SSG

IMAGE PROCESSING AND VISUALIZATION ARCHITECTURE

GRANTS DATA MODEL

SECTION 1: SYSTEM DESIGN

You are designing data infrastructure on the cloud for a company whose main business is in processing images.

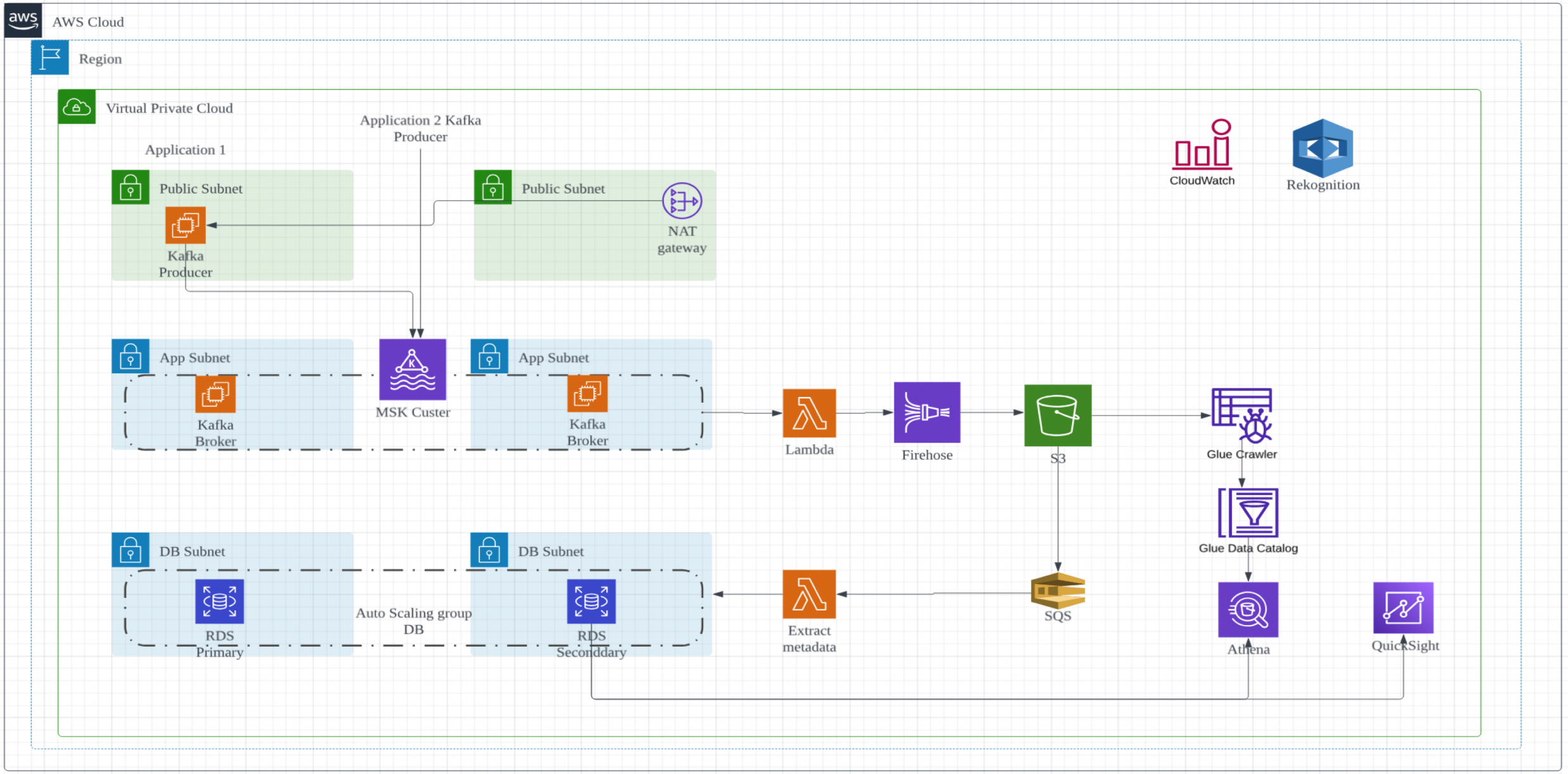
The company has a web application which collects images uploaded by customers. The company also has a separate web application which provides a stream of images using Kakfa stream. The company's software engineers have already written the code to process the images. The company would like to save processed images for a minimum of 7 days for archival purposes. Ideally, the company would also want to be able to have some Business Intelligence (BI) on key statistics including number and type of images processed, and by which customers

Produce a system architecture diagram in powerpoint using any of the commercial cloud provider's ecosystem to explain your design. Please also indicate clearly on assumptions made at any point.

