MAPREDUCE И БАЗЫ ДАННЫХ: КОНКУРЕНЦИЯ ИЛИ КООПЕРАЦИЯ?

С.Д. Кузнецов

Как отмечалось в Клермонтском отчете [1], "... сбор, интеграция и анализ данных больше не считаются расходами на ведение бизнеса; данные – это ключ к достижению эффективности и прибыльности бизнеса. В результате быстро развивается индустрия, поддерживающая анализ данных". Если к концу прошлого века программные средства, пригодные для организации хранилищ данных и выполнения над ними оперативного анализа, можно было пересчитать по пальцам одной руки (IBM DB2, Teradata, Sybase IQ, Oracle, частично Microsoft SQL Server, причем только в DB2 и Teradata поддерживалась массивно параллельная архитектура без общих ресурсов между узлами (sharing nothing) и только в Sybase IQ использовалось поколоночное хранение таблиц (column-based store)), то с начала нового тысячелетия активизировалось направление специализированных аппаратно-программных систем, полностью ориентированных на поддержку хранилищ данных и/или анализа данных (Data Warehouse Appliance или Analytic Appliance; в дальнейшем для соблюдения точности и для краткости я будут обозначать это направление и относящиеся к нему системы аббревиатурой DWAA). Основной целью этого направления являлось и является создание аппаратно-программных средств, которые были бы существенно дешевле средств поддержки хранилищ данных, предлагаемых поставщиками универсальных СУБД, но при этом обеспечивали бы не меньшую, а желательно, большую производительность и масштабируемость при работе со сверхбольшими хранилищами данных.

1. Аналитические параллельные СУБД сегодня

Как отмечается в [2], в действительности направление DWAA появилось еще в 1980-е гг., и соответствующие пионерские продукты были созданы в компании Britton Lee Inc. [3], которая в 1989 г. была сначала переименована в ShareBase Corporation, а затем поглощена компанией Teradata [4], которая к этому времени тоже придерживалась подхода DWAA. Аппаратно-программное решение, основанное на ассоциативной адресации элементов хранения данных, имелось у компании ICL (Content Addressable File Store [5]). Однако на рынке систем поддержки хранилищ данных на основе подхода DWAA с тех пор осталась только Teradata.

Возрождение направления DWAA в начале 2000-х, безусловно, связано с упомянутым выше ростом заинтересованности компаний в сравнительно недорогих и эффективных решениях, направленных исключительно на поддержку хранилищ данных и их анализа. Вокруг этого направления стали возникать софтверные стартапы, первым из которых стала компания Netezza [6], основавшая свое эффективное DWAA-решение на использовании программируемых вентильных матриц (Field Programmable Gate Array, FPGA) и процессоров PowerPC. Использование FPGA в контроллерах магнитных дисков позволяет осуществлять "на лету" первичную фильтрацию данных, а применение PowerPC вместо процессоров Intel (по утверждению компании) позволяет снизить энергопотребление и расходы на охлаждение.

С тех пор появилось еще около десяти новых компаний, ориентирующихся на разработку DWAA с применением (почти всегда) разновидностей массивно-параллельной архитектуры (MPP) "sharing-nothing":

- Vertica Systems [7] – MPP, поколоночное хранение таблиц;
- DATAllegro Inc. [8], недавно поглощенная Microsoft, которая основала на продукте этой компании проект Madison, ставший основой SQL Server 2008 R2 Parallel Data Warehouse [15], – MPP, система основана на использовании СУБД Ingres [16] (тем самым, таблицы хранятся по строкам);
- Greenplum [9] – MPP, система основана на использовании СУБД PostgreSQL [17] (тем самым, таблицы хранятся по строкам);
- Aster Data Systems [10] – MPP, таблицы хранятся по строкам;
- Kognitio [11] – MPP, таблицы хранятся по строкам;
- EXASOL AG [12] – MPP, поколоночное хранение таблиц;
- Calpont Corporation [13] – MPP, поколоночное хранение таблиц, система (InfiniDB) внешне схожа с MySQL;
- Dataupia Corporation [14] – MPP, таблицы хранятся по строкам;
- Infobright [15] – поколоночное хранение таблиц, система основана на MySQL, ориентирована на использование многоядерных процессоров, массивный параллелизм не используется;
- Kickfire [16] – поколоночное хранение таблиц, используется специальная аппаратура, ускоряющая выполнение SQL-запросов, система создана на основе MySQL и не основана на массивно-параллельной архитектуре.

