Highly Scalable Real-Time Analytics with CloudDBAppliance

Boyan Kolev, Oleksandra Levchenko, Florent Masseglia, Reza Akbarinia, Esther Pacitti, Patrick Valduriez

Highly Scalable Real-Time Analytics with CloudDBAppliance[1]

Boyan Kolev, Oleksandra Levchenko, Florent Masseglia,

Reza Akbarina, Esther Pacitti, Patrick Valduriez

Inria and LIRMM, France

The current cloud landscape is getting populated with many applications that are being migrated to the cloud due to its convenience and ease of use. However, there are still a subset of applications that are infrequently seen in the cloud. These are data-intensive and time critical applications. Data intensive applications, in the best case, experience bad performance in the cloud, as the current database infrastructure in the cloud fails to satisfy high loads and does not provide predictable performance. Many critical applications today still run on mainframes due to their high resilience and high performance. Unfortunately, there is no equivalent in the cloud till today.

The CloudDBAppliance project addresses this problem by introducing a cloud appliance for providing a database-as-a-service able to match the predictable performance, robustness and trustworthiness of on-premise architectures such as those based on mainframes. The project introduces innovations across the following areas:

New Efficient Hardware Enabling In-Memory Databases: The new Bullion hardware will provide a high modular server technology with blade and density-optimized servers with a many-core architecture that will provide over 1000 cores and reach 32 TB of memory.

In-Memory Data Management Solutions that Scale up efficiently on many-core architectures:

(1) An ultra-scalable database with fast analytical queries for data-intensive cloud applications both on the operational and the analytical side, an offering non-existing today.

(2) A real-time analytics framework for providing data mining and machine learning functions, implemented in an efficient scale-up streaming solution over the operational data.

(3) An operational data lake for enterprises relying on Hadoop data lake technologies.

CloudDBAppliance provides an innovative solution for providing **real-time analytics** combining the ultra-scalable operational database with an online analytics solution that accesses directly the operational data. The online analytics is based on a highly efficient streaming solution able to scale vertically and linearly data streaming in many-core architectures to 1000+ CPU cores. The data streaming provides algebraic operators and also allows for custom-operators to incorporate data mining and machine learning algorithms.

To provide robust big data analytics on data streams for CloudDBAppliance, data analytics algorithms will be exploited with the focus on boosting their efficiency in the context of massive parallelism on many-core architectures. For this purpose, the algorithms are being implemented on top of the scalable streaming engine. A typical example analytical technique we focus further on is **time series correlation mining**. The time series correlation discovery algorithm is based on a recent work on fast window correlations over time series of numerical data, and concentrates on adapting the approach for the context of a big number of parallel data streams, thus making it highly scalable. The analysis is done incrementally on sliding windows of time series data, so that recent correlations are being continuously discovered in nearly real-time. The algorithm takes full advantage of the new efficient hardware architecture that enables a large in-memory storage model with low access times, which allows intermediate data to be efficiently shared across different streaming operators and across different instances of the same operator.

[1] Work partially funded by the European Commission under the Horizon2020 project CloudDBAppliance, grant agreement No. 732051