並列データ解析処理基盤のI/O性能評価に関する一考察

山田　浩之†　　合田　和生†　　喜連川　優†

†東京大学生産技術研究所

1　はじめに

企業が保持する情報は大規模化しつつあり、データ解析による大規模データからの価値創出への期待は高い。企業における大規模データ解析は、大規模ストレージに接続されたハイエンドサーバ等の集約ステムにより従来から行われてきた。他方、近年では多数のコモディティサーバ、OS、DBMSを一体化したデータウェアハウス・アプライアンスの登場により、並列処理による解析が行われ始めている。データウェアハウス・アプライアンスは、多数のコモディティサーバを利用した並列データベースアーキテクチャを採用し、構造化データ処理に対して優位であることを特徴としている。データ解析技術としては1990年代までの並列データベース技術を応用したものに過ぎないと考えられるが、データ解析システムを安価にパッケージ化することで、大規模データ解析が一般企業へと普及する大きなきっかけとなった。

並列処理による大規模データ解析が普及しているなか、並列データ解析処理基盤のHadoop[1]により、この流れは加速しつつある。HadoopはGoogle社のMapReduce[2]のオープンソース実装であり、非構造化データ処理の柔軟な扱いを特徴としている。また、YahooやFacebookでは数千ノードでの使用実績があるなど、スケーラビリティに対しても優位性が高い。さらに、Cloudera等のディストリビューションにより、その導入や管理コストも大きく低下していることから、ウェブ企業でのデータ解析におけるデファクトスタンダードとなりつつある。各社データウェアハウスベンダにおいても、独自のデータ解析処理基盤とHadoopとの連携が進められていることもあり、エンタープライズ分野においてもその活用が進んでいる。

近年のHadoopによる大規模データ解析においては、I/O帯域を確保するために、各ノードに多数のディスクドライブを搭載する流れが進んでいる[3, 4]。しかし、現状のHadoopは必ずしもI/O処理能力が高くない[5]。本稿では、TPC-Hデータセットを用いた並列解析処理基盤HadoopのI/O性能評価を行うことにより、HadoopのI/O処理能力を検証するとともに、その実行アーキテクチャについて考察する。

An Experimental Study on I/O Performance in Parallel Data Processing Platform
†Hiroyuki YAMADA　†Kazuo GODA †Masaru Kitsuregawa
†Institute of Industrial Science, The University of Tokyo

2　実験環境

実験に用いたハードウェア構成を以下に述べる。Hadoop用サーバとしてIBM System x3850 M2を1台と、データ用ストレージとしてAMS500を用いた。当該サーバは、Intel Xeon X7350を4ソケット、メモリを48GB(DDR2 ECC SDRAM RDIMM)、ローカルHDDとしてSASディスク(10Krpm)を4台搭載している。ローカルHDDはOSとHadoopの一部の管理データの保存用に用いた。また、ストレージは32台のFCディスク(10Krpm)を使用した。2台のFCディスクをLU(RAID0)とし、16LUを構築した。サーバ･ストレージ間は4本の4Gbpsファイバチャネルにより接続されている。

次に、実験に用いたソフトウェア構成を以下に述べる。OSはLinux(2.6.18)、Hadoopはバージョン0.20.2を使用した。HDFSのブロックサイズは128MB、I/Oのバッファサイズ(io.file.buffer.size)は32MBとし、MapReduceにおけるMapタスク数を8とした。実験においてはソフトウェアのレベルで複数のLUを扱う必要があるが、今回はLVMによるストライピングを用いた。

実験データはTPC-Hを用いた。スケールファクタを100とし、LINEITEM表(約75GB)とPART表(約2.3GB)を使用した。実験クエリは、以下の2つの問い合わせを実施した。

- Job1: 1Mapジョブ(map処理は空)
- Job2: 1Mapジョブ(map処理は選択処理)

Job1は、LINEITEM表を入力とし、map処理は何も行わない。つまり、Hadoopのフレームワークを介してLINEITEM表を全スキャンするのみの処理となる。Job2は、LINEITEM表を入力とし、map処理はshipdateカラムに対する選択処理(選択率1%)を行う。選択結果のHDFSへの書き込みは行わない。

3　実験と考察

実験は、1LU, 2LU, 4LU, 8LU, 16LUの5つのディスク構成に対して実施した。図1には、Job1、Job2に対する、各ディスク構成における読み取りのディスク転送レートを示している。比較として、LVMによる論理

