The VLDB Journal (2014) 23:843
DOI 10.1007/s00778-014-0371-0

GUEST EDITORIAL

Special section on data-intensive cloud infrastructure

Ashraf Aboulnaga · Beng Chin Ooi · Patrick Valduriez

Published online: 30 September 2014
© Springer-Verlag Berlin Heidelberg 2014

More and more individuals, companies and organizations are relying on the cloud to store and manage their data, which translates into increasing pressure on the cloud infrastructure. Cloud data can be very diverse, including a wide variety of personal data collections, very large multimedia content repositories and very large datasets. Users and application developers can be in very high numbers, with little DBMS expertise. Data-intensive applications can be very diverse too, with requirements ranging from basic database capabilities to complex analytics over big data. In particular, the pay-as-you-go model makes the cloud attractive for supporting novel large-scale elastic applications.

NoSQL solutions for the cloud, for instance, have traded consistency and transactional guarantees for scalability. However, the grand challenge for a data-intensive cloud infrastructure is to provide ease of use, consistency, privacy, scalability and elasticity, simultaneously, over cloud data. Addressing this challenge requires novel solutions across the spectrum of data management techniques, including massive data storage, elastic parallel query processing, transactions over data replicated at geographically distributed sites, security and privacy, and efficient data loading and access. This special section focuses on recent advances in research and development in data-intensive cloud infrastructures.

The first paper proposes a workload-aware data placement and replication approach for minimizing resource consumption in cloud data management systems. The workload is modeled as a hypergraph, which allows drawing connections to graph theoretic concepts. Using query span, i.e., the average number of machines involved in the execution of a query or a transaction, as the metric to optimize, the authors develop data placement and replication algorithms as well as scalable techniques to reduce the overhead of partitioning and query routing. To deal with workload changes, they also propose an incremental repartitioning technique. The experiment shows significant reduction in total resource consumption for OLAP workloads, and improved transaction latency and overall throughput for OLTP workloads.

The second paper deals with applications such as bioinformatics, time series and web log analysis, which require the extraction of frequent patterns, called motifs, from one very long (i.e., several gigabytes) sequence. It presents ACME, a parallel cloud-oriented system for extracting such motifs.

ACME uses a combinatorial approach that scales to gigabyte long sequences, and is the first to support supermaximal motifs. ACME can be deployed on thousands of CPUs in the cloud and includes an automatic tuning mechanism that suggests the appropriate number of CPUs to utilize, in order to meet the user runtime constraints while minimizing cloud resources usage. The experiments show that, compared to the state of the art, ACME supports 3 orders of magnitude longer sequences, scales out very well and supports elastic deployment in the cloud.

A. Aboulnaga (✉) · B. C. Ooi · P. Valduriez
Waterloo, Canada
e-mail: ashraf@uwaterloo.ca