MapReduce vs. Parallel DBMS vs. IR systems

University of California, Berkeley
School of Information

IS 257: Database Management

Announcement

- Sign up for final project presentations
  - http://doodle.com/poll/46ivgbr5b73yzagr

Outline

- Review
  - BDAS and Spark – Slides from Ion Stoica
- What technologies to use
- What's happening with DBMS – Mike Stonebraker

BDAS and Applications

- The Berkeley Data Analytics Stack – talk from Ion Stoica

What to use?

- An ongoing question in this course, and in application of various big-data technologies is which to use for a particular situation

- The question is what is the “system of the future”

MapReduce and Hadoop

- As you have heard MapReduce-based processing is very powerful and can be used in many situations
- There are, however, latency problems (some largely solved by Spark or Impala) and not all tasks are really as batch-oriented as MR tends to be
- The Hadoop tools provide very powerful and useful abstractions, but still have the batch-oriented character of MR

Advantages of RDBMS

- Relational Database Management Systems (RDBMS) available for almost any system or special need when dealing with almost any data type
- Possible to design complex data storage and retrieval systems with ease (and without conventional programming).
- Support for ACID transactions
  - Atomic
  - Consistent
  - Independent
  - Durable

IR Systems

- There are many commercially available, as well as open source and freely available IR search engines
  - E.g. Apache SOLR and Lucene run much of the website search for major companies today
- These search engines are embracing Big Data approaches
  - E.g. Cloudera Search
- The main limitation for IR systems is that they seldom support data update or transactions well

Advantages of RDBMS

- Support for *very large* databases
- Automatic optimization of searching (when possible)
- RDBMS have a simple view of the database that conforms to much of the data used in business
- Well suited for transactional data with lots of updates, replacements, etc.
- Standard query language (SQL)

But will it be a RDBMS?

- About 10 years ago, Mike Stonebraker (one of the people who helped invent Relational DBMS) suggested that the "One Size Fits All" model for DBMS is an idea whose time has come – and gone
  - This was also a theme of the Claremont Report
- RDBMS technology, as noted previously, has optimized on transactional business type processing
- But many other applications do not follow that model

Will it be an RDBMS?

- Stonebraker predicts that the DBMS market will fracture into many more specialized database engines
  - Although some may have a shared common frontend
- Examples are Data Warehouses, Stream processing engines, Text and unstructured data processing systems
- The latter is basically IR systems

Will it be an RDBMS?

- Streaming data, such as Wall St. stock trade information is badly suited to conventional RDBMS (other than as historical data)
  - The data arrives in a continuous real-time stream
  - But, data in RDBMS has to be stored before it can be read and actions taken on it
    - This is too slow for real-time actions on that data
  - Stream processors function by running "queries" on the live data stream instead
    - May be *orders of magnitude* faster

Will it be an RDBMS

- RDBMS will still be used for what they are best at – business-type high transaction data
- But specialized DBMS will be used for many other applications
- Consider Oracle's acquisions of SleepyCat (BerkeleyDB) embedded database engine, and TimesTen main memory database engine
  - specialized database engines for specific applications

Some things to consider

- Bandwidth will keep increasing and getting cheaper (and go wireless)
- Processing power will keep increasing
  - Moore's law: Number of circuits on the most advanced semiconductors doubling every 18 months
  - With multicore chips, all computing is becoming parallel computing
- Memory and Storage will keep getting cheaper (and probably smaller)
  - "Storage law": Worldwide digital data storage capacity has doubled every 9 months for the past decade

But that was 10 years ago…

- Stonebraker was recently awarded the IEEE "Test of Time" award for that paper
- His acceptance speech summarizes the ideas and updates them for today

- https://www.youtube.com/watch?v=9K0SWs1mOD0