Подход DWAA постепенно проникает и в продукты основных поставщиков SQL-ориентированных СУБД. Как отмечалось выше, разработка компании DATAllegro стала основой массивно-параллельного варианта Microsoft SQL Server (SQL Server 2008 R2 Parallel Data Warehouse), а компания Oracle обеспечивает специализированное массивно-параллельное хранилище табличных данных Oracle Exadata Storage Server [18], позволяющее значительно ускорить работу основной СУБД.

У разных решений категории DWAA имеются свои интересные технические особенности, заслуживающие более грубокого обсуждения, анализа и сравнения. Их можно классифицировать и сравнивать по разным критериям. Однако это не является целью данной статьи. Некоторую попытку такого анализа представляет собой обзор [19]. Значительный рост интереса к направлению DWAA, к специализированным СУБД вообще и к СУБД Vertica в частности вызвала статья [20].

2. **При чем здесь MapReduce?**

В данном случае интересен частный, но очень важный в настоящее время вопрос взаимоотношений технологий массивно-параллельных аналитических СУБД и MapReduce [21]. При рассмотрении этого вопроса контекст DWAA является вполне естественным, поскольку практически все СУБД, созданные на основе подхода DWAA, являются массивно-параллельными без использования общих ресурсов. Эти системы создавались в расчете на использование в кластерной аппаратной архитектуре, и они сравнительно легко могут быть перенесены в "облачную" среду динамически конфигурируемых кластеров.

Поэтому появление "родной" для "облачной" среды технологии MapReduce и в особенности энтузиазм по части ее использования, проявленный многими потенциальными пользователями параллельных СУБД, очень озаботили представителей направления DWAA. Сначала авторитетные представители сообщества баз данных и одновременно активные сторонники подхода DWAA Майкл Стоунбрейкер (Michael Stonebraker) и Дэвид Девитт (David J. DeWitt) старались убедить общественность в том, что MapReduce – это технология, уступающая технологии параллельных баз данных по всем статьям [22-23]. Потом была проведена серия экспериментов, продемонстрировавшая, что при решении типичных простых аналитических задач MapReduce уступает в производительности не только поколоночной СУБД Vertica, но и традиционной массивно-параллельной СУБД с хранением таблиц по строкам [24].

Все приводимые доводы и результаты экспериментов были весьма солидными и убедительными, и вряд ли кто-нибудь из людей, знакомых с обеими технологиями, сомневается в том, что MapReduce не вытеснит параллельные СУБД, и что эти технологии будут благополучно сосуществовать в "облаках" и в среде кластерных архитектур вообще. Однако возникает другой вопрос: а нет ли в технологии MapReduce каких-либо положительных черт, которых не хватает параллельным СУБД? И можно ли каким-либо образом добавить эти черты в параллельные СУБД, сохранив их основные качества: декларативный доступ на языке SQL, оптимизацию запросов и т.д. (Кстати, понятно, что у параллельных СУБД имеется масса положительных черт, которыми не обладает MapReduce, но похоже, что добавление их к MapReduce изменило бы суть этой технологии, превратив ее в технологию параллельных СУБД.)

И на эти два вопроса удалось получить положительный ответ. В нескольких проектах, связанных с направлением DWAA, удалось воспользоваться такими преимуществами MapReduce, как масштабируемость до десятков тысяч узлов, отказоустойчивость, дешевизна загрузки данных, возможность использования явно написанного кода, который хорошо распараллеливается. Сразу следует заметить, что пока ни в одном проекте не удалось воспользоваться сразу всеми этими преимуществами, но даже то, чего уже достигли исследователи и разработчики, позволяет добавить в параллельные СУБД важные качества, которыми они до сих по не обладали.

