Data Processing in the Era of Big Data

Pangfeng Liu

Department of Computer Science and Information Engineering
National Taiwan University

October 3, 2014

Big Data – a New Jargon

Introduction

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications – Wiki[1].
- Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization – Garner[2].

[1] http://en.wikipedia.org/wiki/Big_data

[2] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Introduction

- We can derive more information from a single large data set than many data sets of the same total volume.

The Size Matters

The current limit on dataset is about an exabyte, as Wiki claimed.

- Megabyte 10^6
- Gigabyte 10^9
- Terabyte 10^{12}
- Petabyte 10^{15}
- Exabyte 10^{18}, or $1,000,000,000,000,000,000$.

The Size Matters

How to store an exabyte of data?

- You need one million 1Terabyte disks.
- Price – $2000NT \times 1,000,000 = 2,000,000,000NT$
- Weight – $0.6Kg \times 1,000,000 = 600,000Kg$
- Power – $2W \times 1000000 = 2,000,000W$
- Height – $3cm \times 1000000 = 30km$, this is about 100 times the height of Taipei 101 tower.

Who Needs Big Data?

- Meteorology
- Genomics
- Connectomics
- Complex physics simulations
- Biological and Environmental Research
- Internet search
- Finance
- Business informatics

Where Does Big Data Come from?

- Ubiquitous information-sensing mobile devices
- Remote sensing
- Software logs
- Cameras
- Microphones
- Radio-frequency
- Identification readers
- Wireless sensor networks

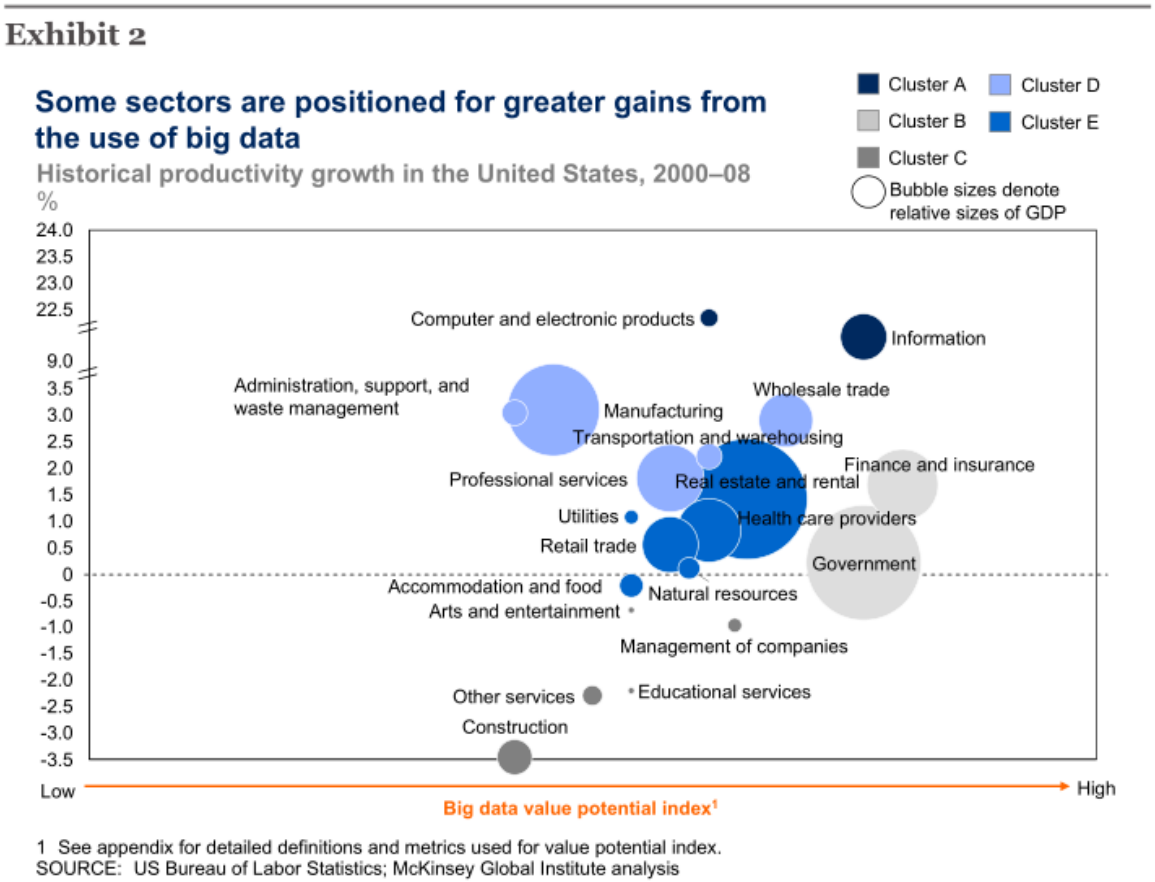
Importance

- Big Data delivers the cost-effective prospect to improve decision-making in critical development areas such as health care, employment, economic productivity, crime and security, and natural disaster and resource management.
- To spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.

Importance

- Big data has the potential to result in a new kind of digital divide: a divide in data-based intelligence to inform decision-making.
- Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, and HP have spent more than 15 billion on software firms only specializing in data management and analytics. In 2010, this industry on its own was worth more than 100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

Big Data Potential Index



[3] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

An Example

Tobias Preis et al. used Google Trends data to demonstrate that Internet users from countries with a higher per capita gross domestic product (GDP) are more likely to search for information about the future than information about the past. The findings suggest there may be a link between on-line behavior and real-world economic indicators[4].

[4] http://en.wikipedia.org/wiki/Big_data

- The study examined Google queries logs in 45 different countries in 2010 and calculated a "future orientation index", which is the ratio of the volume of searches for the coming year to the volume of searches for the previous year.
- They compared the future orientation index to the per capita GDP of each country and found a strong tendency for countries in which Google users inquire more about the future to exhibit a higher GDP.
- The results hint that there may potentially be a relationship between the economic success of a country and the information-seeking behavior of its citizens captured in big data.

Challenge

- Capture
- Storage
- Search
- Sharing
- Analysis
- Visualization

Big Data Characteristics

- Volume – amount of data
- Velocity – speed of data in and out
- Variety – range of data types and sources
- Veracity – the correctness of data

Volume

- Enterprises are awash with ever-growing data of all types, easily amassing terabytes even petabytes of information[5].
- The per-capita data volume roughly doubled every 40 months since 1980, and now we create 2.5 exabyte (2.5×10^{18}) of data every day.
- There is 40% projected growth in global data generated per year, but only 5% growth in IT spending.

[5] http://www-01.ibm.com/software/data/bigdata/

Volume

- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.
- Walmart has more than 2.5 petabytes of customer data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.
- The Utah Data Center constructed by the United States National Security Agency will hold yottabytes (10^{24}) information collected by the NSA over the Internet.

