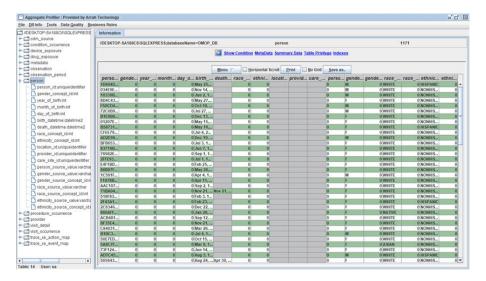# AGGREGATE PROFILER

Data profiling and analysis tool

Links

- ✓ https://sourceforge.net/projects/dataquality/
- ✓ http://www.arrahtech.com/docs/installation_guide.html

License	GNU General Public License version 3.0 (GPLv3
Version	v6.3.0
Last Update	4/27/2020
OS	Windows, Solaris, Linux

System Requirements   JDK / JRE 1.7.0_07 or higher

Description   Aggregate Profiler (AP) also known as "Open Source Data Quality and Profiling", is an open source project, dedicated to Data Profiling and developed in Java. AP supports both traditional database and big data, such as Hadoop or Hive, and it offers statistical analysis, pattern matching, distribution chat, basket analysis, etc. AP also supports data generation, data preparation, data masking features, and address correction for data quality projects. Moreover, this tool offers data validation (metadata profiling, analytical profiling, and structural profiling), and data quality (removing duplicate data, null values, and dirty data)

Features

- Data profiling, filtering, and governance
- Similarity checks
- Data enrichment
- Real time alerting for data issues or changes
- Basket analysis with bubble chart validation
- Single customer view
- Dummy data creation
- Metadata discovery
- Anomaly discovery and data cleansing tool
- Hadoop integration

Connectivity / Supported Data Sources & Formats

- XML, XLS or CSV format, PDF export
- Teiid, Mysql, Oracle,Postgres,Access,Db2,SQL Server certified Big data support - HIVE

Limitations

Aggregate Profiler user interface freezes and is inferior compared to other tools and handling of functions was not intuitive. It will take additional training and understanding of the product to fully use all capabilities.

Performance

Aggregate Profiler successfully loads the 1.3M data but the UI freezes when we try to review the analysis

Feedback from Cystic Fibrosis Team:

Kieran Earlam, Policy and Evidence Manager at Cystic Fibrosis Trust provided the following feedback:

- The aggregate profiler presented difficulties from the start, after installation and running of the windows bat file I was presented with a command prompt window, and the Arrah connection dialog. I attempted to fill out the connection dialog and then was prompted to update my Java, upon updating my JAVA the system crashed and would not reboot.

- Additionally, the Aggregate profiler required specific knowhow/qualifications that I did not meet – "Window User should know how to create DSN or Use existing DSN for MS Access, SQL Server, MySQL, and Oracle on Windows. "

  The supporting document "**Data Quality Tool Assessment Request - Data Custodian Cystic Fibrosis.docx**" contains all feedback provided by the Cystic Fibrosis team.

AGGREGATE PROFILER DATA ANALYSIS

1. Download Aggregate Profiler from "https://www.softpedia.com/get/Internet/Servers/Database-Utils/Aggregate-Profiler.shtml"
2. Open the downloaded folder and start the "runprofiler.bat".
3. Give the new connection details in the below format, and verify the Connection is successful.

4. View the data and the statistics as shown below,

SAMPLE STATISTICS:

visit_occurrence MetaData Information ✕

Menu ▽   ☐ Horizontal Scroll   Print   ☐ No Grid   Save as..

Table	Column	DBType	SQLType	Size	Precision	Radix	Remark	Default	Bytes	Ordinal Pos	Nullable
visit_occurr...	visit_occurr...	uniqueidentif...	Char	36						1	NO
visit_occurr...	person_id	uniqueidentif...	Char	36						2	NO
visit_occurr...	visit_concep...	int	Integer	10	0	10				3	NO
visit_occurr...	visit_start_d...	date	Date	10	0					4	YES
visit_occurr...	visit_start_d...	datetime2	TimeStamp	27	7					5	NO
visit_occurr...	visit_end_da...	date	Date	10	0					6	YES
visit_occurr...	visit_end_da...	datetime2	TimeStamp	27	7					7	NO
visit_occurr...	visit_type_c...	int	Integer	10	0	10				8	NO
visit_occurr...	provider_id	uniqueidentif...	Char	36						9	YES
visit_occurr...	care_site_id	uniqueidentif...	Char	36						10	YES
visit_occurr...	visit_source...	varchar	VarChar	50					50	11	YES
visit_occurr...	visit_source...	int	Integer	10	0	10				12	NO
visit_occurr...	admitted_fro...	int	Integer	10	0	10				13	NO
visit_occurr...	admitted_fro...	varchar	VarChar	50					50	14	YES
visit_occurr...	discharge_t...	varchar	VarChar	50					50	15	YES
visit_occurr...	discharge_t...	int	Integer	10	0	10				16	NO
visit_occurr...	preceding_vi...	uniqueidentif...	Char	36						17	YES

DUPLICATE PERCENTAGE:

person Information ✕

Menu ▽   ☐ Horizontal Scroll   Print   ☐ No Grid   Save as..

