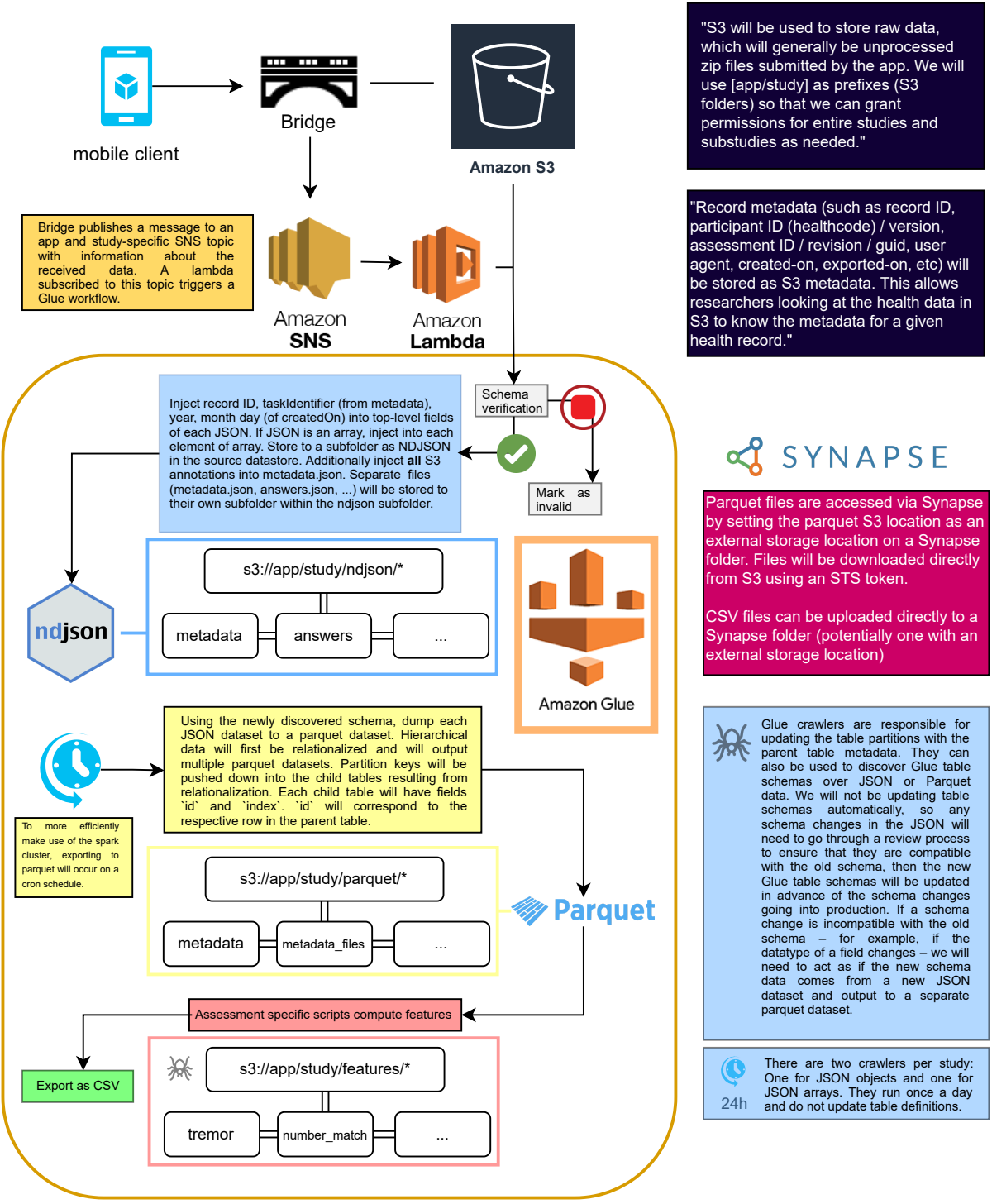
mobile client
Bridge
Amazon S3
"S3 will be used to store raw data, which will generally be unprocessed zip files submitted by the app. We will use [app/study] as prefixes (S3 folders) so that we can grant permissions for entire studies and substudies as needed."
Bridge publishes a message to an app and study-specific SNS topic with information about the received data. A lambda subscribed to this topic triggers a Glue workflow.
Amazon SNS
Amazon Lambda
"Record metadata (such as record ID, participant ID (healthcode) / version, assessment ID / revision / guid, user agent, created-on, exported-on, etc) will be stored as S3 metadata. This allows researchers looking at the health data in S3 to know the metadata for a given health record."
Inject record ID, taskIdentifier (from metadata), year, month day (of createdOn) into top-level fields of each JSON. If JSON is an array, inject into each element of array. Store to a subfolder as NDJSON in the source datastore. Additionally inject all S3 annotations into metadata.json. Separate files (metadata.json, answers.json, ...) will be stored to their own subfolder within the ndjson subfolder.
Schema verification
Mark as invalid
SYNAPSE
Parquet files are accessed via Synapse by setting the parquet S3 location as an external storage location on a Synapse folder. Files will be downloaded directly from S3 using an STS token.
CSV files can be uploaded directly to a Synapse folder (potentially one with an external storage location)
ndjson
s3://app/study/ndjson/*
metadata
answers
...
Amazon Glue
Using the newly discovered schema, dump each JSON dataset to a parquet dataset. Hierarchical data will first be relationalized and will output multiple parquet datasets. Partition keys will be pushed down into the child tables resulting from relationalization. Each child table will have fields `id` and `index`. `id` will correspond to the respective row in the parent table.
To more efficiently make use of the spark cluster, exporting to parquet will occur on a cron schedule.
s3://app/study/parquet/*
metadata
metadata_files
...
Parquet
Glue crawlers are responsible for updating the table partitions with the parent table metadata. They can also be used to discover Glue table schemas over JSON or Parquet data. We will not be updating table schemas automatically, so any schema changes in the JSON will need to go through a review process to ensure that they are compatible with the old schema, then the new Glue table schemas will be updated in advance of the schema changes going into production. If a schema change is incompatible with the old schema – for example, if the datatype of a field changes – we will need to act as if the new schema data comes from a new JSON dataset and output to a separate parquet dataset.
Assessment specific scripts compute features
Export as CSV
s3://app/study/features/*
tremor
number_match
...
24h
There are two crawlers per study: One for JSON objects and one for JSON arrays. They run once a day and do not update table definitions.