



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>


11/10/2018  © Sharma Chakravarthy 1



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

11/10/2018  © Sharma Chakravarthy 3



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!

11/10/2018  © Sharma Chakravarthy 2



Beyond DBMSs

- DBMSs have served enterprise data storage and management for over 3 decades
- 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, Plan generation, cc and recovery in multi-processor environments

11/10/2018  © Sharma Chakravarthy 4

Scalability of DBMSs



- **Data 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 bandwidth or add additional I/O bandwidth!
 - This is not easy!
- **Computing scalability** involves adding more processors and distributing data across more disks
 - However, query optimization, cc, and joins have to be managed for this environment
 - Requires software enhancement; just cannot add processors or disks at will!

11/10/2018



© Sharma Chakravarthy

5

Path to cloud computing



- Main Frames and terminals (70's)
- Desktop/PC revolution
- HPC (High Performance computing)/super computers
- Internet
- Grid computing
- Cloud computing

11/10/2018



© Sharma Chakravarthy

7

Beyond RDBMSs



- I think we have come a **full circle** from the pre-DBMS days to current (NoSQL) 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 types and size
 - Representation alternatives
 - Customizing for a specific representation

11/10/2018



© Sharma Chakravarthy

6

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 had become common place
- Improvements and advances in hardware and software technologies as well!
- Due to Internet accessibility, fluctuations of access and usage became much larger (Internet-scale)

11/10/2018



© Sharma Chakravarthy

8

What is Big Data?



- There is not a consensus as to how to define Big Data
 - "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*
 - "Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population." – *Tera- data magazine article, 2011*
 - "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." – *The McKinsey Global Institute, 2011*

11/10/2018



© Sharma Chakravarthy

9

Big Science vs. Big Business



- Common
 - Need technologies to work with data
 - Use algorithms to mine data
- Big Science
 - Source: experiments and research conducted in controlled environments
 - Goals: to answer questions, or prove theories
- Big Business
 - Source: transactions pertaining to all aspects of a biz
 - Goals: to discover new opportunities, measure efficiencies, uncover relationships

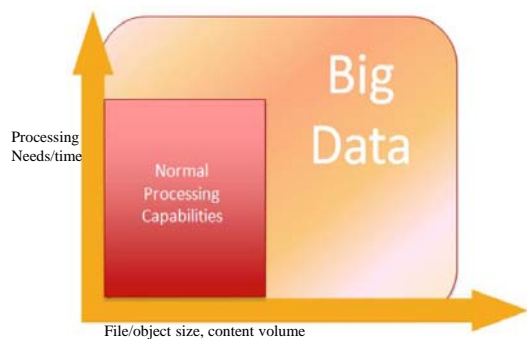
11/10/2018



© Sharma Chakravarthy

11

What is Big Data?



11/10/2018



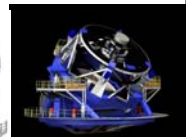
© Sharma Chakravarthy

10

Big Data from Science!



- CERN - Large Hadron Collider
 - ~10 PB/year at start
 - ~1000 PB in ~10 years
 - 2500 physicists collaborating
- Large Synoptic Survey Telescope (NSF, DOE, and private donors)
 - ~5-10 PB/year at start in 2012
 - ~100 PB by 2025
- Pan-STARRS (Haleakala, Hawaii)
 - US Air Force
 - now: 800 TB/year
 - soon: 4 PB/year



11/10/2018





© Sharma Chakravarthy

12

Big Data from different sources!


12+ TBs of tweet data every day







Billions of camera Images, Video uploads, Web usage, Web logs etc.

25+ TBs of data Every day



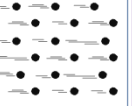









11/10/2018

© Sharma Chakravarthy
13

Characteristics of Big Data


• **4V: Volume, Velocity, Variety, Veracity**

Volume	Velocity	Variety	Veracity
 	 	 	

11/10/2018

© Sharma Chakravarthy
15


Big Data Business sectors

- US Health care
 - \$300B per year
- Europe public sector administration
 - 250B pounds per year
- Global personal location data
 - \$100B+ revenue for service providers
- US retail
- Manufacturing

11/10/2018

© Sharma Chakravarthy
14

Big Data Analytics/Science

- Definition: a process of **inspecting**, **cleaning**, **transforming**, and **modeling big data** with the goal of **discovering** useful information, **suggesting** conclusions, and **supporting** decision making
 - Holistic analysis and science needed for that
- 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**

11/10/2018

© Sharma Chakravarthy
16

Example



- 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 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

11/10/2018



© Sharma Chakravarthy

17

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 and all the cost and headaches that came with that
- Cloud computing is the answer for this problem!

11/10/2018



© Sharma Chakravarthy

19

Traditional approach



- Keep enough spare capacity to deal with the needs of 2 to 3 weeks in a year
 - Expensive
 - Hardware gets old and obsolete
 - Maintain and manage people
 - Software acquisition and maintenance
- It is not easy/possible to outsource this just for 2 weeks
- It is not possible to rent just for 2 weeks
- This is also true for companies maintaining their own IT shops

11/10/2018



© Sharma Chakravarthy

18

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!

