Project Proposal:
Performance Analysis of Hash-Functions and Record Reader

In this project, we will investigate on the performance contribution of hash functions and record reader. The results will be gathered for both Hadoop and Spark and compared thereafter. The preliminary proposed project consists of four parts:

1. Evaluating how big the performance impact of a bad hashing algorithm is
2. Evaluating the record reader and its impact on the overall execution time
3. Evaluating hash functions and how well they perform for different types of data
4. Compare the found results between Hadoop and Spark

1. The Importance of Hashing

To evaluate the impact of hashing we will create custom hash functions. Those hash functions will be varying in their effectiveness and thus perform better or worse. I.e. a bad one will assign the same hash value to all keys whereas a good one will give a distinctive hash value for each different key. This change of precision will be gradually changed and the resulting performance curve analyzed.

2. Impact of the Record Reader

This part will elaborate on the record reader and its performance impact. The goal is to identify how different record reader impact the subsequent performance. A possible parameter to tweak is the data size passed on in a single iteration (i.e. multiple lines of text instead of one).

3. Effectiveness of Hash Functions

In this section of the project different classes of keys will be identified (i.e. numerical key, strings). Thereafter, multiple, commonly used hashing functions will be used for the shuffling stage and the resulting performance will be analyzed.

4. Comparison between Hadoop and Spark

The last part of the project will deal with possible differences in the found performance impacts between Hadoop and Spark. For this purpose, the experiments are implemented and run in both Hadoop and Spark and their results are compared.

Project Team: XTJ

Johannes Vamos	johannes.vamos@rutgers.edu
Tong Wu	tw445@rutgers.edu
Xin Yang	xin.yang@rutgers.edu