長谷川隆之*
Takayuki Hasegawa

データ活用基盤と構成例

Platform for Effective Data Utilization and Configuration Example

要　旨

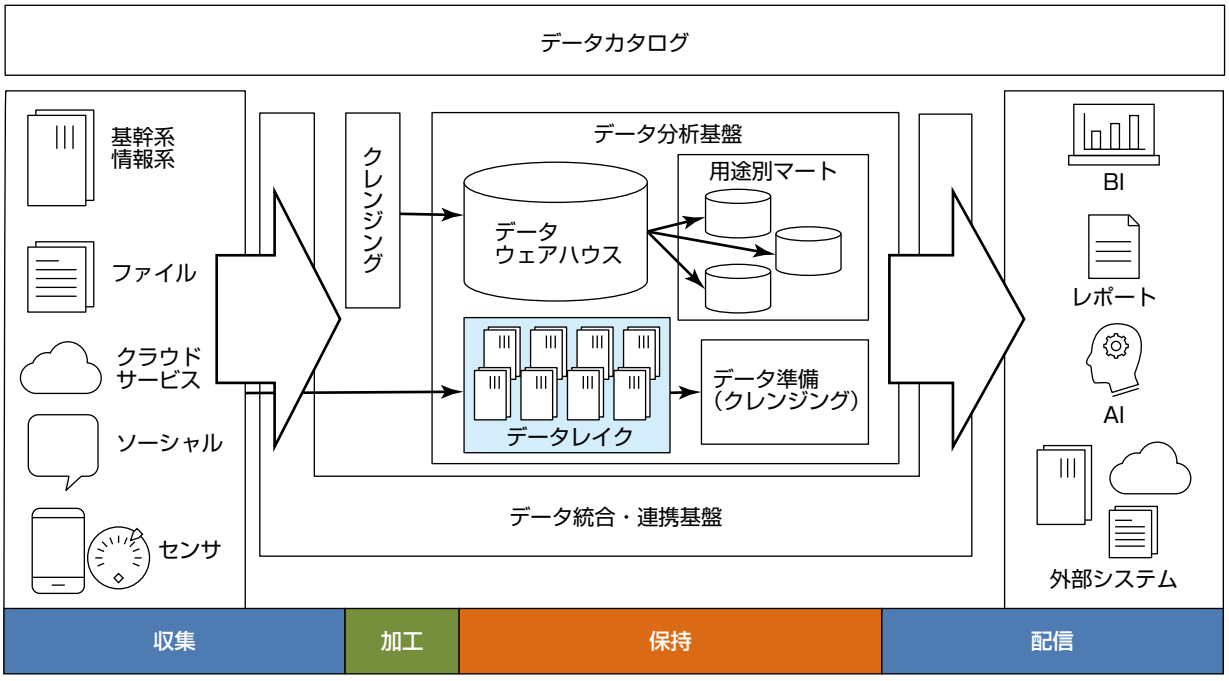
データは“21世紀の石油”とも言われるように，その利活用方法が企業や国の発展の鍵を握ることになる。しかしながら，単にデータを大量に集めることに価値はなく，集めたデータを分析し，様々な情報や知識を創出することで初めて価値が生まれる。また，近年では情報システム部門の特定の人だけではなく，実際の業務に携わる幅広い部門の人が必要な時に速やかにデータを活用できることが望まれるようになってきている。

データの種類と量は年々増えており，データを効率良く収集・分析し，活用することが重要である。

三菱電機インフォメーションネットワーク㈱(MIND)では，データ統合・連携や，データ分析を行う製品を20年近く販売・サポートしており，これらの製品を用いたシステム構築を280件以上(2019年末時点)行ってきた。

この経験から得た結論は，データの活用を促進するためには，①データの意味が明確であり，②データがどこから来たものかはっきりしており，③データの品質が良いことが必要であるということである。これら３点が満たされないと，せっかくデータを収集しても有効に活用できない。

データ活用を促進するためのこれら三つの必要事項を実現するために，MINDでは適切な製品やサービスを選定し，顧客に最適なデータ活用のためのインフラとして“データ活用基盤”を提供する。



