ビッグデータ分散処理基盤における
共有データ二次利用時のセキュリティ制御に関する考察

理学専攻・情報科学コース　　横山　紗妃

1　はじめに

近年、各分野におけるビッグデータ活用に注目が集まっている．購買履歴，気象情報といったデータを収集することにより，収集したデータを解析し，そこから有用なデータを生み出し、そのデータを二次的に利用することが増えている．そういった社会的背景において，ビックデータの安全性が懸念され，同時にセキュリティ保護へのニーズが高まっている．ビックデータの二次利用を考えたとき，一次利用組織は二次利用組織におけるデータの使い方を直接的に管理することは難しい．よって組織間でセキュリティ基準が共有でき，一元管理できるようなアーキテクチャを検討する必要がある．

そこで，本研究ではビッグデータの二次利用に携わる機関にとってセキュリティが確保されたアーキテクチャを検討する．ビッグデータの組織間の受け渡しにおいて Hadoop の HDFS が注目されているが，セキュリティ面における課題もある．その課題を考慮した上でビックデータ流通基盤の構築を目指す．

2　関連研究

関連研究においても，組織間でセキュリティ基準をの水準を統一化し，自動管理するアーキテクチャが提唱されている．セキュリティ標準規格を適用することで企業間でセキュリティ情報の共有や分析の自動化をする方式を検討している [1].

また，Hadoop の HDFS を用いたビッグデータの制御も提案されている．

しかし，Hadoop はセキュリティ上の問題点がいくつか挙げられる．ファイルシステム以上の粒度の細かいセキュリティ制御ができないため管理者によるアクセス制御が困難であり，管理者とその他ユーザが同列扱いであるため，一次利用組織側が一元管理することができない．大事なデータは Hadoop には置かない，管理者以外に全く権限を与えないという対策を取らざるを得ない．

そこで，本研究では，Apache Ranger を適用していくことで，この課題を検討する．

3　Hadoop

図 1 のようにビックデータの受け渡しをするファイルシステムとして Hadoop の HDFS が今注目されている．Hadoop とは大規模データの分類、集計のフェーズから構成される MapReduce 処理による分散処理技術によって実現するオープンソースのミドルウェアである．データ複製が不要でクラスタのリソースを効率よく利用することが可能なファイルシステムの構築が可能になる．

Hadoop はデータ管理や分散処理ジョブの管理を行う NameNode から構成される Hadoop マスタサーバ群と分散処理の実行やデータの保存を NameNode の指示によって行う DataNode からなる Hadoop スレーブサーバ群から構成され，それぞれの NameNode,DataNode はファイルシステムである HDFS を利用する．

データ複製が不要でクラスタリソースを効率よく利用することが可能であり，ビッグデータ流通基盤を検討したとき最適である．しかし，ファイルシステムレベルでのアクセス制御であり，データの匿名化 [2] を考えたとき，それ以上のカラム，テーブルレベルの制御が望まれる．

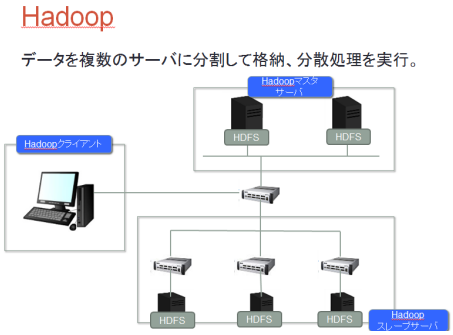

図 1: 分散処理ファイルシステム Hadoop

4　Apache Ranger による制御

本研究では Apache Ranger を検討した．Apache Ranger とは Apache 関連プロジェクトのセキュリティを保護するフレームワークである．[3]．Hadoop 上で動作させることが可能で，HDFS でのデータの受け渡しを想定した場合，管理者にとってセキュリティが確保された環境可能にする．

Hadoop はセキュリティ上の問題点がいくつか挙げられた．ファイルシステム以上の粒度の細かいセキュリティ制御ができないためアクセス制御が困難であり，管理者とその他ユーザが同列扱いのため一次利用組織側が一元管理することができない．機密性の高いデータは Hadoop には置かない，管理者以外に全く権限を与えないという対策を取らざるを得ない．Apache Ranger を適用するとこの問題が解決できる．Apache Ranger とは Apache 関連プロジェクトのセキュリティを保護するフレームワークである．Hadoop 上で動作させることが可能で，HDFS でのデータのやり取りを想定した場合，管理者にとってセキュリティが確保された環境構築を可能とする．(図 4)

Web インタフェースを通して管理者がアクセス定義することを可能にする Ranger Admin Server，ユーザからアクセスがあるとユーザとグループ名を取得して DB に定義する Ranger UserSync Server，リクエスト評価する Ranger plugin から構成される．

Plugin はアクセスがあると，アクセスされるファイル，アクセスタイプ，ユーザ，グループ，時間，IP アドレスなどのコンテクストを収集し，一定期間ごとに管理サーバからキャッシュしたアクセスポリシーをもとにアクセスを評価し，アクセス許可を判断し，アクセスログを取る．

