分散データベースの性能によるデータの入力選択

Selection of Data Input Type According to The Performance of Distributed Database

郭崇　久永忠範　能登大輔　渕田孝康
（鹿児島大学大学院理工学研究科）

1　はじめに

近年，ネットワーク上に存在するデータは莫大な量となっている．これまでのデータベース管理システムの主流であるリレーショナル型データベース管理システム（RDBMS）では，ネットワーク上に偏在する情報を集中管理することはもはや困難となってきている．そこで，新しいデータベース「NoSQL」が注目され始めている．

しかし，各分散データベースはそれぞれ異なる性能を持つ．我々は，先行研究で，実際のビッグデータを用いて分散データベースの性能評価を行い，実行時間と書き込み速度について評価を行った．本研究では，先行研究の測定データを利用して，各分散データベースの書き込み方式の違いに基づき，異なるデータサイズに対して，最適な入力選択方法を提案する．

2　NoSQL

NoSQL とはリレーショナルデータベースを象徴する SQL 言語に依存しないという意味から誕生した言葉で，非リレーショナル，分散，オープンソースと水平スケーラブルの特徴を持つ，非リレーショナルデータベースの総称である [1]．NoSQL としては分散型データベースが注目されており，これには Big Table, Cassandra, Hadoop, HBase, cloudata, Azure Table などがある．

3　先行研究

先行研究では，我々は四つのデータベースを利用した．代表的な NoSQL データベース：Cassandra[2], MongoDB[3]を利用し，そして，比較のためリレーショナルデータベースの MySQL[4]も使った．

インターネット上で公開された各サイズのビッグデータ，例えば：2014 人気ニュース，NASA 地球データなどをダウンロードして，各データベース API に基づいた自作のプログラムを通じて，各データベースに書き込む．そして，書き込み時間と書き込みスピードを収集した．

3　実験

今回の実験は，入力データを行・列それぞれについて一つずつ増やして，各データベースの書き込み時間を収集し，その時間増加の傾向を分析する．

3．1 実験環境

本研究の具体的な内容について以下に述べる：

CPU：Intel Core i3 3.20GHz
Memory：4GB
HardDisk：300GB HDD
OS：Windows 10 Pro
Tool：Java,Eclipse

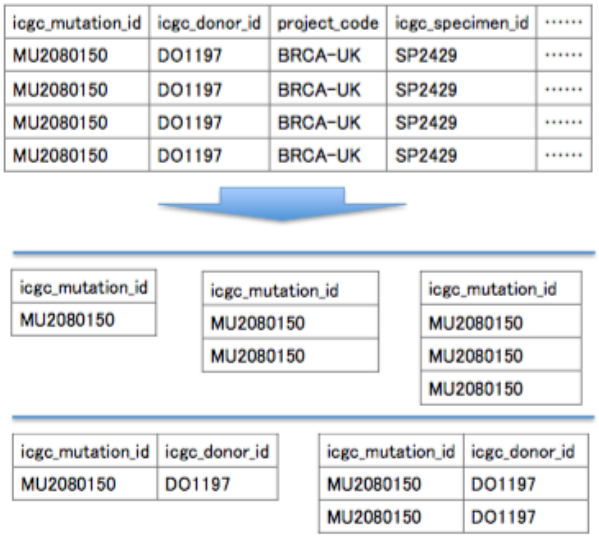

図 1．生成ファイルの示す

今回の実験では生物遺伝子のビッグデータを用いて，より詳細なデータを得るため，新たなプログラムを作った．

今回のプログラムは，一つの大きなデータファイルで各サイズのデータファイルを生成する．具体例を図 1 に示す．カラムと行によって，同じカラム数で違う行数のファイルを生成し，また，同じ行数で違うカラム数のファイルを生成する．そして，生成したデータファイルをデータベースに書き込み，その時間を計測する．さらに，収集したデータを用いて，各データベースに最適な入力方法を検討する．今回は，各書き込みは 15 回に行い，その平均時間を取る．

3．2 結果

MongoDB, Cassandra と MySQL についての実験結果を図 2~5 に示す．

図 2 により，MongoDB の書き込み時間は大体 3200 行のファイル以前は徐々に上昇している．しかし，その後書き込み時間は大幅に上下する．図の中の黒い線は実行時間の傾向である．図 2 ではこの近似曲線はほぼ直線に見える，しかし，拡大図の図 3 により，実行時間の約 3000 行を境にして傾向は指数曲線に近似する．

図 4 は Cassandra の書き込み時間である．Cassandra は 4250 行のファイル以前は指数曲線と同じ傾向で上昇している．そして，4250 行の後急に下がって，また上昇する．さらに，4600 行あたりの時実行時間はまた急に下がって，その後上昇する．今回の実験では Cassandra が 4250 行から 5000 行まで同じ結果を繰り返した．

