

Android 端末を用いた Hadoop ベースのライフログシステムの試作

新垣 智規^{†1} 城間 政司^{†1} 長田 智和^{†1}
谷口 祐治^{†2} 玉城 史朗^{†1}

近年、さまざまなセンサー類が搭載された携帯電話やスマートフォンの普及により、人間の行動をデジタルデータとして保存するライフログに注目が集まっている。記録されたデータの有用性は利用時に判明するものであり、ライフログでは可能な限りすべての行動を詳細に記録することが望ましい。しかし、RDBMS で構成された現行システムでは膨大なデータを扱うことが困難である。そこで、本研究では、ライフログの収集端末として Android OS が搭載された携帯電話を用いて位置情報データの取得し、ログの保存と解析に Hadoop と HBase を利用したライフログシステムの試作を行う。本稿では、その有用性と課題について検討する。

Trial Manufacture of the Hadoop-based LifeLog System using the Android Terminal

TOMONORI ARAKAKI,^{†1} TADASHI SHIROMA,^{†1}
TOMOKAZU NAGATA,^{†1} YUJI TANIGUCHI^{†2}
and SHIRO TAMAKI ^{†1}

The mobile devices and smart phones which have many types of sensors in it, is becoming common. Because of this, life log which stores digital data of human behavior, action, or movements which is called life-log is getting more attentions these days. The availability of the data that is stored in can only be known when actually using it. Therefore, it is desirable to have all the data we can get out of life-log when collecting. However, the existing systems which is constructed using RDBMS does not have the ability to compute floods of data. In this paper, we propose a Life-log system which uses a smart phone with an Android OS in it. This system will get location data and save them. After saving the data, it will be sent and analysed using Hadoop and HBase. we made a prototype system and disqus the availability of the system.

1. はじめに

2011年3月現在、内閣府の消費動向調査によると携帯電話の普及率は90%を超えており¹⁾、もはや携帯電話は現代社会に必要な不可欠なものになっている。カメラ機能やGPS、各種センサーなど、電話機能以外にもさまざまな機能が搭載された携帯電話が増えたことや、PCに近い機能を備えたスマートフォンの登場により、携帯電話は常に持ち歩くもっとも身近なコンピュータといえる。このように、小型の各種センサーが内蔵された携帯電話が発展し普及したことで、人間の行動をデジタルデータとして記憶するライフログに注目が集まっている²⁾³⁾。ライフログに記録された情報の有用性は後になってわかるため、ライフログでは人間の行動すべてを記録していることが望ましい。しかし、現在主流のMySQLやPostgreSQLなどのRDBMSで構成されたシステムでは、ライフログのような人間の行動に関する膨大なデータを扱うことは困難であり、かつ、データベースサーバーが単一であることに起因するさまざまな問題がある。さらに、写真や音声、動画ファイルをライフログとして使用する例もあり、柔軟な利用を想定したアーキテクチャを実現することが望ましいが、RDBMSでこれらのデータを保存することは一般的ではない。

一方、ここ数年でクラウドコンピューティングの発達が著しい。クラウドコンピューティングとは、インターネット上でさまざまなサービスを提供できる仕組みのことを意味し、GoogleやAmazonなどを筆頭に、現在ではさまざまな形態のサービスが存在する。我々は、Googleの各種サービス基盤をもとに開発されたオープンソースソフトウェアであるHadoopにおける、特定のサーバーに依存せず、コンピュータリソースをスケールアウトできる点に注目し、Hadoopをライフログの解析や保存先として利用できないかと考えた。

本研究では、ライフログの収集端末としてAndroid OSが搭載された携帯端末を用いて位置情報データを取得し、ログの保存と解析のためにHadoopとHBaseを利用したライフログシステムを試作する。この試作によりAndroid端末とHadoopベースで構成されたライフログシステムの有用性と課題を検討する。

^{†1} 琉球大学 理工学研究科 情報工学専攻

Graduate School of Engineering and Science, Information Engineering Course, University of the Ryukyus.

^{†2} 琉球大学 総合情報処理センター

Computing and Networking Center, University of the Ryukyus.

