首都大学東京
TOKYO METROPOLITAN UNIVERSITY

分散処理実践セミナー

Apache Spark による MapReduce の基礎

日時 12/12（金）14:40–17:00
場所 2–101 号室＠日野キャンパス

近年ビッグデータという言葉を巷でよく聞くようになりました。ビッグデータを特徴づけている性質は 3Vと呼ばれ、volume（データ量）・velocity（入出力データの速度）・variety（データの種類）がこれまでのデータ処理と異なると言われます。このうち、特にデータ量はウェブデータにとどまらず、ログデータ・センサーデータなど、多種多様なデータが大量かつものすごい速度で増加しており、従来のデータ処理基盤では処理することが困難です。そのためGoogleはMapReduceと呼ばれる巨大なデータ集合に対するプログラミングモデルを2004年に発表し、大きな注目を浴びました。MapReduceはオープンソースのApache Hadoopというプロジェクトでも実装され、Google外ではデファクトスタンダードとして使われています。

しかしながら、Hadoopはバッチ処理（データを貯めて一気に処理する）には向いているものの、動作が遅いためリアルタイム処理には不向きという問題がありました。そこで、インメモリで動作させることで高速化を図る Apache Spark というプロジェクトが産まれました。機械学習においても反復的な処理が必要であるため、Apache Sparkを用いることで Hadoop と比較して100倍以上もの高速化を達成しています。

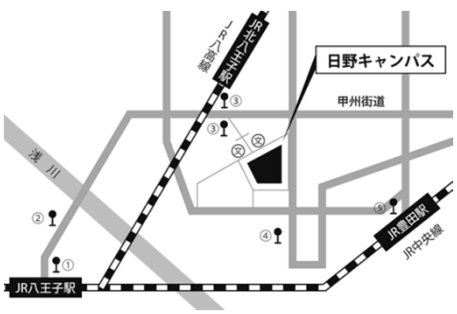
そこで、この演習では、Apache Spark を用いた MapReduce による分散処理の基礎について学びます。MapReduce は汎用的なプログラミングモデルなので、さまざまな言語で実装されており、一般のサーバで扱いが難しいような大規模データセットに対して分散処理が可能です。

多くの分野からの大学院生・学部生のみなさまの参加をお待ちしております。

開催日：	2014年12月12日(金) 14:40-17:00
会場：	２号館１０１室（情報系演習室）@日野キャンパス
申し込み・問合せ先：	小町守（システムデザイン学部情報通信システムコース, komachi@tmu.ac.jp）
注意：	1) 参加を希望する場合は、必ず事前に上記アドレスまで申し込みください。 2) 定員30名に達した場合には、申し込みを締め切る場合があります。 3) 20GB以上の空き容量があり、Java がインストール済みのノート PC を持参してください。 4) プログラミング言語の知識は問いませんが、Apache Spark がサポートしているのは Scala, Java および Python です。

Class 1	MapReduce の基礎（60分）
Class 2	Apache Spark による MapReduce 演習（60分）
講師：	小町守（システムデザイン学部情報通信システムコース准教授）

主催： 首都大学東京・教育改革推進事業「メニーコア・クラウド基盤技術の実践的教育」
後援： システムデザイン研究科情報通信システム学域



日野キャンパスへのアクセス

日野キャンパス内マップ