ハイブリッド・マルチクラウドにおける 非構造データの発見・取得容易化

早坂光雄† 鴨生悠冬† 野村鎮平†

1. はじめに

オンプレミスとクラウドを連携させたハイブリッドクラウド市場が成長している。ハイブリッドクラウド市場における主のユースケースに、オンプレミスにあるデータを，クラウドで AI やデータ分析を行うデータ利活用がある。クラウドは計算資源が豊富なため，特に計算資源を要する生成 AI の普及により，更に，ハイブリッドクラウドへの要望が高くなっている。AI やデータ分析は大量の非構造データであるファイルデータを使用する。そのため，ファイルデータのハイブリッドクラウドにおけるデータ連携容易化が必要である。

AI やデータ分析は，データ分析プロセスである①ビジネス理解②データ理解＆準備③モデル構築④評価⑤デプロイのサイクルを回すことで処理が進む。本処理の内，②データ理解＆準備にかかる割合が 5〜8 割と大きな割合を占めている[1][2]。データ理解＆準備は，データ発見→データ取得→データ理解→データ整形→結合→可視化のステップを踏む。ハイブリッドクラウドになると，データが他拠点に存在するため，最初のステップであるデータ発見とデータ取得が困難となる課題がある。

本稿では，ハイブリッドクラウド環境におけるデータ発見とデータ取得を容易化する方式を提案し，その有効性を示す。

2. 関連研究

他拠点のファイルを発見・取得する方式として，以下の 3 方式がある。

まず、従来方式#1 として、拠点内でクラスタを組む分散ファイルシステムを、拠点間へ拡張する方式がある。しかし，本方式は，ファイルやディレクトリの更新に伴うロックが拠点をまたいだ WAN 越しのグローバルロックになり性能が大きく低下する問題が発生する。

次に、従来方式#2 として、他拠点のデータを自拠点にあるようにファイルの存在だけをみせるファイルの仮想化技術を用いるものである[3]。データの実体は他拠点にあるままであり，仮想化されたファイルへアクセスすると，他拠点からデータを取得するものである。本方式だとデータがある拠点の性能への影響をかなり小さくできる利点がある。しかし、拠点数をが増えてくると，他拠点のデータを自拠点で存在だけみせる仮想化技術のメタデータが大量に増えて容量を消費してしまう。更にスタブファイルの数が多数となり大量のデータから所望のデータを発見することが困難になる。

最後に、データカタログ製品を使用する案がある。データカタログは内部にデータベースを持ち，各拠点のファイルデータのメタデータを本データカタログに集約することで、データ検索を行えるようにするものである。登録するレコードにファイルのロケーションが記載されており，ユーザがそのロケーションにアクセスしてデータを取得する。1 拠点にメタデータを集約する必要があるのと，データカタログという DB などの別のサーバ等が必要となり，コストが多くなる問題がある。

3. 課題

関連研究の議論から、以下が求められる。

課題#1 各拠点の性能低下を回避

課題#2 データ利活用拠点におけるファイルのメタデータ情報による消費容量を削減

課題#3 DB を使用せず、コストを削減

4. データ検索機能組込型ファイル仮想化方式

近年、非構造データを格納するストレージは，ファイルプロトコル(NFS/SMB)やオブジェクトプロトコル(HTTP)の両方をサポートする。これは、ファイルプロトコルで格納したデータを、オブジェクトプロトコル経由で AI やデータ分析にかける要望や，ファイルストレージとオブジェクトストレージを別々に用意し容量の分割損を回避するためやコスト低減などによる。

更に、近年のオブジェクトストレージは，オブジェクトストレージに格納された csv や parque などの半構造データを検索できる REST API を持つ。

そこで、図 1 に示すように、オンプレの各拠点のファイル仮想化が、自拠点に格納されたファイルのファイルリスト(半構造データ)を作成する。各拠点にあるファイルリストを、データ活用をしたい拠点からデータ検索 API で検索することでデータ発見を行い、必要データのみ、ファイル仮想化によりスタブファイルを作成することでメタデータの容量を低減する。本スタブファイルにアクセスすることで他拠点からデータ取得が自動で行われる。これらにより、

