スマートインダストリ向け組織横断データレイク活用のための整合性を考慮したポリシー管理技術の提案

増田　博亮† 大越　淳平† 馬場　恒彦‡

†(株)日立製作所研究開発グループ 〒185-8601 東京都国分寺市東恋ヶ窪 1-280
‡(株)日立製作所産業・流通ビジネスユニット 〒101-0021 東京都千代田区外神田 1-5-1

E-mail: † ‡ {hiroaki.masuda.vh, junpei.okoshi.pc, tsunehiko.baba.gc}@hitachi.com

あらまし　データ分析による業務改善を目的に，異なる組織を横断するデータ利活用が推進されており，セキュアなデータ利活用の実現が重要である．組織毎でデータを管理していたため，組織毎にアクセス制御の記述であるポリシーの定義やコンテキストが異なる．そのため，組織間のポリシーのコンテキストを整合させる必要があるが，組織毎の管理者間で繰り返し議論し手動で整合させていたため，多大なリードタイムを要する問題がある．本研究では，データカタログとメタデータを活用し，異なるポリシー間の整合性を考慮するポリシー管理システムを提案した．異なる組織を横断するデータ利活用のユースケースにおいて，手動でのポリシー管理におけるリードタイムの 93%を削減可能である．

キーワード　データレイク，アクセス制御，ポリシー，データカタログ，メタデータ，コンテキスト

1. はじめに

IoT システムの普及に伴って，様々な事象のデータを収集できるようになり，それら収集されたデータを分析することで，生産効率の向上といった新たな価値の創出が可能となっている．効果の高い分析を行うために，多くのデータの利活用が求められる中で，従来は組織毎で管理されていたデータの横断的な利活用が検討され，複数組織のデータを纏めるデータレイクの構築が注目されている．

一方で，データ利活用のコンプライアンスの策定が進んでいる．データを利活用する際は，コンプライアンスに基づいてアクセスポリシー（以降，本論文ではポリシーと称する）を定義し，アクセス制御を行うことでセキュアなデータ利活用を実現することが求められている．

複数組織を横断するアクセス制御の実現においては，組織毎のアクセス制御の連携を考慮する必要がある[1][2][3]．しかし，従来組織毎でデータを管理していた背景から，アクセス制御を行うコンポーネントの違いによってポリシーの定義が異なる場合や，ポリシーの記述のコンテキストが異なることによる不整合があるため，単純にポリシー間での比較ができず，簡単に連携することができない[4][5]．そこで，データレイクの管理者が，各組織のデータ管理者とヒアリング・議論を行い，各組織の定義する全てのポリシーを把握し，それらとの整合性を考慮した上でポリシーをデータレイク上に定義することで，セキュアなデータの利活用を監視・管理する必要があり，管理工数の増大が問題となっている．

本論文では，データカタログで管理するメタデータを活用し，ポリシー内のコンテキストを異なるポリシーのコンテキストへの解釈と，各組織で定義するポリシーと比較と評価をシステム化し，従来のデータレイク管理者が手動で行っていたポリシー管理プロセスを自動化することで，組織横断でのセキュアなデータ利活用を実現するための膨大な管理工数の削減を目指す．2 章では，異なる組織間でのアクセス制御の統合管理に関する研究について示す．3 章では，セキュアなデータ利活用のユースケースとその課題について示す．4 章では，提案手法であるポリシー管理システムとその特徴機能を示す．5 章では，提案手法を適用した際の評価を示す．6 章で本論文を纏める．

2. 関連研究

異なる組織間における統合的なアクセス制御方式が検討され，各組織がクラウド上で管理するデータを各クラウドベンダの提供するアクセス制御コンポーネントに適応したフレームワークの研究が行われている[6]．

各ベンダのアクセス制御機構に依存せず，ハイブリッド/マルチクラウド環境のデータに対するアクセス制御方式についても研究されている．マルチクラウドのソフトウェアが稼働するサーバ統合環境に対し，アクセス制御の統合管理基盤が提案されている[7]．

