# 分散処理アプリ演習

## 平成 25 年度シラバス

2013 年 1 月 4 日

国立情報学研究所

トップエスイープロジェクト

代表者　本位田　真一

## 1. 科目名

分散処理アプリ演習

## 2. 担当者

（株）NTT データ　基盤システム事業本部　担当者（予定）

## 3. 本科目の目的

本科目では、主に演習を通して、実践的な大規模データの分散処理技術を習得する。

## 4. 本科目のオリジナリティ

NII で構築した学習用クラウドを講義・演習用環境として活用し、実際の業務に役立つ事例を中心とした題材を使用することで、実践的に分散処理アプリケーション開発を体験できる。

## 5. 本科目で扱う難しさ

大規模データを効率的に処理し活用したいという要望が今後益々増えてくると考えられる。しかし、まだ一般的には大規模データの分散処理技術の適用事例を経験する機会が少なく、その技術・ノウハウを身に着けることが難しいと考えられる。

## 6. 本科目で習得する知識・技術

本科目で扱う具体的な分散処理技術は、Hadoop である。Hadoop の構成要素である MapReduce や HDFS の動作の仕組み、MapReduce アプリケーション（抽出、結合、集計・統計等）の実装方法、テスト方法、運用・監視方法、性能チューニング方法、および Hadoop の関連技術である Hive、Pig、HBase の利用方法等について学ぶ。

## 7. 前提知識

本科目の受講生は、以下の項目を習得済みであることが望ましい。

(1) クラウドコンピューティングの基礎

(2) Java プログラミング

このうち、項目(1)はトップエスイー開講科目「クラウド入門」で習得可能である。

## 8. 講義計画

**概要**

第 1 回： Hadoop の概要
- Hadoop クラスタの構成
- Hadoop 環境構築
- HDFS の基礎
- MapReduce の基礎
- Streaming API

第 2 回： MapReduce アプリケーションの概要
- 文献単語解析アプリ（WordCount）の実行

第 3 回： MapReduce アプリケーションの実装
- MapReduce アプリケーション実装の基本
- MapReduce アプリケーション実装の実践的なテクニック
- 代表的な MapReduce の適用領域（集計・統計処理）

第 4 回： MapReduce アプリ演習
- レコメンデーションアプリの実装

第 5 回： Hadoop の動作詳細
- HDFS の詳細
- MapReduce の動作詳細

第 6 回： 高度な MapReduce プログラミング
- カウンター
- ロギング
- SequeceFile や圧縮ファイルの使い方
- 分散キャッシュ
- Secondary ソート
- MapReduce ジョブとしての作法

第 7 回： MapReduce アプリケーションのテスト
- MRUnit の使い方
- MapReduce アプリケーションのテストの考え方

第 8 回： Hadoop の性能チューニング
- MapReduce アプリケーションの性能チューニング
- Hadoop 設定の性能チューニング
- JVM、OS 設定の性能チューニング

第 9 回： Hadoop の運用・監視
- Web UI での MapReduce ジョブ分析

- Ganglia でのリソース情報と Hadoop メトリクスの確認
- MapReduce ジョブ失敗時の復旧手順
- 運用の作法

第 10 回： Hive の概要
- Hive と Hadoop との関係
- Hive の用途
- Hive のアーキテクチャ
- Hive と RDBMS との違い

第 11 回： Hive アプリ演習
- Hive による POS データ分析アプリの実装
- Hive クエリの MapReduce への変換の仕組み
- Hive クエリの高速化テクニック

第 12 回： Pig の概要
- Pig と Hadoop との関係
- Pig と Hive との違い
- Pig Latin の書き方
- Pig によるアプリの実装

