.. \_troubleshooting:

TroubleShooting

==================

[Batch] Regenerating a corrupt version

--------------------------------------

Symptom

~~~~~~~~~~

The Summingbird batch job failed due to the following error:

.. code:: bash

Caused by: com.twitter.bijection.InversionFailure: ...

It typically indicates the corrupt records of the aggregate store (not the other side of the DataRecord source).

The following describes the method to re-generate the required (typically the latest) version:

Solution

~~~~~~~~~~

1. Copy \*\*the second to last version\*\* of the problematic data to canaries folder. For example, if 11/20's job keeps failing, then copy the 11/19's data.

.. code:: bash

$ hadoop --config /etc/hadoop/hadoop-conf-proc2-atla/ \

distcp -m 1000 \

/atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates/1605744000000 \

/atla/proc2/user/timelines/canaries/processed/aggregates\_v2/user\_mention\_aggregates/1605744000000

2. Setup canary run for the date of the problem with fallback path pointing to `1605744000000` in the prod/canaries folder.

3. Deschedule the production job and kill the current run:

For example,

.. code:: bash

$ aurora cron deschedule atla/timelines/prod/user\_mention\_aggregates

$ aurora job killall atla/timelines/prod/user\_mention\_aggregates

4. Create backup folder and move the corrupt prod store output there

.. code:: bash

$ hdfs dfs -mkdir /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates\_backup

$ hdfs dfs -mv /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates/1605830400000 /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates\_backup/

$ hadoop fs -count /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates\_backup/1605830400000

1 1001 10829136677614 /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates\_backup/1605830400000

5. Copy canary output store to prod folder:

.. code:: bash

$ hadoop --config /etc/hadoop/hadoop-conf-proc2-atla/ distcp -m 1000 /atla/proc2/user/timelines/canaries/processed/aggregates\_v2/user\_mention\_aggregates/1605830400000 /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates/1605830400000

We can see the slight difference of size:

.. code:: bash

$ hadoop fs -count /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates\_backup/1605830400000

1 1001 10829136677614 /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates\_backup/1605830400000

$ hadoop fs -count /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates/1605830400000

1 1001 10829136677844 /atla/proc2/user/timelines/processed/aggregates\_v2/user\_mention\_aggregates/1605830400000

6. Deploy prod job again and observe whether it can successfully process the new output for the date of interest.

7. Verify the new run succeeded and job is unblocked.

Example

~~~~~~~~

There is an example in https://phabricator.twitter.biz/D591174

[Batch] Skipping the offline job ahead

---------------------------------------

Symptom

~~~~~~~~~~

The Summingbird batch job keeps failing and the DataRecord source is no longer available (e.g. due to retention) and there is no way for the job succeed \*\*OR\*\*

..

The job is stuck processing old data (more than one week old) and it will not catch up to the new data on its own if it is left alone

Solution

~~~~~~~~

We will need to skip the job ahead. Unfortunately, this involves manual effort. We also need help from the ADP team (Slack #adp).

1. Ask the ADP team to manually insert an entry into the store via the #adp Slack channel. You may refer to https://jira.twitter.biz/browse/AIPIPE-7520 and https://jira.twitter.biz/browse/AIPIPE-9300 as references. However, please don't create and assign tickets directly to an ADP team member unless they ask you to.

2. Copy the latest version of the store to the same HDFS directory but with a different destination name. The name MUST be the same as the above inserted version.

For example, if the ADP team manually inserted a version on 12/09/2020, then we can see the version by running

.. code:: bash

$ dalv2 segment list --name user\_original\_author\_aggregates --role timelines --location-name proc2-atla --location-type hadoop-cluster

...

None 2020-12-09T00:00:00Z viewfs://hadoop-proc2-nn.atla.twitter.com/user/timelines/processed/aggregates\_v2/user\_original\_author\_aggregates/1607472000000 Unknown None

where `1607472000000` is the timestamp of 12/09/2020.

Then you will need to duplicate the latest version of the store to a dir of `1607472000000`.

For example,

.. code:: bash

$ hadoop --config /etc/hadoop/hadoop-conf-proc2-atla/ distcp -m 1000 /atla/proc2/user/timelines/processed/aggregates\_v2/user\_original\_author\_aggregates/1605052800000 /atla/proc2/user/timelines/processed/aggregates\_v2/user\_original\_author\_aggregates/1607472000000

3. Go to the EagleEye UI of the job and click on the "Skip Ahead" button to the desired datetime. In our example, it should be `2020-12-09 12am`

4. Wait for the job to start. Now the job should be running the 2020-12-09 partition.