LOG8430E - Revision - Big Data

1. We develop an application for a commercial company. The company manages contracts and invoices of commercial transactions. The data need to be stored in a secure and persistent way for a long time. The company functions at a global level with a great quantity of transactions. What data model would you choose and why?

- ◯ Key-value, because the invoices should be retrieved by their key
- ◯ Document, because invoices are files and can be searched by content
- ◯ Column-wide, because invoices can have variable structure.

2. Wikipedia is an online encyclopedia where each entry is a page and it can contain links to other entries. Wikipedia also organizes its entries in categories and subcategories to facilitate search. What data model would you choose to implement this system and why?

- ◯ Key-value, because the structure of each page can be different.
- ◯ Document, because every page is an entry and we can search by its content.

- ◯ Graph, because the pages are connected to each other with complex relationships.
- ◯ Wide-column, because this is the closest model to a relational database.

3. We have a monitoring system for cloud resources. In frequent and fast enough intervals, the system sends measurements (CPU, memory, disk, network) for each resource (virtual machine). What data model would you choose and why?

- ◯ Document, measurements for every resource are stored in a file.
- ◯ Key-value, the data is small and it is indexed by resource
- ◯ Wide-column, to account for missing values

4. Databases for biological data have existed for quite some time now. They contain data for genes, proteins, organisms. The entities have attributes, but it is possible to discover new attributes in the future. Various analyses and tools already exist to help us study the natural world. What data model would you use and why?

- ◯ Relational database, every type is a table and the relationships are modeled with foreign keys
- ◯ Graph, to capture the complex relationships between the entities

- ◯ Wide-column, to account for existing tools and new attributes.

5. We want to develop a program to count the words of thousands of documents organized in folders with the first letter of the file's name ("A", "B", "C"...). Describe the MapReduce implementation for this problem (number of workers, map tasks, reduce tasks)

- ◯ One worker per letter, map task counts per document, reduce task aggregates for all documents
- ◯ One worker per letter, map task counts per letter, reduce task aggregates for all letters
- ◯ Assign workers according to word distribution for English letters, map task counts per worker, reduce task aggregates for all workers.

6. We have the CRA database for income taxes. The CRA has multiple data servers around the country. We want to find who paid the most taxes in 2018. Describe the MapReduce implementation for this problem (number of workers, map and reduce tasks).

- ◯ One worker per province, map finds max per province, reduce finds max from all provinces
- ◯ One worker per province, map sorts tax returns per province, reduce sorts tax returns for all provinces
- ◯ Distribute workers according to population, map finds max per worker, reduce finds max among all

workers.

7. We want to find the player with the most average goals per game in the entire history of NHL. Each team keep record of their own statistics. (Do not consider teams that do not exist now). Describe the MapReduce implementation for this problem (number of workers, map and reduce tasks).

- ◯ One worker per team, map finds averages per team, reduce finds max for all players
- ◯ One worker per team, map finds max average per team, reduce finds max for all players
- ◯ Assign workers according to how old a team is, max finds max average per team, reduce finds max for all players.

8. Each University holds all defended theses in their respective library. We want to search in the entire world for theses that contain the term “DevOps” in their title. Describe the MapReduce implementation for this problem (number of workers, map and reduce tasks).

- ◯ One worker per university library, map finds all DevOps titles, reduce aggregates results in a list
- ◯ One worker per university, map finds all DevOps titles, reduce aggregates results in a list
- ◯ One worker per university, map finds all DevOps titles, reduce finds the thesis with most citations.

9. We want a complete lists of all cities, villages and communities of the world order by population in descending order. Census servers in each country contain population data. Describe the MapReduce implementation for this problem (number of workers, map and reduce tasks).

- ◯ One worker per country, map sorts cities per country, reduce aggregates all cities in a sorted list
- ◯ One worker per census server, map sorts cities per census server, reduce aggregates all cities in a sorted list

Done