BI：Business Intelligence，AI：Artificial Intelligence

MINDが提供する“データ活用基盤”

MINDでは，顧客のデータ活用を促進するためのインフラとして，データ活用基盤の提供を行っている。データ活用基盤は，大きくデータの収集から加工，配信を行うデータ統合・連携基盤と，大量のデータを保持して分析するデータ分析基盤から構成される。

*三菱電機インフォメーションネットワーク㈱

1. まえがき

全世界で一年間に生成されるデータ量は2018年の約33ゼタバイトから2025年には約175ゼタバイトに達すると予想されている。

企業内には多くのシステムがあり，各システムが様々なデータを持ち，日々利用されている。また，SNS(Social Networking Service)などでの自社の評判，自社製品やサービスに対する評価，機器のセンサログなど日々発生するデータも企業にとって重要なデータになる。これらを収集して分析し，マーケティングや製品開発・改良，また故障予知・異常検知などのアフターサービスに活用する取組みが広がっている。

データを利活用するためには適切な方法でデータを保持し，必要な時にいつでも利用できる状態にしておくことが肝要で，本稿ではこのような仕組みを“データ活用基盤”と定義する。本稿では，データ活用基盤に求められる要件とMINDでのデータ活用基盤の実現方法について述べる。

2. データ活用

2.1 データの重要性

DMBOK(Data Management Body Of Knowledge：データマネジメントの知識体系)[1]によれば，“データ”とは様々な形式での事実の表現のことで，データが定義，形式，時間枠，関連性などのコンテキストを備えると“情報”になる。さらに認識やパターンの解釈に基づく統合した視点で情報を見ると“知識”になるとしている。データは情報や知識の源泉であり，品質の高いデータを保持して活用することが重要である。

2.2 データ活用とは

データ活用とは，データから有益な情報や知識を得ることである。データを収集・分析し，分析結果を基に仮説を構築・検証する。このサイクルを繰り返すことで，経営判断，業務改善，効果的なマーケティングや製品開発に役立つ有益な情報や知識を得ることができる。

データは日々の企業活動で数多く発生している。これらを収集・分析することで，企業活動のプラスになる情報や知識を得ることが企業にとってのデータ活用である。

2.3 データ活用の課題

データを分析することで有益な情報や知識を得ることができるが，データを収集して分析可能な形に加工するまでのデータの準備には一般的に多くの時間と手間がかかる。収集から分析完了までに費やす時間の80％が分析前に使われているとも言われている。データの準備に時間がかかる理由を次に列挙する。

(1) どこに必要なデータがあるか不明

分析に必要なデータがどのシステムのどの項目か分からない，複数のシステムに類似の項目があるためどれを選択するのが最良なのか分からないといったことが多く発生する。どちらも各データ項目が何を意味しているかを示す仕様書がすぐに参照できない，又は存在しないことが原因である。

(2) データの出どころが分からない

システム間で統合されているデータを用いて経営判断に関わるレポートを作成するような場合に，判断材料になるデータがどこから来たかものなのか分からないとそのデータの正当性を担保できない。

(3) データが不正確・不完全

オペレータがデータ入力する場合は，意図せず誤ったデータを入力してしまうこともあるし，処理を進めるために適切なデータでないことを承知でダミーデータを入力することもある。また，センサから出力されるデータ等では，機器との通信の状況によってデータが欠損する，異常値が記録されるケースも発生する。このようなデータをそのまま用いて分析を行うと，誤った結論を導き出してしまう可能性がある。

2.4 課題解決方法

(1) データ項目の意味を明文化する

各データ項目の意味を第三者にも分かるように，平易な表現で定義し，システム全体での定義を集めてデータ項目辞書を作成する。その際，システム横断で，重複がないか，定義文書の詳しさの程度にずれがないかなどをチェックすることが重要である。

(2) データの出どころを明らかにする

データ統合・連携を行う製品には，データ項目ごとにそのデータがどのシステムのどのテーブルに由来するものか確認できる機能がある。この機能を利用することでデータの正当性を担保できる。