Имеются три подхода к интеграции технологий MapReduce и параллельных СУБД, предложенных и реализованных специалистами компаний Greenplum [25] и Aster Data [26], университетов Yale и Brown [27], а также компании Vertica [28] соответственно, которые можно было бы назвать:

- MapReduce внутри параллельной СУБД;
- СУБД внутри среды MapReduce и
- MapReduce сбоку от параллельной СУБД.

В общих словах, первый подход ориентирован на поддержку написания и выполнения хранимых на стороне сервера баз данных пользовательских функций, которые хорошо распараллеливаются в кластерной (в том числе, в "облачной") среде. Т.е. в данном случае используется преимущество MapReduce по применению явно написанного кода и его распараллеливанию.

Второй подход направлен на использование MapReduce в качестве инфраструктуры параллельной СУБД, в качестве базовых компонентов которой используются традиционные не параллельные СУБД. Применение MapReduce позволяет добиться неограниченной масштабируемости получаемой системы и ее отказоустойчивости на уровне выполнения запросов.

Наконец, при применении третьего подхода MapReduce используется для выполнения процедуры ETL (Extract, Tansform, Load) над исходными данными до их загрузки в систему параллельных баз данных. В этом случае используется преимущество MapReduce в отношении дешевой загрузки данных до их обработки.

3. **Заключение**

Еще пару лет назад было непонятно, каким образом можно с пользой применять возникающие "облачные" среды для высокоуровневого управления данными. Многие люди считали, что в "облаках" системы управления базами данных будут просто вытеснены технологий MapReduce. Это вызывало естественное недовольство сообщества баз данных, авторитетные представители которого старались доказать, что пытаться заменить СУБД какой-либо реализацией MapReduce если не безнравственно, то, по крайней мере, неэффективно.

Однако вскоре стало понятно, что технология MapReduce может быть полезна для самих параллельных СУБД. Во многом становлению и реализации этой идеи способствовали компании-стартапы, выводящие на рынок новые аналитические массивно-параллельные СУБД и добивающиеся конкурентных преимуществ. Свою лепту вносили и продолжают вносить и университетские исследовательские коллективы, находящиеся в тесном сотрудничестве с этими начинающими компаниями.

На сегодняшний день уже понятно, что технология MapReduce может эффективно применяться внутри параллельной аналитической СУБД, служить инфраструктурой отказоустойчивой параллельной СУБД, а также сохранять свою автономность в симбиотическом союзе с параллельной СУБД. Все это не только мешает развитию технологии параллельных СУБД, а наоборот, способствует ее совершенствованию и распространению.

ЛИТЕРАТУРА:

