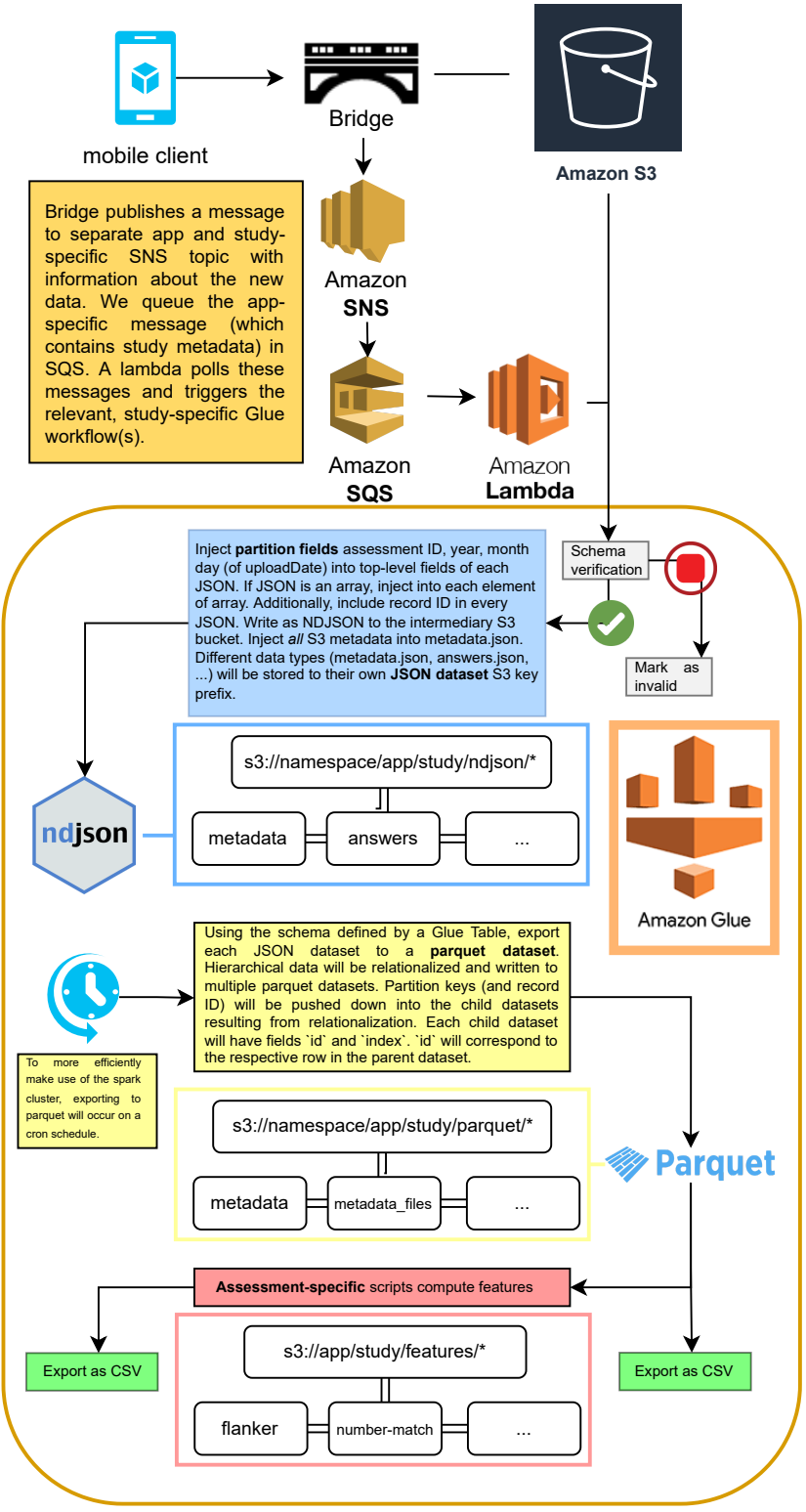
mobile client
Bridge
Amazon S3
Bridge publishes a message to separate app and study-specific SNS topic with information about the new data. We queue the app-specific message (which contains study metadata) in SQS. A lambda polls these messages and triggers the relevant, study-specific Glue workflow(s).
Amazon SNS
Amazon SQS
Amazon Lambda
Inject partition fields assessment ID, year, month day (of uploadDate) into top-level fields of each JSON. If JSON is an array, inject into each element of array. Additionally, include record ID in every JSON. Write as NDJSON to the intermediary S3 bucket. Inject all S3 metadata into metadata.json. Different data types (metadata.json, answers.json, ...) will be stored to their own JSON dataset S3 key prefix.
Schema verification
Mark as invalid
ndjson
s3://namespace/app/study/ndjson/*
metadata
answers
...
Amazon Glue
Using the schema defined by a Glue Table, export each JSON dataset to a parquet dataset. Hierarchical data will be relationalized and written to multiple parquet datasets. Partition keys (and record ID) will be pushed down into the child datasets resulting from relationalization. Each child dataset will have fields `id` and `index`. `id` will correspond to the respective row in the parent dataset.
To more efficiently make use of the spark cluster, exporting to parquet will occur on a cron schedule.
s3://namespace/app/study/parquet/*
metadata
metadata_files
...
Parquet
Assessment-specific scripts compute features
Export as CSV
s3://app/study/features/*
flanker
number-match
...
Export as CSV