MonetDB/DataCell: leveraging the column-store database technology for efficient and scalable stream processing
Liarou, E.

Citation for published version (APA):
Summary

Numerous applications nowadays require online analytics over high rate streaming data. For example, emerging applications over mobile data can exploit the big mobile data streams for advertising and traffic control. In addition, the recent and continuously expanding massive cloud infrastructures require continuous monitoring to remain in good state and prevent fraud attacks. Similarly, scientific databases create data at massive rates daily or even hourly. In addition, web log analysis requires fast analysis of big streaming data for decision support.

The need to handle queries that remain active for a long time (continuous queries) and quickly analyze big data that are coming in a streaming mode and combine it with existing data brings a new processing paradigm that cannot be exclusively handled by the existing database or data stream technology. Database systems do not have support for continuous query processing, while data stream systems are not built to scale for big data analysis. For this new problem we need to combine the best of both worlds.

In this thesis, we study how to design and implant streaming functionalities in modern column-stores which targets big data analytics. In particular, we use the open source column-store, MonetDB, as our design and experimentation platform. This includes exploitation of both the storage/execution engine and the optimizer infrastructure of the underlying DBMS. We investigate the opportunities and challenges that arise with such a direction and we show that it carries significant advantages. The major challenge then becomes the efficient support for specialized stream features such as incremental window-based processing as well as exploiting standard DBMS functionalities in a streaming environment.

We demonstrate that the resulting system, MonetDB/DataCell, achieves excellent stream processing performance by gracefully handling the state of the art stream benchmark, the Linear Road Benchmark. In addition, we demon-
strate that MonetDB/DataCell outperforms state of the art commercial stream management systems as the stream data increase. These results open the road for scalable data processing that combines both persistent and streaming data in an integrated environment in modern data warehouses.