INF 550: Data Science at Scale

Homework 6, Fall 2019

Due: Dec. 3, 2019. 12:00 PM (Noon)

Briefly explain the followings:

1. What is the main purpose of Name Node in HDFS? How does a Name Node try to avoid a single point of failure issue?

2. What is speculative execution (also called backup tasks)? What problem does it solve?

3. You have a web log file (140MB) and want to store it in HDFS. Your Hadoop system has 10 slave nodes. Illustrate how this file is stored in this system. And explain the steps how this file can be accessed later.

4. What is the role of Jobtracker in MapReduce?

5. Explain how HBase distributes its table over multiple machines.

6. What is sparse data and how HBase is handling sparse data?