Course Title

Big Data Emerging Technologies

❖ **Modules**

1. Big Data Rankings & Products
2. Big Data & Hadoop
3. Spark
4. Spark ML & Streaming
5. Storm
6. IBM SPSS Statistics Project

Big Data

HADOOP

Hadoop Related Systems

❖ **Hadoop**

- Hadoop is an open source version of Google's MapReduce
- Hadoop is the name of a yellow elephant toy owned by the son of the Creator of Hadoop Doug Cutting, who decided to use this name because it was meaningless as well as relatively easy to spell and pronounce, and also was not used elsewhere

Hadoop Related Systems

❖ **HBase**

- Java based open source, non-relational, distributed database developed as part of the Apache Hadoop project
- HBase was modeled based on Google's BigTable and runs on HDFS (Hadoop Distributed Filesystem) to enable fault-tolerant storing of large quantities of sparse data

Hadoop Support Systems

❖ Apache Flume

- Provides services to efficiently collect, aggregate, and move large amounts of log data in a reliable distributed way
- Flume uses a simple and flexible data flow streaming mechanism that is fault tolerant by using tunable reliability mechanisms and recovery schemes
- The word “Flume" mean “a human-made channel for water,” which refers to the flexible data flow streaming features of Flume

Hadoop Support Systems

❖ Apache Mahout

- Provides free implementations of distributed and scalable ML (Machine Learning) algorithms that provide collaborative filtering, clustering, and classification of data, and also provides Java libraries for common math operations
- The word “Mahout" means “elephant rider, trainer, or keeper.” This is related to the name Hadoop

Hadoop Related Systems

❖ **HIVE**

- Developed by Facebook to enable its programmers (or those who are not Java experts) to easily write query codes in SQL to emulate on-demand SQL operations on 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 (commodity servers)

Hadoop

❖ HDFS: Hadoop Distributed FileSystem

- HDFS was designed to be optimal in performance for a WORM (Write Once, Read Many times) pattern, which is an 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

HDFS

❖ Blocks

- Files in HDFS are divided into block size chunks ➔ 64 Megabyte default block size
- Block is the minimum size of data that it can read or write
- Blocks simplifies the storage and replication process ➔ Provides fault tolerance & processing speed enhancement for larger files
- Hadoop's default replication factor is 3

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

HDFS

❖ Metadata

- Traditional concept of the library card catalogs
- Categorizes and describes the contents and context of the data files
- Maximizes the usefulness of the original data file by making it easy to find and use

HDFS

❖ Metadata Types

- Structural Metadata
  - Focuses on the data structure's design and specification
- Descriptive Metadata
  - Focuses on the individual instances of application data or the data content

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
- DataNodes send Heartbeat signals to the NameNode every 3 seconds to ensure connection and report operation status

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 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
- In the next lecture, we will compare MapReduce with a traditional RDBMS (Relational Database Management System)

Big Data

References

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