# Batch Publication - Troubleshooting

On this page you can find most identified error messages or issues related to this service. If the answer is not provided, please see the [Data Platform - Support Guidelines](file:///C:\display\DP\Data+Platform+-+Support+Guidelines) page for further steps.

**Troubleshooting** pages are distinct from how-tos and must not be confused with them:

* **Troubleshooting** pages describe solutions for very specific problems.  For instance it shows how to solve a certain error message that you get when using a service in our data platform.
* **How-tos**are recipes to guide you through the steps involved in addressing key use-cases.
* [General issues and tips](#BatchPublicationTroubleshooting-General)
  + [How to read, interpret, and act on error logging](#BatchPublicationTroubleshooting-Howtore)
* [Azure DevOps](#BatchPublicationTroubleshooting-AzureDe)
  + [How to grant permission for a pipeline to build?](#BatchPublicationTroubleshooting-Howtogr)
  + [How to set up the right container for your pipeline?](#BatchPublicationTroubleshooting-Howtose)
  + [How to recognize a publication which uses pubengine v3?](#BatchPublicationTroubleshooting-Howtore)
  + [ValueError: badly formed hexadecimal UUID string](#BatchPublicationTroubleshooting-ValueEr)
* [Glue](#BatchPublicationTroubleshooting-Glue)
  + [Out of memory issues / SparkOutOfMemoryError](#BatchPublicationTroubleshooting-Outofme)
  + [Get more logging if you have unexpected issues in your publication](#BatchPublicationTroubleshooting-Getmore)
  + [Table or view not found / EntityNotFoundException](#BatchPublicationTroubleshooting-Tableor)
  + [Total size of serialized results bigger than spark.driver.maxResultSize](#BatchPublicationTroubleshooting-Totalsi)
  + [Access Denied](#BatchPublicationTroubleshooting-AccessD)
  + [An error occurred while calling o7676.save.: java.lang.StackOverflowError](#BatchPublicationTroubleshooting-Anerror)
  + [An error occurred while calling o877.count. Job aborted due to stage failure: ShuffleMapStage](#BatchPublicationTroubleshooting-Anerror)
* [Pubconfig](#BatchPublicationTroubleshooting-Pubconf)
  + [Installation: Python not recognized](#BatchPublicationTroubleshooting-Install)
  + [Installation: '503 Service Unavailable' error](#BatchPublicationTroubleshooting-Install)
  + [Installation: anl-pubconfig requires Python '>=3.7.9,<3.11'](#BatchPublicationTroubleshooting-Install)
  + [Installation: OSError: [WinError 145] The directory is not empty](#BatchPublicationTroubleshooting-Install)
  + [Installation: OSError: [WinError 3] The system cannot find the path specified](#BatchPublicationTroubleshooting-Install)
  + [Run: No Module named 'anl\_pubconfig'](#BatchPublicationTroubleshooting-Run:NoM)
  + [Token or status missing](#BatchPublicationTroubleshooting-Tokenor)
* [Step function / State Machine](#BatchPublicationTroubleshooting-Stepfun)
  + [An error occurred (MalformedPolicy) when calling the PutBucketPolicy operation: Invalid principal in policy](#BatchPublicationTroubleshooting-Anerror)
  + [Workflow encountered an error: Redshift cluster status 'RESUMING'/'PAUSED' is not allowed](#BatchPublicationTroubleshooting-Workflo)
  + [How to handle 'There is an '@' in where clause after applying SQL filters'](#BatchPublicationTroubleshooting-Howtoha)
* [Workflow API](#BatchPublicationTroubleshooting-Workflo)
  + [How to handle error messages when orchestrating a workflow/publication?](#BatchPublicationTroubleshooting-Howtoha)
* [Known issues when ingesting Publication data back into the datalake](#BatchPublicationTroubleshooting-Knownis)
  + [Ingestion error; invalid decimal for format decimal(p, s)](#BatchPublicationTroubleshooting-Ingesti)
  + [java.io.IOException Failed to delete key](#BatchPublicationTroubleshooting-java.io)
* [Known issues when working with publication data in the Datalab](#BatchPublicationTroubleshooting-Knownis)
  + [Publication data is not visible in Redshift with DBeaver after loading it as an external schema but I can see and query the data in Zeppelin notebook. I can also see the data in CSV files in the file browser of the Datalab.](#BatchPublicationTroubleshooting-Publica)
  + [I see duplicate records in the Datalab when a Publish data from the Datalake.](#BatchPublicationTroubleshooting-Iseedup)

## General issues and tips

### How to read, interpret, and act on error logging

When a error occurs, the logging might not always make the issue immediately clear and you might not find any matching result on our documentation.  
Below are some pointers on how to get more insights on what's going wrong so you can try and tackle the issue on your own!

