CS6350

Jiadao Zou jxz172230

I. What is Big Data

1. Section 1.1 of the paper
    1. **What does the term Big Data (BD) refer to? How is BD different from traditional datasets?**
        - The term of big data is mainly used to describe enormous datasets.
        - Compared with traditional datasets, big data typically includes masses of unstructured data that need more real–time analysis. In addition, big data also brings about new opportunities for discovering new values, helps us to gain an in–depth understanding of the hidden values, and also incurs new challenges, e.g., how to effectively organize and manage such datasets.**
    2. **What challenges have emerged because of the rise of BD?**
        - The main challenge is collecting and integrating massive data from widely distributed data sources.
        - The rapid growth of cloud computing and the Internet of Things (IoT) further promote the sharp growth of data. Such data in both quantity and mutual relations will far surpass the capacities of the IT architectures and infrastructure of existing enterprises, and its realtime requirement will also greatly stress the available computing capacity. The increasingly growing data cause a problem of how to store and manage such huge heterogeneous datasets with moderate requirements on hardware and software infrastructure.
        - In consideration of the heterogeneity, scalability, real– time, complexity, and privacy of big data, we shall effectively “mine” the datasets at different levels during the analysis, modeling, visualization, and forecasting, so as to reveal its intrinsic property and improve the decision making.**
2. Section 1.2 of the paper
    1. **This section presents several definitions and features of BD. Write down in pointwise fashion the features of BD. Pay special attention to the 3V definition proposed by Laney and understand what each term means.**
        - Datasets which could not be captured, managed, and processed by computers within a reasonable time frame.[Hadoop]
        - Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software.[McKinsey & Company]

- Big data was defined challenges and opportunities brought about by increased data with a 3Vs model, i.e., the increase of Volume, Velocity, and Variety.[Doug Laney:3V models]
  - Volume: with the generation and collection of masses of data, data scale becomes increasingly big.
  - Velocity: the timeliness of big data, specifically, data collection and analysis, etc. must be rapidly and timely conducted, so as to maximumly utilize the commercial value of big data;
  - Variety: the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data.
- Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis.[IDC: 4V Models)
  - Volume (great volume)
  - Variety (various modalities)
  - Velocity (rapid generation)
  - Value (huge value but very low density)
- Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizon- tal zoom technologies

3. Characteristics of BD (Chapter 1 of book)
   1. **What is meant by volume of BD. How has it changed over time?**
      - Volume means the size of the data being collected and stored.
      - The sheer volume of data being stored today is exploding. Countless applications and enterprises generate terabytes of data every hour and that number would keep increasing as long as techniques growing. So the volume would grow in a very fast speeding every day.
   2. **How has increased volume created a "blind zone" for organizations?**
      - Organizations are facing massive volumes of data nowadays, but organizations that don't know how to manage this data are overwhelmed by it.
      - As the amount of data available to the enterprise is on the rise, the percent of data it can process, understand, and analyze is on the decline, thereby creating the blind zone
   3. **What is meant by variety of BD? What are the various types of data that large organizations acquire today?**

- Variety means, variety represents all types of data-—a fundamental shift in analysis requirements from traditional structured data to include raw, semistructured, and unstructured data as part of the decision-making and insight process.
- The data includes not only traditional relational data, but also raw, semistructured, and unstructured data from web pages, web log files (including click-stream data), search indexes, social media forums, e-mail, documents, sensor data from active and passive systems, and so on.

4. **How is velocity of data applied to data in motion. What are the advantages of streams computing?**
    - The definition to data in motion: The speed at which the data is flowing.
    - Dealing effectively with Big Data requires that you perform analytics against the volume and variety of data while it is still in motion, not just after it is at rest. E.g. Consider examples from tracking neonatal health to financial markets: IBM define it as an inclusional shift from solely batch insight (Hadoop style) to batch insight combined with streaming-on-the-wire insight.
    - The advantages of streams computes:
        - getting an edge over your competition
        - be able to analyze the data with a very short shelf-life in near real time and find insights in this data
        - get continuously updated results

II. What is the value of Big Data

1. Section 1.3 of the paper and chapter 2 of the book
    1. **Read section 1.3 of the paper and chapter 2 of the book. They list several industries (e.g. US medical industry, retail industry, government operations, public health, etc) that can benefit enormously by using Big Data techniques. Choose any one such industry and do research about Big Data applications in that industry. Write a brief 2-3 paragraph report.**
        - Report for BD in public health:
        - Big public health data sets usually include one or more of (a) measures of participant biology, as in genomic or metabolomic data sets; (b) measures of participant context, as in geospatial analyses; (c) administratively collected medical record data that incorporate more participants than would be feasible in a study limited to primary data collection; (d) participant measurements taken automatically at extremely frequent intervals as by a GPS device; or (e) measures compiled from the data effluent created by life in an electronic world, such as search term records, social media postings, or cell phone

records.

- How can we improve the potential for Big Data to improve health and prevent disease? (a) One priority is that a stronger epidemiological foundation is needed. Big Data analysis is currently largely based on convenient samples of people or information available on the Internet. (b) There also must be a means to integrate knowledge that is based on a highly iterative process of interpreting what we know and don't know from within and across scientific disciplines. This requires knowledge management, knowledge synthesis, and knowledge translation. (c) Another important issue to address is that Big Data is a hypothesis–generating machine, but even after robust associations are established, evidence of health–related utility (i.e., assessing balance of health benefits versus harms) is still needed. Documenting the utility of genomics and Big Data information will necessitate the use of randomized clinical trials and other experimental designs. (d) As with genomics, an expanded translational research agenda (14) for Big Data is needed that goes beyond an initial research discovery.
- The combination of a strong epidemiologic foundation, robust knowledge integration, principles of evidence–based medicine, and an expanded translation research agenda can put Big Data on the right course.