2. 技術概要

2.1 Android

Android とは、OS やミドルウェアとアプリケーションを含む携帯端末のためのソフトウェアスタックである。アーキテクチャは Linux カーネルをベースとし、アプリケーション部分には Java 技術を取り入れたプラットフォームとなっている。スマートフォンに代表される携帯電話に使用されるケースが多いが、最近ではデジタル TV やカーナビゲーションなどの携帯電話以外の組み込み分野にも広がりを見せている。また、加速度センサーやジャイロセンサー、NFC(Near Field Communications) など多数のセンサーを API レベルでサポートしており、ユーザーの周辺環境や状況を常に記録できる環境が整いつつある。

2.2 Hadoop

Hadoop とは、Google の MapReduce や GFS(Google File System) について公開された論文をもとに Apache のプロジェクトとして開発が進められているオープンソースソフトウェアである⁴⁾。Hadoop は、MapReduce を Hadoop/MapReduce、GFS を HDFS(Hadoop Distributed File System) として実装されており、テラバイトやペタバイトを超える大規模なデータを処理するための並列分散プラットフォームを提供する。

2.2.1 MapReduce

MapReduce とは、Google で開発されたプログラミングフレームワークであり、巨大なデータを多数のノードに分散して処理することが可能である。MapReduce では、投入されたジョブを Map 処理と Reduce 処理の 2 つのフェーズで処理することで、多数のマシンを用いる大規模並列計算処理を可能にしている。各フェーズでノードごとに独立した処理を行っているため、単純にノードの追加により処理性能を向上させることができる。

2.2.2 HDFS(HadoopDistributedFileSystem)

HDFS とは、巨大なファイルを分割して複数のサーバー群に格納する分散ファイルシステムである。クラスタ管理を行う NameNode と呼ばれる 1 台のマスターサーバーと、実際に I/O 処理を行う DataNode と呼ばれる多数のスレーブサーバーで構築されている。HDFS には以下のような特徴がある。

- 数百メガバイトから数ペタバイトにおよぶ巨大なファイルを扱うことが可能
- レプリケーションにより耐障害性が高い
- リバランス機能によりスケラビリティが高い

2.3 HBase の特徴

HBase とは、Google の分散データベースである Bigtable のクローンである列ベースの分散データベースである。HBase はマスター/スレーブ型の複数ノードで構成され、ノード間の分散協調には ZooKeeper が使用される。HBase は HDFS 上で構築されることを想定しており、ブロックサイズの小さい書き込み/読み込みといった HDFS の弱い部分を補完するための上位システムである。また、MapReduce 処理の入力元/出力先としても指定できるため、HBase のテーブル情報の内容を使った分散処理も可能となっている。

HBase は、ノードの追加により線形的にスケールすることができるように設計された分散データベースである。テーブル領域が増えるとテーブルを水平分割し、HBase クラスタへ分散する仕組みとなっており、テーブルの書き込み/読み込みの分散を実現している。Yahoo! Cloud Service Benchmark のベンチマーク結果によると、Cassandra および MySQL と比較して以下の傾向がある⁵⁾。

- HBase は Cassandra より書き込みレイテンシが小さく、読み込みレイテンシは大きい
- 範囲スキャン性能はレイテンシ/スループットともに Cassandra より速い
- MySQL に比べ読み込み性能は低い、書き込み性能は高い

HBase は書き込みをメモリーに対して行う設計であるため、その他のデータベースに比べて書き込み性能が高く、読み込みは HDFS から行うため低い結果となっている。同じ BigTable のクローンである Cassandra に比べ、範囲スキャンが可能であり、データの一貫性を維持できるという違いがある。

3. 試作システム

Android 端末はさまざまなセンサー類を搭載しており、さらに小型であるため身につける際に邪魔にならないなど、ライフログの取得に適している。しかし、ライフログ取得となると容量やデータ処理能力が不足するといった問題が発生する。この問題を解決するため、我々はライフログの取得/表示に Android 端末を用い、ログの保存および処理に Hadoop を使ったシステムを試作している。試作システムの実行環境を表 1 および表 2 に示す。Android 端末ではライフログとしてもっとも基本的なデータである位置情報の取得、データの保存および表示を行う。Hadoop 側では Android 端末で取得したライフログデータのファイルを受け取り、HBase へ格納する。格納されたデータはユーザーの命令を受けて、MapReduce 処理を行う。

表 1 Android 端末情報

端末名	Nexus S
CPU	1GHz Cortex A8
メモリ	512MB
バージョン	Android 2.3.6
カーネルバージョン	2.6.35.7

表 2 Hadoop 実装環境

プロセッサ	Intel Xeon L5410 2.33GHz / 4 コア
メモリサイズ	16GB
OS / カーネル	CentOS 5.7 / Linux 2.6.18
Hadoop/HBase	Hadoop 0.20.2/ HBase 0.90.3
台数	4 台

3.1 Android 端末の処理

今回の試作システムでは、位置情報および時刻をライフログとして扱う。Android 端末は小型で常に持ち歩くことが可能であり、各種センサーを搭載されていることからライフログの取得端末として適している。また、Android OS ではネットワーク接続も API レベルで提供されているため、インターネットとの親和性が非常に高く、ライフログの保存先として Hadoop を使えると考えた。