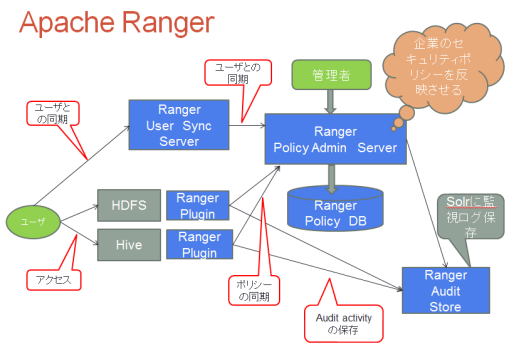

図 2: Apache Ranger

ユーザからのアクセスがあると User Sync Server を通して一次利用組織側の管理者の Ranger Admin Server と同期を取り，同時にユーザ情報はファイルシステム上で動作する Plugin に渡す．Plugin は管理者が設定したポリシーを一定時間おきにキャッシュし，そのポリシーをもとにしてアクセスを許可するかを判断する．[4].

5 想定環境

本研究では，図 3 のような構成で実験を行う．ogl010 に管理サーバを置き，外部からアクセスがあると UserSync Server がユーザと同期を取り，プラグインにキャッシュされた情報をもとにアクセスの可否を判断する環境を想定する．一次利用組織には全アクセス権限を付与し，下位組織に対しては付与可能なアクセス権限のみ許可する．Ranger を導入することにより，一次利用組織の一元管理が可能となる．

表 1: 開発環境

Ranger	Apache Ranger 0.5.0
Hadoop	Hadoop2.7.0
Java	Java 1.8.0
OS	Debian 6.0.4 64bit

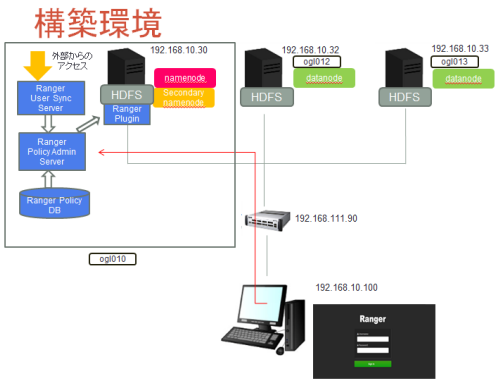

図 3: 実験環境イメージ

6 ヘルスケアアプリデータ

ビッグデータの二次利用例としてヘルスケアアプリデータを検討した．近年医療情報を活用するアプリの開発が進んでいる．アプリから取得できる情報として，氏名，生年月日，緊急連絡先，血液型，アレルギー，既往症がある．収集した情報を k 匿名化を用いることにより，その値を指標として Ranger で管理することでデータの二次利用においてプライバシの保護ができるようなモデルを検討している [2].

一次利用組織においてデータを取得すると，識別子と準識別子に分類する．識別子はデータ分析に用いないため，データの歪みが大きいマスキングを行い，準識別子に k 匿名化を適用する．データの性質により非特定化を進め，Ranger によって一次組織側または利用者が監視することにより，k 匿名化が守られているか管理できるモデルを想定する．

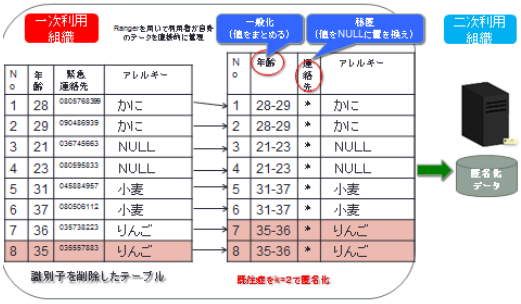

図 4: 匿名化手法

7 まとめと今後の課題

まとめとしては，大規模処理基盤を想定した環境を検討し，ファイルシステムである Hadoop の HDFS 上にセキュリティを保護するフレームワークである Ranger を構築し，その活用を検討した．

今後の課題としては実装を進め，データの統合を考慮した上で，匿名性が守られるようなシステムの構築をし，その性能を評価することである．

謝辞

本研究を進めるにあたり大変有益なアドバイスを頂いた情報セキュリティ大学院大学の後藤厚宏教授に深く感謝致します．

参考文献

[1] 長内　仁，　後藤　厚宏：　「企業間における情報セキュリティ連携アーキテクチャの検討」，SCIS 2014，2014 年 1 月.

[2] Khaled EL Emam, Luk Arbuckle 著，　木村　映善，　魔 狸　監訳，　笹井　崇司　データ匿名化手法，オイラリージャパン，　2015 年 5 月.

[3] Apache Ranger：http://ranger.apache.org.

[4] Securing Hadoop with Apache Ranger https://www.slideshare.net/mobile/Hadoop_Summit/securing-hadoop-with-apache-ranger.

[5] 後藤　厚宏：「社会の安心・安全に向けたビッグデータ処理ネットワークの課題」，Panasonic Technical Journal，Vol.59，No.2，pp4-8，2013 年 11 月.