Information	Count
Total Count	1171.0
Duplicate Pattern	0.0
Duplicate Count	0.0
Duplicate %	0.0
Unique Pattern	1171.0
Unique %	100.0

MISSING VALUES DETAILS:

File   DB Info   Tools   Data Quality   Business Rules

Information

//DESKTOP-SA188C6\SQLEXPRESS

- cdm_source
- condition_occurrence
- device_exposure
- drug_exposure
- metadata
- observation
- observation_period
- person
  - person_id:uniqueidentifier
  - gender_concept_id:int
  - year_of_birth:int
  - month_of_birth:int
  - day_of_birth:int
  - birth_datetime:datetime2
  - death_datetime:datetime2
  - race_concept_id:int
  - ethnicity_concept_id:int
  - location_id:uniqueidentifier
  - provider_id:uniqueidentifier
  - care_site_id:uniqueidentifier
  - person_source_value:varchar
  - gender_source_value:varchar
  - gender_source_concept_id:int
  - race_source_value:varchar
  - race_source_concept_id:int
  - ethnicity_source_value:varcha
  - ethnicity_source_concept_id:i
- procedure_occurrence
- provider
- visit_detail
- visit_occurrence
- trace_xe_action_map
- trace_xe_event_map

//DESKTOP-SA188C6\SQLEXPRESS;databaseName=OMOP_DB   person   1171

Show Condition MetaData Summary Data Table Privilege Indexes

Menu ▽   ☐ Horizontal Scroll   Print   ☐ No Grid   Save as..

person Information

Menu ▽   ☐ Horizontal Scroll   Print   ☐ No Grid   Save as..

Empty Column #	Count	Percentage
0 Empty Column	0	0.0
1 Empty Column	0	0.0
2 Empty Column	0	0.0
3 Empty Column	171	14.602903501280956
4 Empty Column	1000	85.39709649871904
5 Empty Column	0	0.0
6 Empty Column	0	0.0
7 Empty Column	0	0.0
8 Empty Column	0	0.0
9 Empty Column	0	0.0
10 Empty Column	0	0.0
11 Empty Column	0	0.0
12 Empty Column	0	0.0
13 Empty Column	0	0.0
14 Empty Column	0	0.0
15 Empty Column	0	0.0
16 Empty Column	0	0.0
17 Empty Column	0	0.0
18 Empty Column	0	0.0
19 Empty Column	0	0.0

AGGREGATE PROFILER AVAILABLE OPTIONS:

PREDEFINED DATA TYPES:

Standard SQL Type Information

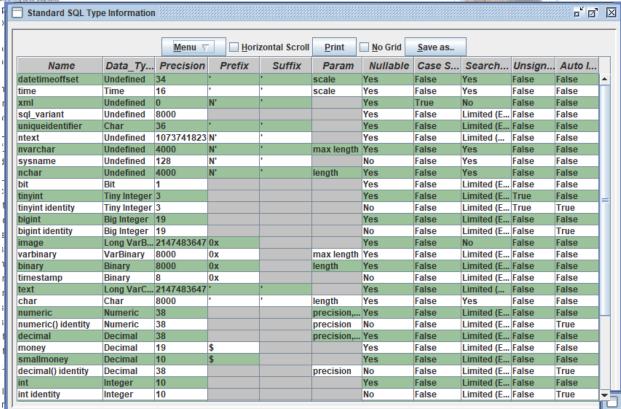| Name	Data_Ty...	Precision	Prefix	Suffix	Param	Nullable	Case S...	Search...	Unsign...	Auto I...
datetimeoffset	Undefined	34	'	'	scale	Yes	False	Yes	False	False
time	Time	16	'	'	scale	Yes	False	Yes	False	False
xml	Undefined	0	N'	'		Yes	True	No	False	False
sql_variant	Undefined	8000				Yes	False	Limited (E...	False	False
uniqueidentifier	Char	36	'	'		Yes	False	Limited (E...	False	False
ntext	Undefined	1073741823	N'	'		Yes	False	Limited (...	False	False
nvarchar	Undefined	4000	N'	'	max length	Yes	False	Yes	False	False
sysname	Undefined	128	N'	'		No	False	Yes	False	False
nchar	Undefined	4000	N'	'	length	Yes	False	Yes	False	False
bit	Bit	1				Yes	False	Limited (E...	False	False
tinyint	Tiny Integer	3				Yes	False	Limited (E...	True	False
tinyint identity	Tiny Integer	3				No	False	Limited (E...	True	True
bigint	Big Integer	19				Yes	False	Limited (E...	False	False
bigint identity	Big Integer	19				No	False	Limited (E...	False	True
image	Long VarB...	2147483647	0x			Yes	False	No	False	False
varbinary	VarBinary	8000	0x		max length	Yes	False	Limited (E...	False	False
binary	Binary	8000	0x		length	Yes	False	Limited (E...	False	False
timestamp	Binary	8	0x			No	False	Limited (E...	False	False
text	Long VarC...	2147483647	'	'		Yes	False	Limited (...	False	False
char	Char	8000	'	'	length	Yes	False	Yes	False	False
numeric	Numeric	38			precision,...	Yes	False	Limited (E...	False	False
numeric() identity	Numeric	38			precision	No	False	Limited (E...	False	True
decimal	Decimal	38			precision,...	Yes	False	Limited (E...	False	False
money	Decimal	19	$			Yes	False	Limited (E...	False	False
smallmoney	Decimal	10	$			Yes	False	Limited (E...	False	False
decimal() identity	Decimal	38			precision	No	False	Limited (E...	False	True
int	Integer	10				Yes	False	Limited (E...	False	False
int identity	Integer	10				No	False	Limited (E...	False	True

LIMITATIONS:

- Aggregate profiler loads the 1.3M and when trying to process the UI freezes as shown below,