MySQL の書き込み時間は図 5 に示す．MySQL の書き込み時間が 1500 行あたり以前は対数曲線の傾向で上昇している．そして，1500 行以後の書き込み時間の

傾向はほぼ直線の状態で増える．

4　まとめ

今回の実験は各データベースの書き込み時間の傾向が得られた．そして，その傾向の曲線を計算した．今回の実験ではデータの行数が 5000 行までと比較的小さい規模だったが，今後の課題とし，もっと大きな行数で実験を行う．また，分散データベースに対してマルチノードで書き込み処理を行った場合の入力性能を評価することがあげられる．

参考文献

［1］ 松本 ゆきひろ:”NoSQL の世界へようこそ”，みてわかる　クラウドマガジン.Vol.1,pp.64-65(2010)

［2］ Java Driver for Cassandra：
http://docs.datastax.com/en/drivers/java/2.0/

［3］ Java MongoDB Driver：
https://docs.mongodb.com/ecosystem/drivers/java/

［4］ JDBC：
https://dev.mysql.com/downloads/connector/j/5.0.html

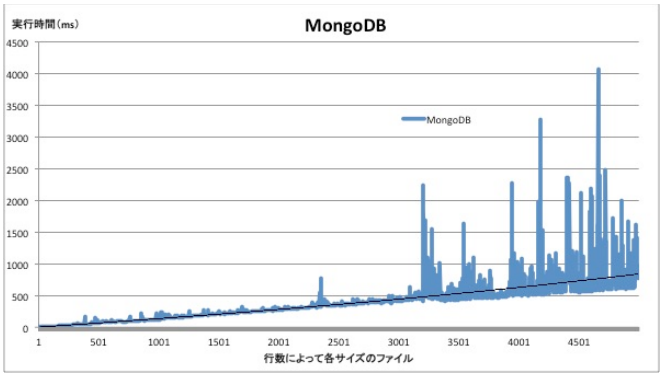

図 2．MongoDB の書き込み時間

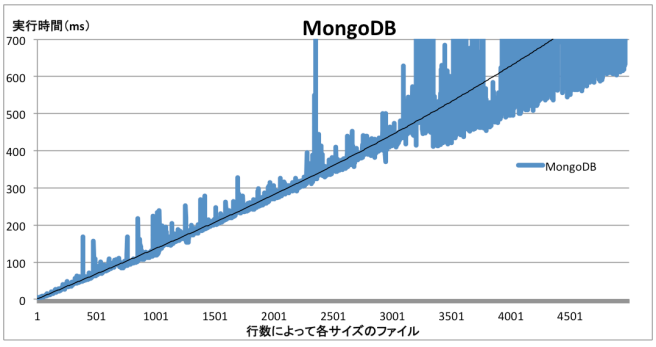

図 3．MongoDB の書き込み時間の拡大図

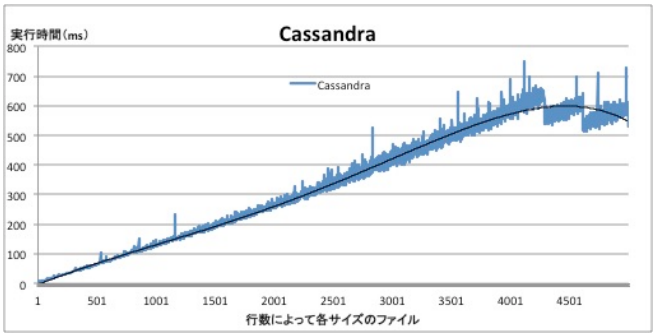

図 4．Cassandra の書き込み時間

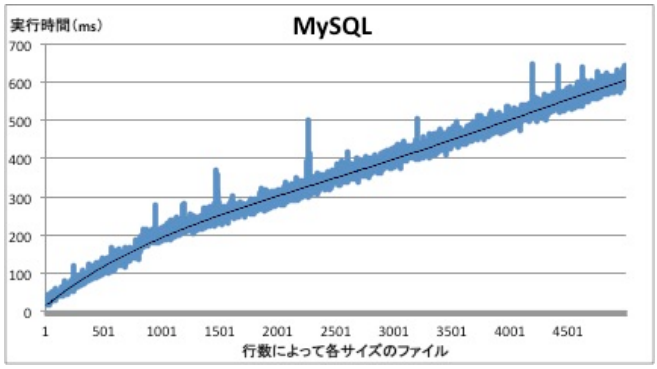

図 5．MySQL の書き込み時間