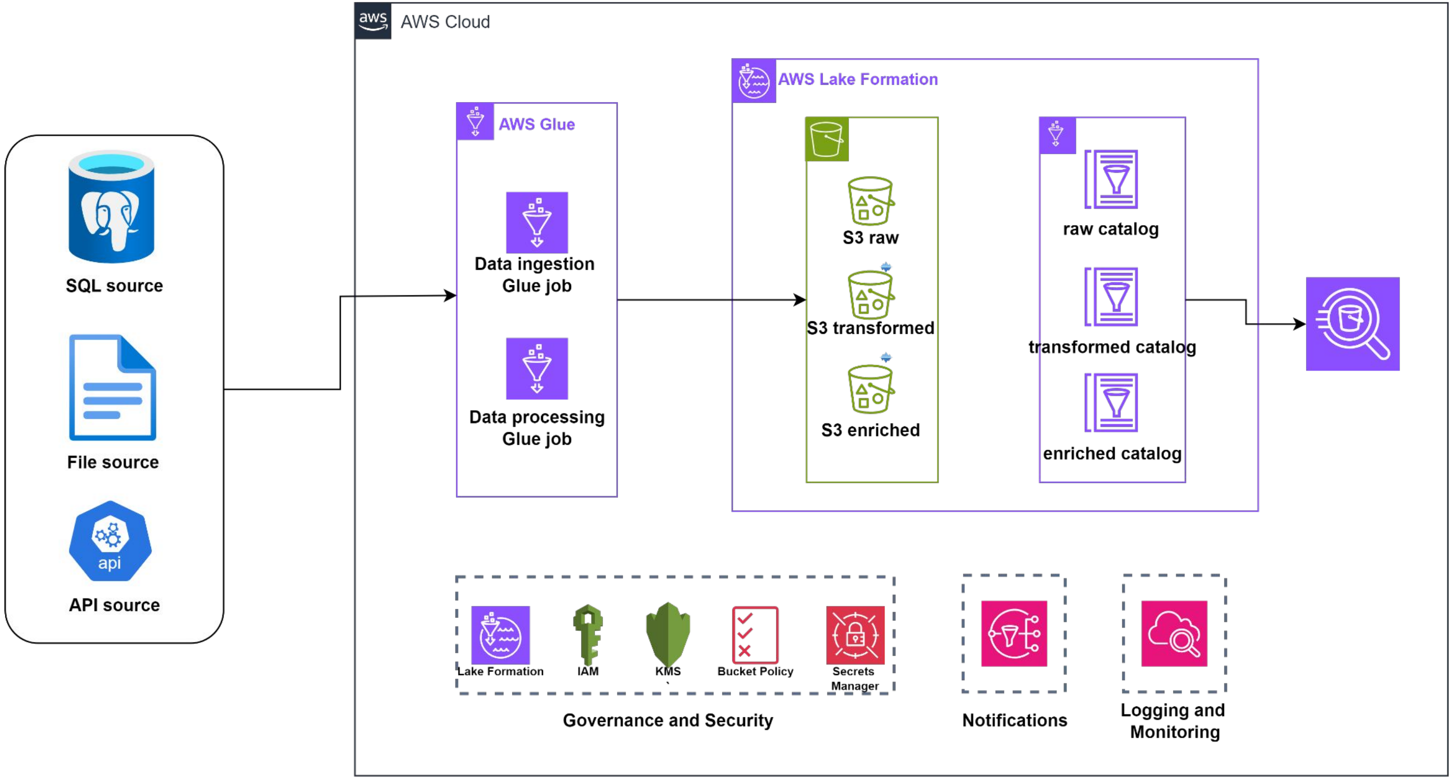
Data ingestion to AWS

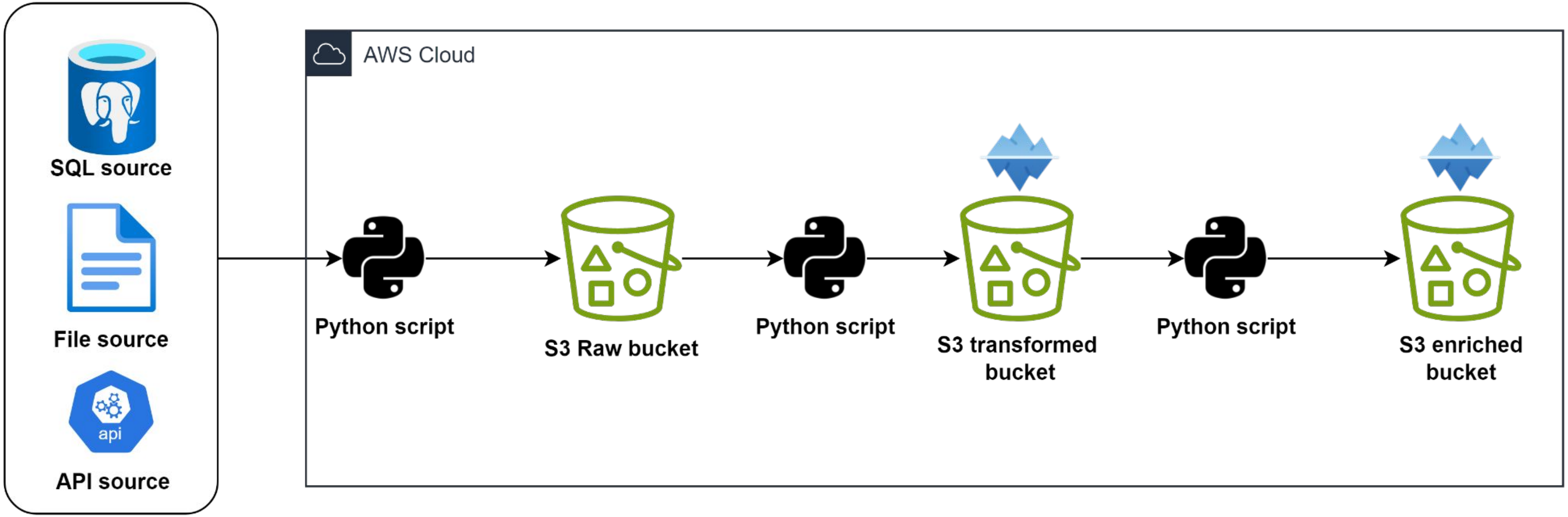
AWS architecture



Services Used

1. **S3**: Forms our data lakehouse. Can scale infinitely
2. **Glue**: Used for Data ingestion, processing. Supports multiple sources and targets. Can scale up and down based on requirement. Managed service
3. **Lake Formation**: Manages Access control of Data Lake resources to users
4. **Glue Data Catalog**: Stores metadata of S3 layers
5. **Athena**: Serverless Query engine to query the data lake
6. **IAM**: Manage User/Role Permissions
7. **KMS**: Encrypt S3 data
8. **Secrets Manager**: Store confidential data like API keys, DB username, passwords
9. **SNS**: Send job run status notifications to stakeholders
10. **Cloudwatch**: Analyse and monitor logs

Data Flow


SQL source
File source
api
API source
AWS Cloud
Python script
S3 Raw bucket
Python script
S3 transformed bucket
Python script
S3 enriched bucket

Data Flow

1. Load data from all sources into the S3 raw layer as part of ingestion. The raw layer structure will be an exact copy of the source, so that this layer can act as a source during backfills. External tables are created here, to query the raw data.
2. Read the raw data, perform transformations and load them to transformed layer as part of processing. External tables are created here, to query the transformed data.
3. Read the transformed layer, perform aggregations, denormalizations, joins based on business rules and load to enriched layer. External tables are created here, to query the enriched data.