(3) 不正確・不完全なデータをなくしてデータの品質を高める

データを分析用にデータウェアハウス(DWH)に取り込む前にデータクレンジングや標準化，マッチングなどによってデータ品質を高める。実際のデータは，必須入力項目に正しい値が入っていない，入力形式が守られていない，同じデータが重複している等の問題を持つことが多い。多くはデータ入力者のミスなどが原因だが，入力ミスをさせ

ない仕組みや日次のバッチなどでデータ不備を自動修正できることが望ましい。

3. データ活用基盤に対するMINDのアプローチ

MINDは，顧客が目指すデータ活用を実現する最適なデータ活用基盤のためのプラットフォームと製品・サービスの提供や，その構築・導入を行っている。また，導入したデータ活用基盤が長期間安定して活用できるように維持するための運用やサポートの提供も併せて行っている。次に，データ活用基盤実現のためにMINDが行うことを述べる。

3.1 プラットフォームと製品・サービスの提供

⑴ プラットフォーム

MINDでは，サーバ，ストレージ，ネットワーク等のプラットフォーム製品や，“CloudMinder”に代表される自社クラウドサービスの販売・提供を行っている。また，プラットフォーム製品やクラウドサービスを顧客のニーズに合わせて選定・構築を行い，多数の導入実績も持っている。このノウハウを活用し，最適なプラットフォームを提供する。

⑵ データ分析基盤の製品・サービス

MINDでは，データ分析基盤の製品として高いデータ圧縮率を実現し，高速な検索が可能なDWH製品である“AnalyticMart”を自社で開発・製造・販売している。日本国内で1,200社以上の利用実績があり，顧客分析やログ分析等の様々なデータ分析で活用されている。

⑶ データ統合・連携基盤の製品・サービス

MINDでは，データ統合・連携，管理分野での業界リーダーであるインフォマティカ社との協業によって，各種データ統合・連携，管理を行う製品・サービスの提供を行っている。社内外のシステムからデータを集め，加工・整形し，データ分析基盤へデータを投入するためにインフォマティカ社のInformatica PowerCenter(注1)を採用し，2000年から日本国内で販売，サポート，導入コンサル，システム構築を提供してきた。これまでに280社以上に導入し，様々な環境上でデータ統合・連携基盤を構築した実績がある。

（注１） Informatica PowerCenterは，Informatica Corp.の登録商標である。

3.2 課題の解決

2. 4節でデータ活用の課題解決方法として，“データ項目の意味を明文化する”“データの出どころを明らかにする”“不正確・不完全なデータをなくしてデータの品質を高める”の３点を挙げた。これらの課題のデータ活用基盤での解決方法を述べる。

⑴ データ項目の意味を明文化する

データ項目の意味を統一された平易な言葉で記述し，これらを集約してビジネス用語辞書を作成する。MINDでは，ビジネス用語辞書のため，インフォマティカ社のPowerCenterや Enterprise Data Catalog(EDC)等を利用してデータの見える化を支援する。

⑵ データの出どころを明らかにする

PowerCenterでは，専用のGUI(Graphical User Interface)ツールを使ってデータの読み込み元(ソース)からデータ配信先(ターゲット)へのデータの流れを定義することで，データの統合・連携を実現する。PowerCenterで構築したシステム間のデータの流れは，マッピング情報として保管される。この情報を利用することで，データの出どころを明確にできる。

⑶ データの品質を高める

MINDがデータの品質向上のために利用するインフォマティカ社の Informatica Data Quality(DQ)(注2)では，次に代表されるデータ品質向上のための各種機能が提供される。

①想定外の形式や未入力項目を持つデータの割合を調べる(データプロファイリング機能)。
②重複データ，不完全データ等を抽出して自動修正する。
③データ修正要否の判断をデータ管理者に照会する。
④①～③の処理をバッチに組み込んで定期的に実行する。