Banyal らは，マルチクラウドのセキュリティとプライバシー問題に対処するための柔軟かつ効率的なアク

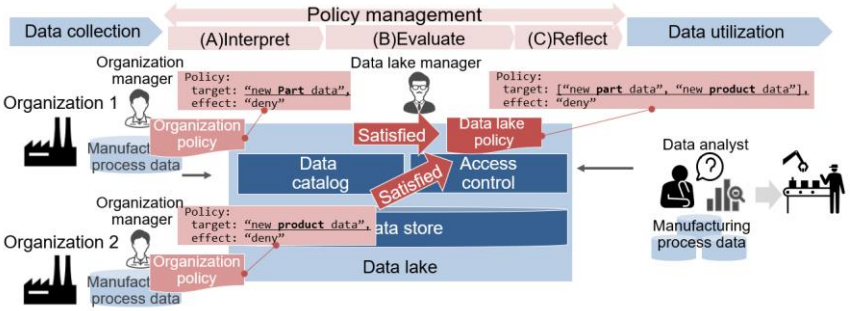

図 1 Secure data utilization flow

セス制御フレームワークを提案している[8]. 提案フレームワークは，静的および動的な信頼要素を用いて信頼値を算出し活用することで不正な操作を保護し，承認されたユーザがアクセス可能としている．

Komninos らは，種々の属性セットに対しアクセス制御ポリシーを紐づけることで，データアクセスを制御するフレームワークを提案している[9]. Li らは，プロキシベースのマルチクラウド環境用に属性ベースのアクセス制御システムを構築し[10]，分散アクセス制御を実現する MACPABE スキームを提案している．

Demchenko らは，組織内のマルチクラウド環境の統合アクセス制御に関して，ゲートウェイ方式を用いたアーキテクチャを研究している[11].

いずれの研究に関しても制御対象のデータに対し，アクセス制御コンポーネントで定義されるポリシーのコンテキストを，異なるアクセス制御コンポーネントのポリシーのコンテキストへの変換に関する研究であるが，変換前後のポリシーの整合性と複数のポリシーの組合せとの整合性の考慮とその管理に関する管理者の工数には触れられていない．

3. 組織横断でのセキュアなデータ利活用の実現と本研究の課題

本章では，組織横断で共有されるデータに対するセキュアなデータ利活用の実現と本研究の課題について述べる．

3.1. セキュアなデータ利活用の実現

組織横断のデータレイクを活用したデータ利活用において，常にデータレイク内の全てのデータに対し，全分析者がアクセスしてよいわけではない．各組織の管理者が，データレイク内のデータに対して，ビジネス上の理由により共有可能な範囲を狭めることを要求する場合がある．例えば，新製品のデータは，一部の分析者のみに共有を限定する等がある．この場合，共有を限定する対象のデータを示すコンテキストでポリシーを定義し，アクセス制御を行う必要がある．

図 1 を用いてセキュアなデータ利活用を実現するプロセスを説明する．データ利活用のサイクルは「データ収集」「ポリシー管理」「データ活用」から構成されるとした．また，セキュアなデータ利活用の実現に重要な「ポリシー管理」は，A)ポリシーの解釈，B)ポリシーの評価，及び C) ポリシーの反映の 3 プロセスから構成されるとした．データレイク管理者は，各組織の管理者が定義するポリシー（以降，本論文では組織ポリシーと称する）を収集し，ポリシー内で記述されるコンテキストから，保護対象となるデータを正確に把握する必要がある．具体には，組織 1 の中で「新部品」というコンテキストで記述されるデータが，データレイクに格納されるデータのどのデータ，あるいはデータレイクにおける別のコンテキストとどう対応するかを把握する必要がある．加えて，データレイク管理者は，データレイク上で定義するポリシー（以降，本論文ではデータレイクポリシーと称する）と組織ポリシーの一つ一つとの整合性を考慮し，整合しているデータレイクポリシーと組織ポリシーを比較評価しアクセス制御コンポーネントに反映することで，アクセス制御を行うことが求められる．