Volume

- Facebook has 50 billion photos from its user base, and Facebook users share 30 billion pieces of contents every month.
- Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.
- The Utah Data Center constructed by the United States National Security Agency will hold yottabytes (10^{24}) information collected by the NSA over the Internet.
- Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

Velocity

- The Large Hadron Collider has 150 million sensors delivering data 40 million times per second, and generates 500 exabytes per day before replication.
- The Sloan Digital Sky Survey (SDSS) collects more astronomical data in its first few weeks than all data collected in the history of astronomy, generating about 200 GB per night.
- Decoding the human genome originally took 10 years to process; now it can be achieved in one week.
- Walmart handles more than 1 million customer transactions every hour.

Variety

- Big data is any type of data – structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more[6].
- New insights are found when analyzing these data types together.

[6] http://www-01.ibm.com/software/data/bigdata/

Veracity

- One in three business leaders don't trust the information they use to make decisions[7].
- How can you act upon information if you don't trust it?
- Establishing trust in big data presents a huge challenge as the variety and number of sources grows.

[7] http://www-01.ibm.com/software/data/bigdata/

The Outlook

- Q: What is our chance doing Exabyte computing?
- A: Not very good, so far.

Our Enemy

8

[8] http://us.123rf.com/400wm/400/400/soify/soify1210/soify121000002/15649042-monster-cartoon.jpg

Our Weapon

9

[9] http://cdn.smosh.com/sites/default/files/bloguploads/cute-weapon-tiny2-b.jpg

Big Data Technology

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times.

Technologies

A 2011 McKinsey report suggests suitable technologies for Big data[10].

- AB testing
- Association rule learning
- Classification
- Cluster analysis
- Crowdsourcing
- Data fusion and integration
- Ensemble learning
- Genetic algorithms
- Machine learning

[10] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

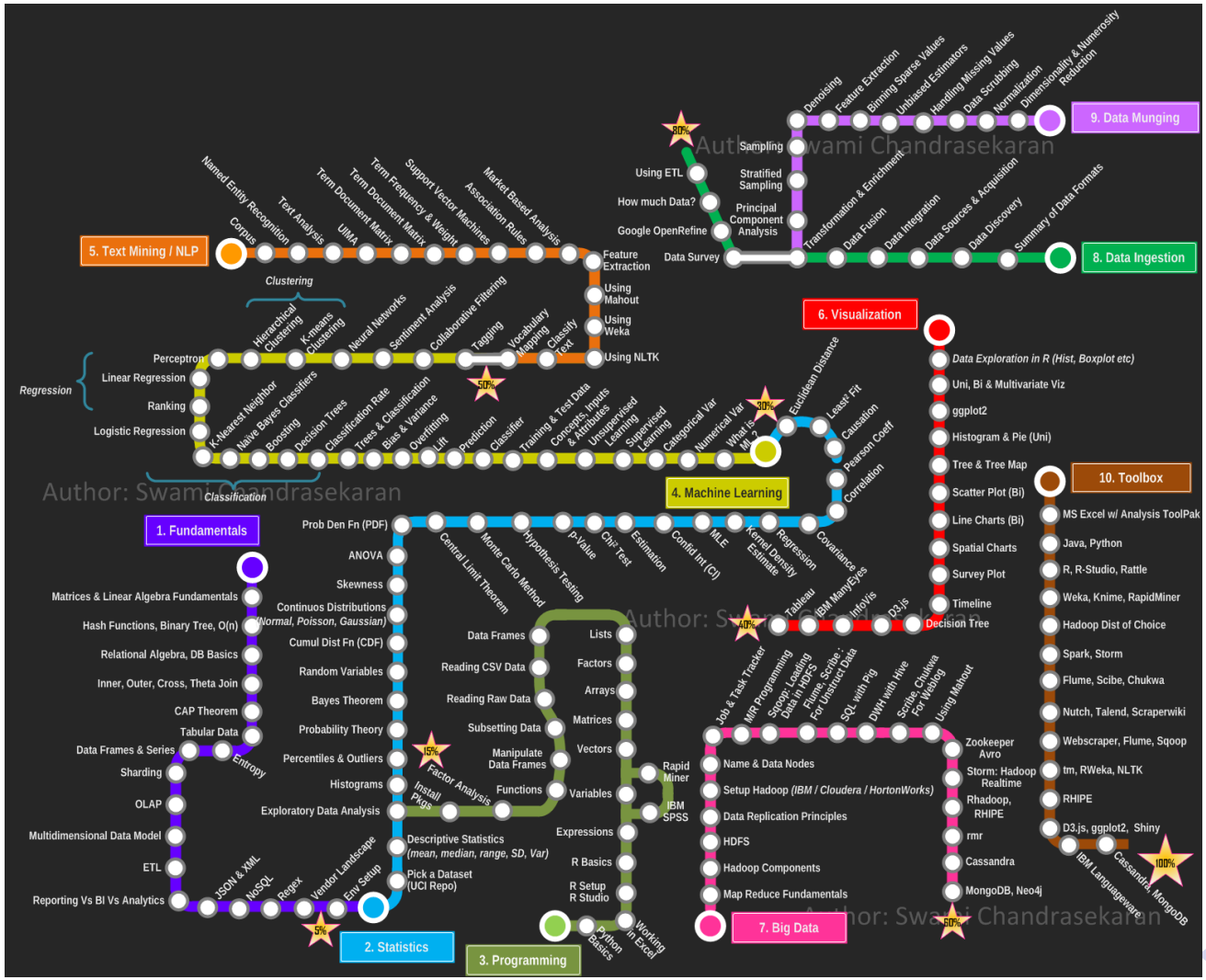
Technologies

- Natural language processing
- Neural networks
- Pattern recognition
- Anomaly detection
- Predictive modeling
- Regression, sentiment analysis
- Signal processing
- Supervised and unsupervised learning
- Simulation
- Time series analysis and visualization

Really??

This is questionable. None of these can even solve the problem of storing the data.

The Road


1. Fundamentals
2. Statistics
3. Programming
4. Machine Learning
5. Text Mining / NLP
6. Visualization
7. Big Data
8. Data Ingestion
9. Data Munging
10. Toolbox
Author: Swami Chandrasekaran

Data Science

1. Fundamentals
2. Statistics
3. Programming
4. Machine Learning
5. Text Mining/Natural Language Processing
6. Visualization
7. BigData
8. Data Ingestion
9. Data Munging
10. Toolbox

Technologies

Additional technologies being applied to big data include.

- Massively parallel-processing (MPP) databases
- Search-based applications
- Data-mining grids
- Distributed file systems
- Distributed databases
- Cloud based infrastructure

Lack of Technologies

Gartner suggest the following to deal with the "volume" issue[12].

- Limiting data collected to that which will be leveraged by the current or imminent business processes.
- Limiting certain analytic structures to a percentage of statistically valid sample data.
- Profiling data sources to identify and subsequently eliminate redundancy.
- Monitoring data usage to determine "cold spots".
- Outsourcing. (You can never beat that)

[12] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

What do we need to deal with Big Data?

