for creating a data lake solution. Please select the correct option from below.

A. You would be parsing all these data and store in AWS RDS so that user can write SQL queries on the data.

B. You would transform this data and then store in the AWS Datawarehouse solution redshift and user can write query to fetch the data and also you can use QuickSight for visualizing these data.

C. You would be storing this data in S3 and create a Data Lake and then you would be using Athena, AWS QuickSight, and AWS Data Glue.

D. You would be storing this data in DynamoDB and create a Data Lake and then you would be using Athena, AWS QuickSight, and AWS Data Glue.

Answer: C

Exp: As this question is specifically asked for creating a Data Lake solution and building an interface from which user can write SQL queries to fetch the data. While data extraction he can use QuickSight to create visualization for the data. Read below to understand more about AWS Athena.

Querying data which is stored in Amazon S3 using Athena: Athena is an interactive query service that makes it easy to analyze data directly in Amazon Simple Storage Service (Amazon S3) using standard SQL. With a few actions in the AWS Management Console, you can point Athena at your data stored in Amazon S3 and begin using standard SQL to run ad-hoc queries and get results in seconds.

Athena is serverless, so there is no infrastructure to set up or manage, and you pay only for the queries you run. Athena scales automatically—executing queries in parallel—so results are fast, even with large datasets and complex queries.

Structured data : CSV, JSON, Columnar Data Format, Parquet, ORC, Avro these all data is considered highly structured data.

Unstructured data: HTML, XML, Word Doc, eMail etc are considered semi structured data.

Using Athena-adhoc you can query both type of data using standard SQL and no need to load or aggregate this data in the Athena.

If you want to create visualization for the data stored in the S3, use the QuickSight. You can use Athena to generate reports or to explore data with business intelligence tools or SQL clients connected with a JDBC or an ODBC driver.

Athena integrates with the AWS Glue Data Catalog, which offers a persistent metadata store for your data in Amazon S3. This allows you to create tables and query data in Athena based on a central metadata store available throughout your AWS account and integrated with the ETL and data discovery features of AWS Glue.

Question-23: You are working with an eCommerce giant named EcoDeco.com your IT team had developed some applications which collects the data from the website clickstream. And this clickstream, collects the data using Kinesis Data Stream which is having various shards created based on the Data Category. It is running since last one year and you recently joined as BigData architect and found from the CloudWatch metric generate that the there are various shards which are either having lot more data and facing performance issues and some shard are getting 20% data from their capacity. You have to improve the performance as well as reduce the cost, how can you achieve the same?

A. You would identify the hot shards based on the CloudWatch metric and split these hot shards based on the hash key.

B. Split the shards which are getting more data than the capacity defined.

C. You would be merging the cold shards which are getting the less data then defined capacity.

D. You would be creating new shards based on the CloudWatch metric and configure the same and delete the existing shard.

E. You would be creating new shards in the different region based on the Cloudwatch metric and then delete the existing shard.

Ans : A, B,C

Exp : Creating new shards is not a good solution rather you should split the existing shard for increasing the capacity of use the shard merging strategy for the shards which are getting data then defined capacity and with this you can optimize the cost and help in reducing the cost. The purpose of resharding in Amazon Kinesis Data Streams is to enable your stream to adapt to changes in the rate of data flow. You split shards to increase the capacity (and cost) of your stream. You merge shards to reduce the cost (and capacity) of your stream.

One approach to resharding could be to split every shard in the stream—which would double the stream's capacity. However, this might provide more additional capacity than you actually need and therefore create unnecessary cost.

You can also use metrics to determine which are your "hot" or "cold" shards, that is, shards that are receiving much more data, or much less data, than expected. You could then selectively split the hot shards to increase capacity for the hash keys that target those shards. Similarly, you could merge cold shards to make better use of their unused capacity.

You can obtain some performance data for your stream from the Amazon CloudWatch metrics that Kinesis Data Streams publishes. However, you can also collect some of your own metrics for your streams. One approach would be to log the hash key values generated by the partition keys for your data records. Recall that you specify the partition key at the time that you add the record to the stream.