Android 端末で実行されるアプリケーションでは、位置情報を取得し、GooleMap の地図データ上に移動経路を表すラインとして重ねて表示する(図 1)。位置情報を書きこむ場合、過去の位置情報と時刻からユーザーの移動速度を計算し、停止・徒歩移動・高速移動の 3 つのステータスのいずれかを設定する。ユーザーのステータスが停止中の場合は赤、徒歩移動中は緑、高速移動中には青のラインになるように実装した。このようにステータスを設定することで、ユーザーの移動状態を視覚的にも判別しやすくなり、Hadoop 側の処理で各ステータスごとに移動距離を計算することができる。

Android 端末で取得し、ファイルに書きだすデータは以下の通りである。

- 端末固有識別番号
- データの取得時間
- 位置情報(緯度, 経度)
- 移動距離
- ユーザーステータス

試作システムでは、ユーザー識別のために Android 端末の固有識別番号利用した。また、同じ形式のファイルを読み込むことで過去の移動経路を表示できるように、時刻およびユーザーの状態を追加している。

3.2 Hadoop 側の処理

今回の構成では、マスターノードとして 1 台、スレーブノードとして 3 台の計 4 台のマシンで図 2 のようにシステムを構成した。Hadoop および HBase はマスター/スレーブ型

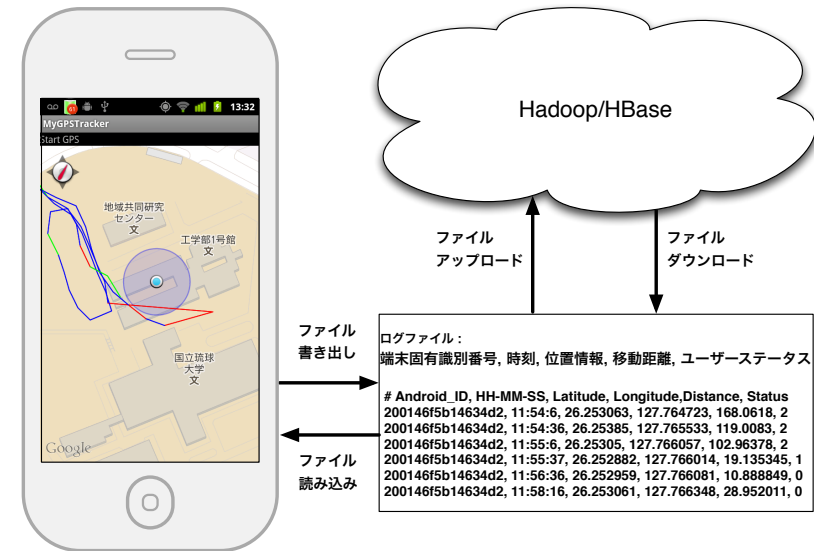


図 1 実装したアプリケーションの動作概要

の実装であり、HBase のテーブルは HDFS 上で構成される。マスターノード、スレーブノードのそれぞれで Hadoop および HBase のデーモンが動作する。マスターサーバーでは、Hadoop/HBase のマスターデーモンである NameNode, JobTracker と HBaseMaster, HBaseRegionServer, Zookeeper が動作し、スレーブサーバーはワーカーデーモンである DataNode, TaskTracker, HBaseRegionServer, Zookeeper がそれぞれ動作する。また、それとは別に NameNode のセカンダリとして SecondaryNameNode をスレーブノードの 1 台で動作させている。Hadoop 側では、Android 端末で取得したライフログファイルを HBase のテーブルへ入力する。HBase では、テーブルの構造が表 3 のようになっており、ROW(行キー)、Column Family(列キー)と Qualifier(修飾子)の 3 つの要素を指定することで格納された値を取得する。また、データは ROW の値で常にソートされて格納される。そこで、今回は ROW としてライフログ取得時の時刻、Column Family として Latitude と Longitude といった各データ名称を設定した。値の入出力を行うためには、Qualifier を一意の値にする必要があり、ユーザー識別のための Android 端末の固有識別番号を利用する。このようにテーブルを設計することで、ROW の値で範囲スキャンが可能となり、ユーザーが指定し