3.2. ポリシー管理プロセスと課題

3.1 節で説明したセキュアなデータ利活用の実現のためのポリシー管理プロセスと課題を述べる．

A) ポリシーの解釈：データレイク管理者は，各組織の管理者から組織ポリシーのヒアリングを行い，組織ポリシーの定義とコンテキストを理解する．データレイク内のデータから，各組織ポリシーのコンテキストが示すデータを確認し，データレイクポリシーでのコンテキストから組織ポリシーのコンテキストを解釈する．

B) ポリシーの評価：データレイク管理者は，A)の解釈の結果からデータレイクポリシーを定義する．データレイクポリシーが各組織ポリシーと整合しているか各組織の管理者にヒアリングを行い，評価する．

C) ポリシーの反映：整合しているデータレイクポリシーをデータレイク内のアクセス制御コンポーネントに反映し，アクセス制御を実施する．

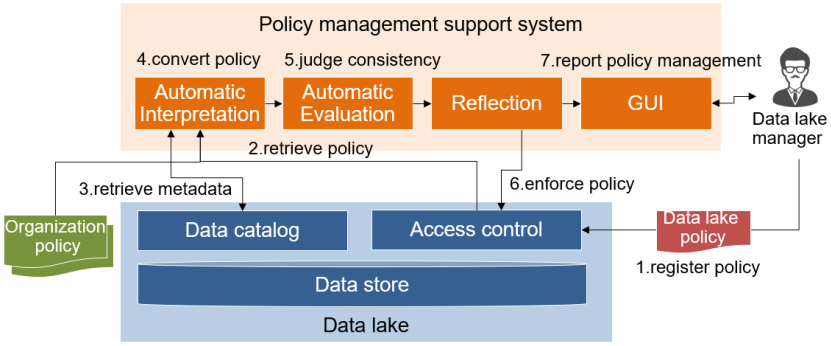

図 2 Policy management support system

A)-C)で定義されるポリシー管理プロセスは，データレイク管理者の人手によって実現される．ここでは，C)のプロセスは，A),B)のプロセスの結果を受けて定義したデータレイクポリシーをアクセス制御コンポーネントに反映する操作のため，管理者間でのヒアリング・議論は発生しないとしている．一方で，A),B)のプロセスは，管理者間のヒアリング・議論を繰り返し実施するにあたり，リードタイムが生じる問題がある．

上記の問題に対し，異なるポリシー間でコンテキストが異なる場合のあるポリシーから異なるポリシーへのコンテキストの解釈と，比較が必要なポリシーの組合せ毎の評価をそれぞれを自動化することで，従来のポリシー管理プロセスにおける A),B)のリードタイムを削減する課題に取り組む．

4. 提案手法

図 2 を用いて本提案手法のポリシー管理支援システムを説明する．本システムは，データレイクポリシーがアクセス制御コンポーネントに定義することを一つの起点として，ポリシーの自動解釈機能が，データレイクポリシーと組織ポリシーを収集し，データレイクポリシーを組織ポリシーのコンテキストで自動的に解釈する．本機能はポリシー管理の A)プロセスの自動化に対応する．次にポリシーの自動評価機能が，整合したデータレイクポリシーと組織ポリシーを比較し自動的に評価する．本機能はポリシー管理の B)のプロセスの自動化に対応する．最後に，評価の結果を GUI 上でデータレイク管理者に通知し，ポリシーの修正と反映の支援を行う．本機能はポリシー管理の C)に対応する．

前述のポリシー管理支援システムの有する 2 つの特徴機能に関しての詳細を説明する．

4.1. ポリシーの自動解釈機能

本機能は，データレイクポリシーを組織ポリシーの定義形式とコンテキストで解釈することで，組織ポリシーとの比較評価を可能とする．

メタデータを活用して，データレイクポリシーと組織ポリシー内で保護対象を記述するコンテキストから，保護対象のデータを抽出する．データレイクポリシーの保護対象とするデータが，組織ポリシーの保護対象とするデータを含むかどうかを判定し，含む場合にはデータレイクポリシーの保護対象を記述するコンテキストを組織ポリシーの保護対象を記述するコンテキストへの解釈できると判定し，コンテキストを変換したポリシー生成する．

