# Big data: Searching in large amounts of data quickly and efficiently

March 1 2013


Computer scientists from Saarbrücken have developed an approach which enables searching large amounts of data in a fast and efficient way. Credit: Bellhäuser - das bilderwerk

Not only scientific institutes but also companies harvest an amazing amount of data. Traditional database management systems are often unable to cope with this. Suitable tools are lacking in information retrieval on big data. Computer scientists from Saarbrücken have developed an approach which enables searching large amounts of data in a fast and efficient way. The researchers will show their results at the trade fair Cebit in Hannover starting on March 5.

The term "big data" is defined as a huge amount of digital information,

so big and so complex that normal database technology cannot process it. It is not only scientific institutes like the nuclear research center CERN that often store huge amounts of data ("Big Data"). Companies like Google and Facebook do this as well, and analyze it to make better strategic decisions for their business. How successful such an attempt can be was shown in a New York Times article published last year. It reported on the US-based company "Target" which, by analyzing the buying patterns of a young woman, knew about her pregnancy before her father did.

The analyzed amount of data is distributed on several servers on the internet. The search queries go to several servers in parallel. Traditional database management systems do not match all use cases. Either they cannot cope with big data, or they overstrain the user. Therefore data analysts love tools which are based on the open-source software framework Apache Hadoop and which use its efficient file system HDFS. Those do not require expert knowledge. "If you are used to the programming language Java, you can already do a lot with it", explains Jens Dittrich, professor of information systems at Saarland University. But he also adds that Hadoop is not able to query big datasets as efficiently as database systems that are designed for parallel processing.

Dittrich's and his colleague's solution is the development of the "Hadoop Aggressive Indexing Library", abbreviated with HAIL. It enables saving enormous amounts of data in HDFS in such a way that queries are answered up to 100 times faster. The researchers use a method which you can already find in a telephone book. So that you do not have to read the complete list of names, the entries are sorted according to surnames. The sorting of the names generates the so-called index.

The researchers generate such an index for the datasets they distribute on several servers. But in contrast to the telephone book, they sort the data according to several criteria at once and store it multiply. "The more

criteria you provide, the higher the probability that you find the specified data very fast", Dittrich explains. "To use the telephone book example again, it means that you have six different books. Every one contains a different sorting of the data – according to name, street, ZIP code, city and telephone number. With the right telephone book you can search according to different criteria and will succeed faster." In addition to that, Dittrich and his research group managed to generate the indexes without any additional costs. He and his group members organized the indexing in such a way that no additional computing time and delay is required. Even the additional storage space requirement is low.

**More information:** Conference Paper: vldb.org/pvldb/vol5/p1591_jens … ittrich_vldb2012.pdf

Provided by Saarland University