Quiz 5 Answers

1. [12pts] For the following questions, clearly circle True or False.

1. Shared-disk parallelism is currently the dominant paradigm of parallel DBMSs.
   **FALSE**

2. It is possible to perform SQL-style joins with MapReduce/Hadoop.
   **TRUE**

3. Data shuffle among workers is a key component of many operator implementations in parallel DBMSs, MapReduce/Hadoop and Spark.
   **TRUE**

4. It is impossible to add fault tolerance to parallel query processing in a parallel DBMS.
   **FALSE**

5. It is impossible to compute VARIANCE as a distributive aggregate similar to AVG and SUM in a parallel DBMS.
   **FALSE**

6. It is impossible to obtain superlinear speedups in the strong scaling setting.
   **FALSE**

2. [9pts] : How does the 2-phase algorithm for parallel hash join improve upon the basic algorithm?

**ANSWER:** Check class notes.

3. [9pts] : Give a concrete example of an analytics task (clearly describe the dataset and the “query”) for which Spark or MapReduce/Hadoop would offer better developability than a parallel DBMSs.

**ANSWER:** Check class notes.