Cloud Computing

Hwajung Lee

Key Reference:
Prof. Jong-Moon Chung's Lecture Notes at Yonsei University

Cloud Computing

- Cloud Introduction
- Cloud Service Model
- **Big Data**
- **Hadoop**
- MapReduce
- HDFS (Hadoop Distributed File System)

Big Data

- **New FLU Virus Starts in the U.S.!**
  - **H1N1 flu virus (which has combined virus elements of the bird and swine (pig) flu) started to spread in the U.S. in 2009**
  - **U.S. CDC (Centers for Disease Control and Prevention) was only collecting diagnostic data of Medical Doctors once a week**
  - **Using the CDC information to find how the flu was spreading would have an approximate 2 week lag, which is far too slow compared to the speed of the virus spreading**

Big Data

- New FLU Virus Starts in the U.S.!
  - What vaccine was needed?
  - How much vaccine was needed?
  - Where was the vaccine needed?
  - Vaccine preparation and delivery plans could not be setup fast enough to safely prevent the virus from spreading out of control

Big Data

- New FLU Virus Starts in the U.S.!
  - Fortunately, Google published a paper about how they could predict the spread of the winter flu in the U.S. accurately down to specific regions and states
  - This paper was published in the journal *Nature* a few weeks before the H1N1 virus made the headline news

Big Data

- **New FLU Virus Starts in the U.S.!**
  - **Millions of the most common search terms and Millions of different mathematical models were tested on Google's database**
    - **Google receives more than 3 billion search queries a day**
  - **Analysis system was set to look for correlation between the frequency of certain search queues and the spread of the flu over time and space**

Big Data

- **New FLU Virus Starts in the U.S.!**
  - **Google's method of analysis did not use data provided from hospitals or Medical Doctors**
  - **Google used Big Data analysis on the most common search terms people use**
  - **Google's system proved to be more accurate and faster than analyzing government statistics**

Big Data

- **Wal-Mart**
  - **Wal-Mart's Data Warehouse**
    - **Stores 4 petabytes (4×10^{15}) of data**
    - **Records every single purchase**
      - **Approximately 267 million transactions a day from 6000 stores worldwide is recorded**

WAL★MART®

Big Data

- **Wal-Mart**
  - **Wal-Mart's Data Analysis**
    - **Focused on evaluating the effectiveness of pricing strategies and advertising campaigns**
    - **Seeking for improvement methods in inventory management and supply chains**

WAL★MART®

Big Data

- **Recommendation System using Big Data**
  - **Based on data analysis of simple elements**
    - **What users made purchases in the past**
    - **Which items do they have in their virtual shopping cart**
    - **Which items did customers rate and like**
    - **What influence did the rating have on other customers to make a purchase**

Big Data

- **Amazon.com**
  - **Amazon.com's Recommendation System**
  - **Item-to-Item Collaborative Filtering Algorithm**
    - **Personalization of the Online Store**
      - ➔ **Customized to each customer**
    - **Each customer's store is based on the customer's personal interest**
      - **Example: For a new mother, the store will display baby supplies and toys**

amazon.com and you're done.™

Big Data

- **Citibank**
  - **Bank operations in 100 countries**
  - **Big Data analysis on the database of basic financial transactions can enable Global insight on investments, market changes, trade patterns, and economic conditions**
  - **Many companies (e.g., Zara, H&M, etc.) work with Citibank to locate new stores and factories**

citibank

Big Data

- **Product Development & Sales**
  - **For example, a Smartphone takes significant time and money to manufacture**
  - **In addition, the duration of popularity for a new Smartphone is limited**
  - **To maximize sales, a company needs to manufacture just the right amount of products and sell them in the right locations**

Big Data

- **Product Development & Sales**
  - **Too much will result in leftovers and a big waste for the company!**
  - **Too less will result in a lost opportunity for company profit and growth!**
  - **Big Data analysis can help find how many smartphones and where the products could be popular based on common search terms that people use ➔ Use this to also estimate how many products could be sold in a certain location ➔ But why is this difficult?**

