GHSUG, Oct. 24, 2014

TransUnion Analytic Team

Database Management with SAS

Robin Zhao
Population Health Research Institute (Current)
TransUnion Canada (2014.01 - 2014.09)

Agenda

- A background of working environment
- Objective of the SAS macro: DB2Space
- Different functionalities
- Structure of the Macro
- More possibilities
- Q&A

TransUnion

A background of working environment

Data source:

DB2 Data Warehouse (Unix environment accessed via ODBC)

Data Warehouse Table Size's:

Tables range from 350GB to 5MB.

Average table has millions of rows with more than tens variables.

Analytics team working space (fixed at 225GB):

A new schema was created in the DB2 environment so that the analytics team could create their our own DB2 tables

TransUnion

Objective of SAS macro: DB2space

Problem:

How to manage limited space (225GB) being used by multiple people and processes?

Solution:

The DB2SPACE macro

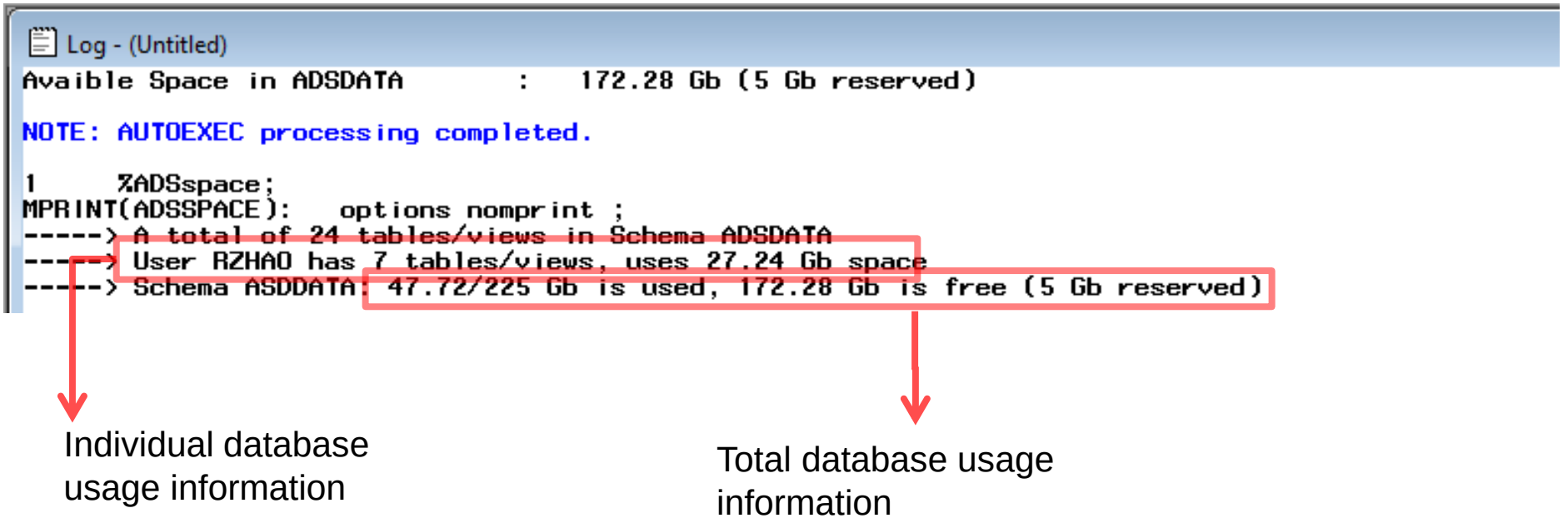
A Marco with 4 functions

1. Query current remaining space
2. Create a SAS table with space statistics
3. Create a detailed database usage report in Excel
4. Generate automatic maintenance emails to database users

TransUnion

Different functionalities

1. Create log message as well as a global macro variable that can be used in your SAS code to control the execution of you data set or program.

