BDCC 19/20
Worksheet #1
Parallel Databases x Map-Reduce
April 22nd, 2020

Based on the following papers:

- [Pavlo et al., SIGMOD 2009] A Comparison of Approaches to Large-Scale Data Analysis
- [Dean and Ghemawat, CACM 2010] MapReduce: A Flexible Data Processing Tool
- [Stonebraker et al., 2010] mapReduce and Parallel DBs: friends or foes?

Answer the following questions:

1) Does it make sense comparing map-reduce approaches with parallel and distributed database systems? Explain your answer.

According to [Stonebraker et al., 2010] map-reduce and Parallel Dbs are two different paradigms that can be complementary.

2) What are the advantages of map-reduce over parallel and distributed databases?

Map-reduce is a programming paradigm. As such, it provides more flexibility for implementation.

3) What are the advantages of parallel and distributed databases over map-reduce?

It provides a tool for data management. The programmer does not need to worry about data organization, partition etc.

4) What kind of operations are allowed in parallel and distributed databases that are not available “out-of-the-box” in map-reduce?

Queries, query optimization, aggregation operators.

5) What kind of operations are allowed in map-reduce that are not available “out-of-the-box” in parallel and distributed databases?

Parallelization constructs, flexible programming constructs, more general purpose programming language.

6) Are there alternatives to Google’s mapreduce? How do they perform?

Hadoop, Spark, Flink, among other tools and languages that provide the map-reduce programming paradigm. (the idea here was that you would perform a search for these kinds of systems and compare performance)

7) Search the web for parallel and distributed databases. In which situations would be interesting to use such solutions instead of using google’s mapreduce solution?

There are various situations where using only a parallel and distributed database would suffice, most of them would be related with SQL-related searches. For general programming, DDBs are not sufficient.