† (株)日立製作所 研究開発グループ
Hitachi Ltd., Research & Development Group

課題 1～3 の解決を行う。

提案方式の詳細を以下の述べる。

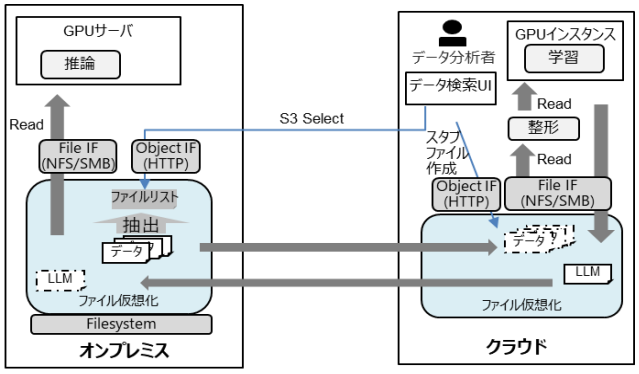

図 1　提案方式

Figure 1　Proposed Method

4.1　ファイル仮想化

ファイル仮想化は、各拠点のファイル・オブジェクトストレージに組み込まれており，全てのデータアクセスは本処理を経由してメディアに格納される。データが格納されるとファイル仮想化は，格納されたデータのファイルパスやファイル名、更新日時などを記載したファイルリスト(半構造データ)を作成する。ファイルパスは，http アクセスできる URL へ置き換えて記載する。

更に，スタブファイル作成パスと参照先 URL を指定する REST API を提供し，自拠点内に NFS や HTTP でのアクセスが可能なスタブファイルを作成する。スタブファイルは拡張属性に参照先データの URL を保持する。

4.2　他拠点データの発見

データ利活用をしたい拠点において、データ検索 UI を用意する。本 UI では，全拠点のファイル・オブジェクトストレージへ HTTP によるアクセスができるものであり，各拠点が用意したファイルリストへ検索を行う S3 Select を実行することで，データ検索を実現する。

4.3　必要なデータのみ取得

データ検索 UI で発見したデータは，UI 上でファイルデータごとに表示される取得ボタンを押すと、自拠点のファイルシステム内に、スタブファイルが作成される。そのスタブファイルには、他拠点データへアクセスする URL を拡張属性に持つ。本スタブファイルに、ファイルプロトコルやオブジェクトプロトコルでアクセスがあると、そのデータを他拠点から取得し、データの中身を確認することができる。更に学習したモデルファイルを同技術で他拠点に戻すことにより他拠点の推論に向けたデータ連携を容易化する。

5.　評価

100GbE につながった 3 台のサーバを用いて、提案方式を実装した。レコード数に対するデータ検索性能の結果を図 2 に示す。図 2 に示すように他拠点データの検索は登録ファイル数が増加しても変わらない性能を提供している。また、図 3 は 1 拠点のファイルシステムに格納されたファイルのメタデータ情報の実測に基づき、拠点数が増えた時の、データ利活用拠点で消費するスタブファイルのメタデータ量を見積もったものである。従来方式#2 では拠点数の増加に伴い消費するメデータデータ量が大きくなる。理由として、データ利活用拠点では、全拠点のスタブファイルを保持する必要があるからである。対して、提案方式は、自拠点のメタデータのみ保持し、他拠点ファイルのデータは必要なもののみ一時的にスタブファイルを作成するため、小さい容量消費で済む。

以上から提案方式の有効性を示した。

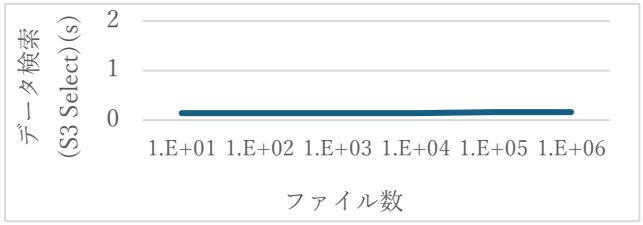

図 2　データ検索性能

Figure 2　Data search performance

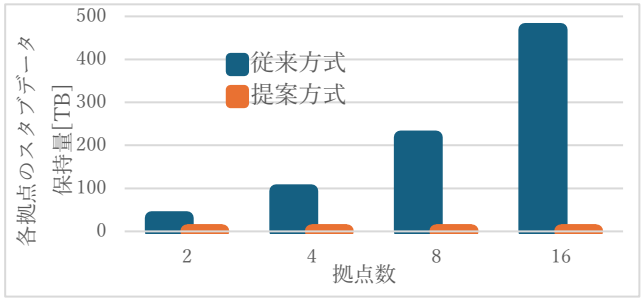

図 3　メタデータ消費量

Figure 3　Consumed capacity due to meta data

6.　おわりに

本研究は，ハイブリッドクラウドにおけるデータ利活用で必要となるデータ発見と取得容易化方式を提案し、その有効性を示した。

今後の課題は、実際のパブクラ環境における評価などが挙げられる。

参考文献

[1] IBM, “IBM SPSS Data Preparation”, 2012.4
[2] For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights, New York Times, 2014.8
[3]K. Matsuzawa, M. Hayasaka, and T. Shinagawa. 2020. Practical Quick File Server Migration. ACM Trans. Storage 16, 2, Article 13 (May 2020), 30 pages.