W4112 Database Systems Implementation,
Spring 2013
Project 1, due 4/3/2013.

In this project, you are asked to evaluate and explain several emerging database technology trends. A list of topics is given below, and each group (of two students) should address all three of the topics.

For each topic, you should:

- Explain the technology, it's motivation and applicability for database system implementation. If there are multiple vendors/approaches then each should be discussed and the differences highlighted. This part should be about 5 pages (see below for formatting guidelines).
- Evaluate the technology. How important is it? Who would care about the benefits it provides? Can the benefits be quantified? Are there mature products based on this technology? Would you invest in this technology if you were a venture capitalist? This part should be about 2–3 pages. I'm asking you to form opinions and to justify them. There may be no "right" answer, particularly when predicting future trends; a well-reasoned response is all that is expected.

Some vendors will make strong claims (e.g., "100X speed improvement"). Assess these claims objectively: are they marketing hype, or is there something truly innovative in the company's solution? Similarly, a research paper could potentially include biases in favor of a preferred or proposed approach (e.g., the system sold by the authors' employers). When reading papers, evaluate and comment on how objective the results are. *Simply quoting claims without some critical analysis is not enough.*

Here are the topics. Example systems/prototypes are included in bold, and some initial references are itemized. These are just a starting point — you are expected to find additional systems/references within the topic. You may also want to consult benchmarking sites, such as www.tpc.org, and other kinds of third-party reviewing sites. Recent research conferences such as ACM SIGMOD and VLDB are also a good source of recent database innovations.

1. To achieve high performance, a number of database vendors and research prototypes are putting the entire database in RAM. What challenges and opportunities does this choice entail? Does in-memory storage help transaction processing? Query processing? What about Durability (the D in "ACID")? **IBM Blink, Exasol, H-Store, RAMcloud**
2. Historically, database vendors chose to store all attributes of each row together; such databases have thus been called "row-stores." Column-stores, in contrast, organize data by attribute, physically storing all values (from many rows) for each column contiguously. When would one choose to use a column store rather than a row-store? Is there a significant performance difference? Is it possible to get the best of both worlds in a single system? **Sybase IQ, Vertica, Vectorwise, Kx Systems, Monet DB**
   - `http://www.vldb.org/pvldb/2/vldb09-tutorial6.pdf`
   - `http://sites.computer.org/debull/A12mar/apollo.pdf`
3. Two recent challenges to the relational database model for database processing are MapReduce and NoSQL. What are these formalisms, and what do they offer that goes beyond the capabilities of relational database systems? What are they missing compared with a traditional relational database system? When would one choose to use such a system instead of a relational database? **Google BigTable, Aster Data, Hive, Tenzing, Hadapt**; **MongoDB, CouchDB, Cassandra**.
   - `http://www.vldb.org/pvldb/vol4/p1318-chattopadhyay.pdf`
   - `http://vldb.org/pvldb/vol5/p1712_avriliafloratou_vldb2012.pdf`

The overall report should be a coherent document with

- An introduction.
- A section for each topic.
- A conclusion. Comparing the topics, which are the most compelling?

- A bibliography. Every source, including web pages, should be cited. The main text should refer to the citations in conventional bibliographic style. It is not sufficient to simply list all references used without indicating which information comes from which reference.

Submission instructions: Your report should be a single-spaced pdf document with 11 point font and normal margins. It should be submitted via courseworks, and also submitted in hardcopy in class on 4/3.

Projects are to be done in teams of two. The TAs will facilitate the matching of students if you have trouble identifying a partner. If for some reason (e.g., a partner drops the class well into the semester) a student ends up working alone, the student is expected to address two (rather than three) of the topics.

CVN students only: We encourage you to find another CVN student to partner with. You should be able to work together reasonably well for this project without face-to-face time. Nevertheless, CVN students have the option of doing this project alone. Students working alone should choose two of the three topics and complete that portion of the project.

Warning. In past semesters, there have been a number of cases where students have plagiarized text and thus violated the Computer Science Department's academic honesty policy. We have used software to automatically identify large segments of copied text. To help guide you, here are three ways that you might write a paragraph about database processing on GPUs:

Plagiarism

Implementations of database operators on GPU processors have shown dramatic performance improvement compared to multicore-CPU implementations. GPU threads can cooperate using shared memory, which is organized in interleaved banks and is fast only when threads read and modify addresses belonging to distinct memory banks. Therefore, data processing operators implemented on a GPU, in addition to contention caused by popular values, have to deal with a new performance limiting factor: thread serialization when accessing values belonging to the same bank.

This is plagiarism because it is taken verbatim from a web source. Changing one or two words, or leaving out a sentence while retaining the rest does not change the fact that unattributed copying has occurred. Plagiarism is cheating and will trigger disciplinary action in line with the department's academic honesty policy.

Cited Use

"Implementations of database operators on GPU processors have shown dramatic performance improvement compared to multicore-CPU implementations. GPU threads can cooperate using shared memory, which is organized in interleaved banks and is fast only when threads read and modify addresses belonging to distinct memory banks. Therefore, data processing operators implemented on a GPU, in addition to contention caused by popular values, have to deal with a new performance limiting factor: thread serialization when accessing values belonging to the same bank." [1]

[1] Ameliorating memory contention of OLAP operators on GPU processors, Evangelia A. Sitaridi, Kenneth A. Ross, DaMoN 2012.

This is not plagiarism because the text is quoted and the source is identified. Nevertheless, text of this form is not likely to get a good grade because it does not demonstrate that you have absorbed the concepts or been able to synthesize an understanding from multiple sources.

Using Your Own Words

Database software can also be written for GPU platforms rather than CPU platforms. Potential performance pitfalls include contention for popular values by concurrent threads, and thread serialization when multiple threads access the same bank of shared memory [1].

[1] Ameliorating memory contention of OLAP operators on GPU processors, Evangelia A. Sitaridi, Kenneth A. Ross, DaMoN 2012.

This is the kind of rewriting you should aim for. The rewritten text is based on the original, but does not quote from it. The rewritten text is actually more direct and informative than the original in the context of a report that is intended to cover many references. Depending on how deeply you wanted to cover this reference, you could go on to describe how contention or banked memory creates a performance hazard.