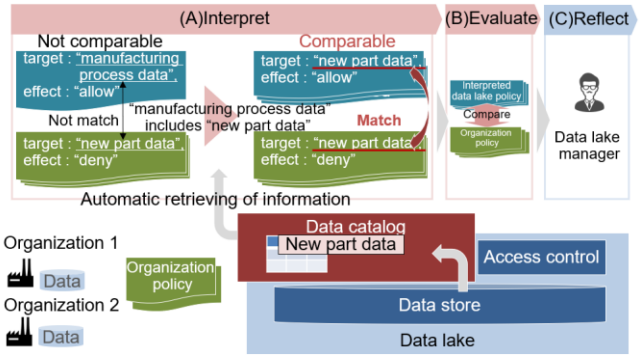

図 3 Automating interpretation function

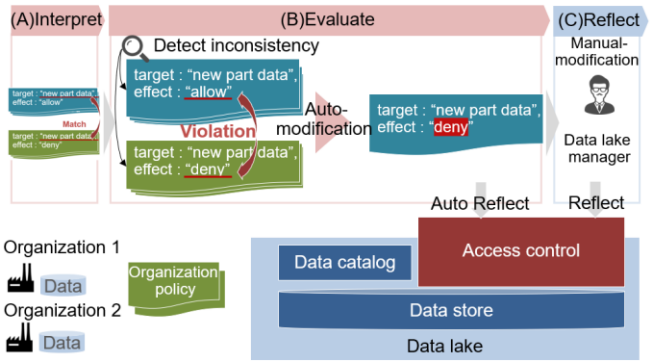

図 4 Automating evaluation function

図 3 を用いて具体例を説明する．データレイクポリシーでは対象のデータを「製造工程データ」というコンテキストで記述しており，アクセスを許可すると定義している．一方，組織ポリシーでは対象のデータを

「新部品のデータ」というコンテキストで記述しており，アクセスは許可しないと定義される．まず，組織ポリシーの「新部品のデータ」とデータレイクポリシーの「製造工程データ」のそれぞれが指すデータ（例えば，DB であればテーブルとカラム，ファイルであればファイルパス等）を，データカタログのメタデータから自動的に収集する.データレイクポリシーの「製造工程データ」が示すデータに，組織ポリシーの「新部品のデータ」が示すデータが含まれる場合，データレイクポリシーの対象のデータを,組織ポリシーの「新部品のデータ」のコンテキストで記述可能として解釈し，自動的に変換する．

4.2. ポリシーの自動評価機能

本機能は，ポリシーの自動解釈機能によって変換したデータレイクポリシーに対して，比較の対応関係にある組織ポリシーを抽出し，比較評価する．

変換され整合しているデータレイクポリシーと組織ポリシーの組合せの数の分，比較評価を繰り返し実施する．1 つ以上の組合せでポリシー間の不一致が判定された場合，当該の組合せにおける変換されたデータレイクポリシーを組織ポリシーと一致するように修正したポリシーを新たに生成し，これを元にデータレイクポリシーへ再度変換を行うことで組織ポリシーと整合したデータレイクポリシーに修正する．修正したデータレイクポリシーを一時的にアクセス制御コンポーネントに反映させることで，データ漏洩を防ぐことを可能とする．また，不一致であったデータレイクポリシーと組織ポリシーの組合せと自動修正内容を，アラートと共にデータレイク管理者に通知することで，データレイクポリシーの修正を促す．

図 4 を用いて具体例を説明する．変換されたデータレイクポリシーと組織ポリシーのコンテキストが揃っているため，アクセスの可否の定義部分を比較し，差異があるかどうかを判定する．変換されたデータレイクポリシーでは「可」である一方で組織ポリシーでは「否」とあるため，不一致があるとして判定する．そして，判定結果と組織ポリシーを元に該当の変換されたデータレイクポリシーのアクセスの可否の定義部分を「否」と修正したものを新たに生成して一時的にデータレイクポリシーとして反映する．

5. 提案手法の評価

5.1. 評価方法

本論文では，提案手法のポリシー管理技術を適用した際のリードタイムの削減効果について，実際のユースケースと比較し小規模なポリシー数での実験を行い所要時間を測定した．その後，測定結果を基に実際のユースケースを想定した試算値との比較を行った．

