分散処理アプリ演習
講座概要

（株）NTTデータ

本講座の概要

- 本講座の目的
  - 本講座では、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。
- 本講座のオリジナリティ
  - 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。
- 本講座で扱う難しさ
  - 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。
- 本講座で習得する知識・技術
  - 本講座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。

本講座の概要(続き)

- 前提知識
  - 本講座の受講生は、以下の項目を習得済みであることが望ましい。
    - クラウドコンピューティングの基礎
      - トップエスイー「クラウド入門」講座で習得可能
    - Javaプログラミング
- 教育効果
  - 本講座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。
- 使用ツール
  - edubase Cloudおよび関連ツール
  - Hadoopおよび関連ツール
- 評価
  - 演習課題レポート、出席日数を総合して評価する。
- 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。

講義計画 1日目

- 第1回：Hadoopの概要
- 第2回：MapReduceアプリケーションの概要
  - 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回：MapReduceプログラミング基礎
- 第4回：MapReduceによるレコメンデーションエンジンの実装
  - 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。

講義計画 2日目

- 第5回：Hadoop動作詳細
- 第6回：MapReduceプログラミング応用
- 第7回：MapReduceアプリケーションのテスト
- 第8回：MapReduceアプリケーションのチューニング
  - 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、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。

講義計画 3日目

- 第9回：Hadoopクラスタの運用
  - Hadoopの運用・監視方法について解説する。
  - 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。
- 第10回：Hive概要
- 第11回：Hive演習
- 第12回：Pig概要・演習
  - POSデータ分析アプリを題材として、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。
  - 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。

講義計画 4日目

- 第13回：HBase概要
- 第14回：HBaseスキーマ設計
- 第15回：HBase演習
  - 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。

参考文献の紹介

- 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）
- Apache Hadoop Webサイト http://hadoop.apache.org/