RECORD LINKAGE PIPELINE AND DATA LAKE

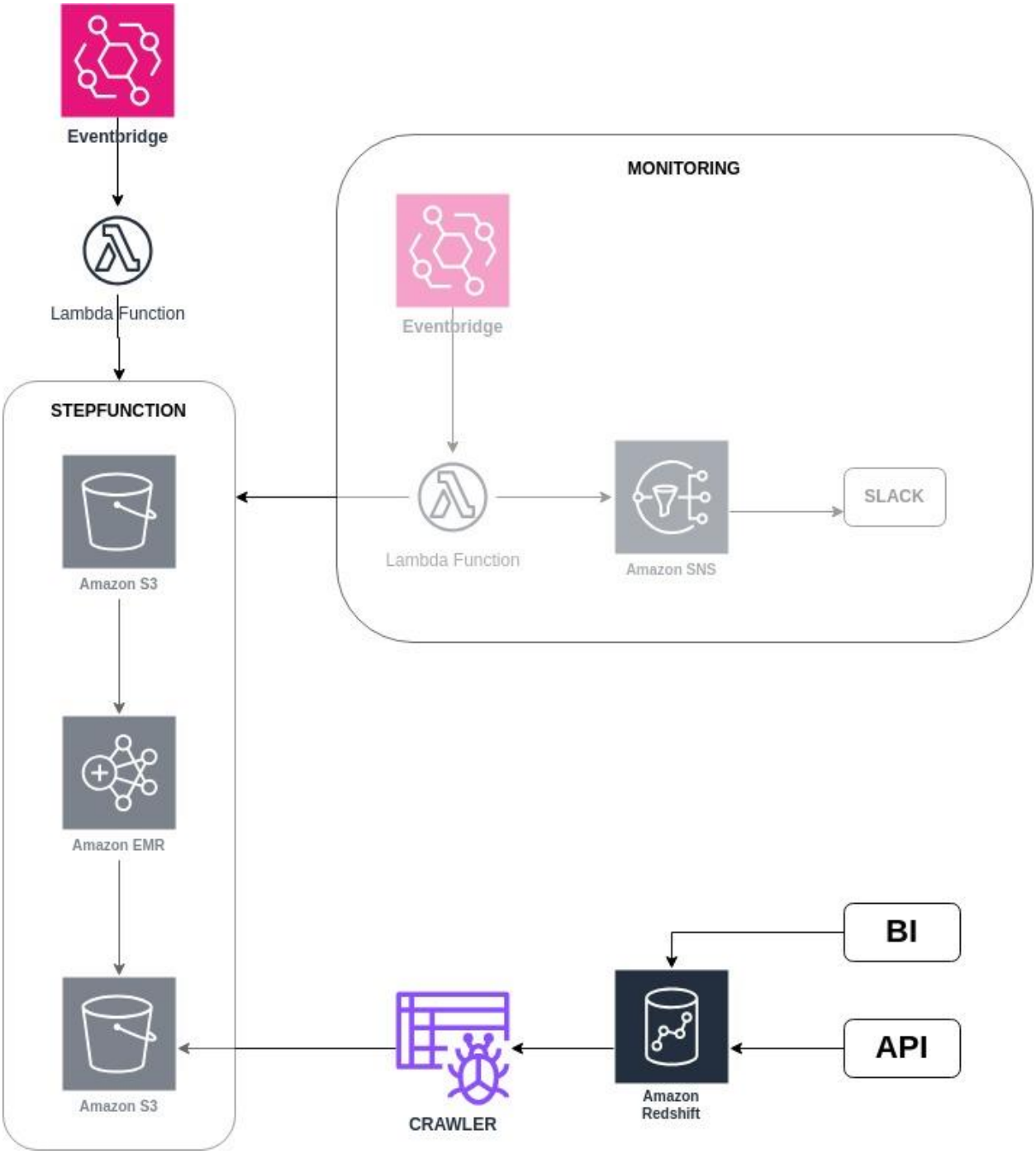
USE CASE:

Product catalogs manually created and uploaded to pos system, creates multiple versions of products and their attributes, making analysis difficult. Record linkage pipeline to dedupe the product data, pipeline and metrics to evaluate the deduping, and monitoring for the pipeline. Result vastly cleaner data forming the foundation of data lake insights then marketed to retailers, distributors, brands, hedgefunds. Focus on the final production pipeline in below description.

NOTE:

metrics assessing deduping (pairwise accuracy metrics + compression for clustering evaluation.)

ARCHITECTURE DIAGRAM:



END TO END DESCRIPTION OF DIAGRAM:

- Eventbridge triggers lambda with python boto script to start stepfunction on schedule.
- Stepfunction creates emr cluster using bootstap file saved in s3 bucket.
- Emr runs pyspark code saved in s3 bucket.
- Pyspark code reads in parquet files exported previously from transactional databases to s3 bucket and dedupes product data. Pyspark code writes deduped product data back to s3 bucket.
- boto3 code at end of pyspark script triggers glue crawler to crawl parquet files and infer schema for glue datacatalog so that redshift has updated schemas for tables to query.
- BI tools and API run queries in redshift spectrum against data stored in s3.
- For monitoring separate eventbridge triggers lambda on schedule to run boto3 code and check for failed, timed out, or exceptionally long running stepfunctions.
- Email with sns to slack to post alert for any stepfunctions meeting criteria.

  NOTE: Deployed eventbridges and lambdas using ci/cd, gitlab, serveless yaml files.