奈良産業大学『産業と経済』第13巻第3号（1998年12月）51—60

≪翻訳≫

バーチャルデータベース技術

Junglee Corporation, "Virtual Database Technology", whitepaper, Jan.30, 1998

杵崎のり子
下崎千代子

1. 序

仮想記憶（Virtual Memory; VM）技術は，コンピュータのディスク装置をメインメモリ拡張に活用するために開発された。仮想記憶が出現する以前は，ほとんどのプログラムはコンピュータのメインメモリの物理的な制約を受けていた。高度な技術を持ったプログラマたちはこの限界近くでプログラムを作動させるために，手の込んだ仕組み——オーバレイ（overlay）やチェインニング（chaining）——を利用した。これらの仕組みはプログラムの論理を複雑にするばかりでなく，エラーを起こしやすい，移植性が悪い，複雑なアプリケーションの開発時間を長びかせるという問題をもっていた。

仮想記憶は，コンピュータアプリケーション作動時，ほとんど無限に提供可能でかつ安価なメモリをメインメモリに置き換えることによって，根本的にコンピューティング技術を一変させた。ほとんどのコンピュータには，メインメモリの容量に限界がある。しかしディスク上の記憶容量の限界は，事実上，無いようなものである。なぜなら補助記憶装置はメインメモリと比べて，けた違いに安いからである。アプリケーションプログラマは，今や，このような記憶装置すべてをコンピュータの物理メモリの一部であるかのように扱うことができる。このように仮想記憶という技術開発は，従来と比べてはるかに複雑なプログラム開発を可能としたのである。

同様に，バーチャルデータベース（Virtual Database; VDB）技術では，外部データ——ワールドワイドウェブ（World Wide Web; WWW）のような——を企業のリレーショナルデータベースシステム（Relational Database; RDBMS）の一部として取り込むことができる。いくつかの推算によると，世界中のデータの90%ほどはリレーショナルデータベース以外のものである。生きたデータはウェブ上やファイルシステム，データベースシステム，従来からのアプリケーション上にばらまかれている。これらのデータ資源はデータの編成方法，使用されている言語，データアクセス方法が異なっている。その多くは素朴な問合せ操作さえもサポートされていない。これらのデータ資源上のデータを結合するためのアプリケーションを作成することは，そ

の異質性ゆえに，複雑であるばかりか，不可能な仕事となっている。

ジャングリー（Junglee）社のＶＤＢ技術は，この“データ分散”問題に対する一つの解決策を提供し，それによって，根本的に企業のコンピューティング方法とワールドワイドウェブを一変させる。ＶＤＢ技術は，多様なデータ資源上に散らばるデータに対して，アプリケーションからの強力な問合せ（query）を可能とする。ＶＤＢはこれらの異種のデータ資源からデータを集め，構造化し，まとめあげる。そして，アプリケーションプログラマに一つの統合的なリレーショナルデータベースシステムを提供する。ＶＤＢ技術はインターネットを企業のアプリケーションで活用する。さらに，それらの“すべての”データを使うことのできる新しい種類のアプリケーション開発へと導いてくれる。

ＶＤＢ技術によって可能となったアプリケーションの実例として，ウェブ上の求人があげられる。職業選択を意義あるものとするために，求職者は，求職機会だけではなく，関連データ――たとえば勤務区域内の住宅，通学区，犯罪発生率に関する情報のような――についての情報をも必要としている。求人情報は，数多くの異なったウェブサイト――会社のホームページ，新聞の広告サイトに代表される集合サイト――上に散らばっている。そこでは，求人リスト内の用語に対するキーワード検索が唯一利用可能な検索方法なのである。

ＶＤＢ技術を用いたアプリケーションを利用することによって，今や求職者はウェブに対して，“サンランシスコから15マイル以内にあり，株価が過去３年間以上少なくとも25％の成長率のある企業で，マーケティングマネジャの職を見つけよ”という問合せを投げかけて，その回答を得ることができる。この簡単な問合せは，多くの企業のウェブ求人リスト，地理的な地図情報を持ったウェブサイトや，企業の公正な価格についての過去の記録を含むウェブサイトにもおよぶ。その問合せによって，住宅価格，校区，犯罪発生率の統計を含めた関連情報の回答をも，それぞれの職（position）ごとに得られるようになっている。

ＶＤＢ技術は特にデータウェアハウスと結びつくと，企業にとって特に有効なものとなる。ＶＤＢ技術を利用することによって，企業は通常利用できない資源（たとえばディレクトリシステム（directory system）のファイル）や外部資源（たとえばワールドワイドウェブ）をデータウェアハウス内に取り込むことができ，このことは重要な意思決定支援アプリケーションの作成を可能とさせる。