ASSUMPTIONS

- AWS could is used in the company.
- Both applications sending kafka messages are deployed in same VPC or else VPC peering/VPN connection has been setup with Kafka cluster.
- Currently only metadata is being extracted such as image size, image dimensions name, storage location along with user and image keys etc.
- Using step function a separate workflow can also be set using lambda function and AWS Rekognition to resize image, detect objects, label/tag images and later stored along with the metadata for further analyses by business users via Quicksigh, Tableau etc.
- All the required policies, permission and execution role has been setup for MSK, lambda and other services used.

ARCHITECTURE DIAGRAM



CONSIDERATIONS

- Event-driven architecture is suggested for cost saving and scalability. Trigger Lambda functions based on events from Kafka topics to process and store metadata.
- Use of auto-scaling groups, lambda, kafka and SQS make the architect fault-tolerant.
- User authentication and authorization in place for business users accessing data via Athena or Quicksight.
- Data is encrypted both at rest in storage and at transit via kafka messages.
- Using RDS currently for metadata storage and BI access but depending on business requirements further complexity and estimated data volume, Redshift can be used as a data warehouse.
- IM8 policies and PII details is taken care of at both infrastructure level and application level.
- SQS and Firehose is used to manage number of lambda trigger at a time. This will help to save the cost and keeping the lambda execution under limit.
- Used Glue catalog and crawler for ad-hock analyses by BI users using Athena.
- Lifecycle management for S3 is set up for 7 days and will be sent to cold storage or can be archive if required later. Data from DB will be archived too.
- Used cloudwatch for proper monitoring and logging.

SECTION 2: DATA MODEL

SkillsFuture Singapore (SSG) drives and coordinates the implementation of the national SkillsFuture movement, promotes a culture and holistic system of lifelong learning through the pursuit of skills mastery, and strengthens the ecosystem of quality education and training in Singapore. We plan and implement initiatives for individuals, employers and training providers.