Big Data

- **Big Data's 4 V Big Challenges**
  - **Volume – Data Size**
  - **Variety – Data Formats**
  - **Velocity – Data Streaming Speeds**
  - **Veracity – Data Trustworthiness**

Big Data

- **Volume – Data Size**
  - **40 Zettabytes (10^{21}) of data is predicted to be created by 2020**
  - **2.5 Quintillionbytes (10^{18}) of data are created every day**
  - **6 Billion (10^{9}) people have mobile phones**
  - **100 Terabytes (10^{12}) of data (at least) is stored by most U.S. companies**
  - **966 Petabytes (10^{15}) was the approximate storage size of the American manufacturing industry in 2009**

Big Data

- **Variety – Data Formats**
  - **150 Exabytes (10^{18}) was the estimated size of data for health care throughout the world in 2011**
  - **More than 4 Billion (10^{9}) hours each month are used in watching YouTube**
  - **30 Billon contents are exchanged every month on Facebook**
  - **200 Million monthly active users exchange 400 Million tweets every day**

Big Data

- **Velocity – Data Streaming Speeds**
  - **1 Terabytes (10^{12}) of trade information is exchanged during every trading session at the New York Stock Exchange**
  - **100 sensors (approximately) are installed in modern cars to monitor fuel level, tire pressure, etc.**
  - **18.9 Billion network connections are predicted to exist by 2016**

Big Data

- **Veracity – Data Trustworthiness**
  - **1 out of 3 business leaders have experienced trust issues with their data when trying to make a business decision**
  - **$3.1 Trillion ($10^{12}$) a year is estimated to be wasted in the U.S. economy due to poor data quality**

Big Data

- New technology is needed to overcome these 4 V Big Data Challenges
  - Volume – Data Size
  - Variety – Data Formats
  - Velocity – Data Streaming Speeds
  - Veracity – Data Trustworthiness

Hadoop

Hadoop

- **Data Storage, Access, and Analysis**
  - **Hard drive storage capacity has tremendously increased**
  - **But the data read and write speeds to and from the hard drives have not significantly improved yet**
  - **Simultaneous parallel read and write of data with multiple hard disks requires advanced technology**

Hadoop

- **Data Storage, Access, and Analysis**
  - **Challenge 1: Hardware Failure**
    - **When using many computers for data storage and analysis, the probability that one computer will fail is very high**
  - **Challenge 2: Cost**
    - **To avoid data loss or computed analysis information loss, using backup computers and memory is needed, which helps the reliability, but is very expensive**

Hadoop

- **Data Storage, Access, and Analysis**
  - **Challenge 3: Combining Analyzed Data**
    - **Combining the analyzed data is very difficult**
    - **If one part of the analyzed data is not ready, then the overall combining process has to be delayed**
    - **If one part has errors in its analysis, then the overall combined result may be unreliable and useless**

Hadoop

- Hadoop

  - Hadoop is a Reliable Shared Storage and Analysis System

  - Hadoop = HDFS + MapReduce + α
    - HDFS provides Data Storage
      - HDFS: Hadoop Distributed FileSystem
    - MapReduce provides Data Analysis
      - MapReduce = Map Function + Reduce Function

Hadoop

- **HDFS: Hadoop Distributed FileSystem**
  - **DFS (Distributed FileSystem) is designed for storage management of a network of computers**
  - **HDFS is optimized to store huge files with streaming data access patterns**
  - **HDFS is designed to run on clusters of general computers**

Hadoop

- **HDFS: Hadoop Distributed FileSystem**
  - **HDFS was designed to be optimal in performance for a WORM (Write Once, Read Many times) pattern, which is a very efficient data processing pattern**
  - **HDFS was designed considering the time to read the whole dataset to be more important than the time required to read the first record**

Hadoop

- **HDFS**
  - **HDFS clusters use 2 types of nodes**
  - **Namenode (master node)**
  - **Datanode (worker node)**