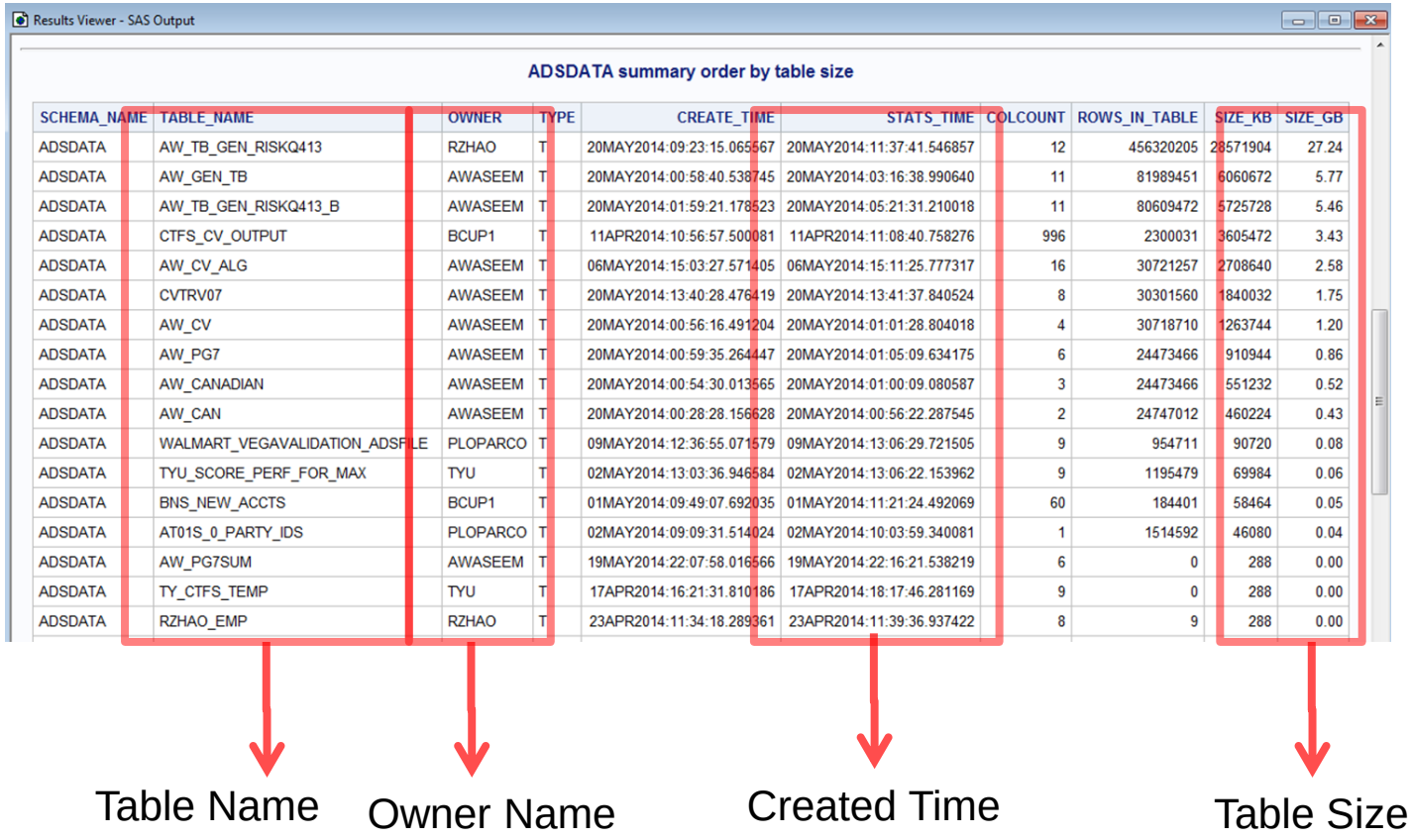


Individual database usage information

Total database usage information

Different functionalities

2. View all table in schema from SAS environment

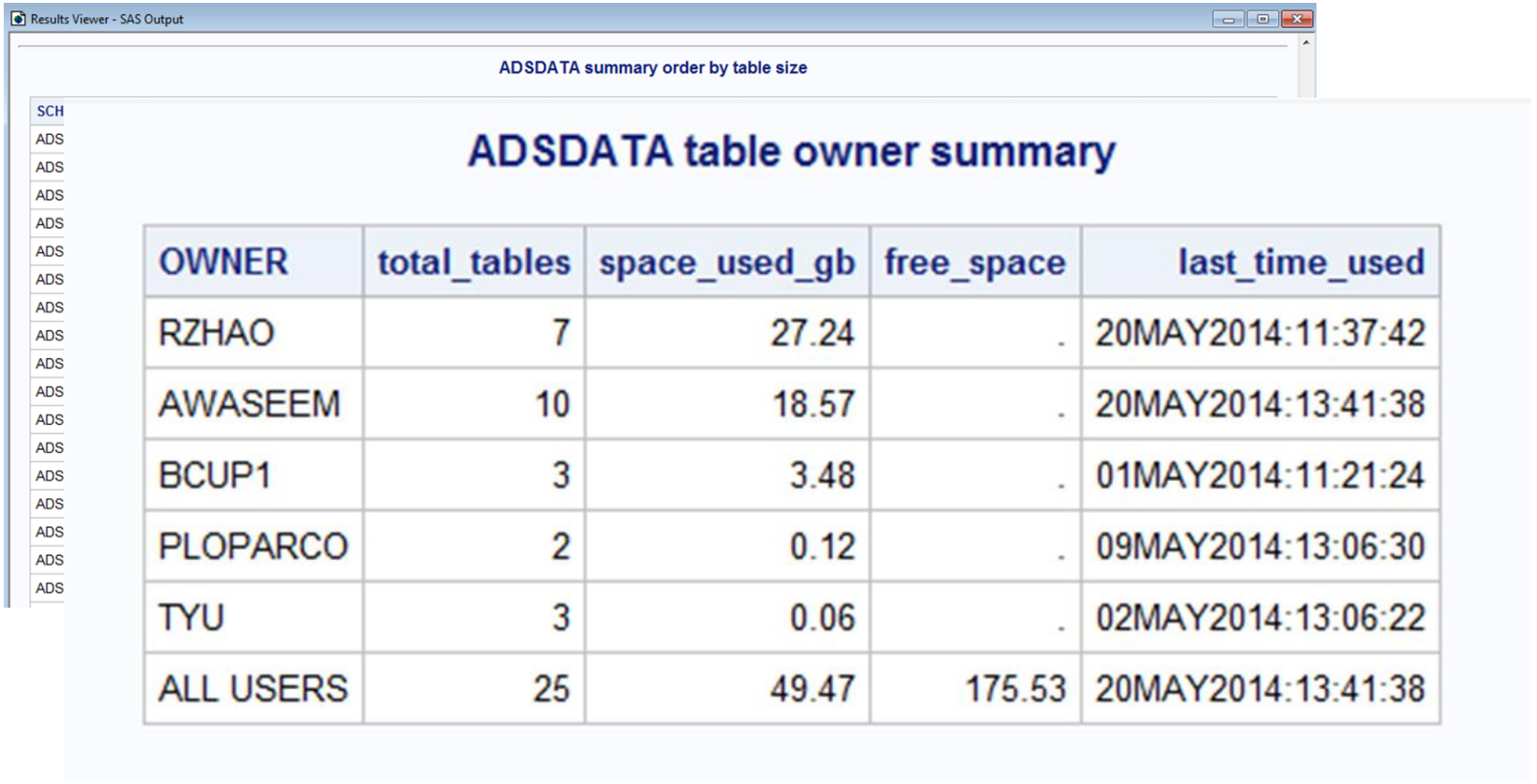


SCHEMA_NAME	TABLE_NAME	OWNER	TYPE	CREATE_TIME	STATS_TIME	COLCOUNT	ROWS_IN_TABLE	SIZE_KB	SIZE_GB
ADSDATA	AW_TB_GEN_RISKQ413	RZHAO	T	20MAY2014:09:23:15.065567	20MAY2014:11:37:41.546857	12	456320205	28571904	27.24
ADSDATA	AW_GEN_TB	AWASEEM	T	20MAY2014:00:58:40.538745	20MAY2014:03:16:38.990640	11	81989451	6060672	5.77
ADSDATA	AW_TB_GEN_RISKQ413_B	AWASEEM	T	20MAY2014:01:59:21.178523	20MAY2014:05:21:31.210018	11	80609472	5725728	5.46
ADSDATA	CTFS_CV_OUTPUT	BCUP1	T	11APR2014:10:56:57.500081	11APR2014:11:08:40.758276	996	2300031	3605472	3.43
ADSDATA	AW_CV_ALG	AWASEEM	T	06MAY2014:15:03:27.571405	06MAY2014:15:11:25.777317	16	30721257	2708640	2.58
ADSDATA	CVTRV07	AWASEEM	T	20MAY2014:13:40:28.476419	20MAY2014:13:41:37.840524	8	30301560	1840032	1.75
ADSDATA	AW_CV	AWASEEM	T	20MAY2014:00:56:16.491204	20MAY2014:01:01:28.804018	4	30718710	1263744	1.20