11/10/2018



© Sharma Chakravarthy

20

Thin client



- This has been around for a long time (not new)
 - The dumb terminals served this purpose at the beginning of network revolution
- DBMSs have thin clients which accept queries and all the processing is done at the server
- Applications that use backend servers and have a front end are also thin clients (both two- and three-tier architectures supported thin clients)
- A browser can be viewed as a thin client through which you can access lots of diverse applications

11/10/2018



© Sharma Chakravarthy

21

Pay as you use



- Sun, at some point, was shipping servers with extra cpus that could be brought into use as needed. You would not be paying for them until you used them!
- Cloud supports pay as you go for individual components such as cpus, storage, etc.

11/10/2018



© Sharma Chakravarthy

23

Grid Computing



- Although a grid can be dedicated to a specialized application, it is more common that a single grid will be used for a variety of different purposes. Grids are often constructed with the aid of general-purpose grid software libraries known as [middleware](#) (e.g., Globus Toolkit, PBS, LSF)
- Systems such as **Condor** pre-date Grid computing (and of course, cloud). They were not only developed to use idle resources in an effective, transparent manner, but also had built-in **recovery** mechanism
 - Developed by people who were database reserachers (at Wisconsin)!

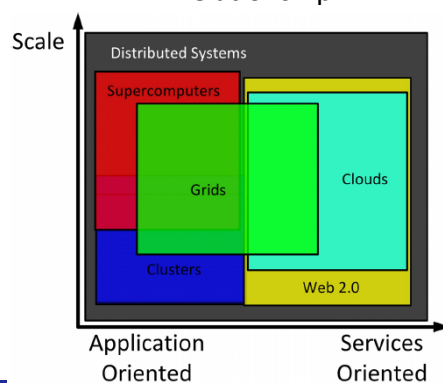
11/10/2018



© Sharma Chakravarthy

22

Relationship



11/10/2018



© Sharma Chakravarthy

24

CC Definitions



➤ Business perspective:

UCBerkeley RADLabs: "cloud computing has the following characteristics: i) the illusion of infinite computing resources, ii) the elimination of up-front commitment from Cloud users, iii) the ability to pay for use as needed"

11/10/2018



© Sharma Chakravarthy

25

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

11/10/2018



© Sharma Chakravarthy

27

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.

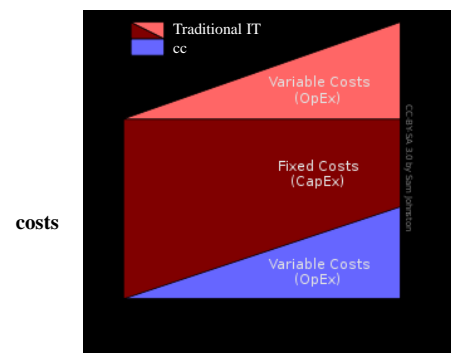
11/10/2018



© Sharma Chakravarthy

26

CC Economics



11/10/2018



© Sharma Chakravarthy

28

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.
-

11/10/2018



© Sharma Chakravarthy

29

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

11/10/2018



© Sharma Chakravarthy

31

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 in the last decade that allows users to store information (or perform processing) without paying any upfront charges. These are typically consumer-oriented services.

11/10/2018



© Sharma Chakravarthy

30

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.

11/10/2018



© Sharma Chakravarthy

32

CC characteristics



5. Since the cloud computing vendor provides services over the web, these are available from any location
6. 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

11/10/2018



© Sharma Chakravarthy

33

IaaS



- Infrastructure as a service (IaaS) involves offering hardware related services using the principles of cloud computing. These could include some kind of storage services (database or disk storage) or virtual servers. Leading vendors that provide Infrastructure as a service are [Amazon EC2](#), [Amazon S3](#), [Rackspace Cloud Servers](#) and [Flexiscale](#).
-

11/10/2018



© Sharma Chakravarthy

35

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)

11/10/2018



© Sharma Chakravarthy

34

PaaS



- Platform as a Service (PaaS) involves offering a development platform on the cloud. Platforms provided by different vendors are typically not compatible. Typical players in PaaS are [Google's Application Engine](#), [Microsoft's Azure](#), Salesforce.com's [force.com](#).

11/10/2018



© Sharma Chakravarthy

36

SaaS



- Software as a service (SaaS) includes a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector. The pioneer in this field has been Salesforce.com's offering in the online Customer Relationship Management (CRM) space. Other examples are online email providers like Google's [gmail](#) and Microsoft's [hotmail](#), [Google docs](#) and Microsoft's online version of office called [BPOS](#) (Business Productivity Online Standard Suite).

11/10/2018



© Sharma Chakravarthy

37

Thank You !



Summary



- Essentially a paradigm/model where capital equipment, management, and maintenance can be rented (instead of owning) on a need basis.
- For raw commodities, such as cpus, storage, and bandwidth, this is straightforward
- In fact, most hosting web sites use a similar model; but they also provide different levels of development tools, software, etc. to make use of the raw commodities.
- However, it would be very useful if you can also support a transparent programming paradigm to use these resources effectively

11/10/2018



© Sharma Chakravarthy

38