（注２） Informatica Data Qualityは，Informatica Corp.の登録商標である。

3.3 導入後の運用

構築したデータ活用基盤は定期的なメンテナンスが重要である。データ量の増大に応じて，CPU，メモリ，ストレージなどのハードウェアリソースを適時増強し，利用製品のパッチ適用や更新を行うことで，セキュリティリスクを低減し，システムの安定稼働を実現する。また，導入製品固有のメンテナンスも必要になる。例えば，PowerCenterでは処理性能を維持するためにリポジトリ内のログの切り詰めやリポジトリで使用しているデータベースの統計情報更新が必要である。これらを定期的に実施する。

さらにDQの機能によって，自動的にデータの重複排除，規則外データの修正，不完全データの補完を行い，データ活用基盤が保持するデータの品質を維持する。

3.4 主 要 製 品

MINDでは，AnalyticMartやインフォマティカ社の製品を用いて，データ活用基盤を実現し，顧客の経営判断のスピードアップや，業務効率向上を支援している。MINDのAnalyticMartとインフォマティカ社のPowerCenter,

Enterprise Data Catalogについて述べる。

⑴ AnalyticMart

AnalyticMartは，大量のデータを圧縮してコンパクトに保管し，高速に検索・集計が可能な分析用データベース(DWH)製品である。規模に応じたスケーラビリティを持っており，標準のインタフェース(ODBC(Open DataBase Connectivity)，JDBC(Java DataBase Connectivity))によってBIツール等の各種製品やシステムとの連携が可能である。ストレージ容量の削減，データ規模に応じたライセンス等，コストパフォーマンスに優れた製品である(**図１**)。

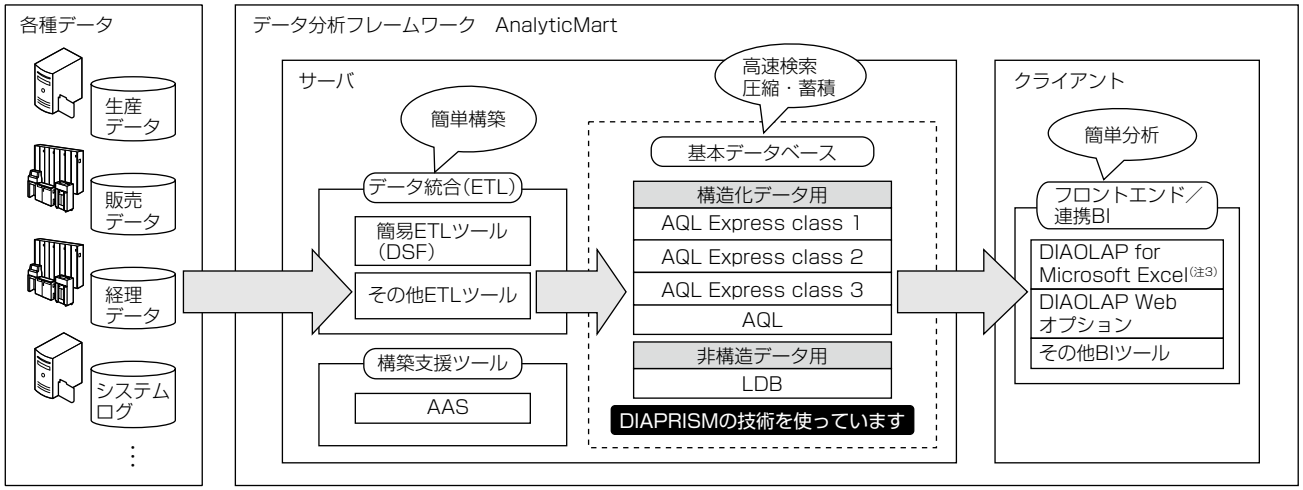
⑵ Informatica PowerCenter

Informatica PowerCenterは，データベース，ファイル，クラウドサービスなどの様々なデータソースからデータを抽出・加工・ロードするデータ統合(Extract, Transform, Load：ETL)製品である。GUI画面から簡単にデータ統合のためのデータ連携(マッピング)を開発できる。Oracle(注4)，SAP(注5)，Salesforce(注6)等の代表的なシステムやクラウドサービスとのコネクタが用意されており，シンプルにシステム間の連携を実現できる。大量データに対応するスケーラビリティや統合先追加に容易に対応できる拡張性も備える(**図２**)。

