云数据存储管理系统评测报告

史英杰 王仲远 王海平 刘兵兵

云计算是当今信息产业最受关注的一种计算模式，在这种模式下，企业和个人可以根据自己的需要购买存储设备和计算能力，而不是花费大量资金购买大规模高性能计算机。作为云计算的一项关键技术，云数据存储和云数据管理为业界带来巨大的潜在商用价值。随着信息产业的发展，企业和公司产生的数据量快速增长，通常数据规模可以达到 TB 甚至 PB 级别。如何管理和分析海量数据是目前很多领域所面临的问题，例如在医疗、通信和互联网领域。云环境是由大量的性能普通、价格便宜的计算节点组成的一种无共享大规模并行处理环境，所以从成本和性能两方面考虑，越来越多的企业更愿意把自己的数据中心从昂贵的高性能计算机转移到共有或私有云环境中。在互联网时代，海量数据的存储和处理操作非常频繁，很多研究者都在从事这方面的研究，也涌现出很多云数据管理系统。下面的 ppt 就介绍了部分当前的云数据管理系统，并对它们的结构、数据模型及数据一致性进行了比较分析。

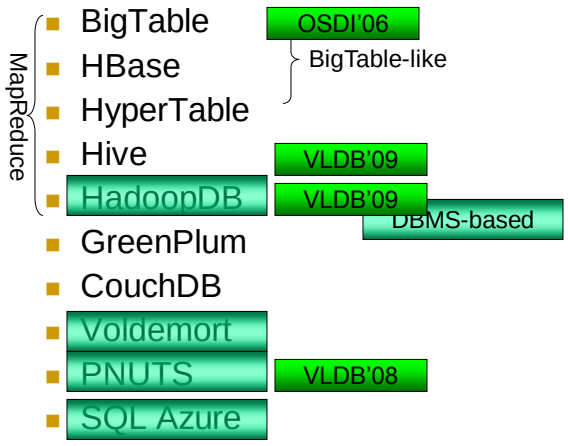
Survey on Data Management in the Cloud

Yingjie Shi

Outline

- Systems surveyed
- Comparison of Systems
- Experiment Benchmark

Systems Surveyed



BigTable – Basic Information

- To manage structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers
- Motivations
  - Scale is too large for most commercial databases
  - Even if it weren't, cost would be very high
  - Low-level storage optimizations help performance significantly

Google

BigTable – Goals

- Fault-tolerant, persistent
- Scalable
  - 1000s of servers
  - Millions of reads/writes, efficient scans
- Self-managing
- Simple!

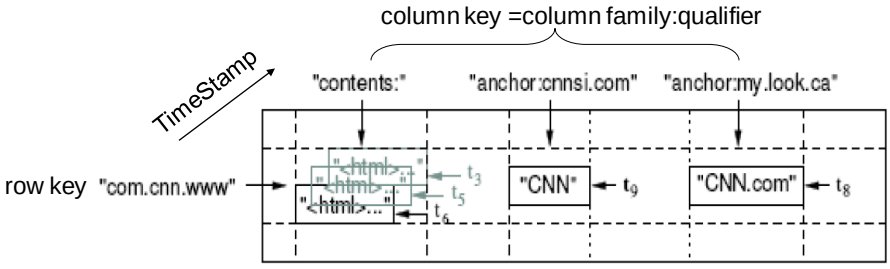
BigTable – Applications

- Based on: GFS(Google File System)
- Applications: Google 地球 Google finance Google maps Google Analytics
- Scale of servers:

No. of tablet servers	No. of clusters
0 … 19	259
20 … 49	47
50 … 99	20
100 … 499	50
>500	12

BigTable – Data Model

- It is a sparse, distributed, persistent multidimensional sorted map.



The map is indexed by a row key, column key, and a timestamp; each value in the map is an uninterpreted array of bytes.

BigTable – Storage

- Column family – oriented storage(key->value)
  - (row:string, column:string, time:int64) ->string