Question-1: You are working in a large Electronic House Hold provider company, which had already installed servers on each electronic device to collect data on every 10 mins, the overall data collected every hour is around 4TB. You need to immediately analyze this data and inform the owner of the device if any functional/non-functional issue found on the device, what is the possible solution you can use from below. Also, assume that devices are spread across entire country.

A. You would be collecting this data in AWS S3 bucket using the Apache Spark.

B. You would be collecting this data using Kinesis Stream in the AWS Redshift cluster.

C. You would be using AWS Snowball Edge as well as AWS Lambda.

D. You would be using AWS EMR service

Answer: C

Exp: As you can see in the question, we need to collect the data from various devices across the country the ideal solution for this is AWS IOT, however this is not given in the option. So we need to choose the solution provided in the option. AWS Snowball Edge can be used to move large amounts of data into and out of AWS, as a temporary storage tier for large local datasets, even it can support local workloads in remote or offline locations. Now question is also asking to analyze the data every hour, which we can do using the AWS Lambda code on the Snowball Edge to perform the tasks such as analyzing data streams or processing data locally.

Question-2: You are working in a company which runs the Video Streaming solution and you wanted to collect all the user data while he watches the video to further analyze his experience and wanted to generate some reports by running SQLs on the collected data in real-time. Which of the following service is well suited for collecting this data and generating the report using SQL?

A. You would be using AWS Kinesis Data Stream

B. You could use the Kinesis Video Stream

C. You would be using only Kinesis Data Firehose

D. You would be using Kinesis Data Firehose and Kinesis Data Analytics

Ans : D

Exp : You may get confused with the Video Streaming application, but here we want to collect the user data and not the streaming videos, hence we have don't have to use Kinesis Video Stream. As we need to collect the data and apply the analytics using SQL to generate the report we can use the Kinesis Data Analytics for analyzing and streaming data using the Kinesis Data Analytics. However, we need to deliver the data to the Kinesis Data Analytics we need to use the Kinesis Data Firehose, which can send the data to any of this AWS services like S3, Redshift, Kinesis Analytics, Elastic Search Service etc.

Question-3: You are working with a company which has an e-commerce website, which continuously generates the web clickstream data which you wanted to store in the Redshift cluster, but before sending to the Redshift cluster you want to process data using your custom complex logic. Which of the following is a better solution for the given requirement?

A. You would be using Kinesis Firehose to deliver the data in the redshift cluster.

B. You would be using Lambda (For custom logic) and Kinesis Firehose to deliver the data to Redshift cluster.

C. You would be using Kinesis Stream and Kinesis Client Library and before sending the data to redshift cluster you would configure the custom logic to process the data.

D. You would be using Amazon SQS to send the data Redshift cluster and using the AWS Lambda you would be processing the data.

Answer: C

Exp: In this question you need to understand the difference between Kinesis Firehose and Kinesis Data Stream. See the below differences to understand further. Both Kinesis Client library and Kinesis Firehose helps in ingesting data in S3, Redshift, Elastic Search, EMR, and AWS Lambda. Then what exactly is the difference in which scenario we should use which one. Few differences are below based on which you have to select correct answer

- Firehose is fully managed, scales automatically and stream needs to be manually managed
- In Kinesis Stream applications are build using the Kinesis Producer Library which put the data into a stream and then process it with application that uses the Kinesis Client Library and using the Kinesis Connector Library which send the processed data to S3, Redshift and DynamoDB etc.
- With the Kinesis Firehose it is simple, where we need to create the delivery stream and send the data to S3, Redshift etc. And you should have Kinesis Agent or API for that.
- Kinesis data stream can keep data for 7 days; hence it can be used as a storage as well. Which helps in custom processing before ingesting data to S3, Redshift or Elastic Search.
- Kinesis data stream is open-ended service at both the end on the producer side you will be configuring data producer to write the data in the Kinesis Stream, and this service will store your data in a continuous manner and able to replay as well, and order would be retained and on the other side, we would be configuring the data consumer to read the data out of the stream and process it with the custom application. Kinesis data stream is a data storage system, it is more

flexible and you can build your custom application as you want, even you have full control how to partition your data, how many shards you want to have for your particular stream.