第 13 回： HBase の概要
- HBase の特徴
- HBase の採用基準・適応領域
- HBase の機能
- HBase のアーキテクチャ

第 14 回： HBase スキーマ設計
- HBase スキーマ設計のポイント

第 15 回： HBase アプリ演習
- MapReduce での HBase データ操作の方法
- twitter ログ解析アプリの実装

**詳細**

第 1～2 回： Hadoop アプリ事例(1) 文献単語解析

- 文献単語解析アプリを題材として、Hadoop（HDFS、MapReduce）の基礎について解説し、演習を行う。

第 3～4 回： Hadoop アプリ事例(2) レコメンデーション

- レコメンデーションアプリを題材として、MapReduce アプリケーションの代表的な適用領域の一つである集計・統計処理について説明するとともに、MapReduce プログラミングの基礎および実践的な実装テクニックについて解説し、演習を行う。
- まず、MapReduce アプリケーション実装の基本として、必要なクラスや設定等を説明する。次に、実践的な実装テクニックとして、Map と Reduce の使い分け、ジョブの分割指針等を解説する。さらに、代表的な MapReduce の適用領域として、集計・統計処理の例であるレコメンデーションについて取り上げ、レコメンデーションアプリを実装する演習を行う。

第 5～12 回： Hadoop アプリ事例(3) POS データ分析

- POS データ分析アプリを題材として、Hadoop の動作詳細、高度な MapReduce プログラミング、テスト方法、性能チューニング方法、運用・監視方法、Hive や Pig によるアプリ開発方法について解説し、演習を行う。
- まず、Hadoop の構成要素である HDFS と MapReduce について詳細な挙動を説明する。Hadoop フレームワークとしてのデータの管理方法や分散処理の仕組みについて第 1 回-第 2 回で説明した内容を掘り下げて解説する。次に、POS データを集計するためのアプリケーションを Java での MapReduce プログラミングにより実装する。この中で、Hadoop の MapReduce フレームワークが提供する各種機能を利用したテクニックについて解説する。そして実装したアプリケーションは、テストやデバッグを経て、分散環境で動作させる。このとき性能に関する観点やチューニングポイントについて説明する。更に、アプリケーションの動作状況を把握するために Hadoop の持つ統計情報を Ganglia にて確認する。
- SQL ライクなクエリ言語をサポートする MapReduce のインターフェイス「Hive」について解説する。MapReduce との関係や RDBMS との違いを解説したのち、POS システムを題材とした演習を行う。さらに Hive との比較として Pig についても解説・演習を行う。

第 13～15 回： Hadoop アプリ事例(4) twitter ログ解析

- twitter ログ解析アプリを題材として、HBase を利用したアプリ開発方法について解説し、演習を行う。
- まず、HBase の概要として、Key-Value ストア、RDBMS や HDFS との比較、HBase の採用基準・適用領域等について説明し、次に、HBase の機能やアーキテクチャ

を解説する。また、HBase のスキーマ設計のポイントについて説明する。さらに、HBase を用いたアプリを実装する演習を行う。

※講義内容は若干変更になる可能性があります。

## 9. 教育効果

本科目を受講することにより、大規模データを処理する実務において、分散処理技術である Hadoop やその周辺技術の適用可否を適切に判断し、それを有効に使いこなすための実践的な技術を身につけることができる。

## 10. 使用ツール

- edubase Cloud および関連ツール
- Hadoop および関連ツール

## 11. 評価

演習課題レポート、出席日数を総合して評価する。

## 12. 実験及び演習

Hadoop の複数の適用事例（文献単語解析、レコメンデーション、POS データ分析、twitter ログ解析）を題材とした演習を用意している。

## 13. 教科書/参考書

- Tom White 著（玉川、兼田訳）「Hadoop 第 2 版」（オライリー・ジャパン）
- 太田、下垣、山下、猿田、藤井著（濱野監修）「Hadoop 徹底入門」（翔泳社）
- Jimmy Lin、Chris Dyer 著（神林、野村監修、玉川訳）「Hadoop MapReduce デザインパターン」（オライリー・ジャパン）
- 「平成２１年度産学連携ソフトウェア工学実践事業（高信頼クラウド実現用ソフトウェア開発（分散制御処理技術等に係るデータセンターの高信頼化に向けた実証事業））事業成果報告書」（経済産業省）