- Cloud Infrastructure
- Extremely large scale database
- Data mining, machine learning
- Domain knowledge

Cloud Infrastructure

- Only cloud computing can provide processing capability for big data.
- Just a simple question – where are you going to place the one million hard disks for exabyte scale database?

Data Center

- A data center, also called a server farm, is a facility used to house computer systems and associated components, such as telecommunications and storage systems.
- It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and security devices.

Data Center

- Continuity – Companies rely on their information systems to run their operations.
- Security – A data center has to offer a secure environment which minimizes the chances of a security breach.
- Integrity – Redundancy of both fiber optic cables and power, which includes emergency backup power generation, to ensure the Integrity of data.

Infrastructure as a Service

- Infrastructure as a Service (IaaS) is the delivery of computer infrastructure (typically a platform virtualization environment) as a service.
- Originally called Hardware as a Service (HaaS)

Amazon Elastic Compute Cloud

- Amazon Elastic Compute Cloud (also known as "EC2") allows customers to rent computers on which to run their own computer applications.
- EC2 allows scalable deployment of applications by providing a web services interface through which a customer can create virtual machines, i.e. server instances, on which the customer can load any software of their choice.

Amazon Elastic Compute Cloud

- Elastic
- Completely Controlled
- Flexible
- Designed for use with other Amazon Web Services
- Reliable
- Secure
- Inexpensive

NoSQL

A NoSQL database provides a simple, lightweight mechanism for storage and retrieval of data that provides higher scalability and availability than traditional relational databases[13].

[13] http://en.wikipedia.org/wiki/NoSQL

Relational Database?

If you want vast, on-demand scalability, you need a non-relational database[14].

- Is that so?
- What are the differences between relational and non-relational databases?
- Is this a sign that relational databases have had their day and will decline over time?

[14] http://readwrite.com/2009/02/12/is-the-relational-database-doomed

Relational Database

- Has been around over 30 years.
- Well studied, well optimized.
- No major changes.
- All of those "revolutions" fizzled out, and none even made a dent in the dominance of relational databases.

Relational Database

For an increasing number of applications, one of these benefits is becoming more and more critical; and while still considered a niche, it is rapidly becoming mainstream, so much so that for an increasing number of database users this requirement is beginning to eclipse others in importance[15].

[15] http://readwrite.com/2009/02/12/is-the-relational-database-doomed

Scalability

- Scalability is the key issue.
- To achieve scalability you need scalable infrastructure.

Database Scalability

- Web 2.0 applications, social networking, and on-line multi-player gaming have become more and more popular.
- These applications typically deal with ever-increasing large amounts of data.
- Deploying these applications on traditional relational database management systems typically suffers limited scalability.

NoSQL Databases

- There are also various NoSQL databases used to manage large amounts of data.
  - BigTable from Google
  - HBase
  - Cassandra from Facebook
  - Dynamo from Amazon

From SQL to NoSQL

- NoSQL databases are not a direct replacement for traditional relational database management systems.
- Many applications require multi-row transaction support.
- Data management tools and many existing applications typically interface with databases using SQL.

Our Works

- SQLMR – a SQL interface for NoSQL[16].
- HSQL – a multi-row transaction system on Hbase[17].
- Kylin – a cloud-based BSP model graph computation engine [18].

---

[16] Meng-Ju Hsieh, Chao-Rui Chang, Li-Yung Ho, Jan-Jan Wu, Pangfeng Liu: SQLMR : A Scalable Database Management System for Cloud Computing. ICPP 2011: 315-324.

[17] Chao-Rui Chang, Meng-Ju Hsieh, Jan-Jan Wu, Po-Yen Wu, Pangfeng Liu: HSQL: A Highly Scalable Cloud Database for Multi-user Query Processing. IEEE CLOUD 2012: 943-944.

[18] Li-Yung Ho, Tsung-Han Li, Jan-Jan Wu, Pangfeng Liu: Kylin: An efficient and scalable graph data processing system. BigData Conference 2013: 193-198.

SQLMR

- A MapReduce-based interface for SQL applications.
- Provide high-performance OLAP processing with SQL syntax.
- A joint work with Institute of Information Science, Academia Sinica.
- `http://otl.sinica.edu.tw/index.php?t=9&group_id=25&article_id=1208`

SQLMR Technology Transfer

HSQL

- A highly scalable database for OLTP applications.
- Built on top of HBase.
- supports many desirable features that OLTP applications require.

Features

- high scalability
- SQL interface
- multi-row transaction support
- secondary index support