⑶ Informatica Enterprise Data Catalog

Informatica Enterprise Data Catalogは，ビジネスユーザーのデータ利活用拡大のために，データの所在，意味，鮮度，管理者等の情報を持ったカタログを作成して管理するためのツールである。各種データベース，ファイル，クラウドサービス，BIツールなどからメタデータを収集・整理し，利用者にカタログ情報を提供する。データ検索，データの系統図の提供，複数データの関連度の提示等の機能を持つ(**図３**)。

(注４) Oracleは，Oracle International Corp.の登録商標である。
(注５) SAPは，SAP SEの登録商標である。
(注６) Salesforceは，Salesforce.com, inc.の登録商標である。



規模，用途に合わせて必要なツールを選択・追加。基本データーベースの選択は必須。セット製品あり。

(注3) Excelは，Microsoft Corp.の登録商標である。
DSF：Data Staging Facility，AAS：AQL Administration Assistant，AQL：Analytical Query Language，LDB：Log DataBase

図１．AnalyticMartの概念図

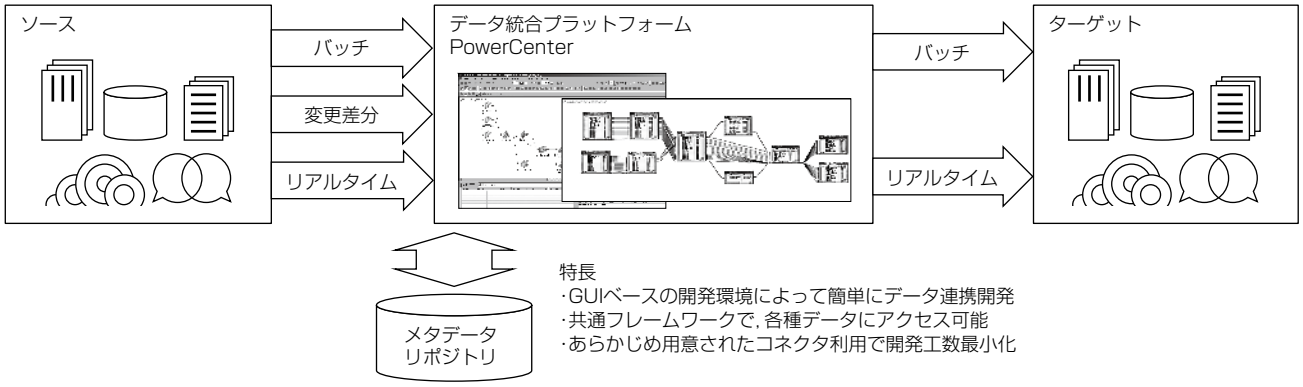
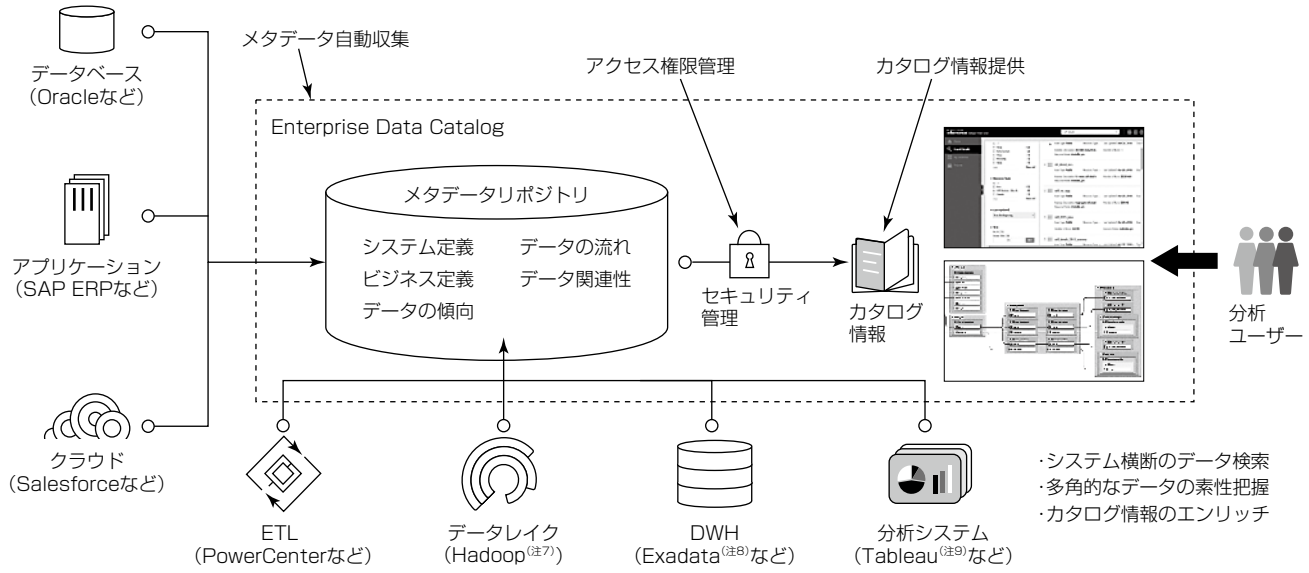


