CS 6350
ASSIGNMENT ____1_____

Names of students in your group:

Jinglin Li (jxl163530)

Number of free late days used: ____0_____

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

1. Book - Understanding Big Data by IBM
https://www.ibm.com/developerworks/vn/library/contest/dw-freebooks/Tim_Hieu_Big_Data/Understanding_BigData.PDF

2. Paper Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, *19*(2), 171-209.
http://mmlab.snu.ac.kr/~mchen/min_paper/BigDataSurvey2014.pdf
or from the ACM Digital Library http://dl.acm.org/citation.cfm?id=2843712

3. Big data analytics in healthcare: promise and potential, Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3

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. Big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values and incurs new challenge like how to effectively organize and manage such datasets.

2. What challenges have emerged because of the rise of BD?

   1. Collecting and integrating massive data from widely distributed data sources.
   2. 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.
   3. In consideration of the heterogeneity, scalability, realtime, 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.

3. 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.

   1. Datasets' volumes that conform to the standard of big data are changing, and may grow over time or with technological advances
   2. Datasets' volumes that conform to the standard of big data in different applications differ from each other.
   3. The increasingly growing data scale
   4. Its management that could not be handled by traditional database technology.
   5. 3V model: increase of Volume, Velocity and Variety.
      a) Volume: data scale becomes increasingly big
      b) Velocity: data collection and analysis must be rapidly and timely conducted/
      c) Variety: various types of data such as semi-structured and unstructured data
   6. Four Vs definition: one more V: Value (huge value but very low density)

4. What is meant by **volume** of BD. How has it changed over time.

   Volume is one of the characteristics of Big Data, which means the immense and massive volumes of data. The shear volume of data being stored today is

exploding. In the year 2000, 800,000 PB of data were stored in the world. We expect this number to reach 35 ZB by 2020.

5. How has **increased volume** created a "blind zone" for organizations?

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.

6. What is meant by **variety** of BD? What are the various types of data that large organizations acquire today?

Variety: With the explosion of sensors, and smart devices, as well as social collaboration technologies, data in an enterprise has become complex, because it includes not only traditional relational data, but also raw, semistructured, and unstructured data from web pages, web log files, search indexes, social media forums, email, documents, sensor data from active and passive systems.
Now, enterprises must be able to analyze all types of data, both relational and nonrelational: text, sensor data, audio, video, transactional and more.

7. How is **velocity** of data applied to data in motion. What are the advantages of **streams computing**?

The speed at which the data is flowing.
With streams computing, you can execute a process similar to a continuous query which related to the location. You will also get continuously updated results, because location information from GPS data is refreshed in real time.

8. 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.

Big Data in Healthcare

Big data is generating a lot of hype in every industry including healthcare. The healthcare industry has generated a large amount of data, driven by record keeping, compliance and regulatory requirements, and patient care. Some academic- or research-focused healthcare institutions are either experimenting with big data or using it in advanced research projects. Reports say data from the U.S. healthcare system alone reaches 150 Exabyte's in 2011.

By digitizing, combining and effectively using big data, healthcare organizations will benefit significantly. Potential benefits include but not limited to:

1. Detecting diseases at earlier stage when they can be treated more easily.
2. Managing specific individual and population health and detecting health care fraud quickly.
3. Developments or outcomes can be predicted and estimated based on vast amounts of historical data

McKinsey estimates that big data analytics enable more than $300 billion in savings per year in U.S. healthcare, two thirds of that through reductions of about 8% in national healthcare expenditures. Clinical operations and R & D are two of the largest areas for potential savings with &165 billion and &108 billion in waste respectively.

It is reported that big data could contribute to the following areas:

Clinical operations and R & D development

Big data could help to build a predictive modeling to lower attrition and produce a faster R & D pipeline in drugs and devices. It also has statistical tools and algorithms to improve clinical trial design and patient recruitment to better match treatments to individual patients, thus reducing trial failures and speeding new treatments to market. In addition, big data analysis enables analyzing clinical trials and patient records to predict, diagnose and treat patients.

Public health

1) Analyzing disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response
2) Faster development of more accurately targeted vaccines, e.g. choosing the annual influenza strains
3) Turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, especially for the benefits of populations.

Evidence-based medicine

Combine and analyze a variety of structured and unstructured data-EMRs, financial

and operational data, clinical data, and genomic data to match treatments with outcomes, predict patients at risk for disease or readmission and provide more efficient care.

Genomic Sequencing

With the development of Human Genome Project, the genomic data in patient becomes more and more important for prediction and treatment. The cost of sequencing an individual's full genome has plunged in recent years. Sequencing will become commonplace and eventually become a commodity lab test. Genomic sequence are huge files and the analysis of genomes generates even more data. Genomics can make advantage of big data. Loading a genetic sequence into a relational database would require a huge Character Large Object (CLOB) or a separate storage just to manage the sequence. With big data, just toss it in the Hadoop cluster, and it is ready for analysis.

