simple random sampling while pulling data from warehouse(oracle engine) using proc sql in sas

Asked 7 years, 4 months ago   Modified 7 years, 4 months ago

Viewed 332 times

0

I need to pull humongous amount of data, say 600-700 variables from different tables in a data warehouse...now the dataset in its raw form will easily touch 150 gigs - 79 MM rows and for my analysis purpose I need only a million rows...how can I pull data using proc sql directly from warehouse by doing simple random sampling on the rows.

Below code wont work as ranuni is not supported by oracle

```
proc sql outobs =1000000;
select * from connection to oracle(
select * from tbl1 order by ranuni(12345);
quit;
```

How do you propose I do it

sampling   sas   oracle-database

Share   asked Aug 4, 2017 at 13:35

Rohan
103 ● 1 ● 8

Better to ask on a programming site as it is not clear that you have a statistical question here. – mdewey Aug 4, 2017 at 13:45

I know but a lot of us statistical folks can benefit from this – Rohan Aug 4, 2017 at 13:48

2   Possible duplicate of simple random sampling while pulling data from warehouse(oracle engine) using proc sql in sas – Allan Bowe Aug 6, 2017 at 10:30

This is no duplicate, because in the refered question, there were only 100 record in the source table. Here sorting should be avoided. – Dirk Horsten Aug 7, 2017 at 20:38 ✎

@DirkHorsten - perhaps you are referring to this question? stackoverflow.com/questions/9868409/… – Allan Bowe Aug 8, 2017 at 6:27

1 Answer

Sorted by: Highest score (default) ⇕

I understand you need a sample of about 700 rows. Let us put that in a macro variable *(For those more familiar with other languagues than SAS, that is someting like a precompiler variable.)*

```
%let required_rows = 700;
```

Now calculate what fraction of the data this is (and hope ORACLE does not scan the whole table). Again I put this

in a macro variable. *(For those familiar with other languagues, yes SAS has techniques to fill in macro variable while executing the code. That is possible because SAS does stepwise compilation.)*

```
proc sql;
    select &required_rows / rows_tbl1
    into :required_fraction
    from connection to oracle
    (   select count(*) as rows_tbl1
        from tbl1 );
```

And finaly retreive about that many records

```
    select * from connection to oracle
    (   select *
        from tbl1
        where DBMS_RANDOM.VALUE <
&required_fraction );
quit;
```

Remarks :

- I did not test this code.

- If you need exactly 700 rows, you might retreive twice the required fraction and randomly sample in SAS