1. Logging can be found in several locations. Depending on the type of error/content of the main error message, some locations are a more useful starting point than others:
   1. For any error message containing "Py4JJavaError", navigate to your Glue job using the PUBCONFIG Role (UENTAWSNLBIDEVPUBCONFIG). After selecting the failing run, click on "Output logs" under the "Cloudwatch logs" header.
   2. When you are in the log stream, click on Display → View in plain text. This way the logs are more readable. After this is done, you can simply search for 'error' using CTRL + F and read through the logs.
   3. You can also view more logs using Cloudwatch. In AWS Glue, click on your glue job run and scroll down to 'input arguments' and copy the value of --log\_group\_name. Go to AWS Cloudwatch → Log groups and search for the log group.
2. The main error message might lack the required information. If this happens, find for any other error, exception, or warning messages using CTRL+F (and/or the built-in search function of the AWS service you're using) with one of the keywords "error", "exception", and "warning".
3. Interpretation of the message and acting upon it can be difficult at times, so here are some tips:
   1. If the message contains "org.apache.spark" somewhere, then it's likely to be a Spark issue. Continue scrolling/searching down through all messages until you get to the "lowest" error messages: those will often contain the most useful information!
   2. If it's a Spark issue, try searching for the string "org.apache.spark.SparkUpgradeException": if this is present, you're probably dealing with a legacy error due to upgrading your Glue version. You can also search for 'caused by'. Please see [How to: Set Glue version](file:///C:\display\DP\How+to:+Set+Glue+version)

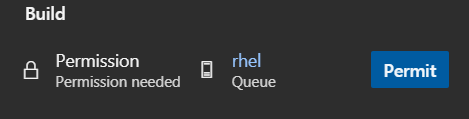
If you still experience issues fixing your error message, contact the OPS team and let us know what you've already found out and tried.

## Azure DevOps

### How to grant permission for a pipeline to build?

The error message is "***This pipeline needs permission to access a resource before this run can continue to Build***". Actually it is not really an error.   
You need to provide permission to start the build and deployment of the publication pipeline.

On the right hand side in the message there is a **view** button. click on **view.** this will open an extra screen:



Click on the blue Permit button to start the Build.

### How to set up the right container for your pipeline?

During a run of your pipeline you might get the following warning:

**The CDK image that is configured in your azure-pipelines.yml should be updated! You cannot use version 2 of the Data Platform packages until you update your CDK image!**

You might also notice that you are restricted to version 1 of our Data Platform automation packages. The issue is that the configuration of your CI/CD pipeline is outdated and you need to update the container settings.

Go to the Git repository associated with this pipeline. Create a feature branch and navigate to the azure-pipelines.yml file in the root of your Git repository. Your azure-pipelines.yml will have a setting for resources, which will look like the following:

 Bron uitklappen

resources:

containers:

- container: cdk

image: swfactory.aegon.com:5001/swf/cdk:node-12.16-python-3.7

repositories:

- repository: templates

type: git

name: datalake-platform/dp-ash-nazg

ref: refs/heads/master

To get rid of the warning and unlock the ability to use higher versions of the Data Platform automation packages, update the image on line 4. The value should be:

[swfactory.aegon.com](http://swfactory.aegon.com):5001/swf/cdk:python-latest

After making the change, the resources setting in your azure-pipelines.yml file should look like the following:

 Bron uitklappen

resources:

containers:

- container: cdk

image: swfactory.aegon.com:5001/swf/cdk:python-latest

repositories:

- repository: templates

type: git

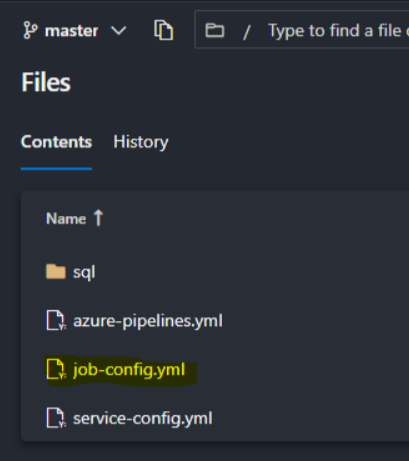
name: datalake-platform/dp-ash-nazg

ref: refs/heads/master

You can now commit this change. Check that a pipeline run from your feature branch will no longer show you the warning. You can then merge this change to your main branch via a pull request.