ADSDATA	AW_PG7	AWASEEM	T	20MAY2014:00:59:35.264447	20MAY2014:01:05:09.634175	6	24473466	910944	0.86
ADSDATA	AW_CANADIAN	AWASEEM	T	20MAY2014:00:54:30.013565	20MAY2014:01:00:09.080587	3	24473466	551232	0.52
ADSDATA	AW_CAN	AWASEEM	T	20MAY2014:00:28:28.156628	20MAY2014:00:56:22.287545	2	24747012	460224	0.43
ADSDATA	WALMART_VEGAVALIDATION_ADSFILE	PLOPARCO	T	09MAY2014:12:36:55.071579	09MAY2014:13:06:29.721505	9	954711	90720	0.08
ADSDATA	TYU_SCORE_PERF_FOR_MAX	TYU	T	02MAY2014:13:03:36.946584	02MAY2014:13:06:22.153962	9	1195479	69984	0.06
ADSDATA	BNS_NEW_ACCTS	BCUP1	T	01MAY2014:09:49:07.692035	01MAY2014:11:21:24.492069	60	184401	58464	0.05
ADSDATA	AT01S_0_PARTY_IDS	PLOPARCO	T	02MAY2014:09:09:31.514024	02MAY2014:10:03:59.340081	1	1514592	46080	0.04
ADSDATA	AW_PG7SUM	AWASEEM	T	19MAY2014:22:07:58.016566	19MAY2014:22:16:21.538219	6	0	288	0.00
ADSDATA	TY_CTFS_TEMP	TYU	T	17APR2014:16:21:31.810186	17APR2014:18:17:46.281169	9	0	288	0.00
ADSDATA	RZHAO_EMP	RZHAO	T	23APR2014:11:34:18.289361	23APR2014:11:39:36.937422	8	9	288	0.00