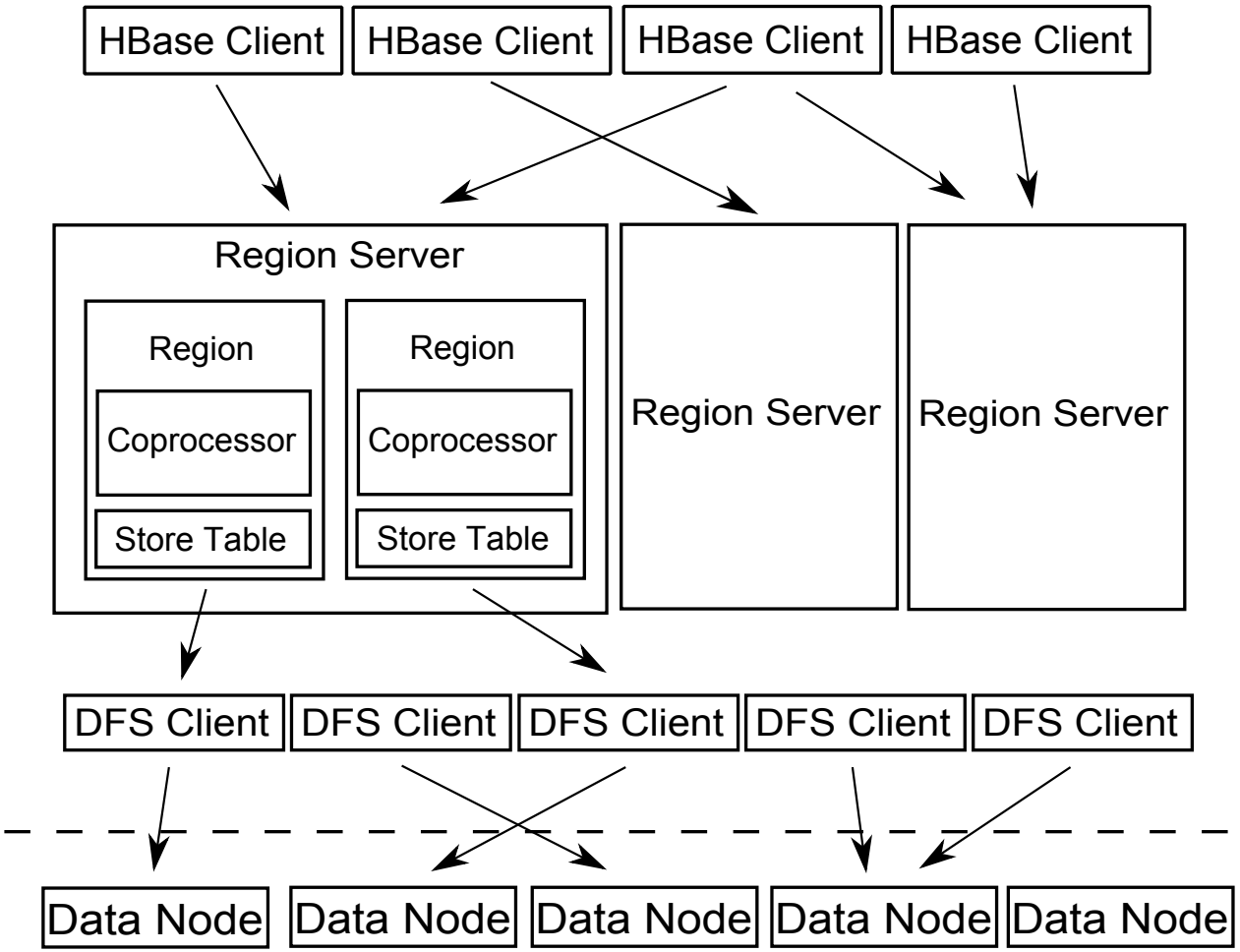
Contributions

- Provide a SQL interface on HBase.
- Support multi-row transactions on HBase.
- Design a distributed secondary indexing scheme for HBase.

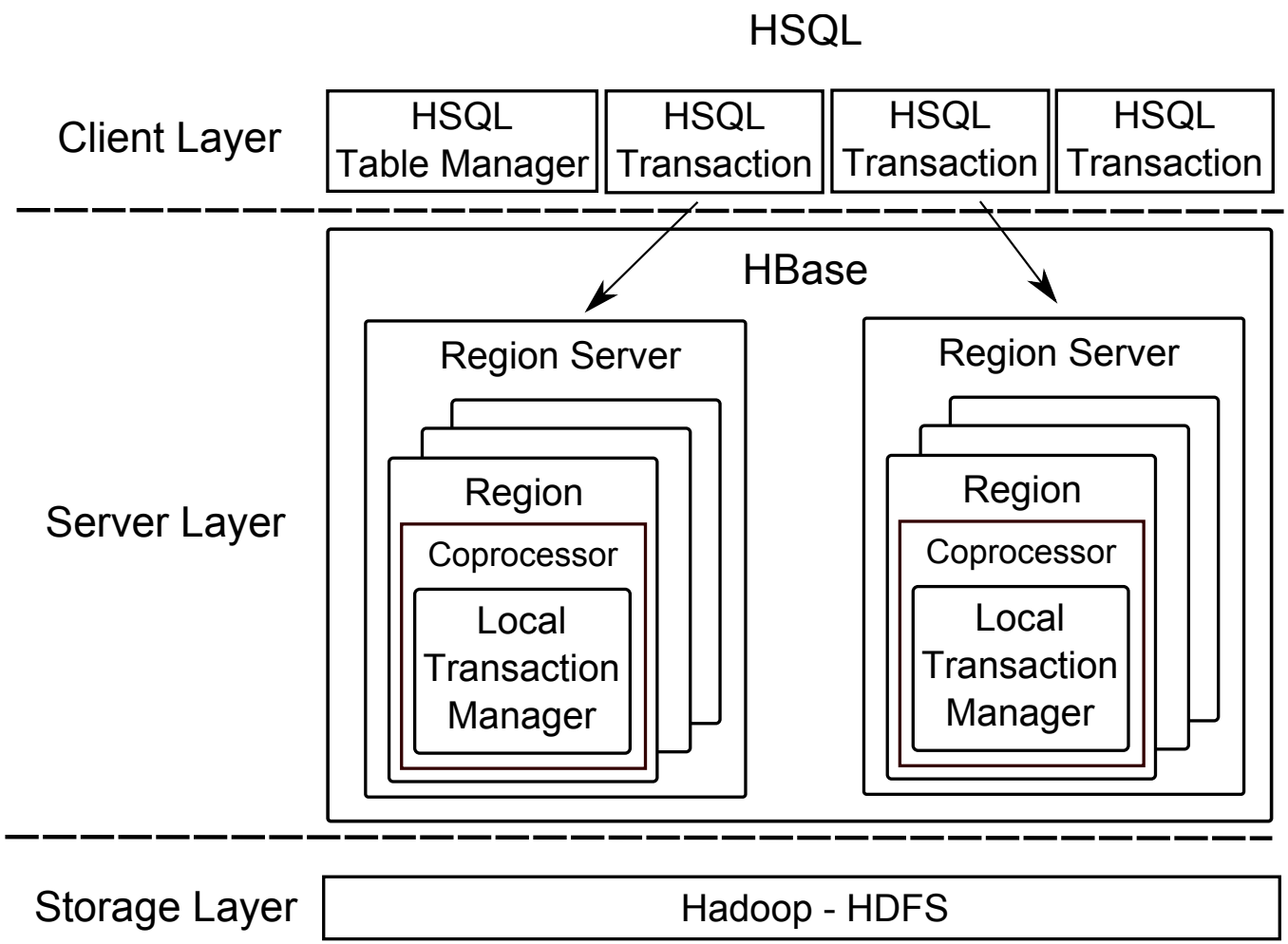
HBase

- a NoSQL database with high scalability.
- designed to host very large tables.
- good at random read/write access.
- built on the Hadoop framework.

