===== ASSIGNMENT ======

One of your clients wants to migrate their on-premise data ingestion solution to AWS cloud. The existing solution is implemented in a single linux server. Thousands of files ranging from size of 1KB to few GBs are fed into the server throughout the day. The incoming files are encrypted and compressed. The solution needs to decrypt and then unzip the file before storing the data into the storage location. The average time to process each file ranges from a few seconds up to 30 minutes. Due to the increase in the number of files and the varying file sizes to be processed, the single linux server is becoming the bottleneck. They expect the number of files to be increased by 10 folds in the next 6 months and looking for a cost-effective, highly available, scalable and secure solution for their data ingestion. How would you design this solution?

Please provide 2 different architecture designs and we will discuss them during the interview.

===== SOLUTIONS ======

We will make a couple of assumptions by asking questions and providing answers to those questions. That will help us slightly cover the “requirement analysis” phase.

1. Is the current system that generates the files also on-premise?
    a. **Yes**. Thus AWS DataSync or S3 API will need to be implemented on the on-premises servers to move the files to S3, along with an eventual one-time file migration to S3.
2. Does the ingestion average time include transformation to be ready for data analysis (SQL, NoSQL, or columnar friendly)?
    a. **Yes**. Thus one of the services: EMR, RedShift, Athena, DynamoDB Kinesis Analytics, Spectrum are needed in the solution.

We will consider building 3 blocks or steps as the implementation of the data ingestion migration (1) Moving files to AWS/S3, (2)Ingest the datafiles(Decrypt,unzip, eventual cleanup), and (3)Process the actual files. Those 3 steps will be considered in both designs. One important point to note in the solution, due to the execution time limit of lambda functions (15 minutes max) and the constraint of file size that could take up to 30 minutes (per the requirements), we also need to use AWS Batch for any big files that require more than 15 minutes of processing.

The solution has been designed under the AWS well-architected pillars below:

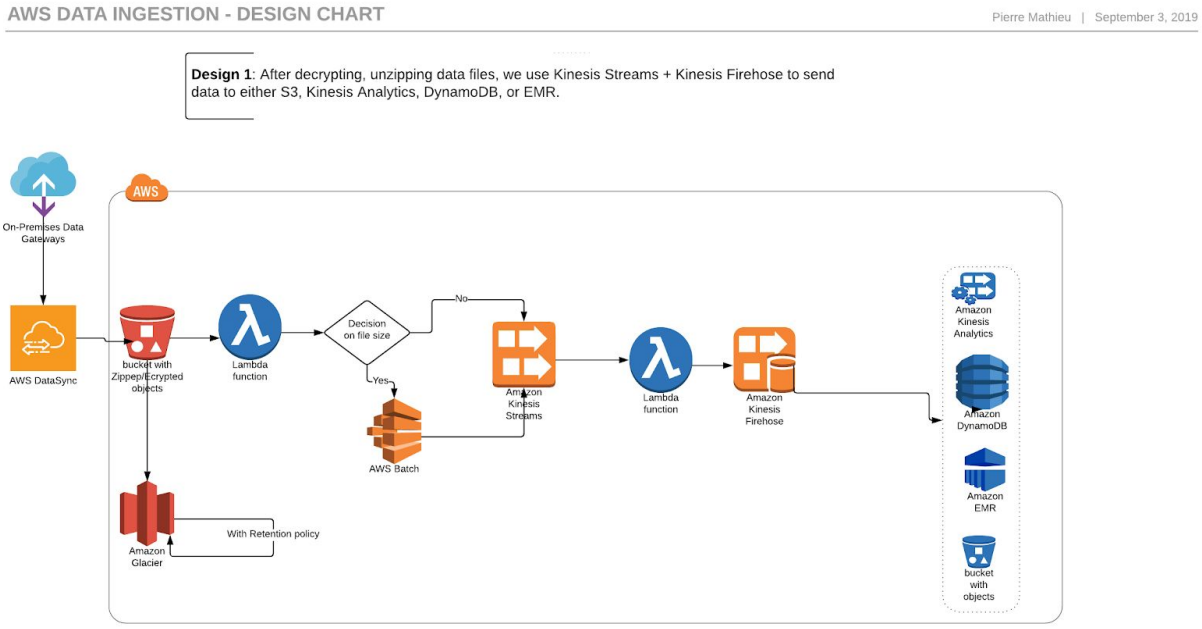
- To ensure the “Operational Excellence” of the provided solution, we will build the entire workload constituted of applications, infrastructure, operations, governance and policies, as code, by using Terraform or Cloudformation to create version-controlled templates of the infrastructure. Also, we will implement CI/CD pipeline to manage the deployment of new code and docker image versions of the AWS/BATCH by running Jenkins or Code Deploy/CodePipeline along with Packer (to rebuild ECR components) .
- To fully control the security of the solution, we will create IAM Roles, assumed roles and policies, and separation of duties by implementing the principle of “least privilege”. Also, proper security groups will be created for

each component. While making sure the solution will be built within a secure VPC with private subnets when it is necessary, all the public accessible services (for instance S3) will use encryption at rest and in-transit during data migration. We will monitor access to direction connections and API calls by enabling CloudTrail and VPC Flow logs for auditing.

- By choosing a serverless architecture (mostly), the system will be highly available where the data will be stored in S3, the code will be running with Lambda function and AWS Batch(with Images stored in ECR), Streaming and Analytic tools will be using serverless architecture.
- The solution is highly scalable as we have chosen services that are easy to scale : S3 unlimited space, Kinesis, Lambda, Glue are serverless, and EMR , AWS/Batch, Redshift are running on scalable clusters.
- For cost effectiveness, a more detailed analysis needs to be done to better understand the client's objectives. However, we know for sure S3 will be very affordable with all the options (S3 Standard Storage, S3 Standard-Infrequent Access, and S3 One Zone-Infrequent Access) where the cost can be as low as $0.01/GB/Month. Also all the AWS service are in the "Pay-for-what-you-use", knowing the number of transactions and file transfer (rate and size) per month will help calculate and readjust the solution to optimize the ratio service/cost with the help of the cost explorer.

Below are the two designs we have provided for the Data Ingestion Migration to AWS.

Design 1: After decrypting, unzipping data files, we use Kinesis Streams + Kinesis Firehose to send data to either S3, Kinesis Analytics, DynamoDB, or EMR.



Design 2: After decrypting, unzipping data files, we use AWS Glue to make data available to query via either Athena or Redshift.