2009 年度上期未踏 IT 人材発掘・育成事業　採択案件評価書

1. 担当PM

加藤　和彦 PM（筑波大学　大学院システム情報工学研究科　教授）

2. 採択者氏名

チーフクリエータ：　藤川　幸一（株式会社シリウステクノロジーズ
グループマネージャー）

コクリエータ　　：なし

3. プロジェクト管理組織

株式会社オープンテクノロジーズ

4. 委託金支払額

4,960,881 円

5. テーマ名

MapReduce 汎用化のための DSL 基盤・実行基盤の開発

6. 関連Webサイト

http://github.com/fujibee/hadoop-papyrus

7. テーマ概要

最近、インターネットの発達に伴ってアプリケーションが処理する情報量はますます増加している。

Google を始めとするインターネットサービス企業などは、多数のマシンを並列に接続

して処理を行う分散フレームワーク環境を開発しており、そのシンプルさとスケーラビリティで注目を集めるのが"MapReduce"という分散処理の仕組みである。

これはそもそも Google の社内開発基盤であったが、近年オープンソースとして"Hadoop"という実装が登場しており、MapReduce 環境を使用する敷居は下がってきた。

しかしそれでも以下の問題点により一般開発者・事業者の利用が難しいのが現状である。

1. 処理プログラム(Mapper, Reducer)の記述制約による開発の困難さ
2. 多数のマシンを利用するため利用マシンクラスタの管理方法の複雑さ

そこで、本プロジェクトではプログラム記述の制約をあらかじめ内包している、Ruby による DSL(Domain Specific Language)基盤を開発し、開発領域(大規模データ処理、インデックス作成、金融工学計算、など)によって DSL を作成できるようにし、参照実装としていくつかの DSL を開発する。

また、MapReduce 環境を利用するためのフレームワークとして、分散ビルド環境として有名なオープンソースツール "Hudson" を利用し、Hadoop を容易に管理できるようにする。

本プロジェクトの成果としては、実行基盤セットの公開、DSL 基盤仕様・実装公開、DSL 基盤を用いた DSL 参照実装(3 領域以上)の公開を目指し、それらを用いた実際のプログラムをフレームワーク上で動作させるところまでを目標とする。

また、各種コミュニティに働きかけ、世界的な普及活動を行う。

8. 採択理由

Google によって提案された MapReduce 技術は、クラウドコンピューティングの基礎技術として大いに注目を集めているが、その利用の敷居は低いとは言えず、実際のプログラミングも簡単ではない。

本提案は、同技術を Ruby 等のスクリプト言語から容易に使えるようにするという提案で、MapReduce 利用の敷居を下げ、クラウドコンピューティング普及を促進することが期待できる。そのアプローチも、提案者のこれまでのアクティビティに基づいたものであり、十分な feasibility を有すると判断できる。

グリーン IT への貢献としては、MapReduce に使用する計算機を動的に増減させることを可能とし、処理負荷に応じて使用する計算機数を調整したり、遊休計算機資源の活用に用いることが出来ると考えられる。

9. 開発目標

以下にあげる機能を開発し、その効果を測定することを目標とした。

・JRuby による Hadoop API Wrapper の実装

MapReduce DSL 基盤を実現するため、Hadoop の機能を JRuby により呼び出すことで Ruby から MapReduce を実行可能にする機能を実装する。

・DSL 基盤とその上で動作する個別 DSL の実装

JRuby により wrap された Hadoop を呼び出すために、特定の領域に限った DSL とその使用サンプルを Ruby により実装する。

・Hudson 上で MapReduce を実行する管理画面の実装

Hudson に上記の Ruby スクリプトを実行する画面をプラグインとして作成し、実行・結果の閲覧を可能にさせる。

10. 進捗概要

当初の予定どおりに進めることができた。

11. 成果

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」を開発した。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した。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した。

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、検証を行った。

Hadoop コミュニティや Hadoop を Ruby から利用している大手サービス提供事業者などでプロジェクトの発表を行い、プロジェクトの知名度向上やユーザ拡大のための活動を行った。

12. プロジェクト評価

Google 社の提案と報告により、大規模データ処理の分野でその有効性が広く知られることになった MapReduce、およびそのオープンソースソフトウェア実装である

Hadoop であるが、応用問題領域においてそれを利用することは簡単ではない。各応用問題をどのようにして、MapReduce の枠組みに落とし込めるかが自明ではなく、高度な知識とスキルが要求されるためである。藤川氏は、問題領域ごとに解法パターンを DSL (Domain Specific Language)として記述できる枠組みを発想し、その設計と実装を行った。Hadoop は Java 上の実装であるが、それを使用して高水準記述を可能にするために JRuby、また、分散実行フレームワークとして Hudson を利用し、Hadoop を容易に管理できるようにした。以上の発想は藤川氏独自のものであり、ソフトウェアの構想力、実装力は素晴らしいものがある。

今回の開発では、DSL としての実証はログ解析を対象として行ったが、より多くの問題領域で提案方式が有効であることが示されれば、提案システムの有効性はより強く検証され、また、広く社会で実用に供していく事になるであろう。

13. 今後の課題

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。