DBSCALE

SYNTHETICALLY SCALING AN ENTERPRISE'S DATABASE

The Hong Kong Polytechnic University | Database Research Group

Contents

This technical white paper is for informational purpose only. WE MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY.

Key Highlights

- Support MySQL 5.x and SQL Server 2008
- May either scale up or down the database, according to a given scale factor
- Preserve the data characteristics of the original database
- Preserve the query characteristics of the application
- Synthetic data well preserve the privacy of original data
- Integrity constraints may be defined to ensure the similarity between generated database and original database

Key Customer Benefits

- Design of convenience: automatically import the original database and also populate the generated database, friendly user interface
- Using up-scaled synthetic database for predicting the scalability of production or application
- Using down-scaled synthetic database for efficient development and debugging
- Using anonymous synthetic database for out-sourced project
- Using anonymous synthetic database for cloud services testing

Introduction

Database scaling refers to generating a database which has similar characteristic of the original database, while the size is different. A tool that can scale up or down an input database is not available in the market yet. DBScale is such a software that takes as input an enterprise's database and a scale factor S, and generates a synthetic database that is similar to the input database but S times its size. It may assist the customer in testing, application debugging, or project out-sourcing, etc.

Why database scaling?

More and more company are now working on a very large database (VLDB). The database may contain billions of transaction records, millions of customer personal information, or gigabytes of posts or comments.

Developing new functions on such a big dataset becomes a problem: running a small query or function on it may costs hours, or even days, while such small tests is always needed during the development.

Also, a company might need to plan their device or resource when expanding their services. To predict the burden or resource requirement remains a challenge.

Database scaling is thus required in these cases. If there is a tool, which may scale up or down the database , i.e. creating a database that is S times the size of the original database, it may be of great help. However, simply copy-and-paste or randomly deleting would not meet the requirement. A special tool would be necessary.

DBScale: synthetically scaling an enterprise's database

DBScale ensures the synthetic databases preserve the data characteristics of the original database and also preserve the query characteristics in the database applications.

A company can use DBScale to create an up-scaled version of its production database, for predicting the scalability of its production environment; or create a down-scaled version of its database, for its developers to debug the programs more easily and effectively. Since the generated databases are synthetic while close to the original data, the company can use them as the test databases in out-sourced projects, or use them to test the performance of different cloud services, without worrying any privacy problems.

For testing and capacity planning

Database of a current successful web service (facebook, flickr, etc) are growing incredibly fast everyday. To fulfill the raised requirement for data processing, service provider must periodically test their service on a bigger database, to confirm if the applications still work smoothly and functionally. A typical copy-and-paste may help to double the size of the dataset, but certain characteristic (some foreign key constraints, for instance) may not be maintained.

DBScale thus plays a role as a tool to create up-scaled (scale factor $S > 1$) synthetic database for capacity planning test. It not only scales up the database, but also preserves the characteristics of the original database. Testing completed with such a DBScaled database may better emulate a real situation, thus the tests results would be very useful for future capacity planning.

For application debugging

A database related application need to be tested with concrete database to verify its functions. Therefore, during development, developers may often test the applications using real data. Working database of a current data system, however, may exceed some gigabytes in size. Testing on such big dataset is too time consuming and not practical.

Instead of testing against such a big production database, one may use DBScale to generate a downscaled version (scale factor $S < 1$) of the database. With this database, developers may run tests faster and ease the debugging tasks.

As stated above, the scaled database preserves most of the characteristic of the original database. Therefore, the test itself provides reliable reference about the functionality of the application.

For privacy preserving

A company may sometimes need to develop a software processing some data. As we mentioned in the last section, developers need to test the software on real data to verify functions. However, the data itself may be sensitive, or confidential, e.g. transaction records or personal information, and it is not likely for the company to reveal them to a

This technical white paper is for informational purpose only. WE MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY.

program developer (especially when the company out-sourcing the project to another company).

Since the synthetic data generated by DBScale preserve the characteristics of the original data, while also preserve the query characteristics of the original application, a company may deliver the DBScaled database instead of original one to the the project team, therefore the (outsourced) team can work on a production-like database without touching the real records in the original database.

Also, with the proliferation of cloud services, many SME are considering to move their applications to the cloud. When evaluating the services (using a free-trial offered by the cloud providers; e.g., Microsoft Azure provides a 30-day free-trail), the SME may want to test the capabilities (e.g., scalability, elasticity, fault-tolerance) of the various candidate cloud service providers before making any commitment. These evaluations would not be that useful unless the benchmark test is carried out with the SME's own DB application as the test driver. This, however, raises a serious privacy concern — the SME does not want to upload their production database to any cloud service provider before the final decision is made.

In this case, the SME can generate a synthetic copy (scale factor S>=1) of their production database with DBScale to carry out the evaluation, so that the company's data privacy is preserved. At the same time, the benchmark result would also be reliable, because DBScale guarantees the synthesized databases mimic the characteristics of the original data and application as much as possible.

Conclusion

Database scaling helps to resize a production database for testing or developing issues. It makes the test or development more efficient. A good database scaling tool also preserves the data and query characteristic of the original data, thus making the test results more reliable.

DBScale assists the customer in testing, capacity planning and software debugging with reliable synthetic datasets. It also preserves privacy; project outsourcing thus becomes easy for the customer.

Call to action

For more information about DBScale, please go to http://dbgroup.comp.polyu.edu.hk

This technical white paper is for informational purpose only. WE MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY.