Beyond RDBMSs

Sharma Chakravarthy
Information Technology Laboratory
Computer Science and Engineering Department
The University of Texas at Arlington, Arlington, TX 76009
Email: sharma@cse.uta.edu
URL: http://itlab.uta.edu/sharma

Presentation Outline

- ***Cloud computing***
- ***Big Data***
- ***Map/reduce***
- ***NoSQL DBMSs***

The above are not the same. Please understand the differences clearly!

Cloud computing can be seen as a better architecture for big data analysis

Map/Reduce is a technique for divide and conquer on a large scale!

NoSQL DBMSs are typically non-relational DBMSs suited for a specific context and may not use SQL!

Acknowledgements

- These slides are put together from a variety of sources (both papers and slides/tutorials available on the web)
- Mostly I have tried to: provide my perspective, emphasize aspects that are of interest to this course, and have tried to put forth a consolidated view of the need for non-relational data processing needs

Beyond DBMSs

- DBMSs have served as enterprise data storage and management for over 2 decades (mainly OLTP)
- Data warehouses have provided additional drill down and OLAP (in contrast to OLTP) and has been integrated well with Relational DBMSs
- Data Mining has been applied on DBMSs and data warehouses
- However, the data and processing needs have evolved since the advent of RDBMSs – Big Data
- Scalability – of both data and processing has not been very easy with the architecture of DBMSs
  - Especially, data distribution, query optimization, CC, and recovery in multi-processor environments

Scalability of DBMSs

- Data (horizontal) scalability is accommodated by adding more disks and/or larger disks
  - This means longer wait for I/O given the same number of processors
  - To overcome impedance mismatch, you need more I/O throughput, i.e., additional I/O bandwidth!
    - This has not been easy!
  - Data partitioning and distribution need to be figured out
- Computing (or vertical) scalability involves adding more processors
  - However, query optimization, cc, and joins have to be managed for this environment
    - Requires software enhancement; just cannot add processors or disks at will!

Beyond RDBMSs

- I think we have a come a full circle from the pre-DBMS days to current situation
  - Things that are the same
    - Deciding appropriate representation
    - Manual or automated optimization
    - Including functionality on a need basis (e.g., atomicity, isolation, durability etc.)
    - Lack of mapping levels
  - Things that have changed
    - Application types
    - Data size and variety
    - Representation alternatives
    - Customizing for a specific representation/application

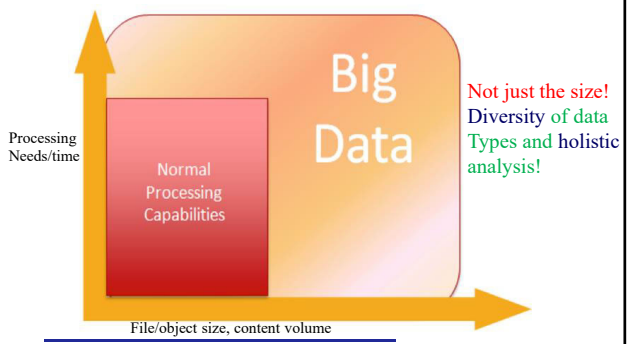
Path to cloud computing

- Main Frames and terminals (70's)
  - Primitive cloud computing, so to say
- Desktop/PC revolution (80's)
- HPC (High Performance computing)/super computers (90's)
- Internet (2000's)
- Grid computing (2000's)
- Cloud computing (2007 onwards)

The cloud

- In 2000's (21st century)
- Evolved from grid computing with improvements
- By this time cluster computing, server farms, and use of commodity components (processors and especially disks) had become common place
- Improvements and advances in hardware and software technologies as well!
- Networking was more robust and bandwidth kept increasing (think of modems with 56K baud rate)
- Due to Internet accessibility, fluctuations of access and usage became much larger (Internet-scale)

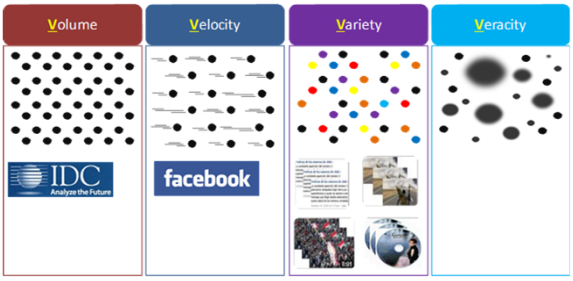
What is Big Data?



Not just the size!
Diversity of data Types and holistic analysis!

Characteristics of Big Data

- 4V: Volume, Velocity, Variety, Veracity



You will see more V's in the literature; but I think they do not characterize data! (e.g., value, variability, validity, vulnerability, volatility, visualization)

Big Data Analytics

- Definition: a process of inspecting, cleaning, transforming, and modeling large amounts of disparate data with the goal of discovering useful information, suggesting conclusions, and supporting decision making (i.e., **holistic analysis**!)
- Connection to data mining
  - Analytics include both data analysis (mining) and communication (guide decision making)
  - Analytics is not so much concerned with individual analyses or analysis steps, but with the entire methodology

Cloud computing

- Consider the official Wimbledon site. The site gets extremely high traffic in the two (to three) weeks when the championship happens. For this two week period, this site will have high server usage.
- For the rest of the year the site will have low traffic and hence most of the resources will be idle
- Spare capacity need to be maintained or leased from somewhere!
- Internet-scale elasticity

Traditional approach

- Keeping enough spare capacity to deal with the needs of 2 to 3 weeks in a year
  - Is Expensive
  - Hardware gets old and obsolete
  - Maintain and manage people
  - Software acquisition and maintenance
- It was not easy/possible to outsource this just for 2 weeks
- It was not possible to rent just for 2 weeks
- IT shops were maintained that were quite expensive!

Drivers

- These situations and needs were a side effect of Internet availability and ubiquitous usage
- Making things available on the web is critical
- Enterprises were maintaining expensive IT shops with all the costs and headaches that came with that
- Cloud computing is the answer for this problem!
  - Could not be done earlier due to technology limitations
    - Reliable networking, bandwidth, manage clusters, privacy, software as a service

Cloud computing: components

- To get cloud computing to work, you need:
  - Thin clients (or clients with a thick-thin switch)
  - Grid computing: ability to link disparate computers to form one large infrastructure, harnessing unused resources (we will see subtle differences later)
  - Utility computing*: paying for what you use on shared servers like you pay for a public utility (such as electricity, gas, and so on).
  - On-demand resource provisioning*: not static provisioning, no need to indicate resource requirements ahead, ability to add and remove computing resources based on needs!

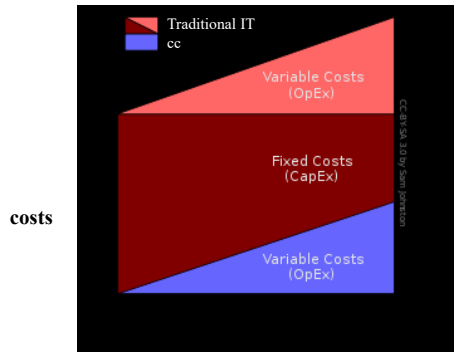
What is cloud computing

- Cloud Computing makes computer infrastructure and services available "on-need" or “on-demand” basis (like utilities).
- The computing infrastructure could include hard disks, development platform, database, computing power, or complete software applications.
- To access these resources from the cloud vendors, organizations do not need to make any large scale capital expenditures.

What is cloud computing

- Organizations need to "pay per use". That is, organizations need to pay only as much for the computing infrastructure as they use.
- The billing model of cloud computing is similar to the electricity payment that we do on the basis of usage.
- Terminology
  - Vendor: cc service provider
  - Organization: cc user

CC Economics



CC applications

- Back to the official Wimbledon site. The site gets extremely high traffic in the two (to three) weeks when the championship happens. For this two weeks period, this site will have high server usage. For rest of the year the site will need to only pay for the reduced usage.
- In general, organizations do not need to bear the cost of computing infrastructure for their peak loads.
- The usage of computing resources that can be increased or reduced on need basis is called elastic computing.

CC Applications

- Hotmail.com was launched in 1996. It is widely considered as the first cloud computing application. The data is stored at the vendor servers, and users could pay incrementally to increase disk space usage. Many other services have emerged since then that allow users to store information (or perform processing) without paying any upfront charges. These are typically consumer-oriented services.
- Also, known as "leveling the playing field" or "barriers to entry"

CC Applications

- Twitter, MySpace, Wikipedia, YouTube, Facebook, LinkedIn, Google docs and blogger all have the characteristics explained above and are examples of cloud computing.
- Companies that provide Hosting services for disk space storage, images, emails are also examples of cloud computing

CC Applications

- Salesforce.com, founded in 1999, was the first successful example of providing software as a service in the business to business (B2B) domain. Salesforce is a CRM tool for sales executives providing features like managing customer details, running promotions etc.
- Google and Microsoft provide development platforms that can be accessed with "pay-per-use" billing model.
- Amazon.com was one of the first vendors to provide storage space and computing resources following the cloud computing model.

CC characteristics

- Since the cloud computing vendor provides services over the web, these are available from any location
- Cloud computing can be ordered online without detailed formal contracts.
- Cloud computing provides a level playing field for smaller organizations. It allows smaller organization access to computing infrastructure without making any significant initial investment.
- As an example Mozy online storage provides online backup using the cloud computing model.
  - $5.99/month for 50 GB
  - Now everyone does it

Types of clouds

- **Location based**
  - Public
  - Private
  - Hybrid
  - Community
- **Service based**
  - Infrastructure as a service (IaaS)
  - Platform as a service (PaaS)
  - Software as a Service (SaaS)

Thank You !