Architecture of HBase

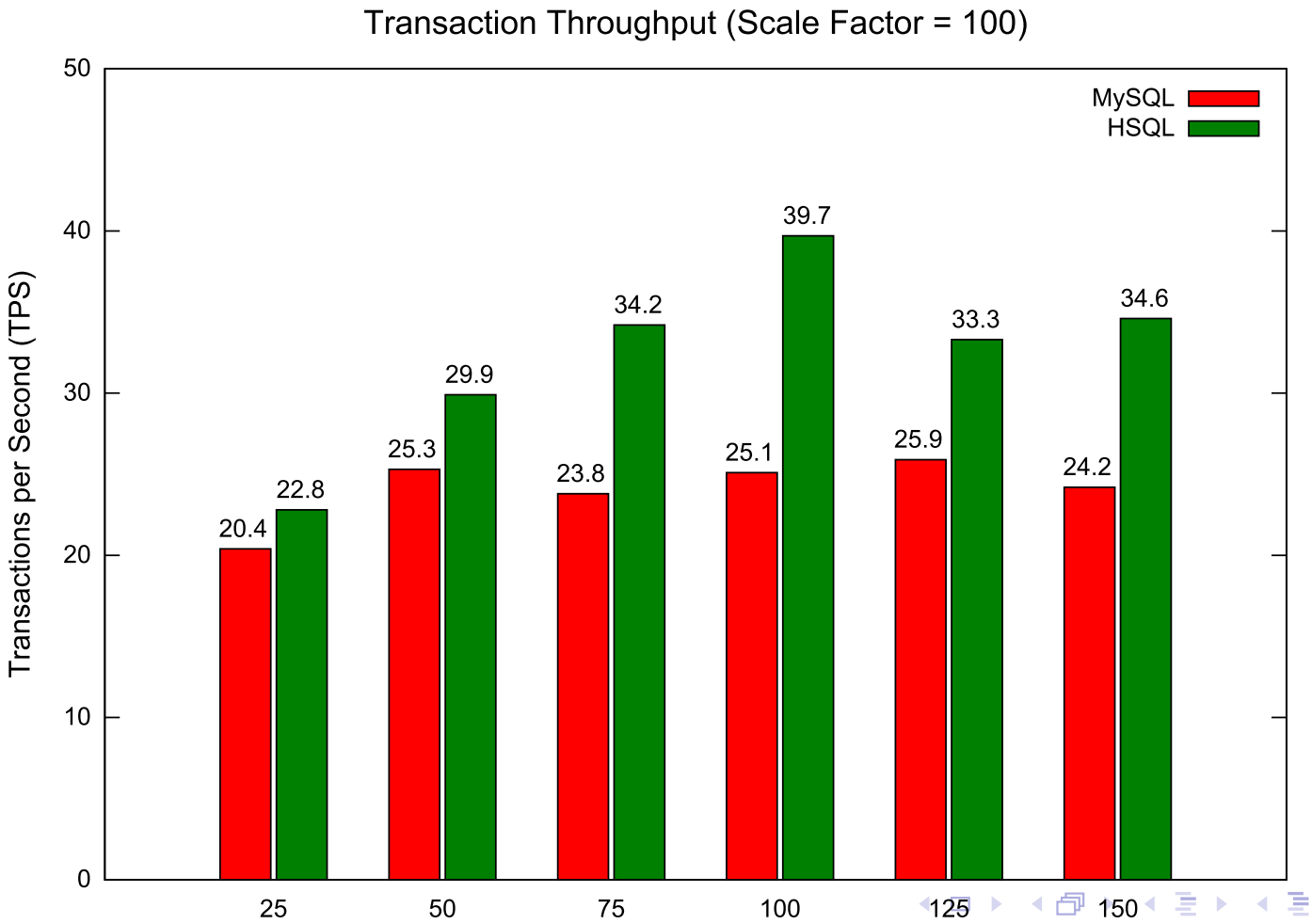


System Architecture of HSQL



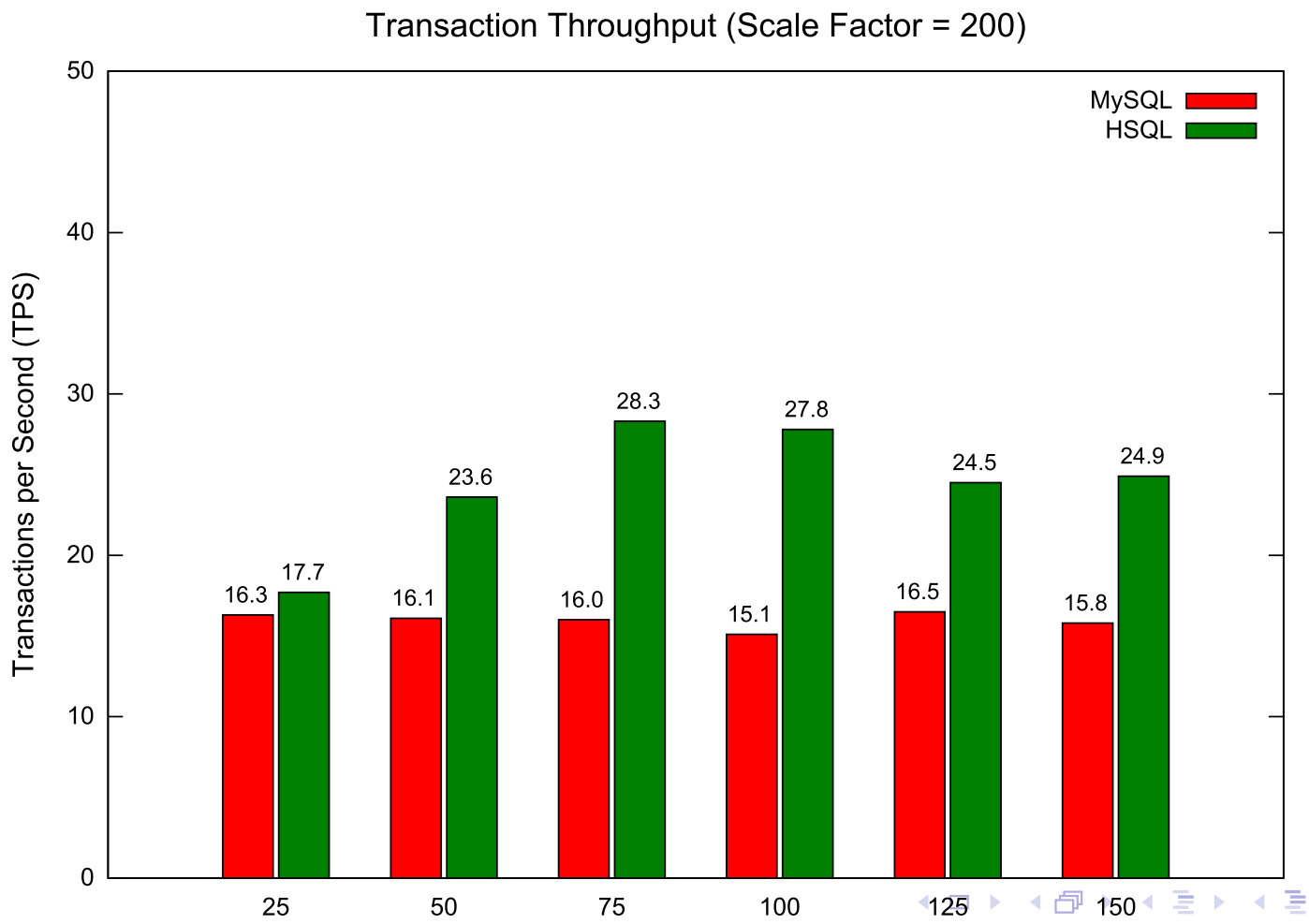
Transaction Throughput

Scale Factor = 100



Transaction Throughput

Scale Factor = 200



HSQL Summary

- HSQL is a highly scalable database for OLTP applications.
- HSQL provides a SQL interface for applications.
- HSQL supports multi-row transactions on HBase.
- HSQL uses a distributed B-tree scheme to improve performance.
- Experiment results indicate that HSQL scale well on large data sets.