TransUnion

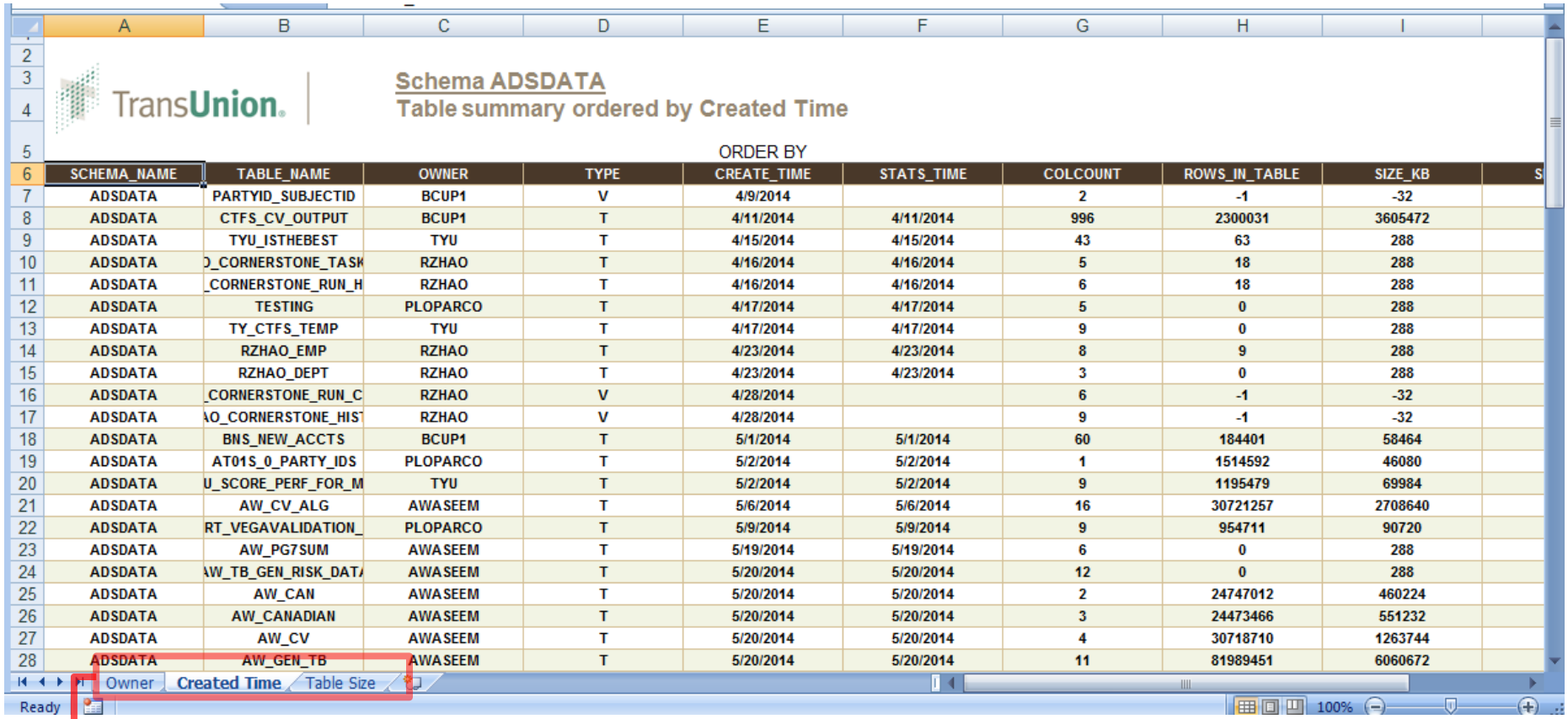
Different functionalities

2. View database usage information divided by each user



Different functionalities

3. Save detailed current database usage information in Excel spreadsheet %DB2space(excel=yes);



TransUnion

Schema ADSDATA
Table summary ordered by Created Time

ORDER BY

SCHEMA_NAME	TABLE_NAME	OWNER	TYPE	CREATE_TIME	STATS_TIME	COLCOUNT	ROWS_IN_TABLE	SIZE_KB
ADSDATA	PARTYID_SUBJECTID	BCUP1	V	4/9/2014		2	-1	-32
ADSDATA	CTFS_CV_OUTPUT	BCUP1	T	4/11/2014	4/11/2014	996	2300031	3605472
ADSDATA	TYU_ISTHEBEST	TYU	T	4/15/2014	4/15/2014	43	63	288
ADSDATA	O_CORNERSTONE_TASK	RZHAO	T	4/16/2014	4/16/2014	5	18	288
ADSDATA	_CORNERSTONE_RUN_H	RZHAO	T	4/16/2014	4/16/2014	6	18	288
ADSDATA	TESTING	PLOPARCO	T	4/17/2014	4/17/2014	5	0	288
ADSDATA	TY_CTFS_TEMP	TYU	T	4/17/2014	4/17/2014	9	0	288
ADSDATA	RZHAO_EMP	RZHAO	T	4/23/2014	4/23/2014	8	9	288
