Hadoopにおけるオートパラメータチューニング手法に向けて

Toward an Automatic Parameter Tuning Method for Hadoop

中下 和則 †　　**小板 隆浩** ‡
Kazunori Nakashita　Takahiro Koita

1　はじめに

近年，情報爆発時代に必要とされている新たなインフラとして，大規模分散処理に適したHadoop[1]が注目されている．Hadoopとは，大規模なデータを効率的に分散して処理できるMapReduce実装のプラットフォームである．Hadoopにはパフォーマンスに影響を与えるパラメータが190個以上存在し，これらのパラメータを適当に特定することが難しい．理由は，パラメータの数が多いため効果のあるパラメータを決めることが困難なこと，またパラメータはアプリケーションの特性によって効果が決まることが挙げられる．仮に，あるアプリケーションに対して適当なパラメータ設定を行い，短時間で，効率よく実行できたとする．そこで違う特性を持つ新たなアプリケーションに対し，同じパラメータを用いても同じような結果を得られるとは限らない．アプリケーションの特性が違うためである．

本研究では，Hadoopのパラメータ値をアプリケーションの実行前にかつ自動的に設定するオートパラメータチューニング手法の実現を目的とする．オートパラメータチューニング手法を目指すにあたって，パラメータを効果のあるパラメータ群と効果のないパラメータ群に分けることとアプリケーションの特性を調査することが重要となる．パラメータチューニングにおいて，アプリケーションの特性が予めわからないため，多数のアプリケーションを調査し，特性をある程度推測することにより，過去の適当なパラメータセットを利用することが可能である．

2　Hadoopのパラメータ

ある特定のアプリケーションに対し，Hadoopを適用し，オートパラメータチューニング手法を設計するための予備評価実験を行う．

本章では，有効なHadoopのパラメータを検討するための方法について議論する．Duke大学のShivnath Babuら[2]は，Hadoopの全パラメータの中の25個がパフォーマンスに大きな影響を与えるとする結果を発表した．ここでのパフォーマンスとは，処理に必要な時間である．これにより，全パラメータの中の25個を対象とする．以下にBabuらがパフォーマンスに影響を与えるとしたその内のいくつかのパラメータを示す．

- mapred.map.tasks
- mapred.reduced.tasks
- io.sort.factor
- io.file.buffer.size
- mapred.job.shuffle.input.buffer.percent

例えば，mapred.map.tasksやmapred.reduce.tasksについて，パフォーマンスへの影響を考える．これらはmapやreduceのタスクの数を指定するためのパラメータである．map，reduceのタスクの数はデフォルトで，それぞれ2個と1個である．実際にデータ処理を行うマシンの台数を増やしても，これらのパラメータによるタスクの制限が存在するため，用意したマシンの台数分の役割を果たしていない．従って，mapred.map.tasksやmapred.reduced.tasksの値はパフォーマンスに与える影響が大きい．

また，io.sort.factorやio.file.buffer.sizeがパフォーマンスに与える影響について述べる．MapReduceは，各reducerへの入力がキーでソートされている．このソートプロセスはシャッフルと呼ばれている．これらのパラメータはこのシャッフル過程で必要になるパラメータである．mapが出力を始める際，つまりmapからシャッフルに移る際，その出力過程は複雑である．単純にディスクに書き込まれるわけではなく，メモリへの書き込みのバッファリングなどのために，前もって多少のソートが行われる．シャッフル過程のパラメータの値を変化させることで，パフォーマンスに大きな影響を与える．

3　予備実験

予備評価実験の概要について述べる．本研究では，Amazon Web Services[3]の一つであるAmazonEC2を用いた分散データ処理を対象とする．AmazonEC2ではインスタンスの数を自由に設定することができる．使用するアプリケーションは，Hadoopに添付されているサンプルアプリケーションである．実験の手順を説明する．インスタンスの数は1,2,4,8とする．前述のパラメータを実際に変化させ，処理に必要な時間を調べる．実験結果から，これらのパラメータが実際に有効なパラメータであるのかを確認する．

4　まとめと今後の課題

本稿では，Hadoopを新しいアプリケーションに適用した際に，短時間で，効率よく実行させるため，オートパラメータチューニング手法の実現を目指すことを示した．今後は，違う特性を持つアプリケーションをいくつか用意し，本稿で得たパラメータに関する知見を元にアプリケーションの特性に基づいた指標を作成することを課題とする．

参考文献

[1] Apache hadoop. http://hadoop.apache.org/.
[2] Shivnath Babu: Towards Automatic Optimization of MapReduce Programs, Proc of the SoCC'10, pp.137- pp.142 , 2010.
[3] Amazon Web Services. http://aws.amazon.com/.

† 同志社大学大学院 工学研究科 情報工学専攻
‡ 同志社大学 理工学部 情報システムデザイン学科