2010 年度情報システム工学科　自主課題研究概要報告

“Hadoop”, “HBase” による分散データベースシステムの構築

計算機ソフトウェア研究室

情報システムコース　3 年　228　清水裕亮

目的

本自主課題では、”Apache Hadoop” と”HBase” を用いて分散データベースを実際に構築することで、分散データベースシステム構成を理解することを目的とする。

要素技術

“Hadoop” とは、Hadoop Distributed File System(HDFS[1])と、Hadoop MapReduce[2] Framework から構成される分散処理のためのオープンソースである。

“HBase” とは、Google が開発した BigTable のクローンを目指して開発され、Hadoop 分散処理環境下で HDFS を利用してデータストアを実現する Apache OpenSource Project である。TB 級の Big Data という条件で最高の性能を発揮する。

開発環境の構築

<サーバー構築>自主課題用の 3 台の PC に CentOS5.5 をインストールし、Apache サーバーを構築する。PC 間は SSH 接続である。サーバー名は tsks とし、3 台の PC を tsks0, tsks1, tsks2 と区別する。

<分散データベースの構築>　Java, Hadoop、HBase をインストールし、設定編集を行う。

実装・運用

HBase 独自の Table /Data 操作を試してみる。実験用に Key-Value[3]データと検索インデックスデータを作成し、そのデータを HDFS 上の HBase に格納する。その検索インデックスデータと Hadoop MapReduce を用いて分散処理を行い実験用 Key-Value データから検索を行う。

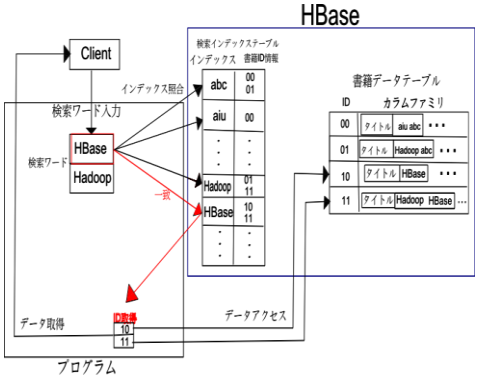

図　検索システムの概念図

考察および反省

Hadoop と HBase を用いて分散データベースを実際に構築することで、分散データベースシステムの仕組みを理解することができた。RDB(リレーショナルデータベース)と比較すると、HBase は write 性能重視で耐故障性が高くつまり、スケールアウトが可能なデータベースといえる。今回の研究では、”Hadoop”, “HBase” を用いた Key-Value データベースの構成についてと、vi コマンドや、Java 言語の学習になった。

今後の課題として、分散アルゴリズムの知識から、各サーバーへのタスクの分散の仕方などについて効率化を図ることが考えられる。

---

[1] HDFS…大規模なファイル大規模なファイルを効率的かつ、安全に保存するための分散ファイルシステム

[2] Hadoop MapReduce…分散コンピューティングモデルに基づくプログラム作成のフレームワークを提供する

[3] Key-Value データ…データはテーブル内でユニークな値である Key 値と、その Key に対するデータを Value として持つ。