The End of an Architectural Era for Analytical Databases

Reynold S. Xin
AMPLab, UC Berkeley
rxin@cs.berkeley.edu

ABSTRACT

Traditional enterprise warehouse solutions center around an analytical database system that is monolithic and inflexible: data needs to be extracted, transformed, and loaded into the rigid relational form before analysis. It takes years of sophisticated planning to provision and deploy a warehouse; adding new hardware resources to an existing warehouse is an equally lengthy and daunting task.

Additionally, modern data analysis employs statistical methods that go well beyond the typical roll-up and drill-down capabilities provided by warehouse systems. Although it is possible to implement such methods using a combination of SQL and UDFs [1], query engines in relational databases are ill-suited for these.

The Hadoop ecosystem introduces a suite of tools for data analytics that overcome some of the problems of traditional solutions. These systems, however, forgo years of warehouse research. Memory is significantly underutilized in Hadoop clusters, and execution engine is naive compared with its relational counterparts.

It is time to rethink the design of data warehouse systems and take the best from both worlds. The new generation of warehouse systems should be modular, high performance, fault-tolerant, easy to provision, and designed to support both SQL query processing and machine learning applications.

This paper references the Shark system developed at Berkeley as an initial attempt [2].

BODY

Data warehouse systems should be modular, flexible, easy to provision, and support machine learning. It's time to rethink the system design.

REFERENCES

[1] J. Cohen, B. Dolan, M. Dunlap, J. Hellerstein, and C. Welton. Mad skills: new analysis practices for big data. *Proceedings of the VLDB Endowment*, 2(2):1481–1492, 2009.

[2] C. Engle, A. Lupher, R. Xin, M. Zaharia, M. J. Franklin, S. Shenker, and I. Stoica. Shark: fast data analysis using coarse-grained distributed memory. In *Proceedings of the 2012 international conference on Management of Data*, SIGMOD '12, pages 689–692, New York, NY, USA, 2012. ACM.

Volume 1 of Tiny Transactions on Computer Science