1. Rakesh Agrawal et al. The Claremont Report on Database Research, http://db.cs.berkeley.edu/claremont/claremontreport08.pdf, 2008 г. Перевод на русский язык: Ракеш Агравал и др. Клермонтский отчет об исследованиях в области баз данных, http://citforum.ru/database/articles/claremont_report/, 2008 г.
2. Curt Monash. Data warehouse appliances – fact and fiction, http://www.dbms2.com/2007/12/03/data-warehouse-appliances-%e2%80%93-fact-and-fiction/, December 3, 2007
3. Википедия. Britton Lee, Inc., http://en.wikipedia.org/wiki/Britton_Lee,_Inc., 2010
4. Teradata Home Page, http://www.teradata.com/t/, 2010
5. C.H.C. Lemg and K.S. Wong. File Processing Efficiency on the Content Addressable File Store, http://www.vldb.org/conf/1985/P282.PDF, Proceedings of the VLDB Conference, Stockholm, 1985, 282-291
6. Netezza Home Page, http://www.netezza.com/, 2010
7. Vertica Systems Home Page, http://www.vertica.com/, 2010
8. DATAllegro Home Page, http://www.datallegro.com/, 2010
9. Greenplum Home Page, http://www.greenplum.com/, 2010
10. Aster Data Systems Home Page, http://www.asterdata.com/, 2010
11. Kognitio Home Page, http://www.kognitio.com/, 2010
12. EXASOL AG Home Page, http://www.exasol.com/, 2010
13. Calpont Corporation Home Page, http://www.calpont.com/, 2010
14. Dataupia Corporation Home Page, http://www.dataupia.com/, 2010
15. Infobright Home Page, http://www.infobright.com/, 2010
16. Kickfire Home Page, http://www.kickfire.com/, 2010
17. SQL Server 2008 R2 Parallel Data Warehouse, http://www.microsoft.com/sqlserver/2008/en/us/parallel-data-warehouse.aspx, 2010
18. Oracle Exadata, http://www.oracle.com/us/products/database/exadata/index.html, 2010
19. Richard Hackathorn, Colin White. Data Warehouse Appliances: Evolution or Revolution?, http://www.beyeresearch.com/study/4639, June 26, 2007
20. M. Stonebraker and U. Cetintemel. One Size Fits All: An Idea whose Time has Come and Gone, http://www.cs.brown.edu/%7Eugur/fits_all.pdf // Proc. ICDE, 2005, 2-11. Перевод на русский язык: Майкл Стоунбрейкер, Угур Кетинтемел. Один размер пригоден для всех: идея, время которой пришло и ушло, http://citforum.ru/database/articles/one_size_fits_all/, 2007
21. Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters, http://labs.google.com/papers/mapreduce.html // Proceedings of the Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004, 137–150
22. Michael Stonebraker, David J. DeWitt. MapReduce: A major step backwards, http://databasecolumn.vertica.com/database-innovation/mapreduce-a-major-step-backwards/, January 17, 2008
23. Michael Stonebraker, David J. DeWitt. MapReduce II, http://databasecolumn.vertica.com/database-innovation/mapreduce-ii/, January 25, 2008
24. Andrew Pavlo at al. A Comparison of Approaches to Large-Scale Data Analysis, http://cs-www.cs.yale.edu/homes/dna/papers/benchmarks-sigmod09.pdf // Proceedings of the 35th SIGMOD International Conference on Management of Data, 2009, Providence, Rhode Island, USA, 165-178. Перевод на русский язык: Эндрю Павло и др. Сравнение подходов к крупномасштабному анализу данных, http://citforum.ru/database/articles/mr_vs_dbms/, 2009
25. Jeffrey Cohen et al. MAD Skills: New Analysis Practices for Big Data, http://db.cs.berkeley.edu/jmh/papers/madskills-032009.pdf // Proceedings of the VLDB'09 Conference, Lyon, France, August 24-28, 2009, 1481-1492. Перевод на русский язык: Джеффри Коэн и др. МОГучие способности: новые приемы анализа больших данных, http://citforum.ru/database/articles/mad_skills/, 2009
26. Eric Friedman, Peter Pawlowski, John Cieslewicz. SQL/MapReduce: A practical approach to self-describing, polymorphic, and parallelizable userdefined functions, http://www.asterdata.com/resources/downloads/whitepapers/sqlmr.pdf // Proceedings of the 35th VLDB

Conference, August 24-28, 2009, Lyon, France, 1402-1413. Перевод на русский язык: Эрик Фридман, Питер Павловски и Джон Кислевич. SQL/MapReduce: практический подход к поддержке самоописываемых, полиморфных и параллелизуемых функций, определяемых пользователями, http://citforum.ru/database/articles/asterdata_sql_mr/, 2010

27. Azza Abouzeid et al. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads, http://www.vldb.org/pvldb/2/vldb09-861.pdf // Proceedings of the 35th VLDB Conference, August 24-28, 2009, Lyon, France, 922-933. Перевод на русский язык: Азза Абузейд и др. HadoopDB: архитектурный гибрид технологий MapReduce и СУБД для аналитических рабочих нагрузок, http://citforum.ru/database/articles/hadoopdb/, 2010
28. Michael Stonebraker et al. MapReduce and Parallel DBMSs: Friends or Foes?, http://database.cs.brown.edu/papers/stonebraker-cacm2010.pdf. // Communications of the ACM, vol. 53, no. 1, January 2010, 64-71. Перевод на русский язык: Майкл Стоунбрейкер и др. MapReduce и параллельные СУБД: друзья или враги?, http://citforum.ru/database/articles/mr_vs_dbms-2/, 2010