### How to recognize a publication which uses pubengine v3?

Since publications using pubengine v2 are no longer supported, it's important to update them to use pubengine v3 if you need our support. You can recognize a publication which uses the pubengine v3 by the presence of a job-config.yml file.



### ****ValueError: badly formed hexadecimal UUID string****

This error will be shown in the 'deploy pipeline' step on Azure DevOps:

File "/\_\_w/17/s/dp\_venv/lib/python3.7/site-packages/anl\_cicd/utils/dp\_accounts.py", line 40, in get\_account\_id   
dp\_account = account\_service.get\_by\_id(uuid.UUID(account\_id))   
File "/usr/local/lib/python3.7/uuid.py", line 160, in \_\_init\_\_   
raise ValueError('badly formed hexadecimal UUID string')

The issue is related to the value of the DEFAULT\_ACCOUNT\_ID in the DP\_ACCOUNT\_INFO Azure DevOps library. The account id should be 32 characters long (in the correct format). The account id will be provided to you when you request an account id (needs to be created first). Please request this (if you don't have one) by contacting one of the Data Platform Consultants.

## Glue

### Out of memory issues / SparkOutOfMemoryError

See this How to page: <https://globalitconfluence.us.aegon.com/x/vc_bJQ>

### Get more logging if you have unexpected issues in your publication

Set the log level to DEBUG: If you are debugging a publication, and logging is not showing enough information, you can increase logging by putting the following in your API call:   {"log\_level": "DEBUG"}. Example:

{  
 "reporting\_date": {  
 "year": "2024",  
 "month": "01",  
 "day": "01"  
 },  
 "log\_level": "debug"  
}

### Table or view not found / EntityNotFoundException

This issue is almost always related to your sql statement: in your FROM statement you refer to a table or view that does not exist in memory. To fix this, make sure you refer to an existing view. For loaded data this is always `database.table` ( ` are backticks not " quotes) and for a previously executed sql file this is the name of the sql file without the ".sql" suffix.

**If the above does not help and you have an error like below, please contact Data Flows In team in the** [**[support] Ingestion channel**](https://teams.microsoft.com/l/channel/19%3aa3cfe51e29b34606b670ec4358cf798a%40thread.tacv2/%255Bsupport%255D%2520Ingestion?groupId=64d3292b-a12a-4ad7-afdb-284a396900dd&tenantId=46e16835-c804-41de-be3c-55835d14dee4) **to check why this table is not present in the Glue catalog**. It could be that the export of this table went wrong or something similar.

An error occurred (EntityNotFoundException) when calling the GetPartitions operation: Database dlk\_prd\_pubengine\_... not found.

### Total size of serialized results bigger than spark.driver.maxResultSize

The error message will be like "Job aborted due to stage failure: Total size of serialized results of XXXX tasks (1024.1 MB) is bigger than spark.driver.maxResultSize (1024.0 MB)".

The maxResultSize defines the max size of the result a worker can send to the driver. If estimated size of the data is larger than maxResultSize given job will be aborted. The goal here is to protect the application from driver loss. It's not possible to increase the maxResultSize for AWS Glue.

To fix this, optimize the SQL query to avoid reaching this condition. You can use the Query Interface in your datalab for this.

### Access Denied

The error message will be like: "An error occurred while calling o81.getDynamicFrame.: [java.io](http://java.io/).IOException: com.amazon.ws.emr.hadoop.fs.shaded.com.amazonaws.services.s3.model.AmazonS3Exception: Access Denied". This means there is no permission to read data from the datalake. This usually means that something went wrong in requesting access from the datalake. To fix this, check the logs of the run-publication-pipeline operator and see if you find anything unusual. Contact us via the support channel if nothing unusual is found.

### An error occurred while calling o7676.save.: java.lang.StackOverflowError

This error occurs if too many levels of transformations occur in your publication. More technically: The JVM will not be able to hold the operations to perform lazy execution when an action is performed in the end. There are a few possible solutions to fix this:

Reduce the amount of transformations used in your publication. Check whether:

* Do you really need to escape those special characters for the given table,
* Do you really need to trim those whitespaces.
* Optimize your SQL query. See the HiveSql manual for tips.
* Do you really need to process increments using the partial record update strategy?