表 1 Measurement results of each process

Process	Average measured value	Value per combination	Scaled value
A) Interpret	1.29(min)	0.107(min)	0.11(day)
B) Evaluate	0.318(min)	0.0265(min)	0.0275(day)
C) Reflect	31.0(min)	3.87(min)	2(day)

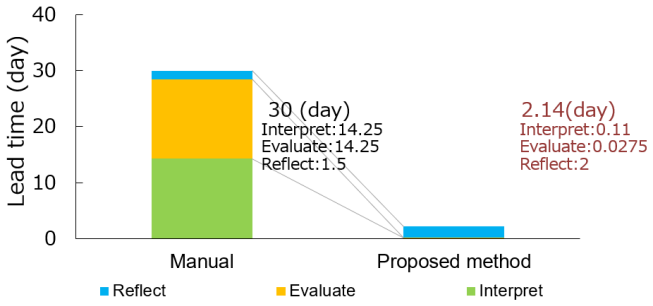

図 5 Evaluation result

【ユースケース】

3 章にて定義した A)-C)のプロセスからなる，2 組織間での組織横断のデータ利活用における整合性を考慮したポリシー管理のデータガバナンスをユースケースとした．

【前提/測定方法】

データレイク管理者相当の被験者 5 人を用いてポリシー管理の A)-C)プロセスのそれぞれでの所要時間を計測し，その計測値の平均を作業時間として算出した値を表 1 の一列目に示す．

データレイク管理者は，事前に用意した 3 種のデータレイクポリシーをアクセス制御コンポーネントに定義するとした．また，異なる 2 組織のそれぞれが定義する 2 種の合計 4 種の組織ポリシーがあり，それら組織ポリシーとの整合性を考慮したポリシー管理の A)-C)に対応する各プロセスを実施するとした．

従来では，特に A),B)において各組織のデータ管理者にヒアリングを繰り返し行い，ポリシーの解釈と評価を行った後でデータレイクポリシーの定義を行うと考えられる．そのため，データレイクポリシーが組織ポリシーに対して不整合となるケースは多くないと予想されるが，本評価実験では検証のため，多くの不整合が生じるケースとして，データレイクポリシーが組織ポリシーの約半数に対して不整合が起こる状態を前提とした．A)-C)の各プロセスにおいては，計測されたプロセス毎の所要時間を基に，ポリシーの 1 組合せ単位の作業時間として算出した値を表 1 の 2 列目に示す．また，C)のプロセスにおいては，3 種の内 2 種のデータレイクポリシーに不整合が生じるようにしており，この時，不整合のあるデータレイクポリシーと組織ポリシーの時の組合せ数は 12 通りの 2/3 となる 8 通りで

あり，この値から整合していない場合の 1 組合せ毎に必要な工数とした．

提案手法の評価を一般的な組織横断でのデータ利活用のユースケースをベースにして行うため，データレイクポリシーが 7 個，組織ポリシーが 100 個存在するとし，データレイク管理者の 1 日の作業時間を 8.25 時間と想定した．前述の想定から，各プロセス毎のデータレイクポリシーと組織ポリシーの組合せとデータレイク管理者の作業時間を基に実験での測定結果を日単位の作業時間に換算した値を表 1 の 3 列目に示す．

従来の A)-C)の各プロセス毎の作業工数の試算においては，A),B)のポリシーの解釈及び評価プロセスでは 1 組合せ毎に 10 分，ポリシーの修正には不整合のポリシーの組合せ毎に 20 分のリードタイムを所要するとして，前述のデータレイクポリシーと組織ポリシー数及び作業リードタイムの想定から，それぞれ 14.25 日を所要するとした．また，同様に C)のポリシーの反映には 1.5 日を所要するとした．