row=row0, column=anchor:cnnsi.com, timestamp=1174184619081 → XXXXXXXXX
row=row0, column=anchor:my.look.ca, timestamp=1174184620720 → XXXXXXXXX
row=row0, column=anchor:my.look.ca, timestamp=1174184617161 → XXXXXXXXX
row=row1, column=anchor:cnnsi.com, timestamp=1174184619081 → XXXXXXXXX
row=row1, column=anchor:my.look.ca, timestamp=1174184620721 → XXXXXXXXX
row=row1, column=anchor:my.look.ca, timestamp=1174184617167 → XXXXXXXXX
row=row2, column=anchor:cnnsi.com, timestamp=1174184619081 → XXXXXXXXX
row=row2, column=anchor:my.look.ca, timestamp=1174184620724 → XXXXXXXXX
row=row2, column=anchor:my.look.ca, timestamp=1174184617167 → XXXXXXXXX
row=row3, column=anchor:cnnsi.com, timestamp=1174184619081 → XXXXXXXXX
row=row3, column=anchor:my.look.ca, timestamp=1174184620724 → XXXXXXXXX
row=row3, column=anchor:my.look.ca, timestamp=1174184617168 → XXXXXXXXX
row=row4, column=anchor:cnnsi.com, timestamp=1174184619081 → XXXXXXXXX
row=row4, column=anchor:my.look.ca, timestamp=1174184620724 → XXXXXXXXX
row=row4, column=anchor:my.look.ca, timestamp=1174184617168 → XXXXXXXXX
row=row5, column=anchor:cnnsi.com, timestamp=1174184619082 → XXXXXXXXX
row=row5, column=anchor:my.look.ca, timestamp=1174184620725 → XXXXXXXXX
row=row5, column=anchor:my.look.ca, timestamp=1174184617168 → XXXXXXXXX
row=row6, column=anchor:cnnsi.com, timestamp=1174184619082 → XXXXXXXXX
row=row6, column=anchor:my.look.ca, timestamp=1174184620725 → XXXXXXXXX
row=row6, column=anchor:my.look.ca, timestamp=1174184617168 → XXXXXXXXX

HBase

- A clone project of BigTable using Java
- Developers: Apache Software Foundation
- Runs on top of Hadoop core
- Production users: Powerset streamy openplaces

YAHOO!

Hypertable

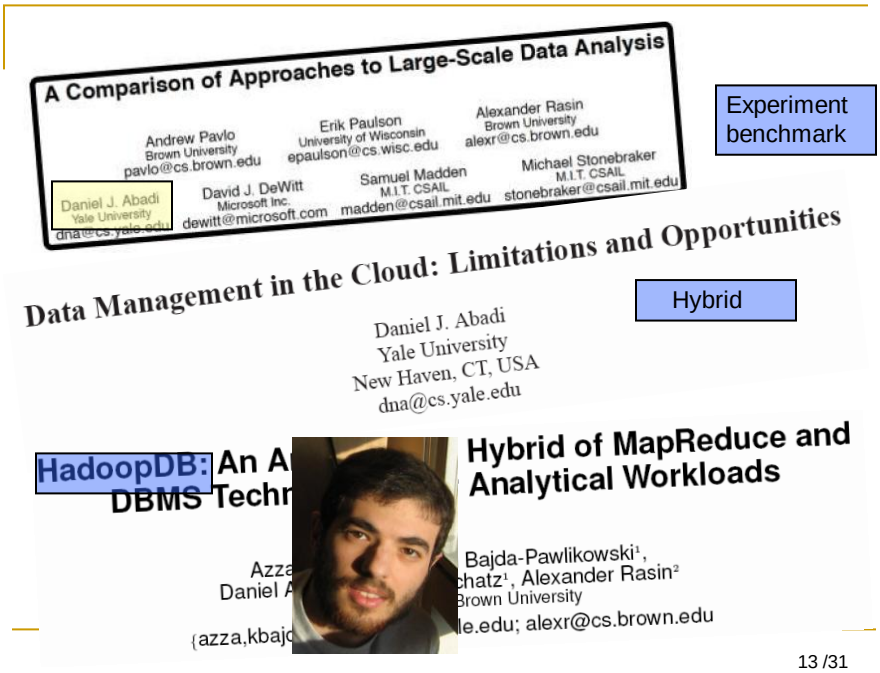
- A clone project of BigTable in C++
- Developers: Zvents Baidu百度 rediff.com
- Runs on top of CloudStore(KFS,Kosmos File System)

BigTable-like VS RDBMS

- Fast Query Rate
  - No Joins, No SQL support, column-oriented database
  - Uses one Bigtable instead of having many normalized tables
- Is not even in 1NF in a traditional view
- Support historial queries

Hive - Basic Information

- A system for managing and querying structured data built on top of Hadoop
  - Map-Reduce for execution
  - HDFS for storage
  - Metadata on raw files
- Key Building Principles:
  - SQL as a familiar data warehousing tool
  - Extensibility - Types, Functions, Formats, Scripts
  - Scalability and Performance



HadoopDB-Philosophy

- Two largest components of the data management market
  - Transactional data management
  - Analytical datamanagement

