User Manual for Esgdata

Esgdata is a kit to generate a great deal of data. It can generate data in parallel on all nodes in your cluster as long as storage space in your cluster is enough. The kit is divided into 2 parts. One is esgdata and the other is esgdata.py. The esgdata written in C language is a core of this kit. Python file esgdata.py will invoke esgdata, so we only need to run esgdata.py to generate data.

Requirements

Python version : 2.6 or above
Excel version : excel files with .xls suffix

Preparation

- Clone the tool from /opt/share.
- edit excel according to your demands.
- confirm your cluster information about the number of nodes and hostnames, and edit ‘nodes.conf’ and ‘dirs.conf’.
- please confirm storage in the cluster is enough.

introduction to esgdata

- -INPUT <s> set input excel file to fetch data types. Only support excel files with .xls suffix.
- -RCOUNT <s> set total row count.
- -FORCE [Y|N] overwrite data files without prompting.
- -RNGSEED <n> set RNG seed, the number usually is abbreviation of date. i.e. 20170815.
- -DELIMITER <s> use <s> as field separator in output files.
- -SUFFIX <s> use <s> as suffix of output files, the default is .dat.
- -PARALLEL <n> total number of parallels is <n> on all nodes.
- -CHILD <n> generate <n>th chunk of the parallelized data.

Example

```
./esgdata -INPUT ./excel_test.xls -RCOUNT 10000 -FORCE Y -RNGSEED 20170815 -DELIMITER ,
-SUFFIX .txt -PARALLEL 5 -CHILD 2
```

The esgdata will obtain data types and other constrains through excel_test.xls, and generate 10000 rows data. If data files with the same name exist, new files would overwrite old files because FORCE option is set Y. Seed number 20170815 is used to generate random data through some algorithms. Finally, output files will separated by ‘,’.

Parallel is 5 that is there are 5 children. 10000 rows data would be divided into 5 children to generate. Each child would invoke an esgdata process to generate 2000 rows data. If current environment contains 3 nodes, children would be assigned by order.

Child 1	Child 2	Child 3	Child 4	Child 5
Node 1	node 2	node 3	node 1	node 2
2000	2000	2000	2000	2000

You can get help information through './esgdata --help'.

Introduction to esgdata.py

- -n NODES a file contains all hostnames or IP address of available nodes to generate data. Default is 'nodes.conf'.
- -d DIRS a file contains directories where generate data. Default is 'dirs.conf'. If the location is not created, the process will automatically create it.
- -p PARALLEL total number of generation parallel that is the same as esgdata -PARALLEL.
- -e EXCEL location of excel location.
- -c RCOUNT total row count for this table.
- -G generate data in linux
- -P PUT put all generated data files to hdfs. Need assign a hdfs location. If the location is not created, the process will automatically create it.
- -C clean all generated data files.
- -T PUTTHREAD number of parallels on each node when put in hdfs

Example generate data

```
python esgdata.py -e ./excel_test.xls -c 10000 -p 5 -G
```

Generate 10000 rows data with 5 parallels according to data types in excel_test.xls. There is not configuration for generated path and nodes. So it depends on 'dirs.conf' and 'nodes.conf'. The best way is to set both configure files in advance.

Introduction to the format of excel

Table name is abstract from sheet name in excel. This kit only parser the first sheet in this excel so please keep your definition of data types at the first place and don't change the format of excel.

Support data types

- int
- bigint
- varchar
- decimal

- date
- time
- timestamp
- interval year
- interval month
- interval day
- interval hour
- interval minute
- interval second
- interval year to month
- interval day to hour
- interval hour to minute
- interval minute to second
- interval day to second

meaning of columns

- No. the number of column
- Size character size when set data type only as varchar.
- Nullable this column whether is null. N is not null and Y is allowed to be null.
- Min/Max set minimum and maximum that you expect.
- Sequence start when a starting number is set, data in this column will be sequence added one each time.
- Note you can make some notes as annotation.
- Content please keep this column empty. This feature doesn’t support currently.
- PK decide if this column set as a primary key. But the feature doesn’t support currently.

Varchar

This type will generate a string selected form a text. You can set size of varchar depending on your demands.

You also can set nullable as N or Y. The default nullable is N.

Int

This type will generate positive integer randomly. Its maximum is 2147438647 and minimum is 0. You also can set min/max by yourself. And you can set nullable as N or Y. The default nullable is N. If you need negative integers, please set min/max to limit range of negative integers.

If you set sequence start number, this column will generate a sequence started from the number and add one each time. You would better not set nullable as N. The sequence type doesn’t support min and max. negative sequence number is also available.

Bigint

This type will generate positive integer randomly. Its maximum is 9223372036854775807 and minimum is 0. You also can set min/max by yourself. Nullable setting can be used to generate null value. The default nullable is N. If you need negative integers, please set min/max to limit range of negative integers.

Decimal

This type will generate float randomly. You must set precision and scale according to decimal rules in our SQL Reference. And you can set min/max as you expect. If you set min and max with different scales, the scale will be decided by the maximal scale between the two and ignore values in precision column and scale column if you set them.

Example

Precision = 5, scale = 1, min = 1.2233, max = 99.55

In this case, precision 5 and scale 1 would be ignored because we have set min and max. And the maximum of scale is 4 decided by min which contains maximal scale. So the max actually equals 99.5500. Therefore maximum of precision is 6.

Recommended maximal precision is 18.

Date

This type will generate date randomly from 1999-01-01 to 2017-01-01. Its format is 'YYYY-MM-DD'. You also can set min and max according to your demands. But the date format must be followed. Nullable setting is available for date type.

Time

This type will generate time randomly in 24-hour time system. The format of time is 'HH:mm:SS.ssssss'. You can set precision for time. The precision of time is the number of digits behind the decimal point. Values of precision are from 0 to 6. If you don't set precision, the default is 0. This type support to set min/max that must obey the format strictly and nullable. If you set min and max with different precision, the precision will be decided by the maximal one between the two and ignore the value in precision column.

Timestamp

This type will generate timestamp randomly combined date and time in 24-hour time system. The format of timestamp is 'YYYY-MM-DD HH:mm:SS.ssssss'. You can set precision for timestamp. The precision of timestamp is the number of digits behind the decimal point at time filed. Value of precision is from 0 to 6. If you don't set precision, the default is 0. This type support to set min/max and nullable. The min/max must obey the format strictly. If you set min and max with different precision, the precision will be decided by the maximal one between the two and ignore the value in precision column.

Interval

- Interval type can support
- YEAR(leading-precision),
- MONTH(leading-precision),
- DAY(leading-precision),

- HOUR(leading-precision),
- MINUTE(leading-precision),
- SECOND(leading-precision, fractional-precision),
- YEAR(leading-precision) TO MONTH,
- DAY(leading-precision) TO HOUR,
- HOUR(leading-precision) TO MINUTE,
- MINUTE(leading-precision) TO SECOND(fractional-precision),
- DAY(leading-precision) TO SECOND(fractional-precision).

Maximum of leading-precision is 18. The default leading-precision is 2. Fractional-precision, the number of digits of precision after the decimal point, is only needed in SECOND field. The maximum of fractional-precision is 6. The default fractional-precision is 6. You can set leading-precision at precision column and set fractional-precision at scale column.
All interval types are not supported for setting min/max. Nullable is available to set.