Tommaso Carraro
student number: 015147592

Big Data Management: Homework 5

1. B
2. C
3. C
4. A
5. C
6. D
7. D
8. B, C, E, F
9. D
10. B
11. In CAP, the term consistency refers to the consistency of the values in different copies of the same data item in a replicated distributed system. This means that all the nodes that have a copy of a replicated data item should have the same copy of that data item, in fact a replicated data item should have the same values of all the other copies, even if they are stored in remote nodes. This means that if a copy of that specific data item has been modified in a node it should be modified even in the other nodes where it is present. In ACID, the term consistency refers to the fact that a transaction will not violate the integrity constraints specified on the database schema. This means that if the database schema is in a consistent state (integrity constraints are respect), after the execution of a transaction the consistency must be preserved.
12. A data lake is a system or repository of data stored in its raw format. It is usually a single store of all enterprise data. A data lake allows to read data based on different data models because it uses the schema-on-read approach. Since it uses the schema-on-read approach it is more flexible and less expensive to manage than the data warehouse.

    A data warehouse is a central repository of integrated data from one or more disparate sources. The main characteristic of the data warehouse is that it uses a schema-on-write approach and this means that data have to follow a specific structure that has to be defined before the data storing. It is obviously less flexible and more expensive for the big data respect to the data lake because it has a fixed structure.

    Finally, the data mart is a subset of the data warehouse and it is usually oriented to a specific business line or team. This means that it is possible to subdivide the data warehouse into multiple data marts, where the information are better organized.