データレイク

佐藤 華和子
Kanako SATO

1　はじめに

近年，AI や機械学習，IoT などが盛んとなっており，ビッグデータを扱う機会が増加している．ビッグデータとは，様々な種類や形式の膨大なデータであり，イメージやオーディオ，ビデオなどの非構造化データも含む．しかし，従来の行と列で管理するリレーショナルデータベース（RDB: Relational Data Base）では，非構造化データを取り扱うことが容易ではない．また，膨大なデータを取り扱うため，処理が遅く，多くの時間を要する．したがって，膨大なデータを蓄積し，分散処理により処理速度の向上が可能なデータ管理システムとして，データレイクに注目が集まっている．

2　データ管理

2.1　データウェアハウス

従来のデータ管理はデータウェアハウスで行うことが多い．データウェアハウスとは様々なデータを時系列に保管するデータベースである．データウェアハウスは RDB の一つであり，行と列でデータを管理する．しかし，行と列で管理するデータベースは非構造化データを扱うことが容易ではない．また，蓄積するデータ構造を要件に応じて事前に定義するため，定義していないデータの蓄積を行うことはできない．

データウェアハウスに蓄積可能なデータは蓄積を行う．次に，データウェアハウスに蓄積したデータから特定の利用目的に合ったデータのみ抽出し，データマートに蓄積し分析を行う．データマートとは，目的別に構築されたデータベースである．

2.2　データレイク

データレイクとはビッグデータを扱うためのストレージである．センサのログや GPS 情報など，様々なデータを生データのまま蓄積するストレージシステムである．また，利用者が必要なデータ形態に合わせて効率的に参照可能なストレージシステムである．データレイクは非構造化データを扱うことが可能である．しかし，データウェアハウスの代わりにデータレイクを利用するわけではない．Fig. 1 にデータレイクを用いたデータ管理方法を示す．

データレイクとデータウェアハウスの関係は，まず，データレイクに生データを蓄積しておく．次に，データを利用する際にデータレイクからデータウェアハウスへ必要なデータを抽出する．この時，データウェアハウスに対応するデータの形態でデータレイクからデータを抽出することが可能である．

データレイクは，データの場所や属性を示し，参照しやすいようにデータ蓄積時にデータ整備を行うという特徴がある．データ整備を行うために，メタデータとデータカタログを提供する．メタデータとは，格納データの作成者や作成日，ソースなどの情報を示したものである．データカタログは，データの格納場所を示したものである．また，SQL と類似の命令文で参照や抽出が可能であるという特徴もある．

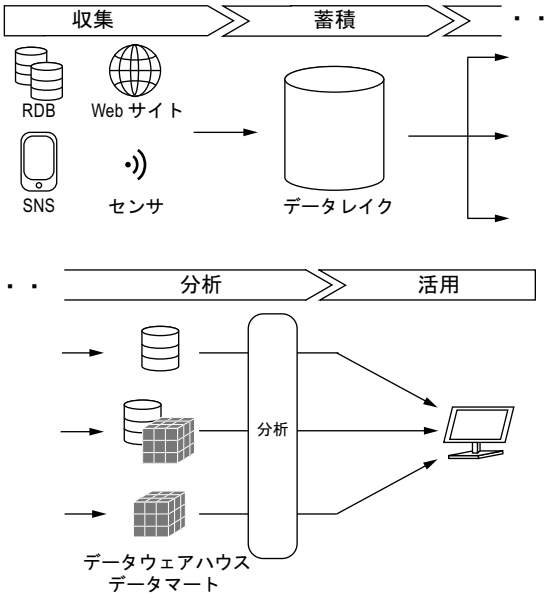

Fig.1　データレイクを用いたデータ管理

データレイクは概念であるため，実装方法はさまざまである．提供されているデータレイクのサービスのひとつに Microsoft の提供している Azure Data Lake がある．

3　活用事例

3.1　データ格納

Azure Data Lake は，Microsoft が提供しているデータカタログサービスである Azure Data Catalog と連携してデータ管理を行っている．格納したいデータが生成された場合，メタデータとデータ格納場所を Azure Data Catalog に登録する．メタデータの他に，自分で検索しやすいようにタグを設定することも可能である．また，格納データのサンプルを登録することも可能である．サンプルを登録することで参照時に実際のデータまでアクセスする必要がなくなるため，処理速度が向上する．

3.2 データ検出

データ検出には Azure Data Catalog に登録されている情報を用いる．検出方法は大きく分けて，キーワード検索と検索構文による検索がある．キーワード検索は，キーワードを入力するとすべてのプロパティにおいてキーワードを含んでいるものを検出する．また，検索構文による検索では，ブール演算子や比較演算子を用いることで，作成日での絞り込みや，複数のキーワードの検索などを行うことが可能である．

Azure Data Catalog を用いてデータ検出を行っているため，新しいデータの追加，データの内容の変更や格納場所の変更があった場合は適切には反映を行う必要がある．

3.3 データ抽出

Azure Data Lake は U-SQL という命令文を用いてデータ抽出を行うことが可能である．U-SQL とは，ビッグデータ処理言語であり，あらゆる規模のデータ処理を可能にするために，宣言型の SQL と命令型の C#を組み合わせた言語である [1]．リスト 1 に，U-SQL のコードを示す．

リスト 1　U-SQL

```
1.  @searchlog =
2.   EXTRACT UserId   int,
3.           Start    DateTime,
4.           Region   string,
5.           Duration int?
6.   FROM "/Samples/Data/SerchLog.tsv"
7.   USING Extractors.Tsv();

8.  @rs1 =
9.   SELECT Start, Region, Duration
10.  FROM @searchlog
11.  WHERE Start >= DateTime.Parse("2012/02/16");

12. OUTPUT @rs1
13.  TO "/output/SearchLog-transform-datatime.csv"
14.  USING Outputters.Csv();
```

SearchLog.tsv から EXTRACT という命令文を用いて，ユーザ ID を int 型，開始日時を DateTime 型，地域を string 型，というようにスキーマを定義して抽出し，searchlog に格納している．次に SELECT 命令により Start，Region，Duration の列を抽出を行う．SELECT 命令で抽出されたデータの中からと WHERE 句により Start が 2012/02/16 以降のデータを抽出し，rs1 に格納している．最後に，OUTPUT 命令により抽出したデータを SerchLog-transform-datatime.csv に出力する．データの出力先は，AzureDataLake の場合，通常 Microsoft が提供しているデータウェアハウスのサービスである Azure Data Warehouse である．以上の EXTRACT, SERECT, WHERE により，Azure Data Lake に格納されているデータを抽出し変換してデータウェアハウスに出力している．

4　データレイクのメリット

データレイクを用いるメリットとして，データの一元管理が可能であることと，分散処理により処理が高速化されていることを挙げることができる．

従来，各基幹システムがデータを個別に蓄積していた．データを利用する際は，各基幹システムにアクセスする必要があり，手間となっていた．しかし，データレイクを利用することで，すべてのデータがデータレイクに蓄積されているので，データレイクにアクセスするだけでデータを収集することが可能となった．

また，データレイクは膨大な量のデータであるビッグデータを扱うために処理速度が必要である．処理速度向上のために分散処理を用いてデータレイクを構成することが一般的である．したがって，分散処理のフレームワークである Hadoop を利用して構成することが主流である．

5　クラウドサービス

データレイクの構築に必要な機器を揃えるには，PB（ペタバイト）クラスの大量のデータを保存するストレージや，分散処理を行うことができる大量のサーバなどが必要となる．これらを実現するには膨大な資金や時間，専門的技術が必要となる．クラウドサービスのクラウドストレージを利用することで安価に大量のデータを保存可能になる．

6　今後の展望

今後ますますビッグデータの市場は拡大していくと考えられる．ウェアラブル端末のブームがあり，IoT が拡大することが予想されているからである．家電や車，住宅など様々な身の回りのものがインターネットにつながり，やり取りされる情報が激増すると考えられる．

また，今後多くの企業がビッグデータ分析を行い，企業の利益につなげていくと考えられる．その中でも，特にマーケティングへの活用が期待できる．従来は，購買データ分析では今まで何が売れたか，どれだけ売れたかなどの分析は行えたが，なぜ売れたのか，今後どれくらい売れるのかという分析を正確に行うことは容易ではなかった．これまで活用していた購買データや顧客データのほかにも，脳波やその日に合ったイベントなど様々なデータを分析して，購買動機の理解や行動予測を行い，効率的な在庫管理，商品開発などに活用することを目指す [2]．

ビッグデータ分析が一般化する中で，データレイクが有効活用することが期待される．

参考文献

1) U-SQL を使ってみる—Microsoft, https://docs.microsoft.com/ja-jp/azure/data-lake-analytics/data-lake-analytics-u-sql-get-started ，参照 Apr.26, 2018
2) 夢展望、ＡＩ・ビッグデータ技術活用の最先端マーケティング技術開発でシナジーマーケティング社と協業—MORNINGSTAR, https://www.morningstar.co.jp/msnews/news?rncNo=1799147 ，参照 Apr.26, 2018