ADSDATA	RZHAO_DEPT	RZHAO	T	4/23/2014	4/23/2014	3	0	288
ADSDATA	_CORNERSTONE_RUN_C	RZHAO	V	4/28/2014		6	-1	-32
ADSDATA	AO_CORNERSTONE_HIST	RZHAO	V	4/28/2014		9	-1	-32
ADSDATA	BNS_NEW_ACCTS	BCUP1	T	5/1/2014	5/1/2014	60	184401	58464
ADSDATA	AT01S_0_PARTY_IDS	PLOPARCO	T	5/2/2014	5/2/2014	1	1514592	46080
ADSDATA	U_SCORE_PERF_FOR_M	TYU	T	5/2/2014	5/2/2014	9	1195479	69984
ADSDATA	AW_CV_ALG	AWASEEM	T	5/6/2014	5/6/2014	16	30721257	2708640
ADSDATA	RT_VEGAVALIDATION_	PLOPARCO	T	5/9/2014	5/9/2014	9	954711	90720
ADSDATA	AW_PG7SUM	AWASEEM	T	5/19/2014	5/19/2014	6	0	288
ADSDATA	AW_TB_GEN_RISK_DATA	AWASEEM	T	5/20/2014	5/20/2014	12	0	288
ADSDATA	AW_CAN	AWASEEM	T	5/20/2014	5/20/2014	2	24747012	460224
ADSDATA	AW_CANADIAN	AWASEEM	T	5/20/2014	5/20/2014	3	24473466	551232
ADSDATA	AW_CV	AWASEEM	T	5/20/2014	5/20/2014	4	30718710	1263744
ADSDATA	AW_GEN_TB	AWASEEM	T	5/20/2014	5/20/2014	11	81989451	6060672