Device/remote monitoring

Capture and analyze in real-time large volumes of fast-moving data from in-hospital and in-home devices, for safety monitoring and adverse event prediction

Patient profile analytics

Apply advanced analytics to patient profiles (e.g. , segmentation and predictive modeling) to identify individuals who would benefit from proactive care or lifestyle changes, for example, those patients at risk of developing a specific disease (e.g., diabetes) who sould benefit from preventive case.

Many U.S. healthcare institutions have benefited from the development of big data. For example, Columbia University Medical Center analyze "complex correlations" of streams of physiological data related to patients with brain injuries. The goal is to provide medical professionals with critical and timely information to aggressively treat complications. The advanced analytics is reported to diagnose serious complications as much as 48 hours sooner than previously in patients who have suffered a bleeding struck from a ruptured brain aneurysm. The University of Michigan Health System standardized the administration of blood transfusions with big data analytics research. This resulted in a 31% reduction in transfusions and $200,000 reduction in expenses per month. Another example of big data analytics in healthcare is the Department of Veterans Affairs' (VA) use of applications on its very large dataset in an effort to comply with "performance-based accountability framework and disease management practice". In one very famous example, California-based Kaise Permanente associated clinical data with cost data to generate a key data set, the analytics of which lead to the discovery of adverse drug effects and subsequent withdrawal of Vioxx from the market. Researchers at the Johns Hopkins School of Medicine discovered that they could use data from Google Flu Trends to predict sudden increases in flu-related emergency room visits at least a week before warnings from the CDC. Also reported is an application developed by

IBM that predicts the likely outcomes of diabetes patients using patients' panel data linked to physicians, management protocols and the overall relationship to population health management averages.

In the future we'll see the rapid, widespread implementation and use of big data analytics across the healthcare organization and the healthcare industry. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will be taken into consideration.

9. Read section 1.5 of the paper and summarize in your own words the challenges of developing and managing Big Data applications.

1. Data representation: Data representation is very important for the effective data analysis. However, datasets are in variety of types including both structure and unstructured dataset, like audio, media, and so on. Therefore, efficient data representation enable efficient operations on different datasets.
2. Redundancy reduction and data compression: It is effective to reduce the indirect cost of the entire system because the potential values of data are maintained.
3. Data life cycle management: Because the values of data depend on data freshness, a data importance principle should be developed to filter the valuable data.
4. Analytical mechanism: the analysis of big data process masses of data with different types. Non-relational databases have advantages in processing of unstructured data instead of traditional RDBMSs. However, we should find a compromising solution between RDBMSs and non-relational databases.
5. Data confidentiality: big data service providers could not effectively maintain and analyze huge datasets due to limited capacity.
6. Energy management: The increase of data volume, processing, storage and so on consume more and more electric energy.
7. Expendability and scalability: The analytical algorithm must be able to process increasingly expanding and more complex datasets.
8. Cooperation: experts in different fields are needed to cooperate for analysis of big data because it is an interdisciplinary research. So the big data network architecture mush be established.

10. What factors should you take into account when using distributed storage for Big Data?

1. Consistency
2. Availability
3. Partition Tolerance

Fill in the blanks / Short answer questions:

1. Hadoop is top level __Apache__ project written in __Java__ programming language.
2. Hadoop was inspired by __Google's work on its Google(distributed) File System (GFS) and the MapReduce programming paradigm__ .
3. Hadoop is different from transactional systems in the following ways:

   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-though response times, real-time warehousing, or blazing transactional speeds; it is about discovery and making the once near-impossible possible from a scalability and analysis perspective.

4. Two parts of Hadoop are:
   A file system (the Hadoop Distributed File System) and a programming paradigm (MapReduce).

5. Why is redundancy built into Hadoop environment?

   It helps 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.

6. The three pieces of Hadoop project are:

   Hadoop Distributed File System (HDFS), the Hadoop MapReduce model, and Hadoop Common.

7. How is it possible to scale Hadoop cluster to hundreds of nodes?

   Data in a Hadoop cluster is broken down into smaller pieces (blocks) and distributed throughout the cluster. The map and reduce functions can be executed on smaller subsets of your larger datasets, and this provides the scalability that is needed for Big Data processing.

8. Each server in a Hadoop cluster uses ___inexpensive_____ (inexpensive / expensive) disk drives

9. What is data locality. What does it achieve?
   Data locality is the location where the data to be processed is stored. It achieves the data which is to be processed.

10. What are the benefits of breaking a file into blocks and storing these blocks with redundancy?

    To achieve availability as components fail. In addition, this redundancy allows the Hadoop cluster to break work up into smaller chunks and run those jobs on all the servers in the cluster for better scalability. You can also get the benefit of data locality, which is critical when working with large data sets.

11. The default size of a block in HDFS is ____64_____ MB.

12. What are the advantages of large block sizes in HDFS?
    It greatly reduce the amount of metadata required by the NameNode.

13. What is a NameNode in HDFS? What are its functions?
    NameNode is a special server which keeps track of all the data files in HDFS, such as where the blocks are stored and more.

14. All of NameNode's information is stored in____memory_____ (disk / memory).