**Contents**

[Extracting Job Statistics and loading to snowflake 1](#_Toc97718025)

[Finding JobID 2](#_Toc97718026)

[Configuring Max Job Runs 3](#_Toc97718027)

[Database DML Scripts 5](#_Toc97718028)

[Database validation scripts 6](#_Toc97718029)

# **Extracting Job Statistics and loading to snowflake**

Streamsets Control Hub (SCH) displays status of only last 50 runs of any given job. There is no provision to look back at jobs statistics farther than last 50 runs, hence there is a need to build a custom solution that captures latest job stats on a periodic basis and store them, from where historic data can be retrieved in an easier fashion. This solution facilitates going back in history and tracking job stats for troubleshooting & other purposes. This document illustrates the implementation of solution that involves extracting latest jobs status through a python script via SCH API, loading extracted data into snowflake staging tables through streamsets pipelines & loading staging data into historic table through stored procedure. Below is detailed description of the steps.

1. A python script is developed to extract history and load data into 2 csv files – One file to contain summary statistics and other to contain detailed statistics for each run. Please note that the script can extract history of only one job at a time. To extract history for multiple jobs the script needs to be run multiple times, one per job.

The script resides on streamsets data collector server (sdc) and can be accessed/run with sdc account. Max. job runs configured on SCH determines the number of runcounts that will be extracted by the script. ([Click here for instructions to configure max job runs](#_Configuring_Max_Job)) Below are the details.

Python script location on svn - svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/python\_scripts/print\_job\_history\_tocsv.py

Python script name – print\_job\_history\_tocsv.py

Python script location on sdc - /home/streamsets/pysdk\_scripts

Syntax to run the script - python3 /home/streamsets/pysdk\_scripts/print\_job\_history\_tocsv.py <jobid> ([Click here for instructions to find jobid](#_Finding_JobID))

Csv files location - /home/streamsets/JobHistoryStatistics

Summary csv filename - <YYYYMMDDHHMI>\_<jobid>\_SummaryStats.csv

Detailed csv filename - <YYYYMMDDHHMI>\_<jobid>\_DetailedStats.csv

(Note that the YYYYMMDDHHMI corresponds to the date & time when the python script ran and created those files)

1. 3 Streamsets pipelines are developed to extract job statistics and load them into snowflake tables. Streamsets/snowflake entity names used in the document are for illustration purposes only and may change with time & environments. Below are the details
   1. Pipeline1 to run the python script for desired jobs whose history needs to be tracked. The scripts that are run in this pipeline correspond to Queue Interval Aggregate load jobs in DEV environment. Cron scheduler stage is configured within this pipeline

Pipeline Name – SCH\_DEV\_JobStats\_Extract\_QIA

Job Name – Job for SCH\_DEV\_JobStats\_Extract\_QIA

Schedule – Runs on the 5th min every hour

* 1. Pipeline2 to load data from \*SummaryStats.csv into corresponding snowflake stage table and archive csv files. This pipeline is wrapped into a job and is scheduled via SCH scheduler

Pipeline Name – SCH\_DEV\_JobStats\_Summary\_Load\_FromCSV

Job Name – Job for SCH\_DEV\_JobStats\_Summary\_Load\_FromCSV

Schedule – Runs on 15th min every hour

Archive files location - /home/streamsets/JobHistoryStatistics/Archive

Snowflake destination table - "PUREINSIGHTS\_DEV"."SCH\_STATS"."SCH\_JOBSTATISTICS\_SUMMARY\_STG"

* 1. Pipeline3 to load data from \*DetailedStats.csv into corresponding snowflake stage table and archive csv files. This pipeline is wrapped into a job and is scheduled via SCH scheduler

Pipeline Name – SCH\_DEV\_JobStats\_Detail\_Load\_FromCSV

Job Name – Job for SCH\_DEV\_JobStats\_Detail\_Load\_FromCSV

Scheduler – Runs on 15th min every hour

Archive files location - /home/streamsets/JobHistoryStatistics/Archive

Snowflake destination table - "PUREINSIGHTS\_DEV"."SCH\_STATS"."SCH\_JOBSTATISTICS\_DETAIL\_STG"

1. A stored procedure is developed to load history tables from respective stage tables. Only that staging data that doesn’t exist or needs update on history tables will be loaded. A task has been configured to define the schedule of stored procedure.

Stored proc name: sp\_populate\_sch\_history\_table

Script location on svn: svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/sp\_populate\_sch\_history\_table.sql

Schedule – Runs on the 25th min every hour

Task name: create\_task\_schedule\_dev

Script location on svn: svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/create\_task\_schedule\_dev.sql

History of last 10 scheduled task runs can be tracked via below query.

select \*

from table(information\_schema.task\_history(

scheduled\_time\_range\_start=>dateadd('hour',-10,current\_timestamp()),

result\_limit => 10,

task\_name=>'sch\_stats\_load\_hist\_tbl'));

# **Finding JobID**

1. Login to SCH and open a job
2. Browse through **Job Status** > click on **show additional Info**. JobID field is displayed as highlighted below

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

# **Configuring Max Job Runs**

1. Open SCH
2. Browse through **Main Menu** > **Administration** > **Organizations**. Click on the gear icon as highlighted
3. **Maximus Number of Job Runs** entry determines number of job runs that can be extracted by python script. Please note that although GUI allows entering more than 100, 100 is the max that can be extracted from API that the python script uses.

Graphical user interface, text, email

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

# **Database tables creations Scripts**

svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/1.create\_tbl\_sch\_jobstatistics\_summary\_stg.sql

svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/2.create\_tbl\_sch\_jobstatistics\_detail\_stg.sql

svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/3.create\_tbl\_sch\_jobstatistics\_summary.sql

svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/4.create\_tbl\_sch\_jobstatistics\_detail.sql

svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/5.create\_tbl\_sch\_jobstats\_error\_log.sql

# **Database tables integrity checks**

svn://svn-staging.maximus.com/dev1d/maxdat/Genesys/sch\_stats/db\_scripts/data\_integrity\_checks.sql