III. Challenges of Big Data

1. Section 1.5 of the paper
    1. **Read section 1.5 of the paper and summarize in your own words the challenges of developing and managing Big Data applications.**
        - Data representation: Find an efficient data representation to make data more meaningful for computer analysis and user interpretation while keep a clean structure and the original value of data.
        - Redundancy reduction and data compression: Reduce the indirect cost of the system on the premise data of trivial value.
        - Data life cycle management: Due to the limitation(computing resource & disk space) of storage system, proper principles should be developed to decide how to deal with the continuously massive incoming (save or discard) for keeping the freshness of data.
        - Analytical mechanism: To find a compromising solution between RDBMSs and non–relational databases to meet the requirement of scalability, expandability, multi–usability and velocity.
        - Data confidentiality: Rely on a third party for processing the big data would increase the potential safety risk without proper preventive measures.

- Energy management: At the time of keeping expandability and accessibility of big data system, system-level energy consumption control and management mechanism are necessary.
- Expendability and scalability: The algorithms should be able to handle expanding datasets as the tasks grows complex in the future.
- Cooperation: The whole system should be able to utilize different kinds of data and user friendly by people in various vocations.

IV. Storage for Big Data

1. Section 4.2 of the paper
    1. **What factors should you take into account when using distributed storage for Big Data?**
        - Consistency: Multi server storage system may have server failures and inconsistency among different copies of the same data.
        - Availability: To make the entire system not be seriously affected to accomplish customer's requirement when server failures happen.
        - Partition Tolerance: The distributed storage system should be able to tolerate a certain level of network failures.
          Since all of the three factors could not be satisfied at the same time, we have CA, CP, AP to meet either two of them
2. Chapter 4 of the book
    1. Fill in the blanks / Short answer questions:
        1. **Hadoop is top level _____ project written in _____ programming language.**
            - Apache,Java
        2. **Hadoop was inspired by _____ .**
            - Google File System
        3. **Hadoop is different from transactional systems in the following ways:**
            - Unlike transactional systems, Hadoop is designed to scan through large data sets to produce its results through a highly scalable, distributed batch processing system.
            - Hadoop is not about speed-of-thought response times, real-time warehousing, or blazing transactional speeds; it is about discovery and making the once nearly impossible possible from a scalability and analysis perspective.
            - The Hadoop methodology is built around a function-to-data model as opposed to data-to-function
        4. **Two parts of Hadoop are:**

- a file system (the Hadoop Distributed File System)
- a programming paradigm (MapReduce)

5. **Why is redundancy built into Hadoop environment?**
    - Not only is the data redundantly stored in multiple places across the cluster, but the programming model is such that failures are expected and are resolved automatically by running portions of the program on various servers in the cluster. Due to the redundancy, it's possible to distribute the data and its associated programming across a very large cluster of commodity components. It is well known that commodity hardware components will fail, but this redundancy provides fault tolerance and a capability for the Hadoop cluster to heal itself. This allows Hadoop to scale out workloads across large clusters of inexpensive machines to work on Big Data problems.

2. Components of Hadoop:
    1. **The three pieces of Hadoop project are:**
        - Hadoop Distributed File System (HDFS)
        - the Hadoop MapReduce model
        - Hadoop Common
3. Hadoop Distributed File System:
    1. **How is it possible to scale Hadoop cluster to hundreds of nodes?**
        - Data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout the cluster. In this way, the map and reduce functions can be executed on smaller subsets of your larger data sets, and this provides the scalability that is needed for Big Data processing.
    2. **Each server in a Hadoop cluster uses _____ (inexpensive / expensive) disk drives.**
        - inexpensive
    3. **What is data locality. What does it achieve?**
        - Data Locality:MapReduce tries to assign workloads to these servers where the data to be processed is stored.
        - Map Reduce use this technique to achieve higher performance cause tasks could be ran in separate nodes paralleled (instead of the principle that using a storage area network (SAN), or network attached storage (NAS), Hadoop could avoid extra network communication overhead).
    4. **What are the benefits of breaking a file into blocks and storing these blocks with redundancy?**
        - That technique makes Hadoop has good built–in fault tolerance and fault compensation capabilities.

5. **The default size of a block in HDFS is _____ MB.**
    - 64
6. **What are the advantages of large block sizes in HDFS?**
    - For larger files, a higher block size is a good idea, as this will greatly reduce the amount of metadata required by the NameNode.
    - The expected workload is another consideration, as sequential access patterns (not random reads) will perform more optimally with a larger block size.
7. **What is a NameNode in HDFS? What are its functions?**
    - All of Hadoop's data placement logic is managed by a special server called NameNode.
    - This NameNode server keeps track of all the data files in HDFS, e.g. where the blocks are stored. All of the NameNode's information is stored in memory, which allows it to provide quick response times to storage manipulation or read requests.
8. **All of NameNode's information is stored in _____ (disk / memory).**
    - memory