Data Science Toolbox Question Sheet

11.1 Parallel Infrastructure

Daniel Lawson

Block 11

1. Show that a streaming algorithm for the standard deviation $\hat{s}_n^2 = d_n^2/(n-1)$ where $d_n^2 = d_{n-1}^2 + (x_n - \overline{x}_n)(x_n - \overline{x}_{n-1})$ follows from the definition of the standard deviation, $\hat{s}_n^2 = \sum_{i=1}^n (x_i - \overline{x}_n)^2/(n-1)$.
2. In big data, people often talk about Volume, Velocity, Variety, and Veracity. Define each and explain the circumstances under which they might be a problem.
3. Describe the HDFS file system at a high level, being sure to give the role of the namenode and datanodes. How does it tolerate faults?
4. Describe the Map/Sort/Reduce framework and explain the importance of each of these steps, as well as the role of the keys in each.
5. Explain how parallelism in the map stage is achieved and describe any limitations.
6. Give two circumstances in which the reduce step is inefficient. How can each be prevented?
7. Explain what an **immutable** data object is, and why it is used in spark.
8. Explain what a transformation is, and how it can lead to efficient parallel computation in spark.