openIDL - Technical Backlog

This page covers the backlog for openIDL.

	Date	Item	Description
1	18-JAN-21	Extraction pattern - tech	What is the tech for the extraction pattern? map/reduce, optimized for scale, GraphQL or others? Context: 1. The extraction pattern model currently uses a map reduce function in MongoDB. This locks us into MongoDB and uses a closed environment without access to the outside world we'll need for correlating other data like census. The extraction capability must be reimagined. 2. Suitable for simple data source layout. Can't do cross lookup, validation or reference data checks 3. Currently works for mongoDB only. Won't work for other sources, example AAIS' data lake (Hadoop/Cloudera) 4. GraphQL seems like a strong candidate. POC is required to validate the hypothesis 5. Discuss any other candidates that should be considered
2	18-JAN-21	How to assert data integrity	How to assert data integrity? Context: 1. Do we expect same answer if the question is asked more than once? Within the parameters of use case, scope, time (duration) etc. For example: total premium for all Auto policies from 1/1/2021 to 12/31/2021 for purposes of stat reporting 2. Do we expect a different answer if the question is asked more than once? Within the parameters of use case, scope, time (duration) etc. For example: new data that has been committed to a past period 3. HLF response: see https://hyperledger-fabric.readthedocs.io/en/v0.6/API/CoreAPI.html#chaincode 4. Solution option: A checksum after record is locked & written to the chain, store the acknowledgement (chaincode hash) from the HLF to a control DB and map it to a record set
3	18-JAN-21	How to assert data quality	How to run technical and business validation on data and certify the data? Context: 1. Technical rules: may include JSON schema validation, format, cardinality check 2. Business rules: enum check, *field-to-transaction-to-record-set-to-dataset* validations, reference data 3. Error threshold: calculate the error threshold. NAIC allows for up to 5% error rate 4. Timing: when should validation be applied? As the data arrives into the HDS (stagingcore), or just before extraction (time pressure, lost opportunity with time that could be used to rectify errors), timeouts on the extraction API etc.
4	18-JAN-21	Common Rule Set	Is it possible to provide a common set of rules that can be used by all carriers against their data before making it available to the extraction? Context: 1. Assumption: rules are standardized across openIDL members for a given use case. Example, rules related to Auto stat reporting
5	24-JAN-21	Is the HDS data source permanent or temporary?	When the question (extraction request) is asked, the HDS (interfacing API) is expected to be available to respond. Context: 1. HDS API is required to be available when the request is due. For example, end of a quarter or end of Feb for an annual stat report 2. The type of HDS stack is driven by the use case. For example, stat reports are due annually, AIPSO reports are due quarterly etc. 3. There is no requirement for the HDS API to be available outside of scheduled activity. It is optional for members to have the HDS API available outside of scheduled activity 4. Consideration is to be given for the HDS to be available for internal systems to *write* to the HDS; or allow write to a scheduled batch event (no daily or more frequent writes) 5. There is delineation between the HDS data source and HDS API. The HDS data source & HDS API can be decoupled. Storage costs can be optimized by holding data in low cost storage (like AWS Glacier), and shift it to immediate access storage (like AWS S3) for the HDS API to be able to query and respond in reasonable time

6	18-JAN-21	How do we support calculating the error rate?	Reporting practice allows for an error rate of up to 5%. Context: Options to consider are 1. to process the entire (large to very large) dataset in-memory 2. break up (shard) the data set and apply validation at a more manageable layer (see data set hierarchy below) and write to a control DB 3. others... For example, for annual stat reporting, the data set is presented in this hierarchy (illustrative) 1. Time period a. LOB i. Entities and sub-entities 1. Months/Duration and/or ZIP codes a. Policy and Loss group i. Policy and Loss sub-group 1. Records making up the sub-group a. Field
7	18-JAN-21	Reference data validation	Where to host reference data service? Within member's enterprise or within node? When to apply reference data validation: 1. at rest: when the data is at rest in the HDS. This would mean that the data is washed and prepared for extraction. Recommended 2. at extraction: when the data is being queried. Not recommended 3. after extraction: at the Analytics node. Not recommended Placement of reference data validation may be different depending on the scenario. Reference data implementation options: 1. lookup tables with list of values/enums 2. APIs. For example, USPS ZIP code validation API 3. Should APIs be looked up at runtime or in the background (call an API and localize the reference data)
8	18-JAN-21	Reference data lookup services/APIs	Which APIs to look up? For example, USPS state/zip validation, Carfax etc. We can be opportunistic with this and certainly allows for MVP, tactical and strategic solutions
9	18-JAN-21	Reference data lookup services/APIs - pricing model	Who pays (assumption - whoever owns the data pays), and how to charge the consumers (via assigned accounts, via centralized billing account prorated to consumption etc.). Who signs the vendor contracts
10	18-JAN-21	Separating the Hyperledger Fabric Network from the data access	Can a carrier participate in the network from a hosted node without putting the data there? That is, can we give a carrier access to the network without them having the data access portion hosted in the same node. The HLF runtimes are not required to run in the carrier, and only a simple api is made available for extraction.
11	18-JAN-21	Simplify the technical footprint	Can we simplify the architecture so that there are not so many technologies required?
12	18-JAN-21	Hosted nodes	Should we consider hosted nodes for the HLF network instead of requiring all carriers who desire data privacy to host the network?
13	19-JAN-21	HDS vs Interface Spec for nodes	Do we need HDS to be uniform or only have uniform interface spec and communicate via service?
14			
15			