If all these options do not help, then contact the [Data Flows Out support team](https://teams.microsoft.com/l/channel/19%3a956548bd147c4caa94e441027a5c1b99%40thread.tacv2/%255Bsupport%255D%2520Realtime%2520Publication?groupId=64d3292b-a12a-4ad7-afdb-284a396900dd&tenantId=46e16835-c804-41de-be3c-55835d14dee4).

### An error occurred while calling o877.count. Job aborted due to stage failure: ShuffleMapStage

This error its mainly related to memory allocation. Check the amount of memory you have configured for executors. After increasing the memory of executor the job might run successfully. Another thing you can check is the joins in the query: try to split the joins as much as possible.

If these options do not help, then contact the [Data Flows Out support team](https://teams.microsoft.com/l/channel/19%3a956548bd147c4caa94e441027a5c1b99%40thread.tacv2/%255Bsupport%255D%2520Realtime%2520Publication?groupId=64d3292b-a12a-4ad7-afdb-284a396900dd&tenantId=46e16835-c804-41de-be3c-55835d14dee4).

## Pubconfig

### Installation: Python not recognized

Make sure you have Python installed on your Aegon Laptop. You can install Python on your Aegon laptop using the Python installer from <https://www.python.org/downloads/windows/> and make sure to deselect the "install for all users option".

### Installation: '503 Service Unavailable' error

Make sure you install from the Aegon network. Artifactory is only reachable from the Aegon network. So the installer doesn't work in the Data lab for instance.

### Installation: anl-pubconfig requires Python '>=3.7.9,<3.11'

You are using an unsupported version of Python for the installation. Make sure you install a supported version. In this case a Python version between 3.7.9 and below 4.0

### Installation: OSError: [WinError 145] The directory is not empty

Sometimes an old installation is not wiped correctly during a re-installation or update. To fix this, manually remove the %HOME%/.pubconfig folder and retry installing.

### Installation: OSError: [WinError 3] The system cannot find the path specified

The error message will be like: [WinError 3] The system cannot find the path specified: 'C:\\Program Files (x86)<\\Microsoft Visual Studio\\>2017\\Professional\\VC\\Tools\\MSVC\\14.16.27023\\PlatformSDK\\lib'.

Your environment misses the Visual Studio Platform SDK, which is needed to build python libraries on Windows. To fix this:

* Install using any ONE of these choices:
  + Microsoft [Build Tools for Visual Studio](https://visualstudio.microsoft.com/thank-you-downloading-visual-studio/?sku=BuildTools&rel=16)
  + Alternative link to Microsoft [Build Tools for Visual Studio](https://visualstudio.microsoft.com/downloads/#build-tools-for-visual-studio-2019).
  + Offline installer: [vs\_buildtools.exe](https://aka.ms/vs/16/release/vs_buildtools.exe)
* Select: Workloads → C++ build tools
* Install options: select only the “Windows 10 SDK” (assuming the computer is Windows 10). Optionally, if you want to use MSVC cl.exe C/C++ compiler from the command line, additionally select the C++ build tools, which takes an additional 2 GB disk space

### Run: No Module named 'anl\_pubconfig'

The error message will be like: Traceback (most recent call last):  
  File "C:\Users\xxxxx\.pubconfig\bin\pubconfig", line 12, in <module>  
    from anl\_pubconfig.console.main import main  
ModuleNotFoundError: No module named 'anl\_pubconfig'

This error message means that pubconfig cannot find the pubconfig module. The cause is that something went wrong during installation. To check if this is the case, go to the %HOME%/.pubconfig/lib folder. If it only shows something like the following image, then the package has not been installed properly.



To fix this:

* Make sure you install the pubconfig from a supported shell type (Windows Powershell/Bash).
* Make sure you are using a supported Python version (check installation instructions).
* Make sure you don't get an error during the installation. If you do get an error, troubleshoot that error.

### Token or status missing

The following error can occur when an outdated dp-version is used:

**Exception in step: 7\_callback**

Exception: Did not get an OK response when doing the callback, instead we got [400], full response was [b'

{

"status": 400,

"message": "Token or status missing!"