5.2. 評価結果

図 5 に，各プロセスの所要リードタイムに関する試算値の比較結果を示す．異なる 2 組織間での組織横断でのデータ利活用におるポリシー管理に関して，提案手法のポリシー管理システムの適用前後で A)-C)のプロセス全体の試算値のリードタイムの 93%の削減が確認できた．特に提案手法の特徴機能が対応する A)，B)プロセスにおいて，異なるコンテキストを持つポリシー間の解釈とポリシー間の評価の自動化により，従来のポリシー管理において問題であったリードタイムの削減に大きく寄与出来ていることが確認できた．一方で，C)のプロセスであるポリシーの反映に関しては従来と比較して所要時間が大きくなる結果となった．これは，被験者のフィードバックから各ポリシーの評価の結果と修正の必要な個所に関するユーザへの提示を改善する必要がある．システム内において異なるコンテキストのポリシーの解釈を行った部分をユーザが把握できる範囲で提示することが必要であることが確認できた．

6. 結論

スマートインダストリ向けの組織横断データレイク利活用ケースにおいて，セキュアなデータ利活用の実現のためのポリシー管理に関するリードタイム削減のため，ポリシー管理の支援が必要である．本論文では，アクセス制御の記述であるポリシーに着目し，異なる組織間で，異なるコンテキストを持つために不整合しているアクセス制御のポリシーをデータカタログで管理するメタデータを利用することで，相互に意味解釈と変換を行うことで整合させ，比較と評価を自動化するポリシー管理支援システムを提案した．提案システムでは，データ管理者および運用者のアクセス制御管理に関する工数の試算を行い，比較評価の結果 93%のリードタイム削減効果を見込んでいる．

本論文では，実際のユースケースと比較して小規模な環境での実験としたため，実際のユースケースに則した環境での実験検証が必要である．また，C)プロセスのポリシーの反映に関して，ユーザビリティを考慮した GUI 及び表示内容に関して引き続き検討していく必要がある．

参 考 文 献

[1] Craig A. Lee, "Cloud federation management and beyond: Requirements, relevant standards, and gaps.", IEEE Cloud Computing 3.1 (2016): 42-49.

[2] Ioram S. Sette, David W. Chadwick, and Carlos A.G. Ferraz, "Authorization policy federation in heterogeneous multicloud environments.", IEEE Cloud Computing 4.4 (2017): 38-47.

[3] Peng Zhao, Lifa Wu, Zheng Hong, and He Sun, “Research on multicloud access control policy integration framework.”, China communications 16.9 (2019): 222-234.

[4] Sabrina Kirrane, Alessandra Mileo, and Stefan Decker, "Access control and the resource description framework: A survey.", Semantic Web 8.2 (2017): 311-352.

[5] Federica Paci, Anna Squicciarini, and Nicola Zannone, "Survey on access control for community-centered collaborative systems.", ACM Computing Surveys 51.1 (2018): 1-38.

[6] 尾崎稜太, 飯島正, "組織間情報アクセス制御ポリシーのための FCA によるロール自動調整とルールの設計洗練化支援.", 情報システム学会 全国大会論文集 14 (2018): S2-D1.

[7] 芦野佑樹, 中江政行, "統合アクセス制御モデルの標準化について.", 第 73 回全国大会講演論文集 2011.1 (2011): 261-262.

[8] Rohirash Kumar Banyal, Vijendra Kumar Jain, and Pragya Jain, “Dynamic Trust Based Access Control Framework for Securing Multi-Cloud Environment.”, In Proceedings of 2014 International Conference on Information and Communication Technology for Competitive Strategies (2014): 1-8.

[9] Nikos Komninos, Aisha Kanwal Junejo, “Privacy Preserving Attribute Based Encryption for Multiple Cloud Collaborative Environment.”, In Proceedings of 2015 IEEE/ACM International Conference on Utility and Cloud Computing (2015):595-600.

[10] Mukesh Singhal, Santosh Chandrasekhar, Tingjian Ge, Ravi Sandhu, Ram Krishnan, Gail-Joon Ahn, and Elisa Bertino, “Collaboration in Multicloud Computing Environments: Framework and Security Issues.”, *Computer* 46.2 (2013): 76-84.

[11] Yuri Demchenko, Canh Ngo, Cees De Laat, and Craig Lee, "Federated access control in heterogeneous intercloud environment: Basic models and architecture patterns.", In Proceedings of 2014 IEEE International Conference on Cloud Engineering (2014): 439-445.