Automated Exploratory Data Analysis of Databases

Diego Arenas @darenasc

04 November 2021



Diego Arenas

- Computer scientist; MSc in Data Science, University of Edinburgh;
 and EngD © in Computer Science, University of St Andrews.
- 15+ years of experience working in data science and data engineering projects
- Co-host of the podcast "escuchAI" about AI (http://escuchai.com)
- Personal webpage https://darenasc.github.io

Agenda

- Motivation
- Data Exploration
- What is aeda
- How to use aeda
- Demo
- Use cases
- Q&A

Do you know the content of all the databases in your organisation?

Data exploration

- Important initial phase in any data related project
- Analyse the content of a data source
- Making sense of the data
- Some data sources have hundreds of tables
- Data quality
- Data governance

Why automate the data exploration phase?

- Series of *repetitive* steps we follow to try to understand the data we are working with
- Time of manual data exploration is often linear 10 new tables will take
 10x additional time to process
- Multiple data sources?
- Standardised way to explore data sources
- We want to see trends, distributions of data, data quality assessment, etc.
- Working with big tables can be optimised from GB to KB of data and waiting times

What is aeda?

- AEDA stands for Automated Exploratory Data Analysis
- It uses two databases: a
 Source database and a
 Metadata database
- Extracts metadata from the source and creates a data catalogue of metadata

Database	SOURCE	METADATA
MySQL SOURCE / METADATA	$\overline{\checkmark}$	$\overline{\checkmark}$
✓ MariaDB SOURCE / METADATA	V	V
Postgres SOURCE / METADATA	V	V
MS SQL Server SOURCE / METADATA	V	V
SQLite3 METADATA		V
Snowflake METADATA		V

What metadata?

Basic 1. information	Server name, table catalog, table schema, table names, column names, ordinal position of the columns, and column data types.
(2.	number of tables in the database. number of rows per table. number of columns per table.
Project size $<$ 3.	number of rows per table.
4.	number of columns per table.
	number of unique values per column.
Discrete data type \implies 6.	number of null values per column.
Continuous data type \Longrightarrow 7.	frequency number per data value per column.
Datetime data type \Longrightarrow 8.	timewise aggregation of the data.
Continuous data type \Longrightarrow 9.	univariate summary statistics for all the numeric data types:
	mean, standard deviation, variance, maximum, minimum,
	percentiles (0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95,
	0.975, 0.99), interquartile range, range, kurtosis, and skewness.

How to use aeda?

```
[my-source-database]
db_engine = <A-SUPPORTED-DB-ENGINE>
host = <IP-OR-HOSTNAME-SOURCE-DATABASE>
schema = <SCHEMA-SOURCE-DATABASE>
catalog = <CATALOG-SOURCE-DATABASE>
user = <SOURCE-USER>
password = <SOURCE-PASSWORD>
port = <SOURCE-PORT>
[my-metadata-database]
ab_engine = <A-SUPPUKTED-DB-ENGINE>
host = <IP-OR-HOSTNAME-METADATA-DATABASE>
schema = <SCHEMA-METADATA-DATABASE>
catalog = <CATALOG-METADATA-DATABASE>
user = <METADATA-USER>
password = <METADATA-PASSWORD>
port = <METADATA-PORT>
metadata database = yes
```

From the terminal python aeda_.py explore my-source-database my-metadata-database

Data exploration

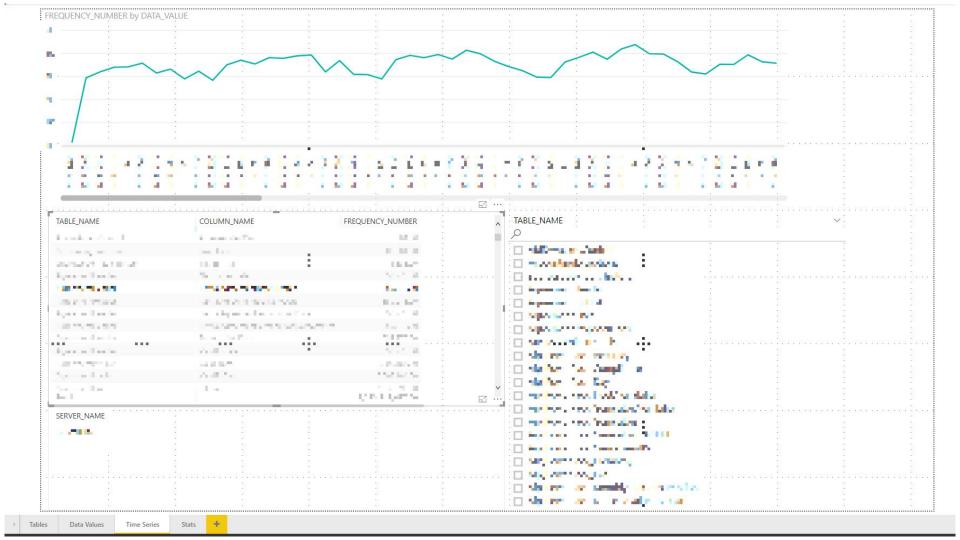
GUI

streamlit run aeda app.py

Demo

- Creation of the connections to databases
- Running an exploration
- Examine the data