Owner | Created Time | Table Size

Ready

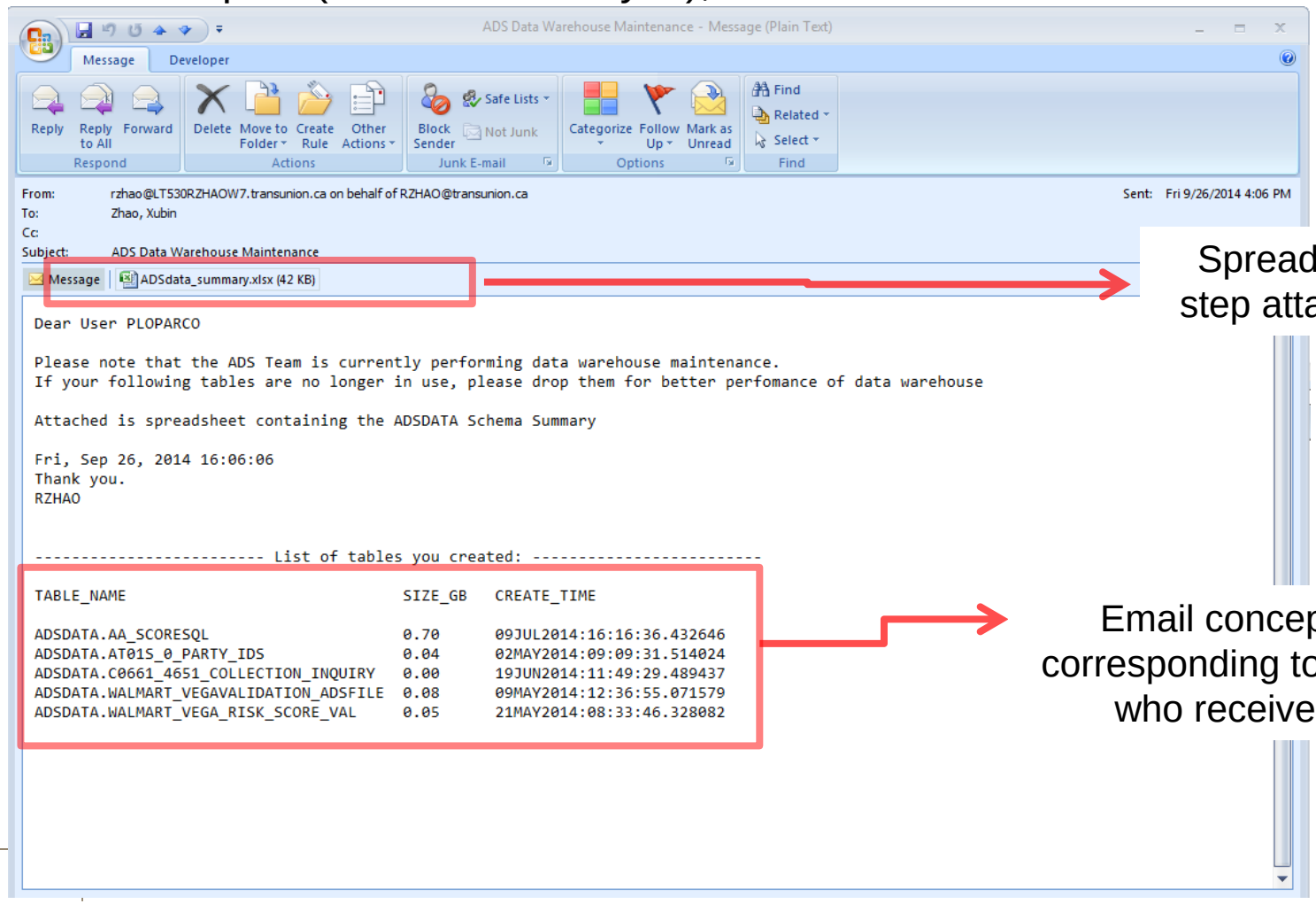
Output stats to Excel template

TransUnion

Different functionalities

4. Generate automatic maintenance emails to all database users

%DB2space(maintenance = yes);



Spreadsheet from last step attached with email

Email concept will be corresponding to the person who received email

TransUnion

Main Structure of the Macro

1. Query database administrational information

```
/*----------------------------------------------------------------------------*/
/* Load Qurey in Macro                                                        */
/*----------------------------------------------------------------------------*/
%macro SQLCode1;

select a.tabschema as schema_name,
       a.tabname as table_name,
       a.owner, a.type, a.create_time, a.stats_time, a.colcount,
       card as rows_in_table,
       (float(a.fpages)*PAGESIZE/1024) as size_Kb,
       decimal(float(a.fpages)*PAGESIZE/(power(1024,3)),8,2) as size_Gb
    from syscat.tables a,
         syscat.tablespaces b
    where a.TBSPACEID=b.TBSPACEID and
          tabschema='ADSDATA'
    order by a.owner, a.tabname

%mend sqlcode1;


/*----------------------------------------------------------------------------
/* Run Qurey and creat SAS dataset with results
/*----------------------------------------------------------------------------*/

proc sql;
    connect to DB2 (USER=&ODBC_User password= "&ODBC_pw." database=&ODBC_PROD );

      create table out_data.adsdata_spacesummary as
          select * from connection to db2
      (%sqlcode1);

      disconnect from db2;
quit;
```

The tables sizes are estimated with database administration information, this code is for DB2