Hadoop

HDFS: Namenode

- Manages the filesystem namespace
- Maintains the filesystem tree and the metadata for all the files and directories in the tree
- Stores on the local disk using 2 file forms
  - Namespace Image
  - Edit Log

Hadoop

- **HDFS: Datanodes**
  - **Workhorse of the filesystem**
  - **Store and retrieve blocks when requested by the client or the namenode**
  - **Report back to the namenode periodically with lists of blocks that were stored**

Hadoop

- **MapReduce**
  - **MapReduce is a program that abstracts the analysis problem from stored data**
  - **MapReduce transforms the analysis problem into a computation process that uses a set of keys and values**

Hadoop

- **MapReduce System Architecture**
  - **MapReduce was designed for tasks that consume several minutes or hours on a set of dedicated trusted computers connected with a broadband high-speed network managed by a single master data center**

Hadoop

MapReduce Characteristics

- MapReduce uses a somewhat brute-force data analysis approach
- The entire dataset (or a big part of the dataset) is processed for every query
  ➔ *Batch* Query Processor model

Hadoop

MapReduce Characteristics

- MapReduce enables the ability to run an ad hoc query against the whole dataset within a scalable time
- Many distributed systems combine data from multiple sources (which is very difficult), but MapReduce does this in a very effective and efficient way

Hadoop

- Technical Terms used in MapReduce
  - Seek Time is the delay in finding a file
  - Transfer Rate is the speed to move a file
  - Transfer Rate has improved significantly more (i.e., now has much faster transfer speeds) compared to improvements in Seek Time (i.e., still relatively slow)

Hadoop

- **MapReduce**
  - **MapReduce gains performance enhancement through optimal balancing of Seeking and Transfer operations**
    - **Reduce Seek operations**
    - **Effectively use Transfer operations**

References

- V. Mayer-Schönberger, and K. Cukier, *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.
- T. White, *Hadoop: The Definitive Guide*. O'Reilly Media, 2012.
- J. Venner, *Pro Hadoop*. Apress, 2009.
- S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “Big Data, Analytics and the Path From Insights to Value,” *MIT Sloan Management Review*, vol. 52, no. 2, Winter 2011.
- B. Randal, R. H. Katz, and E. D. Lazowska, "Big-data Computing: Creating revolutionary breakthroughs in commerce, science and society," *Computing Community Consortium*, pp. 1-15, Dec. 2008.
- G. Linden, B. Smith, and J. York. "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76-80, Jan/Feb. 2003.

References

- J. R. GalbRaith, "Organizational Design Challenges Resulting From Big Data," *Journal of Organization Design*, vol. 3, no. 1, pp. 2-13, Apr. 2014.
- S. Sagiroglu and D. Sinanc, “Big data: A review,” *Proc. IEEE International Conference on Collaboration Technologies and Systems*, pp. 42-47, May 2013.
- M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171-209, Jan. 2014.
- X. Wu, X. Zhu, G. Q. Wu, and W. Ding, “Data Mining with Big Data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
- Z. Zheng, J. Zhu, and M. R. Lyu, “Service-Generated Big Data and Big Data-as-a-Service: An Overview,” *Proc. IEEE International Congress on Big Data*, pp. 403– 410, Jun/Jul. 2013.

References

- I. Palit and C.K. Reddy, “Scalable and Parallel Boosting with MapReduce,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 10, pp. 1904-1916, 2012.
- M.-Y Choi, E.-A. Cho, D.-H. Park, C.-J Moon, and D.-K. Baik, “A Database Synchronization Algorithm for Mobile Devices,” *IEEE Transactions on Consumer Electronics,* vol. 56, no. 2, pp. 392-398, May 2010.
- IBM, What is big data?, http://www.ibm.com/software/data/bigdata/what-is-big-data.html [Accessed June 1, 2015]
- Hadoop Apache, http://hadoop.apache.org
- Wikipedia, http://www.wikipedia.org

Image sources

- Walmart Logo, By Walmart [Public domain], via Wikimedia Commons
- Amazon Logo, By Balajimuthazhagan (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons