Why Data Generation

Most software developers and database researchers use syn- thetic data to test the correctness and performance of their work. Such test datamust be realistic and correct in terms of size and distributions to be useful, e.g., see the TPC bench- marks [1].

This paper describes a platform for the generation of realistic synthetic RFID data that can facilitate the development and testing of data mining tools for supply chain management.

Generating Synthetic Complex-structured XML Data :

Synthetically generated data has always been important for evaluating and understanding new ideas in database research.

In this paper, we describe a data generator for generating synthetic complex-structured XML data that allows for a high

level of control over the characteristics of the generated data.

Synthetic data generators allow us to generate large volumes of data with well-understood characteristics. We can

easily vary the characteristics of the generated data by varying the input parameters of the data generator.

Of course, while

experimenting with synthetic data is an ideal way to explore the behavior of different solutions on data with different

characteristics, an additional validation step may be necessary to ensure that the conclusions drawn from synthetic data

extend to real world applications.

Creating Realistic, Scenario-Based Synthetic Data for Test and Evaluation of Information Analytics Software :

Our distinctive contribution is that we embed known ground truth in a test data set, so that tool developers and others will be able to determine the effectiveness of their software and how they are progressing in their support for information analysts

A Synthetic Data Generator for Clustering and Outlier Analysis :

Syntheticdatagenerationisaninterestingtopicindatamining.Inmanyresearch areas,asetofstandarddatasetisessentialinevaluatingthequalityofaproposed technique. Methods of generating datasets for diﬀerent purposes can be quite diﬀerent. Our work concentrates on the generation of test instances for clustering and outlier analysis algorithms. There are still much room for improving the current data generating system.

synthetic

(of a proposition) having truth or falsity determinable by recourse to experience

Realistic:

representing things in a way that is accurate and true to life.

**Weighted distributions to mimic real value occurrence rates**

RowGen can also create realistic test data that reflects the way real data occurs in production, or in nature. The default [distributions](http://www.iri.com/blog/test-data/distribution-test-data-values-rowgen/) in RowGen are linear, bell curve, weighted, and weighted with items. RowGen users can also define their own distributions should the test data need to adhere to a different frequency pattern.

Bicycle crash data : Bicycle crashes occurs more frequently outside of winter time in both EU and USA

The six months from may to October accounts for 62% of all bicycle crash fatalities in Europe.

Lowest no. of fatalities occur during winter and conversely the greatest percentage of fatalities occur in summer.

<http://nbda.com/articles/industry-overview-2015-pg34.htm> :

The bicycle industry is a seasonal business that can be impacted by unusual weather, particularly late-arriving Spring weather. It is also an industry that relies on discretionary spending so it is seriously impacted by economic conditions. Overall, industry sales of bicycles seem to be stable at around 15 to 20 million bicycle units annually, plus parts, accessories and service, which historically is a very healthy number for the industry. National trends related to the green movement, environmental sustainability, the need to address health problems related to inactivity, and higher gas prices also bode well for the future of human-powered transportation

The discussion begins by showing how to sequentially generate and populate a table. This algorithm is then generalized to one that generates each partition in parallel. To make the discussion concrete, the following sections take generating the TPC-A ACCOUNTS table as a running example [TPC]. The table will have one billion hundred-byte records (.1 TB) partitioned among the 1000 discs as described in the data definition Program (1) above. Each of the one thousand discs will store 100 MB of data as a B-tree [Knuth]. Since B-trees are only 69% full at equilibrium, each disc will use 150 MB of storage to hold it's B-tree. This is well below the 1 GB capacity of small discs. The remainder of the disk space stores data from other tables.