Large Scale Graph Computation

- MapReduce has been proven to be efficient for a specific class of large scale data processing
  - does not perform well on graph data processing
- Google propose Pregel, utilizing *Bulk Synchronous Parallel* (BSP) model for large scale graph processing

BSP Model on Graph Computation

- Vertex centric, iterative computation model
  - User implements compute function which targets at a single vertex
  - Resemble to map and reduce function in Map-reduce model
- Computation consists of a sequence of iterations, called supersteps
- The execution of compute function are synchronized between supersteps

BSP Model Execution Flow

- At first, all vertices are set to active state
- In each iteration, a compute function is invoked on each active vertex to
  1. read messages sent to it in previous iteration
  2. modify its vertex value according to messages
  3. send messages to other neighboring vertices (activate other vertices)
  4. optionally vote to halt computation (become inactive)
- If all vertices are in inactive state, then end computation

Giraph

- An iterative graph processing system
- Originated as the open-source counterpart to Pregel from Google
- Used at Facebook to analyze social graph

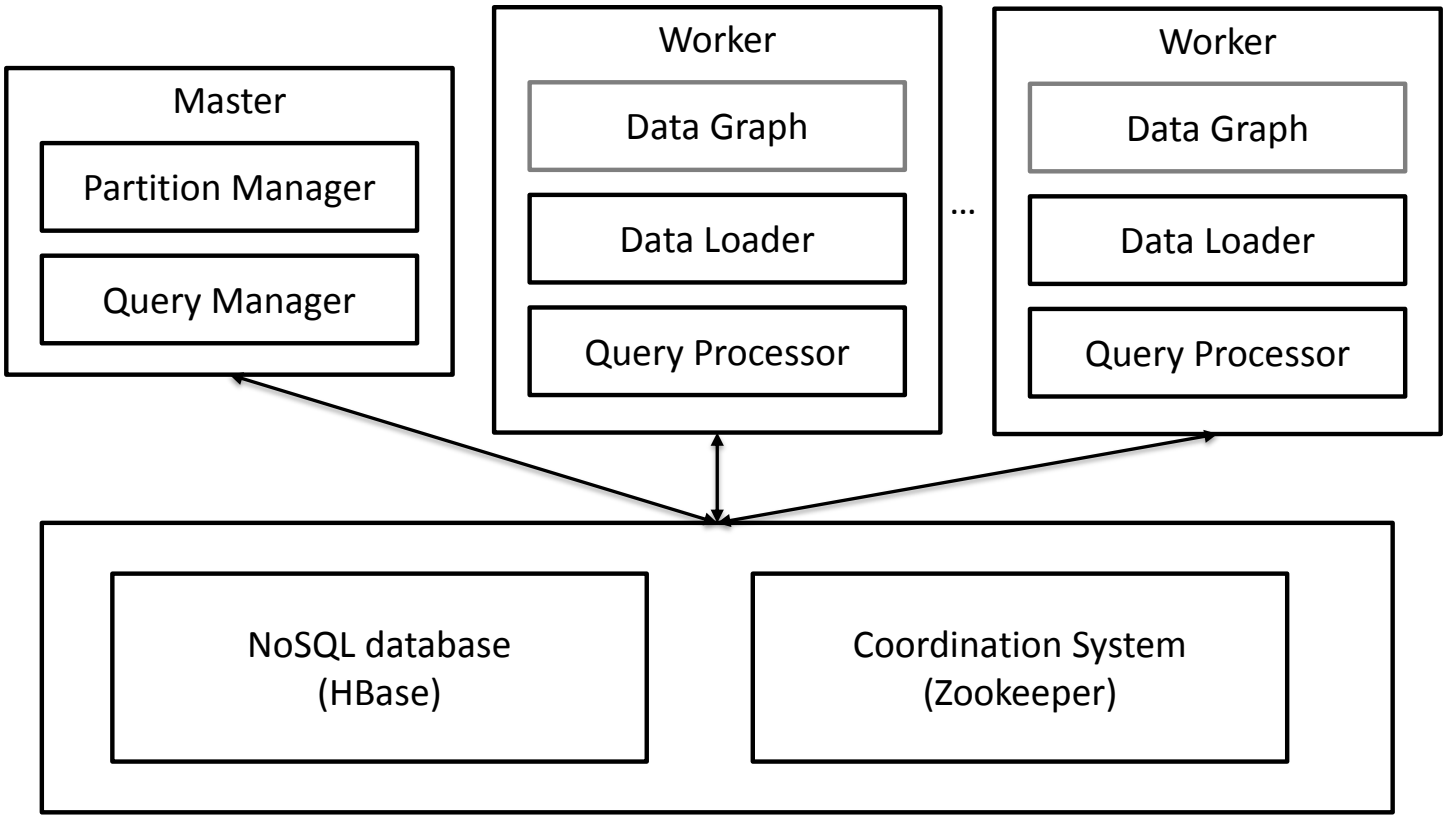
Hama

- A BSP computing framework on top of HDFS
- Designed for massive scientific computations such as matrix, graph and network algorithms

Kylin

- An efficient and scalable graph data processing system
- Highly optimized for processing large scale graphs
- Cooperates with HBase to achieve scalable data manipulation