}

']

To fix this:

* Make sure that the ***service-details.dp-version*** parameter in the **service-config.yml** is either:
  + Not set at all
  + Matches at least 1.17.1

## Step function / State Machine

### An error occurred (MalformedPolicy) when calling the PutBucketPolicy operation: Invalid principal in policy

The error message will be like:

"Failed to trigger security lambda! Because of error: {'errorMessage': 'An error occurred (MalformedPolicy) when calling the PutBucketPolicy operation: Invalid principal in policy'"

This sometimes happens when triggering a publication. The reason is that the bucket policy got malformed. This usually happens when many processes try to update the policy at the same time. There are mechanisms in place to prevent this from happening and fix it when it does happen.

**To fix this: wait around 1 hour. The self-healing mechanism should recover it.**

### ****Workflow encountered an error: Redshift cluster status 'RESUMING'/'PAUSED' is not allowed****

This error occurs when the Redshift cluster attached to the datalab is not in the status 'ACTIVE': these clusters are inactive during the night and any call to them will result in this error.

**To fix, either:**

* **Make sure to run the Glue jobs between 05:30 and 20:30 UTC+0** to prevent trying to contact a paused Redshift cluster.  
  Check the service-config.yml  in the AzDo Repo and make sure the expression in schedule-cron  is valid compared to the 'ACTIVE' period of the Redshift cluster.
* **Remove the line operator-name: create\_external\_schema from service-config.yml** in the AzDo Repo if the external schema already exists.  
  This line makes the publication try to connect to Redshift each Publication run. Once the external schema has been created once and the S3 location does not change, the pre-existing schema will be updated even when this line is removed.

### How to handle 'There is an '@' in where clause after applying SQL filters'

#### Issue

In the anl-pubset-logs, you can encounter the following error:

There is an '@' in where clause after applying SQL filters.

This indicates missing filter(s).

SQL: ...SQL filters: [...]

#### Cause

There are two main causes that might be the reason you get this error:

1. The SQL filters passed down in the API call do not match the used Safe SQL Filters.
2. There are unexpected Carriage Return/CR (\r) sequences in the parsed SQL string.

#### Actions

Make sure to follow the next steps before contacting the support team:

1. Read the documentation [How to: Create a fit for purpose data set using scenarios and SQL filters - Data Platform - Global IT Confluence Production (aegon.com)](https://globalitconfluence.us.aegon.com/display/DP/How+to%3A+Create+a+fit+for+purpose+data+set+using+scenarios+and+SQL+filters)
   1. Verify that your repository is correctly configured.
   2. Verify that your API call is correct: **SQL Filters need to match your SQL**
   3. Verify that your SQL is correct: **it can only contain Safe SQL Filters!**
2. If everything is correctly configured, but you still get the errors:
   1. Check if there are Carriage Return/CR (\r) sequences in the parsed SQL string:
      1. Look in the same logging for the [[SUMMARY]] section and search for \r occurences in the ErrorMessage
      2. If a \r is directly against the WHERE clause, **add a whitespace after the WHERE clause:** "WHERE" → "WHERE "

If the above does not help, then contact the [Data Flows Out support team](https://teams.microsoft.com/l/channel/19%3a956548bd147c4caa94e441027a5c1b99%40thread.tacv2/%255Bsupport%255D%2520Realtime%2520Publication?groupId=64d3292b-a12a-4ad7-afdb-284a396900dd&tenantId=46e16835-c804-41de-be3c-55835d14dee4).

## Workflow API

### How to handle error messages when orchestrating a workflow/publication?

|  |  |  |
| --- | --- | --- |
| **Error** | **Reason** | **Solution** |
| 403 | Access Denied | Make sure that the right credentials are set |
| 500 | Internal server error | Most likely your endpoint or json body is not formatted properly. Make sure your body is a valid JSON and that you have all the required keys. If that didn't work, ask a team mate or else contact support and provide a screenshot of your request body and the response |
| 504 | Gateway timeout | Contact support |

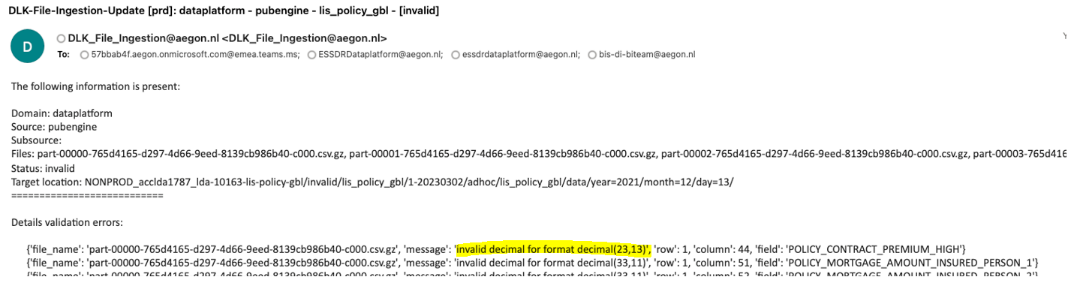
## Known issues when ingesting Publication data back into the datalake

