

EVL Generation Microservice

EVL Data Generation Microservice provides fast, automated and cost-effective method for data generation. Having proper test environment is a must in many areas like application development, implementing ETL processes, or stress testing. Often the tests can't be done on existing production data so data simulation is the only way how to achieve the goal. The simulated data have to comply with the real-life data patterns and data volumes should be close to the expected peaks.

EVL Generation advantages

- Configuration via excel or csv files pre-filled by "reading" the existing data structures
- Automatic random data generation based on data types
- Customization of data pattern based on filled-in parameters like min-max intervals, null values probability, string ranges ...
- Ability to include custom made data generation functions
- Extremely fast generation of vast amount of data by using low-level IO techniques
- Parallel running of jobs and workflow monitoring
- Low implementation and operating costs

EVL Microservices are built on top of the core EVL software and retain its flexibility, robustness, high productivity, and ability to read data from various sources; including csv files, databases–Oracle, Teradata, SQL Server, etc–and Hadoop streaming data like Kafka and Flume.

EVL Generation functions

Following methods, functions and their variations are available on several data types. All functions return NULL when input is NULL.

Data	Null	Min	Max	Pattern	Description
String					E.g. pattern like [a-zA-Z0-9]
Date				Pattern	Values between 1970-01-01 and 2199-12-31
Number	30%	Min	Max		Intervals for integer and decimals and probability of null values

EVL Generation project

An anonymization project consists of following steps:

1. unzipping EVL distribution and defining a few variables and paths
2. filling-in an excel or csv file defining source type (e.g. csv, Oracle ODBC ...), entity and attribute names and generation parameters to be applied
3. automatic generation of EVL jobs for each entity
4. running EVL jobs in a batch or individually
5. monitoring and tuning

Example

Following example shows a simple implementation of EVL Data Generation

Variables

```
# project directory
CONFIG_FILE_DIRECTORY=$HOME/Project/Generation/
```

```
# configuration file name
CONFIG_FILE=generation-config.csv
```

```
# jobs directory
CONFIG_EVD=evd/config.generation.evd
```

```
# global default parameters
CONFIG_GEN_DEAFULTS=val/config.evl.gen
```

Run

```
# generating evl jobs from the config file
evl run/generate_jobs.evl
```

```
# running the test generation job
evl run/generation.test.evl
```

Src	Entity	Attribute	Data type	Null	Min	Max	Pattern	Description
csv	TEST1	ID	int		1	5000		Setting number interval, no null allowed
csv	TEST1	ACC	int	20%				20% can be null
csv	TEST1	Note	string	70%			[a-zA-Z0-9,-]	allowed characters
csv	TEST1	Sex	string		1	1	['M','F']	list of values
ORCL	TEST2	ID	int		1	300000	100+	list of values: 100, 200, 300...
ORCL	TEST2	Postcode	number	5%	5	5		Postcode must be 5 digits, 5% can be null
ORCL	TEST2	Text	string		3	80		Mandatory minimal text