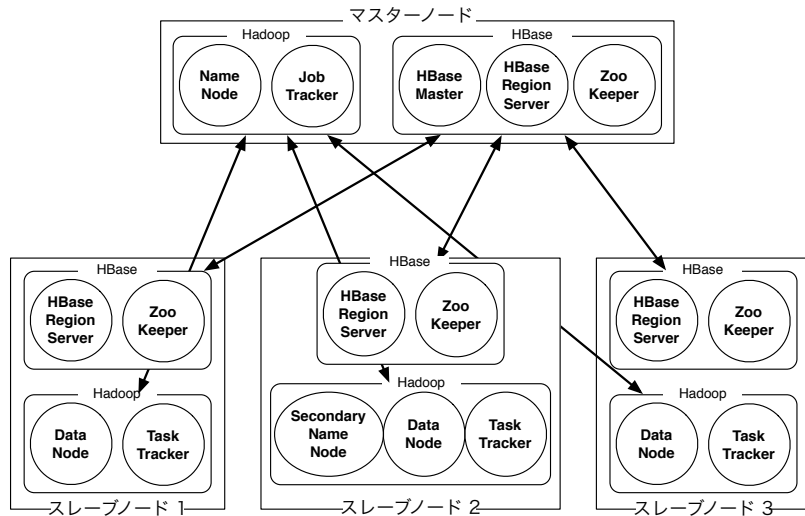


図 2 Hadoop/HBase の構成

た範囲の日時のデータを対象としたピンポイントな処理が実現できる。

表 3 HBase のテーブル設計とデータ例

ROW	Distance		Latitude	Longitude	Status
2011-9-20_15:13:45	8fb554bd5b600987	0.0	26.253723	127.765603	0
2011-9-20_15:13:55	8fb554bd5b600987	19.048351	26.253552	127.765625	0
2011-9-20_15:14:18	8fb554bd5b600987	10.628251	26.253332	127.765807	1

3.3 考察

今回の試作システムでは、Android 端末と Hadoop/HBase を利用してライフログを扱った。ライフログとは、人間の行動をデジタルデータとして記録することであり、生活の中のさまざまな行動の記録を積み重ねることで、データ間の関連性やユーザーの行動パターンをモデル化することが可能になる。そのため、ライフログはデータを継続的に取得する必要があり、そのためには「ユーザーがライフログを取得するために努力しないこと」が重要であると考えられる。本試作システムでは、ライフログの取得のために Android 端末上でアプリケーション実装することで、ユーザーが意識せずに移動経路データを取得できる。このこと

から、「ライフログを取らないといけない」という負担感をユーザーに与えることなく、継続的に続けることが容易になると考える。また、Hadoop/HBase を利用することで、ユーザーが多数になった場合にも膨大なデータを保存することが可能となり、テキストファイル以外のさまざまなファイルの保存先としても利用することができる。さらに、HBase では入力される値のデータ型宣言やデータサイズの指定を必要とせず、Column Family 以外はユーザーが自由に設定可能である。

4. まとめと今後の課題

本研究では、ライフログの収集端末として Android OS が搭載された携帯端末を用い、ログの保存と解析に Hadoop と HBase を利用したライフログシステムの試作を行っている。Android 端末では取得した位置情報から、移動経路を Google Map 上へ表示およびログファイルのアップロード機能を実装し、サーバー側では Hadoop/HBase を利用した MapReduce による移動距離を求める簡単なサンプルを実装した。

今後は、Web ブラウザでの表示に対応することで Android 端末以外からもライフログを閲覧できるように実装する予定である。さらに、Android 端末にて取得するライフログの種類を増やし、Twitter や Facebook, FourSquare といった既存のサービスとの連携を図ることで、位置情報などの現実のライフログ情報とインターネット上のライフログ情報を結びつけることができると考える。このことで、位置情報と時間のような自動的に取得される「現実世界のライフログ」と、ブログや SNS でユーザーが自発的に発信する「インターネット上のライフログ」の両方を取得し、相互の情報を関連させたより詳細なユーザーの行動を記録するライフログシステムを実現したい。

参考文献

- 1) 内閣府 消費動向調査 主要耐久消費財等の普及率（一般世帯），入手先<<http://www.esri.cao.go.jp/jp/stat/shouhi/shouhi.html>>
- 2) 小塚宣秀：ユビキタスネットワーク技術の研究開発 ~ ケータイ de ライフログ ~，入手先<<http://www.kddi.com/business/oyakudachi/square/labo/003/index.html>> (2008).
- 3) i コンシェル，入手先<<http://www.nttdocomo.co.jp/service/customize/iconcier/>>.
- 4) Apache Hadoop, available from <<http://hadoop.apache.org/>>.
- 5) Yahoo! Cloud Service Benchmark, available from <http://research.yahoo.com/Web_Information_Management/YCSB>.