Moved to cloud

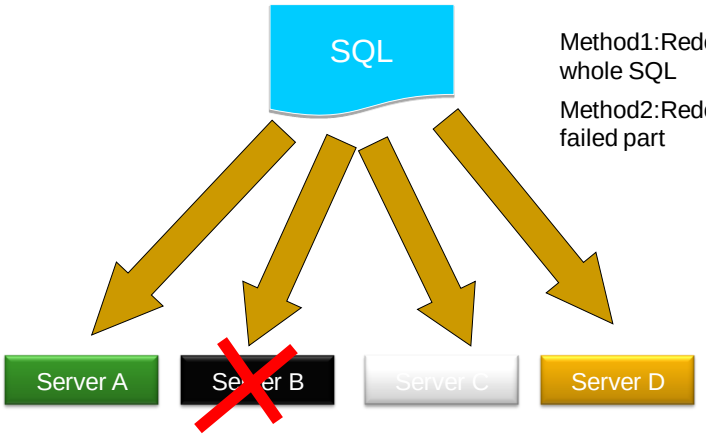
- Two technologies used for data analysis in a shared-nothing MPP architecture
  - Parallel database
  - MapReduce-based system

HadoopDB-Philosophy

	Scalability	Tolerance	High Performance
Parallel database	✗	✗	✓
MapReduce	✓	✓	✗
What we want	✓	✓	✓

Scalability:1000 nodes
High Performance:Queries on structured data

Query Tolerance

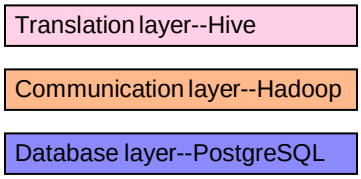


Method1:Redo the whole SQL

Method2:Redo the failed part

HadoopDB-Philosophy

- Goals
  - Performance
  - Tolerance
  - Scalability
  - Flexible query interface
- Design idea
  - Multiple, independent, single-node databases coordinated by Hadoop

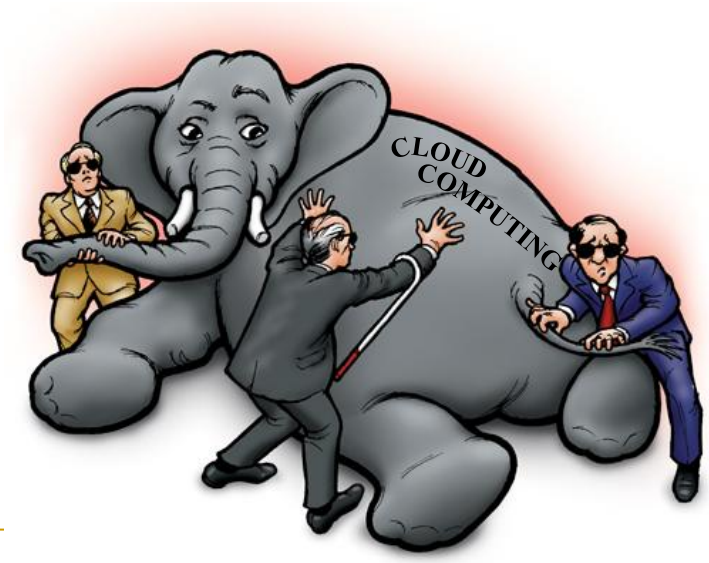


PNUTS

- Developer: YAHOO!
- Applications: Social network, advertising application
- Application characteristic:
  - Scalability
  - Geographic scope
  - Fast response requirement
  - High availability
  - Simplified query needs
  - Relaxed consistency needs

SQL Azure

- A relational database service on the Windows Azure Platform that is built on SQL Server technologies
- Objects can be created on SQL Azure:
  - Tables
  - Indexes
  - Views
  - Stored Procedures
  - Triggers



Outline

- Systems surveyed
- Comparison of Systems
- Experiment Benchmark

Characteristic of Cloud Database

- Performance
- Scalability
  - Ability to scale by adding resources with minimal operational effort and minimal impact on system performance
  - Performance increases with the scale of the system extends
- High Availability and Fault Tolerance
- Ability to run in a heterogeneous environment
- All applications are read-only or read-mostly

Summary of Applications

- Data Analysis
  - Internet Service
  - Private Cloud
- Web Applications
  - Some operations that can tolerate relaxed consistency

BigTable HBase HyperTable Hive HadoopDB…

PNUTS

Architecture

MapReduce-based

BigTable HBase Hypertable Hive

- scalability
- fault tolerance
- ability to run in a heterogeneous environment
- data replication in file system
- a lot of work to do to support SQL

DBMS-based

SQL Azure PNUTS Voldemort

- easy to support SQL
- easy to utilize index, optimization method
- bottleneck of data storage
- data replication upon DBMS

Hybrid of MapReduce and DBMS

HadoopDB

- sounds good
- Performance?