Kylin System Architecture



Kylin Optimization

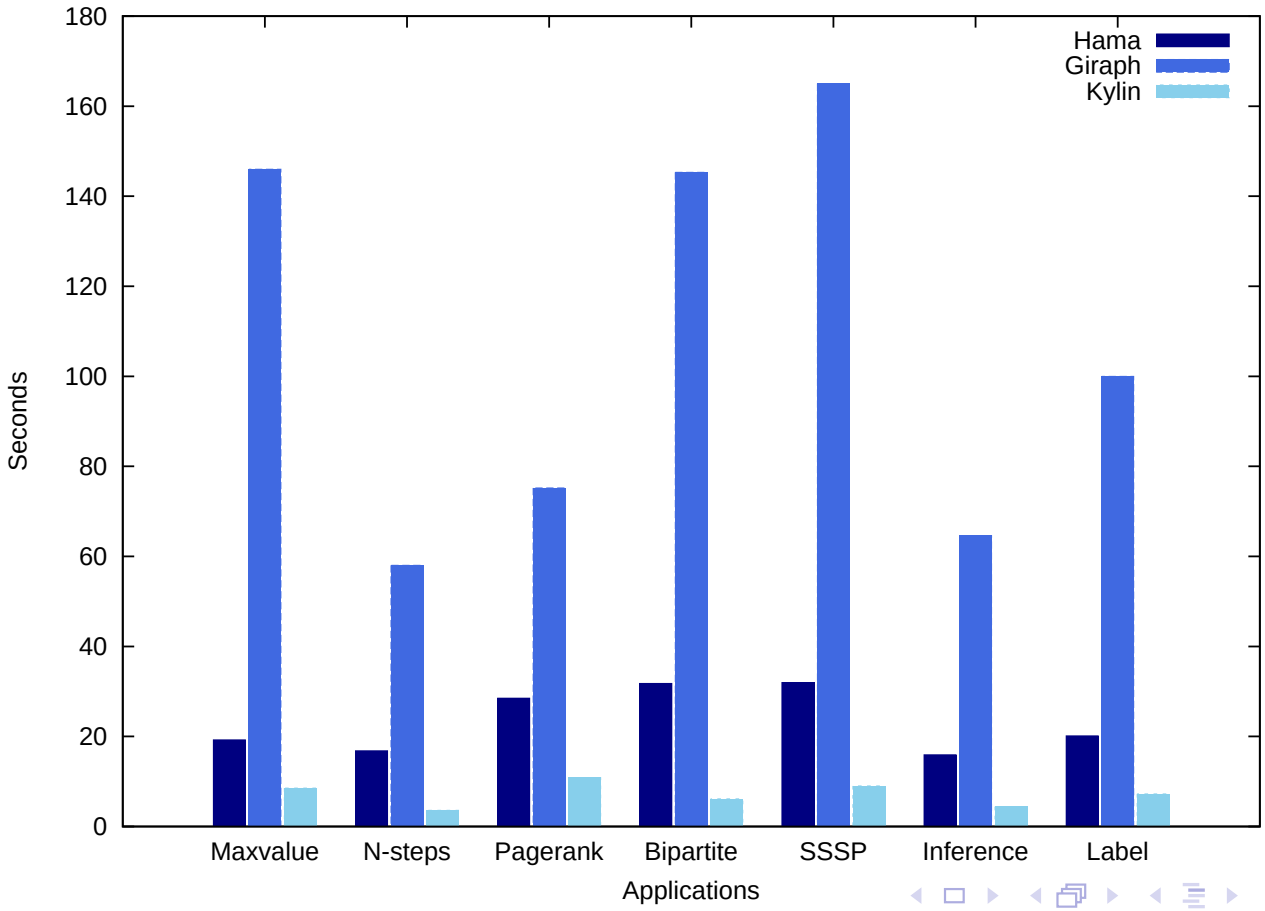
- Pull Messaging
  - Applied to algorithms which requires all neighboring data in order to do computation
- Lazy Vertex Loading
  - Applied to sub-graph query
- Vertex Weighted Partitioning

Dataset

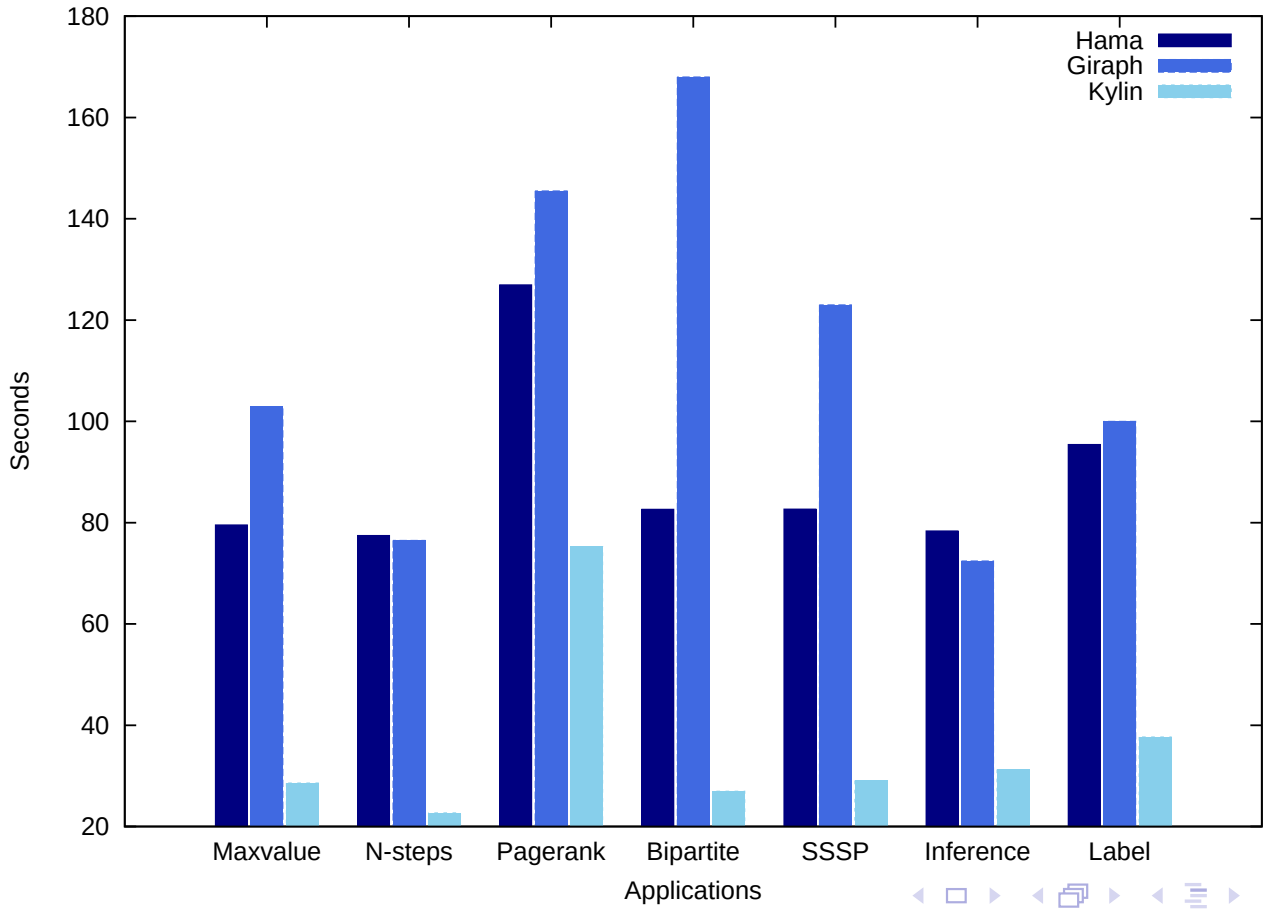
- Four real social networks as the input data

Social Networks	Nodes (millions)	Edges (millions)
Orkut	3.07	117.26
Flicker	1.86	15.97
LiveJournal	5.28	49.4
YouTube	1.16	3.01

