MapReduce汎用化のためのDSL基盤・実行基盤の開発

―MapReduce の力をみんなに―

１．　背景

最近のウェブサービスや企業内アプリケーションでは、大量データ処理の必要性が増している。しかし、例えば全ウェブページのテキストデータの量は、400TB と言われており、それらをハードディスクから読み出すだけでも、2000 日以上かかる計算となる。そこで、並列分散処理フレームワークが必要で、特に有名なのが Google の MapReduce という処理方式である。Hadoop はオープンソースの MapReduce フレームワークで、誰でも利用可能だが、一般の開発者が利用するためには、処理を、Map 処理、Reduce 処理へと適切に分解する必要があること、Java で記述すること、複数種類のサーバプロセスを理解し使いこなすこと、などが要求され、敷居が高さから利用があまり進んでいない。

２．　目的

MapReduce フレームワークである Hadoop の敷居の高さを解消し、専門の技術者ではなくても大規模分散データ処理を可能にすることが目的である。それにより、大量データに埋もれた新しい発見や価値のあるサービスの開発を可能にする。

３．開発の内容

本プロジェクトでは、Hadoop でログ解析などを行うときに、それぞれの開発者が Map 処理、Reduce 処理に分解する必要がないように、Ruby による特定の分野の処理に特化した言語(DSL: Domain Specific Language・領域特化言語)を記述できるフレームワーク「Hadoop papyrus」を開発する。それを、オープンソースとして公開し、簡単に利用することができるようにする。さらに、Hadoop のサーバ構成を簡単に実現するために、Hudson という分散ビルドシステム上で、Hadoop papyrus を実行できるようにする。

また、開発したもののパフォーマンスが十分かどうかを、Amazon EC2 上で最大 50 台での Hadoop クラスタによる Wikipedia 日本版全データ(約 4.5GB)の解析処理を行い、検証する。

４．従来の技術（または機能）との相違

MapReduce 処理を Ruby で記述する技術としては、Hadoop streaming というツールが存在するが、Hadoop streaming は Map 処理、Reduce 処理をそれぞれ記述する必要があり、さらに決まった形式のデータのみしか処理できないため、本プロジェクトは MapReduce を意識しない点や、データ形式を柔軟に扱える点などで優れている。

また、Hadoop の公式サブプロジェクトである Hive、Pig というプロジェクトでは Map 処理、Reduce 処理を意識せずに Hadoop に対して処理を行うことができるが、Hive

はSQL的な記述で大規模データを扱うツールで、Pigはデータの集計処理に特化する独自言語を提供しているというものである。よって、Hadoop papyrus のようにいろいろな領域に適応できる処理を書けるというわけではない。また、独自言語の拡張性という意味でも、Rubyを利用するHadoop papyrusのほうが優れていると考えられる。

５．期待される効果

大規模データ処理が必要な開発者に対して、MapReduce の知識や経験がなくとも、Hadoop を利用しての大規模データ処理をより容易に行うことができるようになる。特に、Map 処理と Reduce 処理へ既存の処理を分割する部分が通常の開発者には難しいが、Hadoop papyrus はそのような部分をフレームワーク側が用意している。そのため、それぞれが行いたい処理の記述のみに集中できる。つまり、フレームワーク内にMapReduce処理のノウハウが凝縮されているので、それを利用する開発者はその恩恵に与るということである。

これにより、大規模なデータ処理を行うための敷居が低くなり、今まで眠っていたアプリケーションログやアクセスログなどを解析するニーズが生まれることが想定される。それにより、新しいサービスを作成し、価値のある情報を得ることなどが容易になる。

６．普及（または活用）の見通し

本プロジェクトはgithubサイト(http://github.com/)にてオープンソースとして公開されており、成果物もrubygemsという開発者に利用しやすい形式で配布している。Hadoop papyrus を公開しているgemcutterという配布サイト(http://rubygems.org/)では、成果物の2つのモジュール(hadoop-papyrus, jruby-on-hadoop)が、プロジェクト終了時に合わせて700回以上ダウンロードされている。また、プロジェクト後半では積極的にHadoop papyrus の発表を行った。Rubyコミュニティのサポートで有名な楽天技術研究所にてプロジェクトの発表を行い、その結果楽天技術研究所フェローであるRuby開発者のまつもとゆきひろ氏に対してプレゼンを行い、高評価を得られた。もちろん、クリエータの所属するシリウステクノロジーズでも、ログ分析等の業務に利用しており、成果を上げている。

これらの成果が認められ、プロジェクト期間外ではあるが2010年3月8日にヤフージャパンにて開催されるHadoop hack nightというイベントのパネリストとして招待された(http://gihyo.jp/event/2010/hadoophn)。

今後の課題としては、Hadoop papyrus の普及のために技術ドキュメントの整備や開発コミュニティの育成が必要である。開発項目としては、基盤コンポーネントの速度改善のための開発やログ解析以外のDSLの作成が挙げられる。

７．クリエータ名（所属）

藤川　幸一（株式会社シリウステクノロジーズ）

（参考）関連ＵＲＬ http://github.com/fujibee/hadoop-papyrus