NoSQL を用いた大規模データ解析処理の高速化

川森 善紀†　井上　潮†
†東京電機大学大学院　工学研究科

1. はじめに

近年，ビッグデータを解析し，結果を経営戦略などに役立てる取り組みが行われている．実際に解析を行う現場では，複数のデータを組み合わせて解析を行うことが大半であるが[1]，解析に用いられるシステムは独自のものが多い．本研究では，NoSQL と呼ばれる大規模なデータを高速に処理できるデータベースと、複数の NoSQL をストレージとして利用できる Apache Spark に着目し、低コストかつ高速に複数データのビッグデータ解析を行えるシステムの実現を目指す。

2. 本研究の目標

本研究目標は，対象となるビッグデータに対して，図 1 のように適切な NoSQL にデータを割り当て高速化を図ることと，複数種類のデータまたは単体でのリアルタイム解析を行う汎用的なシステムの構築である．その為，一般的にデータ解析で利用されているフレームワークなどを利用して環境を構築した後に，多種類のデータへの対応や既存環境との連携など，多様な環境へ対応できるよう機能拡充を行うことで汎用化を目指す．

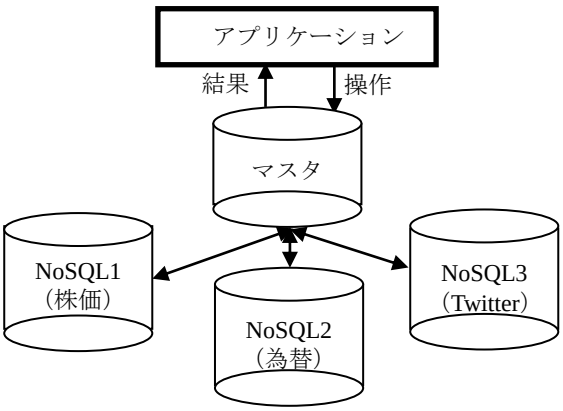

図 1 データ利用イメージ

3. 使用データの検討

本研究ではシステムの実装と動作速度の評価を目標としているため，典型的なビッグデータを用いた既存の解析を実装する．その為，使用データは為替データや株価のような通貨データ及び Twitter 等の典型的なビッグデータを用いる．

ビッグデータの一般的な定義として量，速度，多様性という性質がある．為替データは，通貨ペアやレートと言ったデータで構成される．このデータはリアルタイムで更新が行われ，通貨ペアごとにデータ量が増加する．このことから，為替データは量，速度，と言った特徴を満たす．Twitter のデータは，ID とつぶやきと呼ばれる 140 字以内の文章が紐付いたデータとなる．このデータも為替データと同様に早い更新速度を持ち，多いときでは秒間 14 万件のつぶやきが行われることもある．このことから，Twitter のデータもビッグデータにおける量，速度，と言った特徴を満たす．

4. 現在の進捗

現在，6 台の PC で Hadoop を用いた分散環境を構築した．また，上記環境において各 PC 上で Spark を動作させ，SparkPI と呼ばれる付属のテストプログラムが実行可能であることを確認した．また，為替データは 3 種類の通貨ペアを 5 秒ごとに更新し，ペアごとに約 48 万件のデータを収集した．

5. 今後の進め方

初めに為替データの解析システムを完成させる．次に，Twitter のデータを追加して，複数の NoSQL を同時に動作させ，統合的に利用できるシステムに発展させる．また，システムの評価方法や，最適な NoSQL を判断する方法としてベンチマークの利用なども検討する．

参考文献

[1] 総務省（2015）「情報通信白書」，
<http://www.soumu.go.jp/johotsusintokei/whitepaper/h27.html>(参照 2017-2-9)