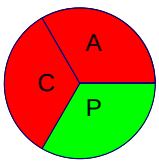
Data Model

- Big Map Model
  - BigTable,HBase,Hypertable
- Simple Relational Data Model
  - Hive, PNUTS, SQL Azure and HadoopDB

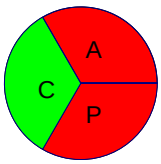
It depends on the real application!

Consistency

- Two kinds of consistency:
  - strong consistency – ACID(Atomicity Consistency Isolation Durability)
  - weak consistency – BASE(Basically Available Soft-state Eventual consistency)

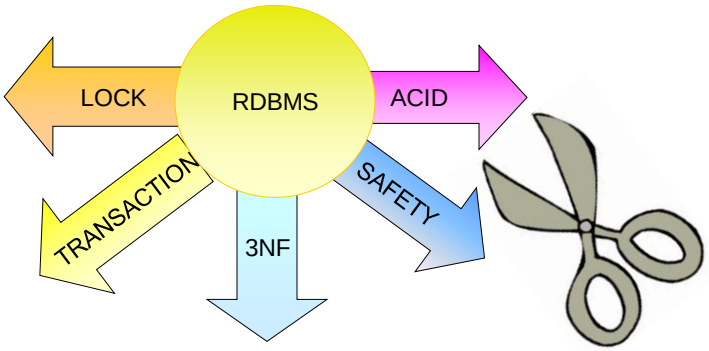


BigTable,HBase,
Hive,Hypertable,HadoopDB



PNUTS

A tailor

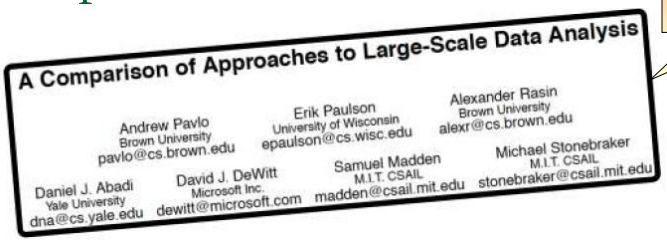


Outline

- Systems surveyed
- Comparison of Systems
- Experiment Benchmark

Experiment Benchmark

MapReduce
DBMS



A Comparison of Approaches to Large-Scale Data Analysis

Andrew Pavlo, Brown University, pavlo@cs.brown.edu
Erik Paulson, University of Wisconsin, epaulson@cs.wisc.edu
Alexander Rasin, Brown University, alexr@cs.brown.edu
Daniel J. Abadi, Yale University, dna@cs.yale.edu
David J. DeWitt, Microsoft Inc., dewitt@microsoft.com
Samuel Madden, M.I.T. CSAIL, madden@csail.mit.edu
Michael Stonebraker, M.I.T. CSAIL, stonebraker@csail.mit.edu

A Benchmark for Hive, PIG and Hadoop[1]

Yuntao Jia, Zheng Shao

July 12th, 2009

HadoopDB
Hadoop
DBMS

6. BENCHMARKS

In this section we evaluate HadoopDB, comparing it with a MapReduce implementation and two parallel database implementations, using a benchmark first presented in [23][4]. This benchmark consists of five tasks. The first task is taken directly from the original MapReduce paper [8] whose authors claim is representative of common MR tasks. The next four tasks are analytical queries designed to be representative of traditional structured data analysis workloads that HadoopDB targets.

Experiment Benchmark

- Tasks:
  - Data Load
  - Grep Task
  - Selection Task
  - Join Task
  - Aggregation Task
- Data
  - Grep
  - UserVisits → Structured data
  - Rankings
  - Documents → Unstructured data

References

- Daniel J. Abadi. Data Management in the Cloud: Limitations and Opportunities. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering
- Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff and Raghotham Murthy. Hive-A Warehousing Solution Over a MapReduce Framework. VLDB 2009
- Azza Abouzeid, Kamil BajdaPawlikowski, Daniel Abadi, Avi Silberschatz, Alexander Rasin. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. VLDB 2009
- Armando Fox, Eric A. Brewer. Harvest, Yield, and Scalable Tolerant Systems. Proceedings of the The Seventh Workshop on Hot Topics in Operating Systems 1999
- J. Hamilton. Cooperative expendable micro-slice servers (cems): Low cost, low power servers for internet-scale services. In Proc. of CIDR, 2009.
- J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In OSDI, 2004.
- Brian F. Cooper, Raghu Ramakrishnan, Utkarsh Srivastava, Adam Silberstein,Philip Bohannon, HansArno Jacobsen, Nick Puz, Daniel Weaver and Ramana Yerneni. PNUTS: Yahoo!'s Hosted Data Serving Platform. VLDB2008.