B-010

オブジェクトストレージ向け SQL アクセス方式の提案と評価
Proposal and Evaluation of SQL Access Method for Object Storage

大越 淳平† 近藤 伸和† 渡辺 聡† 馬場 恒彦†
Jumpei Okoshi Nobukazu Kondo Satoru Watanabe Tsunehiko Baba
†株式会社 日立製作所 研究開発グループ

1. 序論

近年，金融分野において改竄防止技術や冗長化技術を搭載したエンタープライズ向けオブジェクトストレージがアーカイブストレージとして活用されている[1]。オブジェクトストレージは，ディレクトリ構造でデータをファイルとして管理するファイルストレージと比較し，データサイズやファイル名の制約が緩い，HTTP 経由でデータの読み書きが可能であるなど，アーカイブ用途に適した特長を有している。

各金融機関は，蓄積されたアーカイブデータを，与信審査，マーケティング，不正取引監視などに活用している。例えば，与信審査においては，勘定系システムで生成された取引データをアーカイブストレージに格納し，格納された取引データと各取引主体（個人，企業など）の属性情報（年齢，年収，売上など）を組み合わせたデータ分析を行うことで取引主体の与信審査を行っている（図 1）。

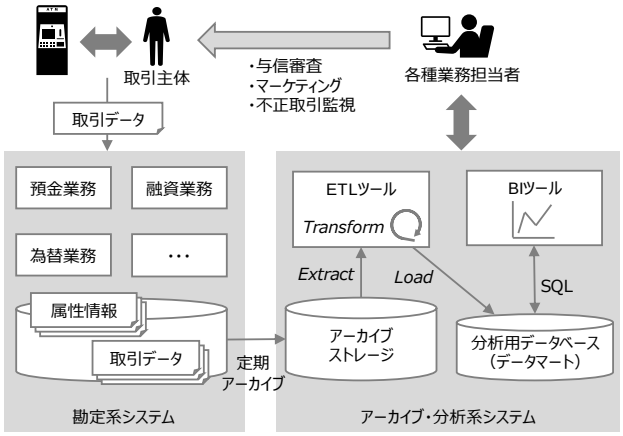

図 1 アーカイブデータの利活用例

しかし，アーカイブストレージとして用いられるオブジェクトストレージの API（Application Programming Interface）の提供機能は，基本的なデータの追加・削除にとどまっている。このため，BI（Business Intelligence）ツールと連携した高度なデータ分析のためには，ETL（Extract Transform Load）ツールと連携した ETL 処理により，データ分析の目的に応じたデータベースであるデータマートの構築が必要となる。データマートの構築は，一般に，データ分析においてデータ準備と呼ばれ，全工数の 6 割程度を占める[2]ことから，アーカイブデータ利活用におけるボトルネックとなっている。本研究の目的は，前述のボトルネックを解消し，低工数でオブジェクトストレージに格納されたデータの分析が可能なアクセス方式を提案することにある。

2. SQL アクセス方式の提案

SQL は，データ分析において最も汎用的に用いられる言語の一つであり[3]，SQL IF（InterFace）を備えることで幅広いデータ分析に対応できる。本研究では，BI ツールとデータマートの IF が SQL であることに着眼し，オブジェクトストレージに格納されたデータを，データマートの構築なしで SQL にてアクセス可能とするアクセス方式を検討した。

提案する SQL アクセス方式のシステム構成を図 2 に示す。提案システムは，オブジェクトストレージ，ディスクベース SQL エンジン，およびインメモリ SQL エンジンにより構成される。オブジェクトストレージと各 SQL エンジン間の IF は，オブジェクトストレージの IF においてデファクトスタンダードとなっている S3[1]を採用する。

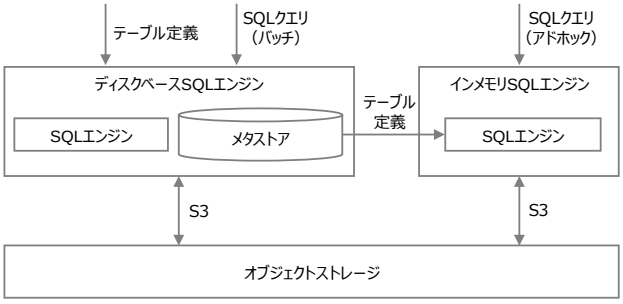

図 2 システム構成

本システムの特長として，ディスクベース SQL エンジンとインメモリ SQL エンジンの 2 つを備えていることが挙げられる。大容量のデータ処理が必要となるバッチ系の処理はディスクベース SQL エンジンに，応答性の要求されるアドホック系の処理はインメモリで高速にデータを処理するインメモリ SQL エンジンに，それぞれ割り当て