### Ingestion error;  invalid decimal for format decimal(p, s)

If (especially for Finance) you get an error email after your Publication tries to write back/ingest into the datalake, containing anything like:

invalid decimal for format decimal(p,s)

with p=precision, s=scale, this is probably due to a Spark issue. See also screenshot below.  
The cause is a known Spark problem: when a column of type Decimal and scale > 6 contains a value of 0, Spark converts it to a floating point number internally. See also <https://issues.apache.org/jira/browse/SPARK-25177>.  
The output csv file will then contain a 0e-13 while the type is Decimal. The datalake ingestion process will not accept this number because it expects proper Decimal values, and the ingestion fails.  
 **You can work-around this issue by casting the Decimal column to a string in your SQL. Alternatively you can make sure the scale <= 6.**

Example File Ingestion Error:  


(for internal reference see [JVI-46](https://globalitjira.us.aegon.com/browse/JVI-46) - Issue details worden opgehaald... STATUS )

### java.io.IOException Failed to delete key

If you get an error message like this in your publication log:  
  
Py4JJavaError: An error occurred while calling o382.save. : [java.io](http://java.io).IOException: Failed to delete key: dataplatform-pubengine/valid/life\_coverage\_gbl/37-20230301/adhoc/life\_coverage\_gbl/data

you bump into an issue which is solved in Pubengine 1.26.0. Please upgrade your publication to use this, or a later, or the latest version.

## Known issues when working with publication data in the Datalab

### Publication data is not visible in Redshift with DBeaver after loading it as an external schema but I can see and query the data in Zeppelin notebook. I can also see the data in CSV files in the file browser of the Datalab.

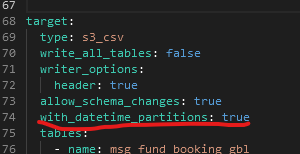
**Problem:**  
I can't see or query data in Redshift when I publish data to a Datalab and load it as an external schema from a Glue catalog (even though loading the data as an external schema does not through an error). Using and querying the data in a Zeppelin notebook works just fine.

When I query the newly created external schema in Redshift I get an error message like:

SQL Error [500310] [XX000]: [Amazon](500310) Invalid operation: index < data\_.a.size  
Details:   
-----------------------------------------------  
error: index < data\_.a.size  
code: 1000  
context: rapidjson::GenericValue<Encoding, Allocator>& rapidjson::GenericValue<Encoding, Allocator>::operator[](rapidjson::SizeType) [with Encoding = rapidjson::UTF8<>; Allocator = rapidjson::MemoryPoolAllocator<>; rapidjson::SizeType = unsigned int]  
query: 17011839  
location: document.h:1302  
process: padbmaster [pid=20070]  
  
**To fix this:**

1. Add the following options to your job-config.yml file of your publication in the 'writer\_options' section under 'Target' (see documentation here: [job-config definition](file:///C:\display\DP\job-config+definition))
   * escape\_special\_characters: True
   * trim\_whitespace: True
2. Post a request in batch publication support channel to delete your existing Glue catalog of the publication (that you use to load the data in Redshift external schema)

### I see duplicate records in the Datalab when a Publish data from the Datalake.

**Problem:**  
Publishing data to the Datalab (or other target in the Dataplatform) shows duplicate records.  
Usually this problem is caused by configuring a publication with the option: with\_datetime\_partitions: true  
  
  
This options writes every run with it's own date partition keys (Year, Month, Day) to the target and doesn't overwrite the older runs. So in the Datalab you get data of all the previous publication runs that have different Year, Month and Date partition keys (that look like a column in the data). This appears like the data contains duplicate records but essentially they are two (or more) snapshots of the data of different runs of the publication. If the data doesn't change between the runs they look like duplicate records.

**To fix this:**

1. Delete all the publication data from Datalab (or have it deleted by D&R support)
2. Set the option "with\_datetime\_partitions" to the value "false".
3. Run the publication again.

If you do need this option to be set to TRUE for some reason then you can filter your data in your queries on Year, Month or Date(or a combination of two of these partition keys). Due to limitation of Redshift setup it is  not possible to filter on ALL THREE at the same time.