Main Structure of the Macro

2. Output simple database usage

```
proc sql noprint;
   select count(*), sum(size_gb),
          &adsdata_total_space -&adsdata_space_delimiter -sum(size_gb) into :ntable, :totalsize, :leftsize
   from out_data.adsdata_spacesummary;
quit;

proc sql noprint;
   select count(*), sum(size_gb) into :user_ntable, :user_totalsize
   from out_data.adsdata_spacesummary
   where owner = "&ODBC_User";
quit;



%put -----> A total of %QTRIM(&ntable) tables/views in Schema ADSDATA;
%put -----> User &ODBC_User has %QTRIM(&user_ntable) tables/views, uses %QTRIM(&user_totalsize) Gb space;
%put -----> Schema ASDDATA: %QTRIM(&totalsize)/&adsdata_total_space Gb is used, %QTRIM(&leftsize) Gb is free
            (&adsdata_space_delimiter Gb reserved);
```

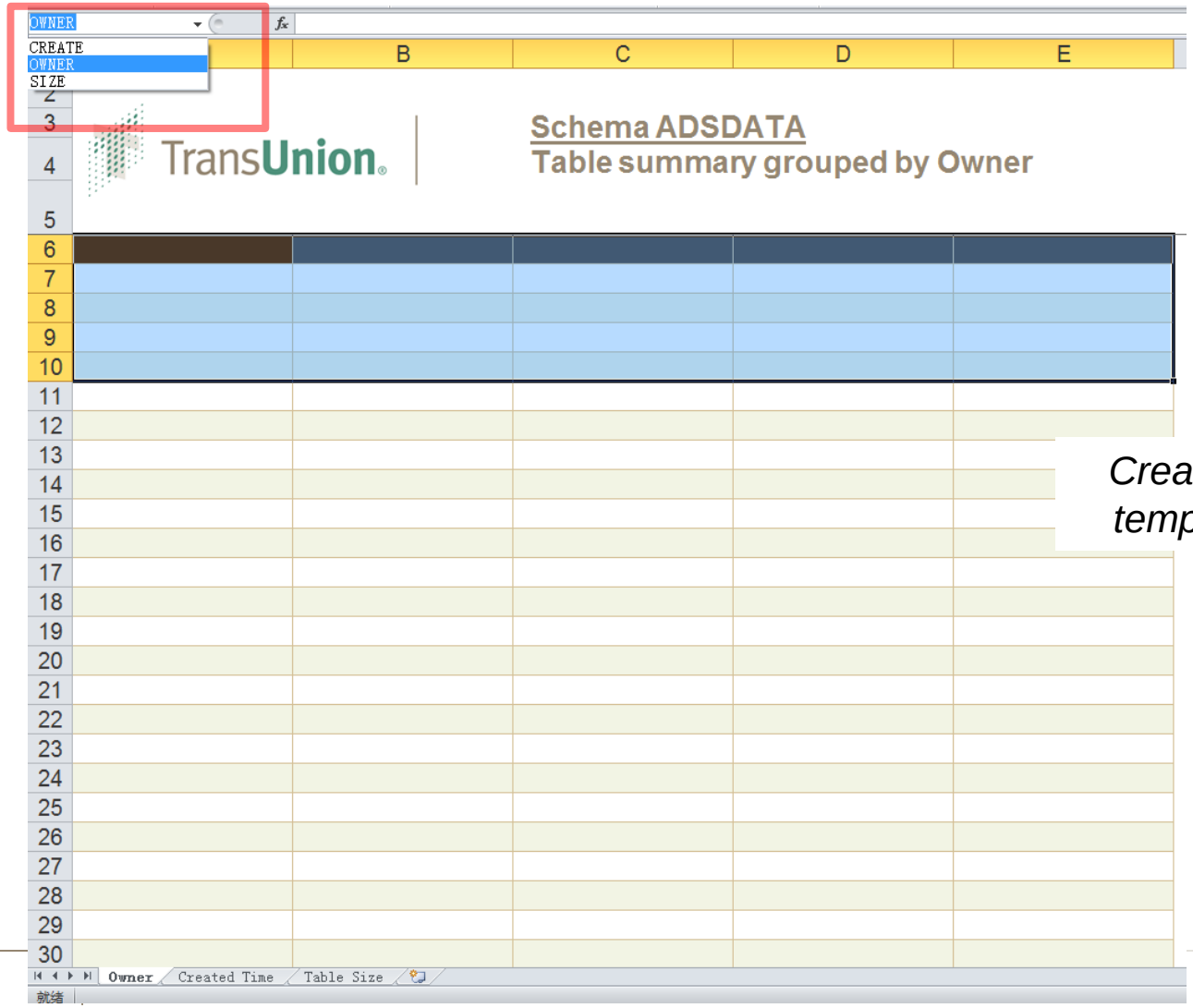
Log - (Untitled)

```
Avaible Space in ADSDATA        :   172.28 Gb (5 Gb reserved)

NOTE: AUTOEXEC processing completed.

1     %ADSspace;
MPRINT(ADSSPACE):   options nomprint ;
-----> A total of 24 tables/views in Schema ADSDATA
-----> User RZHAO has 7 tables/views, uses 27.24 Gb space
-----> Schema ASDDATA: 47.72/225 Gb is used, 172.28 Gb is free (5 Gb reserved)
```