図２．PowerCenterの概念図



（注7） Hadoopは，The Apache Software Foundationの登録商標である。
（注8） Exadataは，Oracle International Corp.の登録商標である。
（注9） Tableauは，Tableau Software Inc.の登録商標である。

図3．Enterprise Data Catalogの概念図

4．データ活用基盤の構成例

顧客の社内外システムのデータ統合にPowerCenter，データ品質維持にInformatica Data Quality，構造化データ分析用DWHとしてAnalyticMartを用いたデータ活用基盤の構成例を**図4**に示す。この構成例のように，最近ではデータ活用基盤をクラウド上に構築するケースも増えてきており，MINDではクラウドの提供や構築も合わせて提供を行っている。

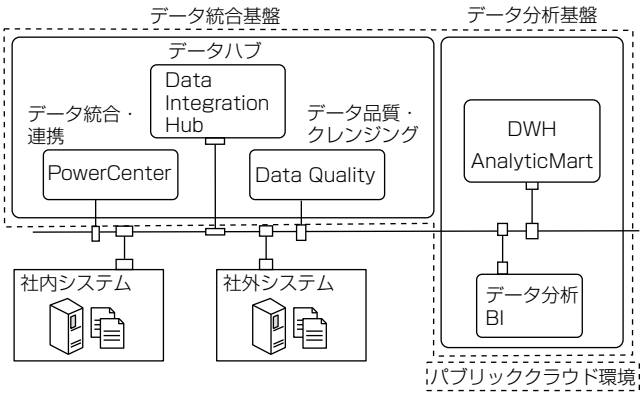


図4．データ活用基盤の構成例

5．む　す　び

近年注目されているデータ活用のためのインフラとしてMINDが提供しているデータ活用基盤の機能や構成例について述べた。

MINDでは，クラウド，データセンター，ネットワークや各種プラットフォーム製品等のITインフラ製品やサービスの提供を行ってきた。また，20年以上にわたり自社製データ分析製品AnalyticMartや，インフォマティカ社のデータ統合製品の販売，構築，サポート等のデータ活用ソリューションを展開してきた。

これら実績をベースに，近年高まるデータ活用ニーズに向けて，データ活用基盤実現のための製品・サービスの強化や拡充を今後も図っていく。

MINDは，データを活用して新たな価値や事業を創出するためのインフラソリューションの提供によって，顧客のデジタルトランスフォーメーション(DX)に貢献していく。

参 考 文 献

(1) DAMA International，ほか：データマネジメント知識体系ガイド 第二版，日経BP社（2018）