Schema Change Detection

Existing Pipeline

mparticle-client-events
sub_sessions_f
Cloud Pub/Sub

Existing Sessions_f
Pipeline
Cloud Dataflow

staging.sessons_f
BigQuery

Legends

Notification

Parse Library/Get Latest Jar

f(n) Process: Parse Msg
(Scala: Class path reflection)

System Variable

JAR v1
JAR v2
JAR vn

Validate
Messages
Cloud Dataflow
(Scala)

Data Engg.
Repo
Event Schema
(git)

Failed
Messages
Cloud Pub/Sub

Failed Msg
Counter

Fetch Latest
Schema
(Python)

Failed messages
Cloud Storage

Trigger

Sleep & Trigger

No

Compare Schema
(Python)

different?

Yes

Fetch latest jar
from JFrog
(Python)

JAR

Test Failed
Messages
Cloud Dataflow
(Scala)

More records
than threshold
pass?

Data Engg.
Repo
Event Schema (git -
previous version)

Threshold

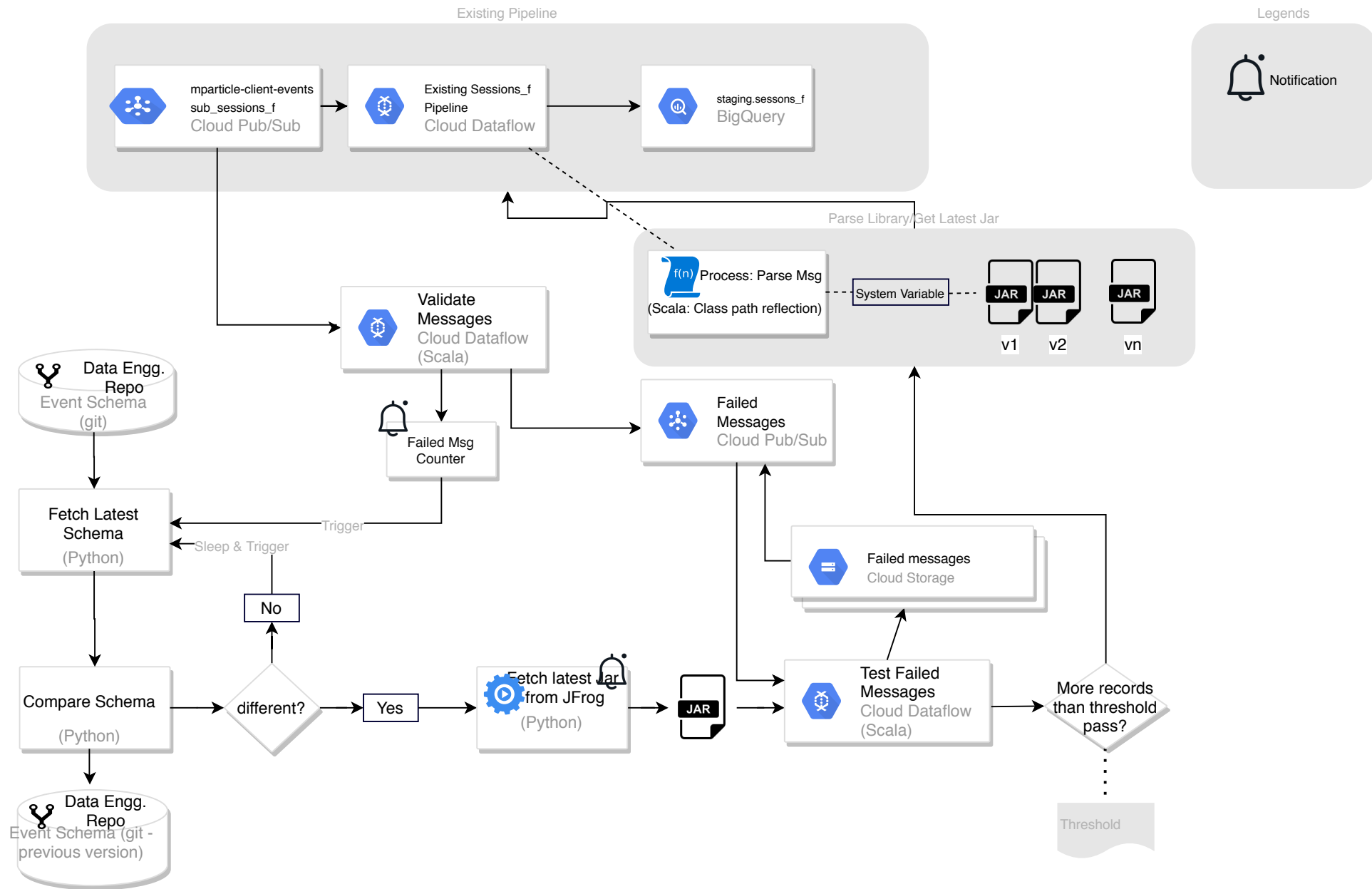## Validate Messages

1. Decrypt Protobuf and validate if the record conforms schema or is there any breaking change?

2. Check for new field (we are researching if/how we can implement it)

Compare Between Two Schemas

1. Parse the proto files

2. Convert to two dataframes - old and new proto file

3. Merge two dataframes like full outer join between two dataframes

4. Find the difference in attributes

It checks for the following changes:

1. If new field is added

2. If datatype is changed

3. If enum is changed

In the end it can return a flag which can trigger the compile process

Dynamically Downloading Latest Jar

Using Reflection we may load the class and get the method at run time