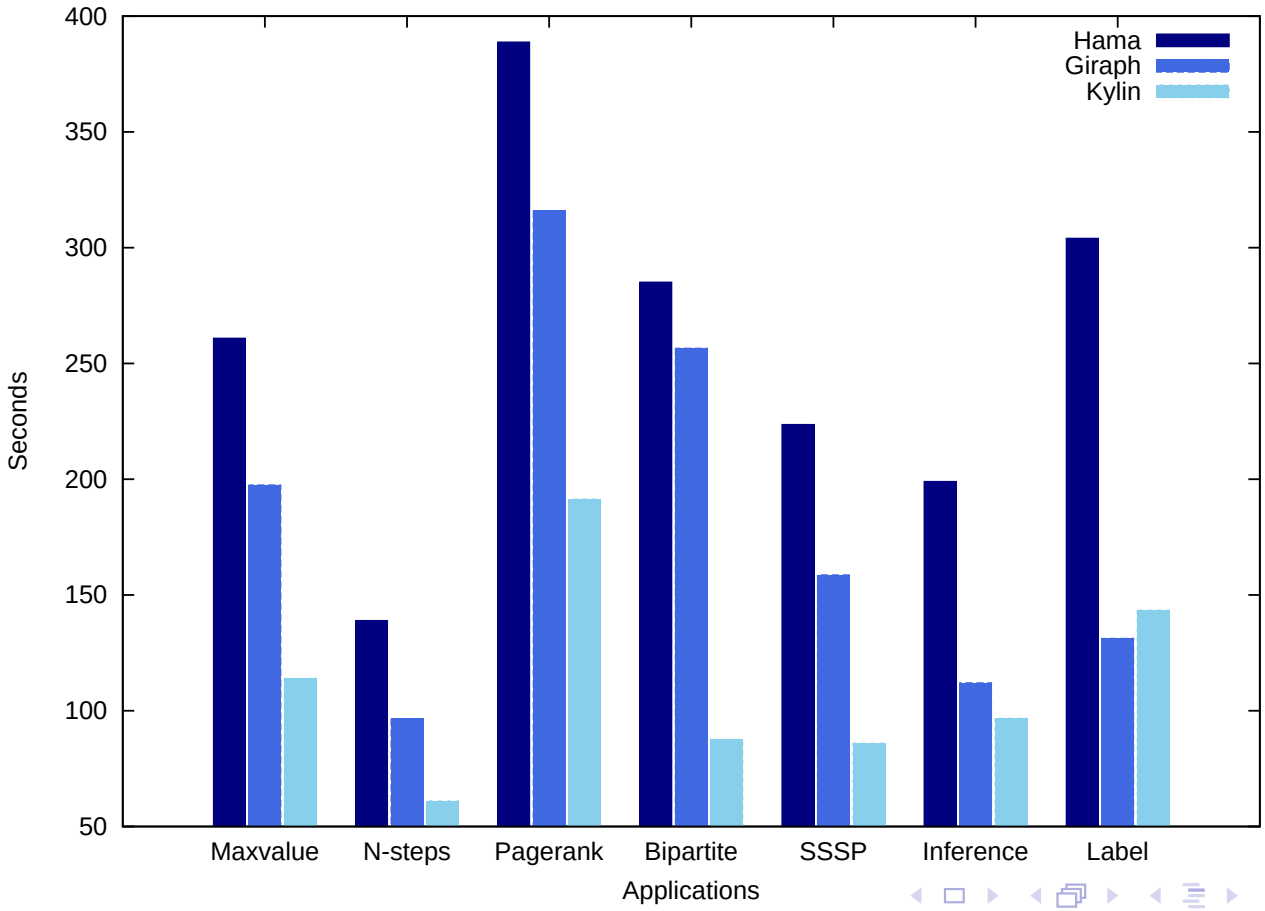
Overall Performance



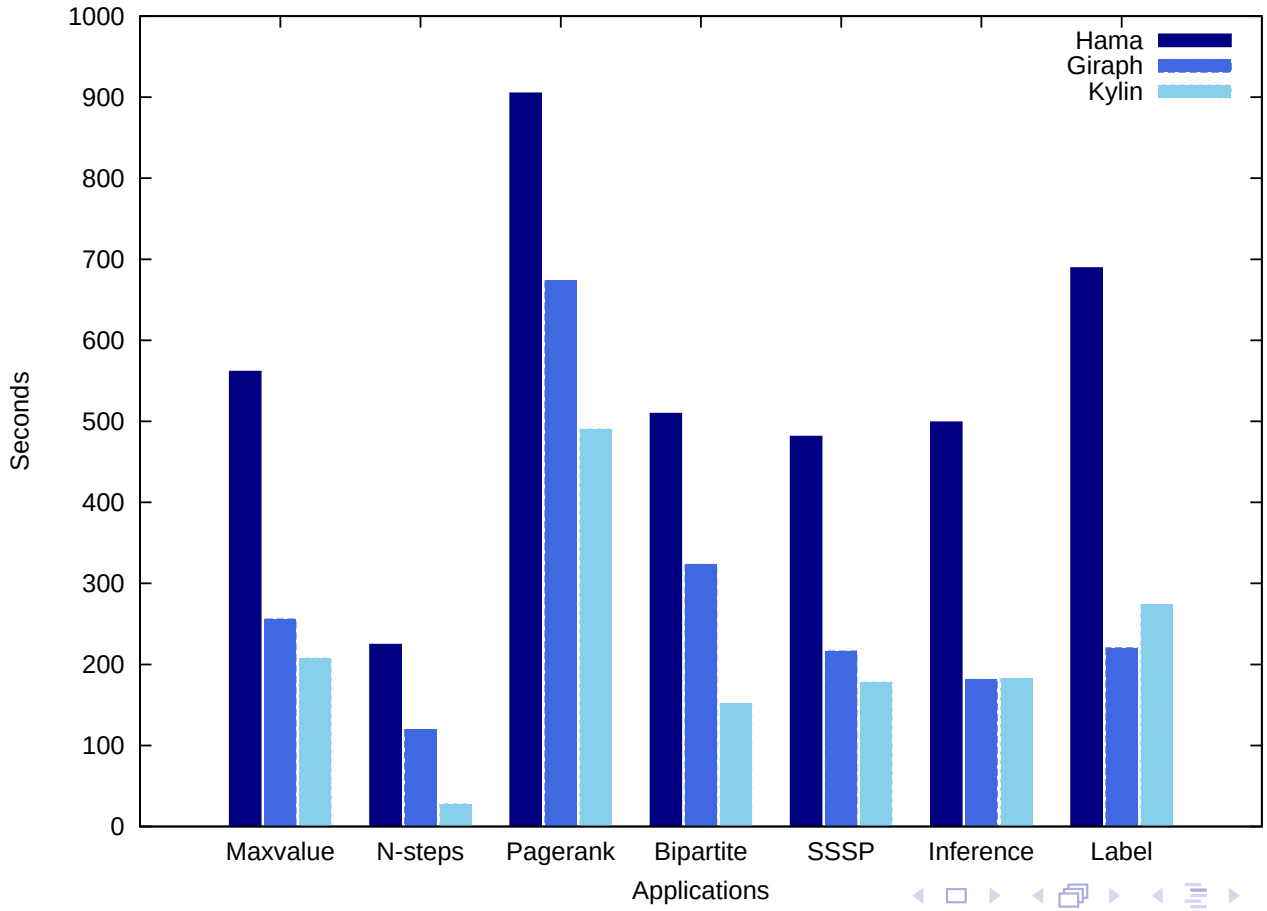
Overall Performance



Overall Performance



Overall Performance



Conclusion

- Scalability is the key issue.
- Cloud infrastructure is essential.
- Data is “big” only when it reaches a scale we cannot process it with traditional IT infrastructure.
- NoSQL will be crucial because of its stability.
- We still have a long way to process exabyte data set.