(Blank)		10.202	*	O.00 int		12	10
	Tracket -	W. M.	100000		5bn	14.80K	823
Track military a toy of the control of the			111111	0.00 bigint 0.00 bigint	2011	14.0UN	023
10.00.0				0.00 bigint	N ROWS	N COLUMNS	N TABLE
J 30 00 00 00 00	Total Comment	harana a	2010 BT 2010			IN_COLUMNS	IN_IADLE
8 8	he bedrelle delle	Secretary III	801.71	0.00 int	N_ROWS by TABLE_NAME		1
	The Control of the Control	20 010000000		0.00 int		22	
TABLE NAME	The second of th	AND THE RESIDENCE OF THE PARTY.	O-20-2	0.00 Int	to said		1.4
TABLE_NAME :		(2.12)	40.00	0.00 bigint	N/ N/ Page 1		24 6 6
ρ	BALL A TURN	Transaction	- LIB.	0.00 bigint			0.44
(Blank)		PA AL	0.000	0.00 datetime		_	
□ MATA Malay	Department	201-100-00	7.77.00	0.00 datetime	The second secon		
□ MatCombuters/Seeders	TO THE PARTY OF TH	II.	2414712	0.00 Int	Tr. Trib. Act	4.5	
□ Millioned	manage spenger person			0.00 bigint	4 14 4	1.2	Ž.
□ Secondarios	The second second	-	100000	0.00 bigint	11 45 14.7	* Sec.	
in the possible place of the same			WE KIND	0.00 bigint	Red S. Property	111%	
 terposeting broudons 	and a mode	P1 04	10000000	0.00 datetime		1 41	:
□ N/VIII/NIN/VIII/	INCHES NO	IN ON	14.5	0.00 datetime		- 4	
☐ M. L. Anteresp., Clinco.		NAME OF TAXABLE	1 4 1 2 2 2	0.00 datetime		1.11	
□ NOOMINE PROTECTION	and the second	PT AN	979.715	0.00 datetime		111	
□ Marion	11 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		PER DATE:	0.00 bigint	The state of the s		
The State Code History	Table 1 a	2012/2014	WELL BUTTON	0.00 datetime	But the property and	index 1	
Classification	10 mm - 10 Million	THE REAL PROPERTY.	Acres 200 (C)	0.00 datetime		-4.75	:
TI NE AssessMillered	LOWNER .	No section	200420	0.00 int	20 1 10 Th American	100	(8)
The State of the s	Could be a production	ALPRES P	Mark Park	0.00 bigint	New Yorks (Sec.	15.0	
□ total extraction is bloken.	way or belief an	Congress of the Control	F 1 1 1 1 1 1 2	0.00 int	The state of the same	8 AV	
indian discons	Francis VIII	A TAX SERVE	Table 10 * at	0.00 int	Last Special	****	
□ hyddi none	Commence of the same	and of solutions 1	1000	0.00 int	When Land	24.44	
Maria Constant	CANADA IN IN	IPS AND	9.14 (19.45)	0.00 datetime		1.14	:
Million March 1 III	CONTRACTOR MANAGEMENT	MANUFACTURE CO.	AT 101 AT AT	0.00 int		0.00	:
Special colors have	and the same of th	TO MA	VM = 1, 575	0.00 datetime			
□ Manager and	Description of the second	41	The state of the s	0.00 float	I make a	0.00	
Suppose Jim Mills Sci Coll	Colt of the Section	North Control	A11.100.00	1.00 bigint	Park Million 1	1184	
	1-1	To discount	and a fell fill	0.00 int	Later Artist Co.	1.50	
- Species Colombia	100 mg (100)	¥	23 S E 1 47	0.00 float	Management of the last	2010	
Distriction with the	The second second second	Section 19 and 19	THE PARTY	0.00 bigint	And the second	1906	
Special State Special		Control of the section	9000 5111	0.00 bigint	V		
 Suprama V consideration 				0.00 int >	0.0bn	1 S 2 S	2.70



Use cases

- 1. New project and requires inspection of the new data sets
- 2. Creation of dev database from a live database
- 3. Data quality inspection
- 4. New CDC or ETL processes from new systems
- 5. Machine learning project
- 6. Find data visible in applications
- "Searching for a needle in a haystack", tracing data values in data sources.
- 8. Link data by value domain
- 9. Rapid growing database of IoT devices?
- 10. Multiple data sources to explore?
- 11. ...

Excellent libraries in Python

- Pandas Profiling https://github.com/pandas-profiling/pandas-profiling
- Soda SQL https://github.com/sodadata/soda-sql
- Sweeviz https://github.com/fbdesignpro/sweetviz
- Datapane <a href="https://github.com/datapane/datapan
- Great expectations
 <a href="https://github.com/great-expectations/great-expectatio
- Superset https://github.com/apache/superset
- Metabse https://github.com/metabase/metabase/

Questions?

Get in touch https://calendly.com/darenasc/

Link to the library:

https://github.com/darenasc/aeda



Metadata schema

There are 6 tables in the metadata schema.

Table name	Description			
columns	Each row is a column of a table			
tables	Each row is a table of a database			
uniques	Similar to the columns table adding the number of rows and null values			
data_values	Each row is a unique data value of a column			
dates	Similar to data_values for date related columns			
stats	Each row presents statistics of a numeric column of a table			