- **Kinesis Firehose**: It is an open ended only one side, you configure the you configure data producer to continuously push data into the Firehose Delivery Stream and on the other side you don't read the data from FireHose delivery stream and you don't write any application for that. Firehose automatically deliver the data to your destination like S3, Elastic Search, Redshift cluster etc.

Question-4: You are working in an e-commerce company where each click on the website should be captured and needs to be analyzed in the real time using the SQL. However, you have to make sure that the evens which are generated in real time should be copied in the multiple availability zones, and you always need to keep track what has be processed till now, in case of failure you need to start again and should not process already processed clicks, which of the following option can help in achieving such requirement?

A. You would be using AWS S3 and Redshift cluster

B. You would be using EMR with the Spark streaming

C. You would be using Kinesis Firehose and DynamoDB

D. You would be using Kinesis DataStream and DynamoDB

E. You would be using Kinesis FireHose and Kinesis Data Analytics

Ans :D

Exp: As you can see in the question it is asking following important points

- Real time data should be available in Multiple Availability zone, which can be achieved using Kinesis data stream, because by default it keeps the data in 3 AZ in a AWS region.
- Generating SQL report can be done using Kinesis Data Analytics
- In case of failure we should be able to restart the stream from where it failed, it can be achieved using the DynamoDB. Because it stored the cursor in DynamoDB and in case of failure it uses the cursor to find the last processed events and move further.

Question-5: You are working in an eCommerce company, who wanted to analyze the customer behavior in real time, your data scientists have given the format in which they are looking for the data, so that they can execute the SQL query on the data. To covert click stream data in specific format you have already written some custom code, which of the following is a good solution for implementing this requirement?

A. You would be using Kinesis Data Firehose and Redshift Cluster as destination

B. You would be using Kinesis Data Stream and DynamoDB as a destination and in between you would be using your custom logic to transform the data in required format.

C. You would be using Kinesis Data Stream and Amazon RDS as a destination and in between you would be using your custom logic to transform the data in required format.

D. You would be using Kinesis Data Stream and Kinesis Data Analytics and in between you would be using AWS Lambda which would have your custom logic in between to transform your data in required format.

Answer: D

Exp: As we need to collect the data in real-time, hence we can use either Kinesis Data Stream or Kinesis Firehose. However, we need to apply custom logic, so we can use the AWS Lambda to submit that custom logic and transform the data before submitting to the Kinesis Data Analytics. Because using the Kinesis Data Analytics data scientists can apply the SQL queries on the received data in real time. And the result of SQL queries can be directly saved in designated storage like S3

Question-6: You are working with a company which are having daily batch jobs to analyze the data of the website user, they are getting millions of users on their website on daily basis. And in the batch job they are doing some data cleanup, transformation and finally generate well formatted data. On this data some SQL queries are issued as part of the batch job to generate the reports. It is not necessary you would be running this batch processing on daily basis, sometime you may not run this batches, however whenever you run the data volume is around 4TB, which of the good solution for implementing this requirement

A. You would be using AWS Kinesis Firehose, Kinesis Data Analytics and AWS Lambda

B. You would be using S3 as a data storage and then create a EC2 cluster to process the data from the S3 bucket.

C. You would be storing this data in the AWS DynamoDB and AWS Lambda on the data whenever required.

D. You would be creating EMR cluster, and whenever jobs needs to be initiated you would be adding tasks nodes and as soon as processing finished you would be removing the task nodes.

Answer: D

Exp: EMR cluster is a good solution for processing huge volume of data. It is basically a Hadoop cluster where you can add or remove nodes as need basis, even if you don’t need EMR cluster then terminate it and whenever needed again you would be spin up again, as more tasks needs to be completed then add more tasks nodes and when the tasks or batch finished remove the task nodes. With the EMR, you don’t need to guess your future requirements or provision for peak demand because you can easily add or remove capacity at any time.