Survey Proposal

Title: Distributed System based OLAP

Authors: Hao Wang(haow4@illinois.edu), Wang Xi(wangxi2@illinois.edu)

Abstract:

As the volume of data soars recent years, OLAP has to handle more massive multidimensional data than before. With the maturity of distributed computing technology, it provides us an effective way to help OLAP tool handle enormous data. This survey focuses on the combination of OLAP and distributed systems. Based on related papers and existing projects, we explore the development status quo of the application of distributed systems on OLAP.

Motivation:

Since distributed system based OLAP is highly discussed in the industry, many developers contribute to implement it. We are interested in existing distributed analytics engines and try to understand related technology.

Related topics:

(1) Zhang, Yan-Song, and Shan Wang. "OLAP query processing method oriented to database and HADOOP hybrid platform." U.S. Patent No. 9,501,550. 22 Nov. 2016.

(2) Arres, Billel, Nadia Kabbachi, and Omar Boussaid. "Building olap cubes on a cloud computing environment with mapreduce." *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*. IEEE, 2013.

(3) Han, Hyuck, et al. "Cloud-aware processing of mapreduce-based olap applications." *Proceedings of the Eleventh Australasian Symposium on Parallel and Distributed Computing-Volume 140*. Australian Computer Society, Inc., 2013.

Specific perspective:

The database technology and the Hadoop technology are combined, and the storage performance of the database and the high expandability and high availability of the Hadoop are combined; the database query processing and the MapReduce query processing are integrated in a loosely-coupled mode, thereby ensuring the high query processing performance, and ensuring the high fault-tolerance performance.

Differences:

Comparing with related topics, we analyze the distributed system based OLAP from a macroscopic perspective. Since our topic is mainly about the utility of combination of OLAP and Hadoop, we accomplish this survey based on related papers and an existing application called Apache Kylin (an open source Distributed Analytics Engine designed to provide SQL interface and multi-dimensional analysis on Hadoop supporting extremely large datasets, original contributed from eBay Inc.) Finally, we will compose survey for layman to understand distributed system based OLAP more easily.

Key contribution:

Researcher	University	Research Lab	Publish Papers
Jiawei Han	UIUC	GOOGLE	SIGMOD

Alfredo Cuzzocrea	MIT	USNA	IEEE-*COMPSAC*
Hung-chih Yang, Ali Dasdan	CMU	YAHOO	Scientific and Statistical Database Managemen
Kuznecov Sergey	Russian Acad. of Sci. Moscow	Microsoft Research	IEEE-Software Engineering
Stratos Idreos	Harvard	DASlab	Scientific and Statistical Database Management
Mohammad Sadoghi	Purdue	ExpoLab	arXiv 2016

Weekly plan:

Week	plan	Content section	progress(%)
13th Feb. - 17th Feb.	Read paper1, paper2	Introduction_1	5%
20th Feb. - 24th Feb.	Read paper3, paper4	Introduction_2	10%
27th Feb. - 3rd Mar.	Read paper5, paper6	Introduction	15%
6th Mar. - 10th Mar.	Experimental data	Comparison	30%
13th Mar. - 17th Mar.	Conclude exps.	Critical thinking	40%
20th Mar. - 24th Mar.	Analysis exps.	Analysis of model	50%
27th Mar. - 31st Mar.	Conclude exps.	Critical thinking	60%
3rd Apr. - 7th Apr.	Design comment	Development history survey	70%
10th Apr. - 14th Apr.	Design comment	Technologies applications survey	80%
17th Apr. - 20th Apr.	Design comment	Finish survey report	100%

Feasibility: the more published papers and researches on OLAP, the more convenient for us to finish this survey. However, it seems hard to collect enough proofs and talks on this topics yet. Difficulty: hard; Time-cost: large.

Reference:

(1) Chaudhuri, Surajit, and Umeshwar Dayal. "An overview of data warehousing and OLAP technology." *ACM Sigmod record* 26.1 (1997): 65-74.

(2) Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.

(3) Yang, Hung-chih, et al. "Map-reduce-merge: simplified relational data processing on large clusters." *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007.

(4) Cuzzocrea, Alfredo. "Analytics over big data: Exploring the convergence of datawarehousing, OLAP and data-intensive cloud infrastructures." *Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual*. IEEE, 2013.

(5) Abelló, Alberto, Jaume Ferrarons, and Oscar Romero. "Building cubes with MapReduce." *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*. ACM, 2011.

(6) Sergey, Kuznecov, and Kudryavcev Yury. "Applying map-reduce paradigm for parallel closed cube computation." *Advances in Databases, Knowledge, and Data Applications, 2009. DBKDA'09. First International Conference on*. IEEE, 2009.

(7) Shvachko, Konstantin, et al. "The hadoop distributed file system." *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*. IEEE, 2010.

(8) Doddavula, Shyam Kumar, and Arun Viswanathan. "System and method for implementing online analytical processing (olap) solution using mapreduce." U.S. Patent Application No. 14/559,642.

(9) Sergey, Kuznecov, and Kudryavcev Yury. "Applying map-reduce paradigm for parallel closed cube computation." Advances in Databases, Knowledge, and Data Applications, 2009. DBKDA'09. First International Conference on. IEEE, 2009.

(10) KUMAR, Sandeep; KUMAR, Dr. Sanjeev. Online analytical processing (OLAP). International Journal of Research and Engineering, [S.l.], v. 4, n. 1, p. 6-9, jan. 2017. ISSN 2348-7860. Available at: <http://digital.ijre.org/index.php/int_j_res_eng/article/view/239>. Date accessed: 27 feb. 2017.