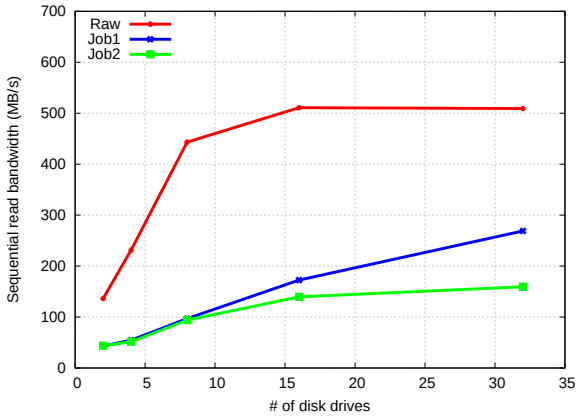

図 1: Job1, Job2 におけるディスク転送レート

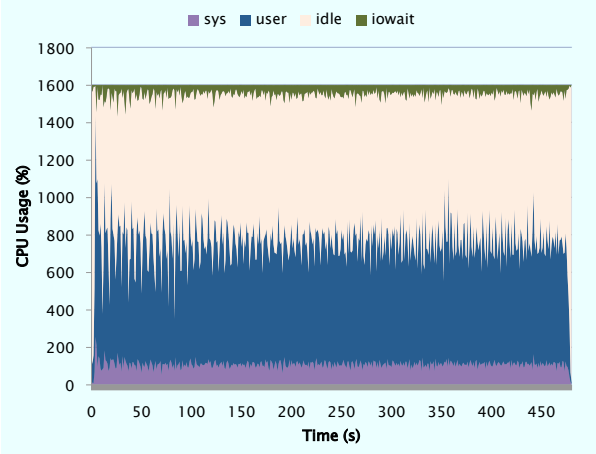

図 2: 32 台のディスクドライブ使用時の Job2 における CPU 使用率の内訳

ディスクドライブを Raw デバイスとして扱った場合の、単一ストリームによる読み取りディスク転送レートを示している。Raw によるディスク転送レートがディスクドライブ 16 台以上で飽和しているのは、ストレージ側のコントローラの性能によるものである。全体として、Hadoop の実行におけるディスク転送レートは、ディスクドライブ本来の性能よりも大幅に下回った結果となっていることがわかる。これは、Hadoop の I/O 処理に伴う CPU 負荷が高いため、同期的な読み取り I/O では十分な I/O を発行できていないことが原因だと考えられる。実行トレースから得られた CPU 使用率を見ると、ディスクを 2 台使用した場合で平均約 160%*前後となり、I/O 処理量の増加に伴い CPU 使用率も増加し、32 台のディスク使用時で平均約 700%前後となっている。つまり、16 台以降でほぼ性能が飽和しているのは 8 つの Map タスクが各々1 つの CPU コアを 100%近くまで使用していることが原因であると考えられる†。図 2 に 32 台のディスク使用時の Job2 の CPU 利用率の内訳を示す。Hadoop の I/O 処理の CPU 使用率が高い原因は、汎用的な入力機構や Java による実装、実装の洗練度が低いことなどが考えられるが、詳細な原因の検証は別稿に譲りたい。

実験の結果、現状の Hadoop は必ずしもディスクドライブがもつ I/O 帯域を有効に活用できないという問題を確認した。また、図 1,2 からもわかるように、CPU リソース及び I/O 帯域に余裕がある状況下においても性能が飽和してしまうという問題があることが明らかになった。

*16CPU コアを最大限に使用する場合を 1600%とした時の割合となる

†Map 処理はブロックごとにプロセスの起動を行うため、各 Map タスクが使用する CPU コアは固定されない。よって、8Map タスクで全体の半分程度の CPU 使用率を使っていることはこのように解釈される。

4 まとめ

近年の並列データ解析処理基盤による大規模データ解析においては、I/O 帯域を確保するために、各ノードに多数のディスクドライブを搭載する流れが進んでいる。本稿では、そのような背景を鑑み、TPC-H データセットを用いた Hadoop の I/O 性能評価を実施した。実験の結果、現状のアーキテクチャでは CPU リソース及び I/O 帯域に空きがあるにも拘わらず性能が飽和してしまうという問題がある。多数の CPU コアを最大限に活用するようにデータ処理アーキテクチャの改変が求められ、今後、当該問題の解決に向けて取り組んでいきたい。

参考文献

[1] Hadoop, http://hadoop.apache.org/ .

[2] Jeffrey Dean, Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters” Proc. of OSDI, pp. 137–150, 2004.

[3] Alex Loddengaard, “Cloudera ’ s Support Team Shares Some Basic Hardware Recommendations”, http://www.cloudera.com/blog/2010/03/clouderas-support-team-shares-some-basic-hardware-recommendations/, 2010.

[4] “Greenplum Data Computing Appliance”, http://www.greenplum.com/products/greenplum-hd .

[5] “The MapR Distribution for Apache Hadoop”, http://www.mapr.com/resources/maprforapachehadoopwp .