[1] https://docs.aws.amazon.com/AmazonS3/latest/API/

ることで各 SQL エンジンを使い分ける。これにより，幅広い SQL クエリに対応することが可能となる。オブジェクトストレージに格納されたデータの構造を定義するテーブル定義は，ディスクベース SQL エンジン内部のメタストアに格納され，インメモリ SQL エンジンと共有することで二重管理を避ける構成となっている。

3. フィージビリティ検証

本研究では，提案システムのアーカイブデータ利活用への適用を想定し，(1) エンタープライズ向けオブジェクトストレージを用いたシステム実装，(2) ベンチマークを用いた SQL コンプライアンス（各クエリに対する実行可能性），の 2 観点のフィージビリティを検証した。

(1) の実装に関しては，図 2 に示したシステム構成を日立製オブジェクトストレージである Hitachi Content Platform（HCP）[2]をベースに，ディスクベース SQL エンジンである Hive[3]とインメモリ SQL エンジンである Impala[4]の組み合わせにより実装し，単純な SQL クエリが実行可能であることを確認した。

(2) の SQL コンプライアンスに関しては，意思決定支援システム向けのデータベースにおける代表的なベンチマークである TPC-H[5]を採用し，表 1 に示す環境で検証を行った。その結果，TPC-H の全クエリが実行可能であることを確認した。

表 1 検証環境

CPU	Intel Xeon 3.60GHz
メモリ	8GB
TPC-H	1GB (ScaleFactor=1)

以上の検証結果より，エンタープライズ向けオブジェクトストレージと OSS（Open Source Software）ベースの SQL エンジンの組み合わせにより，アーカイブデータの利活用に代表される典型的なデータ分析に対応可能なシステムが構築可能であることを検証できた。

4. 評価

提案手法の評価として，金融機関における与信審査のユースケースを想定し，ETL ツールを用いてデータマートを構築してデータ分析を行う従来手法と，直接 SQL アクセスによりデータ分析を行う提案手法とで，分析工数の観点で削減効果を定量的に評価した。

データ分析は，CRISP-DM[4]の定義によれば，ビジネス・データ理解，データ準備，モデリング，評価・展開のプロセスをたどる。想定ユースケースの与信審査で用いられる勘定系システムは，一般に RDB で構築されている。このため，ETL 処理による構造化が不要であり，データ準備工数の大部分が単純なデータの移行によるデータマートの構築であると想定できる。以上の議論により，提案手法によりデータ準備工数の大部分を削減でき，最大でデータ分析工数の 60%を削減できると結論できる（図 3，各プロセスの工数は文献[2]による）。

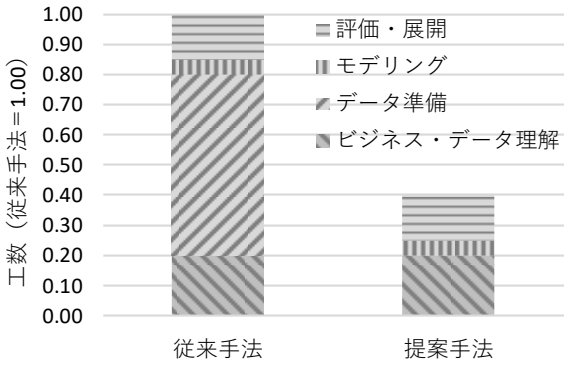

図 3 データ分析工数の削減効果

5. 結言

本研究では，オブジェクトストレージに格納されたデータを，データマートの構築なしで SQL によりアクセス可能とする SQL アクセス方式を提案し，実機を用いた実装と SQL コンプライアンスのフィージビリティ検証を行った。また，金融機関の与信審査を想定したユースケースにおいて，データ分析工数を 60%削減可能であることを明らかにした。

参考文献

[1] Hitachi Vantara, “Rabobank Leads with Hitachi Content Platform (HCP) - Case Study,” 2018.

[2] D. Pyle, “Data Preparation for Data Mining,” Morgan Kaufmann Publishers, Inc., 1999.

[3] Stack Overflow, “Developer Survey Results 2017,” https://insights.stackoverflow.com/survey/2017.

[4] SPSS, “CRISP-DM 1.0,” 2000.

[2] https://www.hitachivantara.com/en-us/products/cloud-object-platform/content-platform.html

[3] https://hive.apache.org/

[4] https://impala.apache.org/

[5] http://www.tpc.org/tpch/