CSIE59830/CSIEM0410 Big Data Systems, Spring 2021
Final Exam

(This is a limited time open-book, open-Web take-home exam. Upon receiving the exam file, edit the file with your answers, convert it into a PDF file, and send the PDF file to me through email by 12:30pm. You are allow to refer to class notes, textbooks, or even search the Web. However, you have to answer all questions **on your own** without discussion with anyone else.)

ID: ___________________ Dept:_________________ Name: ________________

1. **(15%)** For each description of the problem, specify the big data system/tool(s) which is(are) best for solving the problem and **briefly** explain your choice.

Description of the problem to be solved	**Which big data system/tool(s) to use and why?**
The CDC (Taiwan Centers for Disease Control, 衛生福利部疾病管制署) wants to provide a new service of reflecting the current status of the COVID-19 pandemic by combining the coronavirus updates sent by all hospitals in real-time.	
The 7-11 headquarter wants to analyze seasonal trends in product sales for all 7-11 retail stores in Taiwan to make better predictions on future demand and improve inventory management.	
The United Daily News(聯合報) wants to convert all of its past newspaper articles into a scalable database that can be queried and searched.	
The NDHU Environment Protection Project wants to provide real-time analytics of air quality data from IoT sensors installed around the campus and neighboring areas.	
The Federal Bureau of Investigation (FBI) needs to analyze user relationships and group activities on Twitter to uncover potential terrorist organizations.	

2. **(15%)** Answer the following questions **briefly** and to the point.

(a) What is lineage in Spark? Where and when do we use it? Why is it useful?

(b) What is (are) the main reason(s) why Spark is faster than Hadoop MapReduce on iterative jobs?

(c) Why structured streaming is considered more advanced than Spark streaming?

3. **(20%)** Consider the problem of footprint tracking and contact tracing of COVID-19 confirmed cases. Assume that we have N locations, M users and a temporal threshold T. Each user is equipped with an APP to automatically detect and upload all relevant events. More specifically, when a user visits a location or meets with another user, an event is recorded and uploaded to the CDC database with time, location ID, and user ID(s). Upon the confirmation of a particular case, we want to print out the travel history of the case in the past two weeks as well as a list of users that may have been exposed to the coronavirus. The travel history is a list of locations ordered by visit time. The list of users include all users who have direct contact with the confirmed case or visit the same location as the confirmed case within the threshold T. Both are for the past two weeks.
   (a) Propose a big data framework that is best for solving the problem. Explain your choice.
   (b) Select an appropriate tool within the framework to solve the problem.
   (c) Write program segments to generate the travel history and list of users with the tool of your choice.

(* Design your own input/output formats. Explain it clearly. *)

(Empty page for answering the question)

4. **(20%)** Consider the problem of analyzing YouTube user behavior. Assume that you have a list of YouTubers each having one or more channels. Also given is a list of Subscribers each may subscribes to zero or more channels. Solve the following problems with **Neo4j**.

   (a) Briefly describe how would you represent the problem with Neo4j.

   (b) Define and construct the YouTuber-Subscriber database with the Cypher query language.

   (c) Write a query to print out the list of top 50 most-subscribed YouTube channels.

   (d) Write a query to print out the list of top 50 most-subscribed YouTubers based on the sum of subscribers of all channels of the same YouTuber without consideration of the overlap of subscribers among different channels.

(* Design your own input/output formats. Explain it clearly. *)

(Empty page for answering the question)

5. **(20%)** Answer the following questions about HBase.

   (a) Summarize the data model of HBase.

   (b) HBase was built on top of other Apache technologies. What are these technologies?

   (c) What are the fault tolerance mechanisms built into the HBase system? Briefly describe the main purposes of each mechanism.

   (d) Briefly describe the role of ZooKeeper in HBase.

6. **(20%)** Briefly answer the following questions about NoSQL, NewSQL and distributed SQL.

(a) What are the differences between NewSQL and NoSQL?

(b) What is eventual consistency? What guarantees do we get from eventual consistency? When and why do we use it?

(c) What are the purposes of using consistent hashing for partitioning in DynamoDB?

(d) Given an existing set of configuration values for sloppy quorum, how do you modify the value(s) to increase the availability? How about increasing consistency? Can we get both of them?

(e) Give MongoDB statements to create a student collection with three student documents.

(f) What is the main technology that enables Google Spanner to be a Distributed SQL system? How is the technology used in Spanner? Why is it important?

(Empty page for answering the question)