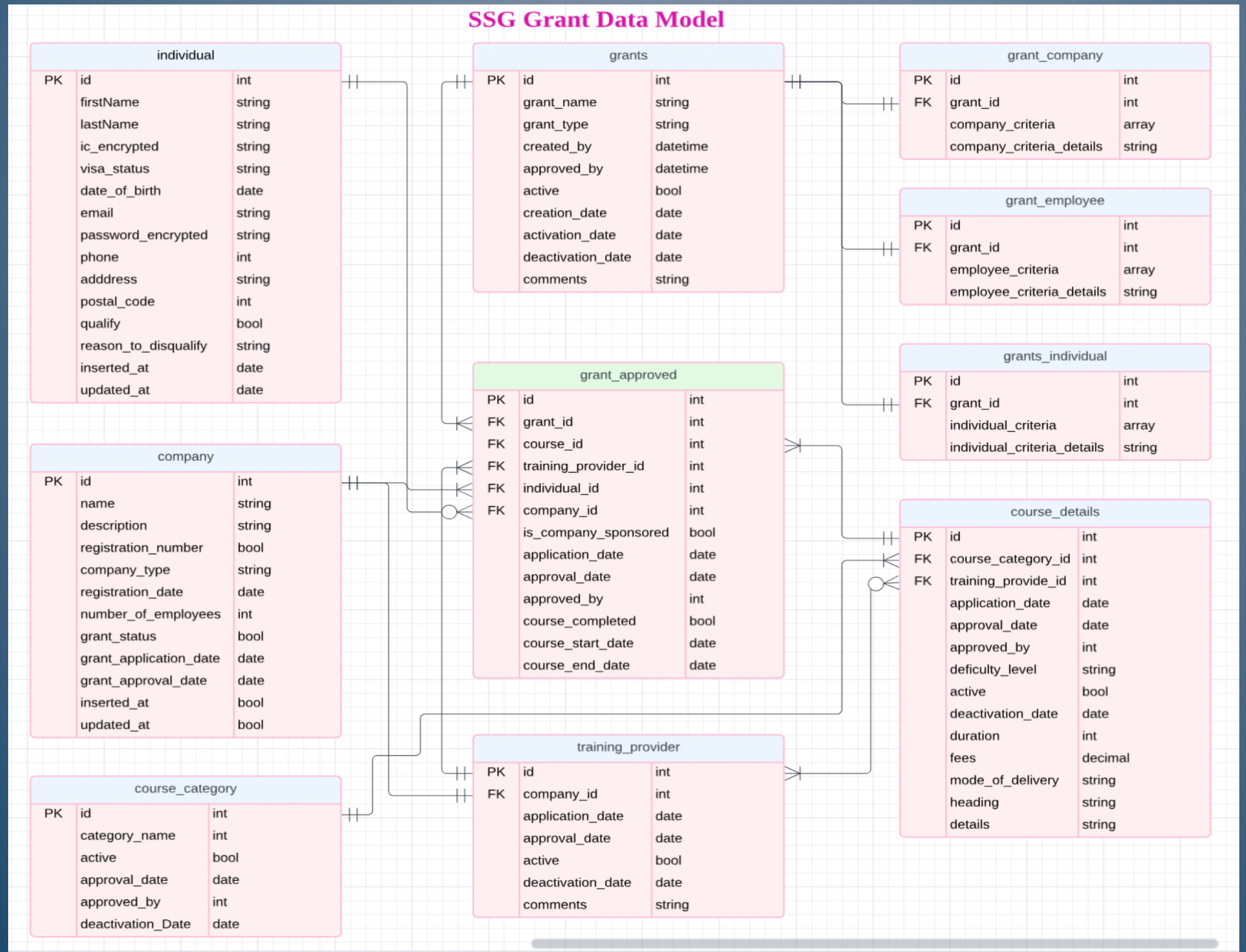
Individual may attend courses and consume SSG grants.
Employers may send their employees for training and be subsidised for doing so.
Training providers run courses and provide training.

Please propose a logical data model for SSG's business which will help to achieve the outcomes listed above.

1. Please indicate key entities and the properties of such entities
2. Please indicate any relationships between the entities and propose suitable keys in order to map relationships

Please also explain the thought process behind your data model (why these entities and properties are key), indicate what considerations would be important and any potential pitfalls or areas to watch out for.

GRANTS - LOGICAL DATA MODEL



CONSIDERATIONS

- SSG employee table (as an approver) is not added in data model.
- Referential data such as mode_of_delivery, company_type etc. is not added.
- An employee is also an individual therefore a separate employee table not mentioned but can be differentiated while enrolment if individual is sponsored by a company or not.
- Grants requirements is captured in separate tables for each company, employee and individual and the details will be captured as a key-value pair in JSON format such as age, visa status, funding type, amount etc.
- Companies, grants, courses and training provides are approved and removed time-to-time therefore their status and relevant action dates are captured accordingly. Rest of the properties in tables are self-explanatory however tables and properties could be added more and get more complex based on the business requirements.

POINTS TO LOOK FOR

- As employee can switch companies or could be self-employed due to which grants may vary with time. Therefore grant and employee details is relevant at the time of each application. Maintaining history in this scenario is very important.
- Master data, referential data, proper quality checks should be in place.
- There are tables containing PII information therefore relevant columns must be encrypted or masked. Proper authentication and authorization must be setup on both database level and application level between SSG data owners, data contributors and data stewards as per IM8 guidelines.
- There could be circular dependency as - an individual could be an employee, a company could be a training provide, SSG grant approvers(employees) could be an individual who also take courses.
- In future, with more people enroll in multiple courses, a proper lakehouse setup is important which is not only optimize reads by allowing to create data marts for specific BI users but also provide a raw and unstructured data(if any) for data science team for further analysis and predictions.