TransUnion

Main Structure of the Macro

3. Create Excel spreadsheet



Create an empty Excel template to insert data

TransUnion

Main Structure of the Macro

3. Create Excel spreadsheet

```
/*-------------------------------------------------------------------------*/
/* Creat copy of Template for loading                                      */
/*-------------------------------------------------------------------------*/
%if  &excel = YES or &email = YES or &maintenance = YES  %then %do;

%let rc =%sysfunc(fileexist("&outputroot.\ADSdata_summary.xlsx"));
%if &rc=1 %then %do;
    %let rc1=%sysfunc(filename(rawfile, &outputroot.\ADSdata_summary.xlsx));
    %let rc2=%sysfunc(fdelete(&rawfile));
%end;

%sysExec copy "C:\PRODSAS\MPRODSAS\ADSdata\Templates\ADSdata_template.xlsx" "&outputroot.\ADSdata_summary.xlsx" ;

/*-------------------------------------------------------------------------*/
/* Create Excel output                                                     */
/*-------------------------------------------------------------------------*/
libname xls pcfiles path="&outputroot.\ADSdata_summary.xlsx";

proc datasets lib = xls nolist;
    delete CREATE SIZE OWNER;
quit;

proc sql;
create table xls.CREATE as
%SQLCode2;
create table xls.SIZE as
%SQLCode3;
create table xls.OWNER as
%SQLCode4;
quit;

libname xls clear;
%end;
```

Make a copy of Excel template

Insert data to selected area in excel

Main Structure of the Macro

4. Email automatic maintenance emails

```
/*-------------------------------------------------------------------------------*/
/* Create Database Maintenance Email                                             */
/*-------------------------------------------------------------------------------*/

%if  &maintenance = YES  %then %do;

proc sql;
create table out_data.adsdata_owners as
select owner, count(*)+1 as count_tables from out_data.adsdata_spacesummary group by owner;
quit;

proc sql noprint;
select count(*) into :count_owner from out_data.adsdata_owners;
quit;

%do i=1 %to &count_owner;
data _null_;
      set out_data.adsdata_owners;
      if _n_= &i then do;
      call symput ('owner',trim(left(owner)));
      call symput ('count_tables',trim(left(count_tables)));
      end;
run;

proc sql;
create table out_data.adsdata_part as
select * from out_data.adsdata_spacesummary where owner = "&owner";
quit;
```

 |

Main Structure of the Macro

4. Email automatic maintenance emails

```
/* SET OPTIONS STATEMENT                                                    */
/*-------------------------------------------------------------------------*/
options emailsys=SMTP emailhost=dchexchange.transunion.ca emailport=25;
/*-------------------------------------------------------------------------*/
/* GENERATE EMAIL MESSAGE                                                   */
/*-------------------------------------------------------------------------*/
Filename mailbox email;
DATA _NULL_;
FILE MailBox TO=("&emailto")
             CC=('')
             FROM="&emailfrom"
             REPLYTO=''
             SUBJECT="&emailsubject"
             ATTACH="&outputroot.\ADSdata_summary.xlsx"
             NOTITLES RECFM=V lrecl=200;
    put "Dear User &owner";
    put " ";
    put "Please note that the ADS Team is currently performing data warehouse maintenance.";
    put "If your following tables are no longer in use, please drop them for better perfomance of data warehouse";
    put " ";
    put "Attached is spreadsheet containing the ADSDATA Schema Summary";
    put " ";
    put "%sysfunc(date(),weekdate17.) %sysfunc(time(),time8.0)";
    put "Thank you.";
    put "&ODBC_User. ";
    put " ";
    put " ";
    put "------------------------ List of tables you created: ------------------------";
    put " ";
    put "TABLE_NAME                              SIZE_GB    CREATE_TIME      ";
    put " ";
```

More possibilities

1. Imply to more databases, such as Oracle, MYSQL, SQL server, and more

2. Create a historical dialog of database summary

Acknowledgement

Thank Peter Loparco for leading, feedback& suggestions and rehearsal.

Thank Justin Jia for suggestions and encouragements.

Thank Leanne Dyal for supporting.

Q&A

Robin Zhao

Email: robin.zhao@phri.com

Thank you