Big Data Management (BDMA & MIRI Masters)

Final Exam

Full name: _____

1. (20%) Imagine you want to design a system to maintain data about citizens and doubt whether to use HBase or MongoDB. Precisely, for each citizen, it will store: their personal data (with pID), their city data (with cID) and their employment data. We also know the workload (i.e., queries and frequency of execution) is as follows:

   - Q_1: Average salary in Barcelona (50% frequency) – information obtained from the set of city and employment data.
   - Q_2: Average weight for young people (less than 18 years old) (45% frequency) – information obtained from the set of personal data.
   - Q_3: Number of VIP persons (5% frequency) – information obtained from the set of personal data.

   Discuss your choice of technology and data model. Clearly specify the structure of the data (i.e., tables/collections, keys, values, etc.), trade-offs and assumptions made.

Solution:

2. (20%) In relational algebra, the antijoin operator ($\triangleright$) is defined as the complement of the semijoin on the primary keys (PKs). Formally, assuming A and B are the PKs of R and S, respectively, then $R \triangleright S = R \setminus R \ltimes_{A=B} S$. Provide the MapReduce pseudo-code implementation of the antijoin operator. Assume the existence of the operator $\oplus$ to concatenate strings, $prj_{att}(s)$ to project attribute att from the tuple s, and $input(s)$ to decide the origin (i.e., R or S) from a tuple s.

Solution:

3. (30%) Consider three files containing the following kinds of data:

   Employees.txt
   EMP1,CARME,400000,MATARO,DEPT1,PROJ1
   EMP2,EULALIA,150000,BARCELONA,DEPT2,PROJ1
   EMP3,MIQUEL,125000,BADALONA,DEPT1,PROJ3

   Projects.txt
   PROJ1,IBDTEL,TV,1000000
   PROJ2,IBDVID,VIDEO,500000
   PROJ3,IBDTEF,TELEPHONE,200000
   PROJ4,IBDCOM,COMMUNICATIONS,2000000

   Departments.txt
   DEPT1,MANAGEMENT,10,PAU CLARIS,BARCELONA
   DEPT2,MANAGEMENT,8,RIOS ROSAS,MADRID
   DEPT4,MARKETING,3,RIOS ROSAS,MADRID

   Provide the ordered list of Spark operations (no need to follow the exact syntax, but just the kind of operation and main parameters) you would need to obtain the departments with at least one employee which have all of their employees assigned to the same project. The result must include department number. Save the results in "output.txt". In the previous example, the result should be "DEPT2".

Solution:

4. (15%) Assume we ingest a stream with an event every time a ticket is sold at a theater. Precisely, the stream has the structure $(movieID, theaterID, timestamp, price)$. Next, we ingest the following ordered set of events:

   - Event 1: $(m_3, t_4, 12h, 10\$)$
   - Event 2: $(m_1, t_2, 13h, 17\$)$
   - Event 3: $(m_2, t_4, 14h, 11\$)$
   - Event 4: $(m_4, t_1, 15h, 8\$)$
   - Event 5: $(m_1, t_3, 16h, 9\$)$
   - Event 6: $(m_3, t_4, 17h, 5\$)$
   - Event 7: $(m_6, t_1, 18h, 15\$)$
   - Event 8: $(m_5, t_2, 19h, 12\$)$
   - Event 9: $(m_7, t_5, 20h, 17\$)$
   - Event 10: $(m_1, t_1, 21h, 11\$)$

Which theaters would be considered heavy hitters (using the approximate method) considering a required frequency of 33%? Provide a detailed answer (i.e., describe the process).

Solution:

5. (15%) What problem do Data Lakes solve?

Solution: