Wisconsin Benchmark Data Generator: To JSON and Beyond

Shiva Jahangiri
Advisors: Michael J. Carey, Johann-Christoph Freytag
University of California, Irvine

Motivation

Benchmarks, as one of the greatest assets in evaluating DBMSs, need to evolve with the advancements in data, systems, etc.

Synthetic Data:
- Customizable
- Unrealistic
- Does not capture the complexity and relationships of real data

Real Data:
- Not very customizable for testing specific features
- Captures real-world data's features

Motivation: A data generator for customizable and scalable synthetic data capable of capturing features of real data

Background

Wisconsin Benchmark:
- Defined in 1980 's by DeWitt et al
- It was initially a single-user micro-benchmark for measuring individual query performance

Debit-Credit: A multi-user OLTP benchmark spearheaded by Jim Gray

Multi-User Wisconsin Benchmark: Did not attract significant attention due to other competitors such as Debit-Credit

Today's benchmarks: TPC-x family, YCSB, other well-known ones.

Attribute Name	Range	Order	Comment
unique1	0..(MAX-1)	random	unique, random order
unique2	0..(MAX-1)	random	unique, sequential
two	0..1	cyclic	(unique1 mod 2)
four	0..3	cyclic	(unique1 mod 4)
ten	0..9	cyclic	(unique1 mod 10)
twenty	0..19	cyclic	(unique1 mod 20
onePercent	0..99	cyclic	(unique1 mod 100)
tenPercent	0..9	cyclic	(unique1 mod 10)
twentyPercent	0..4	cyclic	(unique1 mod 5)
fiftyPercent	0..1	cyclic	(unique1 mod 2)
unique3	0..(MAX-1)	cyclic	unique1
evenOnePercent	0,2,4,...,198	cyclic	(onePercent * 2)
oddOnePercent	1,3,5,...199	cyclic	(onePercent * 2)+1
stringu1		random	candidate key
stringu2		cyclic	candidate key
string4		cyclic	

Approach

Our data generator is derived from Wisconsin benchmark

Reasons:
- Attribute distributions and relation structures are easy to understand and to control
- Scalable
- Highly customizable

JSON Specifications & Records

- **JSON specifications:** Easy to use, especially for semi-structured data

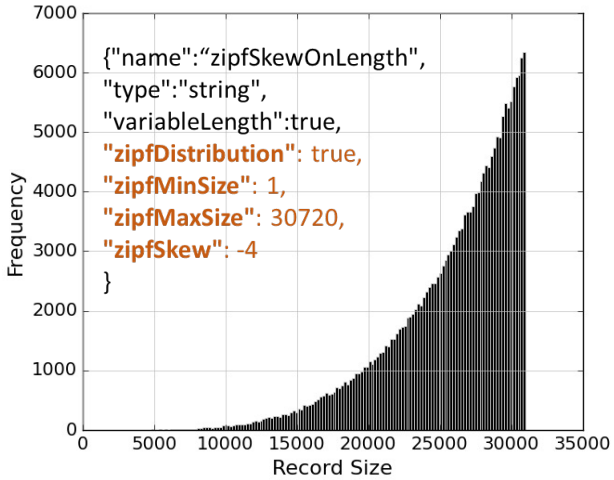
```
[ { "name":"unique1","type":"integer","order":"random" },...]
```

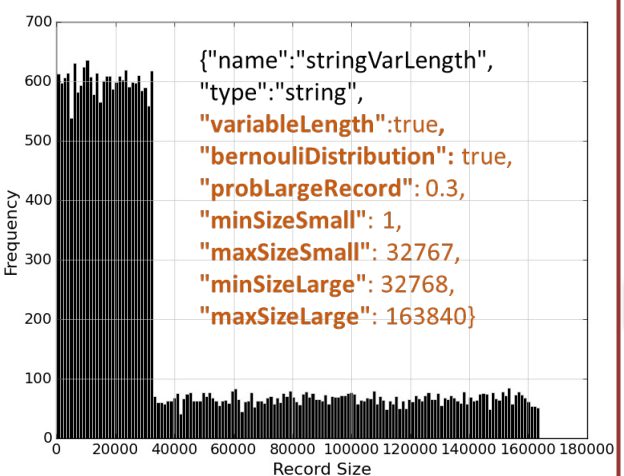
- **JSON records**: JSON is the input & data model of many semi-structured DBMSs

```
{"unique1":2, "four":2, "ten":2, ...,"string4":"CAAAXXX..X"}
{"unique1":4,"four":0, "ten":4, ..., "string4":"EAAAXXX..X"}
```

Variable Record Lengths

- Originally string attributes had a fixed length, with "X" character padding out to the desired length
- Several knobs are now provided to control the length and distribution of length of records
- Distributions: Zipf, Uniform, Normal, and Gamma





Real Words & Hex Strings

- Strings are generated by concatenating real words instead of random or repeating characters
- **Importance:** Reduces the impact of data compression, closer to real-world strings
- Hex strings with fixed and variable lengths (Uniform, Normal, and Gamma distributions) are also supported

```
{ "name":"stringu5",
"type":"string",
...
"variableLength":true,
"HEX":true,
"gammaDistribution": true,
"shape":1.5,
"scale":1.0}
```

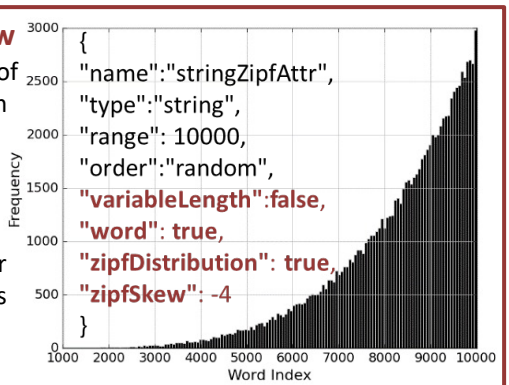
Nullable & Missing Attributes

- Capability to define an attribute as sometimes being null or missing
- Provides knobs to control the distribution of null values and/or missing attributes
- **Importance:** Semi-structured DBMS, missing from previous benchmarks

```
{ "name":"integer1",
....
"optional":true,
"missings":0.1,
"nullable":true,
"nulls":0.3,
...}
```

Attribute Value Skew

- An important feature of real data, missing from other benchmarks
- Normal, Gamma, Zipf distributions are supported
- **Importance:** Useful for attribute skew analysis in joins and other operators



Contributions & Future Work

Data generator used in several recent research and publications for:
- Large and scalable analytical frameworks
- Memory management involving variable-sized records
- Join skew analysis

Future work: Support for nested array-valued fields in one relation that are related to field values in other relations