2. 技術アーキテクチャ

2.1 バーチャルデータベース

ジャングリー社独自の特許出願中であるＶＤＢ技術は，異種異質な情報資源を一つのリレーショナルデータベースシステムのごとくに動かせる。図１は，後ほど“書籍ＶＤＢ”として紹介する簡単なＶＤＢの動作時の様子である。このＶＤＢは２つの書店（アマゾン・コム社（Amazon.com）とパウエルズブックス（Powell's Books））とニューヨークタイムスブックレビュー

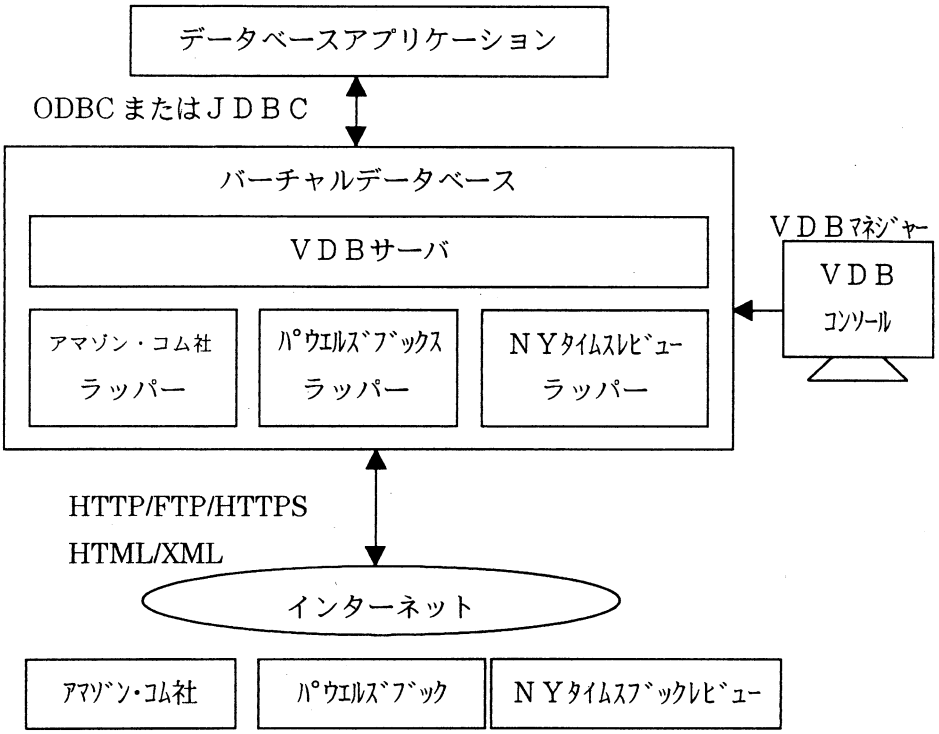

図１　ＶＤＢの動作時の様子

(New York Times Book Reviews) のコンテンツを統合させて，“書籍”と“書評”という２種類のテーブルからなる統合スキーマ (unified schema) を提供している。データベースアプリケーションはこの統合スキーマ上で動き，ＪＤＢＣあるいは ODBC API によるＳＱＬ問合せをおこなう。アプリケーションそのものはデルフィ (Delphi)，パワービルダー (PowerBuilder)，ビジュアルベーシック (Visual Basic)，あるいは類似したジャバツールキット (Java toolkits) などのような標準的なＲＡＤツールを使って作られている。

ＶＤＢはＶＤＢサーバを通してアクセスされる。また，ブラウザベース (browser-based) のＶＤＢコンソールによって管理されている。ＶＤＢには，個々の外部データ資源ごとに，これらのデータ資源とＶＤＢサーバとを結びつける“ラッパー (wrapper)”を内蔵している。ラッパーはウェブサイトのような勝手な形式の外部データ資源をある一つの RDBMS のように動作させる一方，ＶＤＢサーバはこれらの別々のリレーショナルデータベースをひとつのＶＤＢに統合する。

図２は，動作時の個々のラッパーを表している。ラッパーは大概ＨＴＴＰやＨＴＭＬを使って，ウェブサイトと接触している。ラッパーはフォーム (forms)，クッキー (cookies)，認証 (authentication) のようなＨＴＴＰプロトコルに関連した問題に対処している。ラッパーは，JDBC API によってアクセスされることから，クライアントは JDBC API を用いてＳＱＬ問合せをおこなう。この場合，ラッパーにむけて出されたＳＱＬ問合せは，結局，アマゾン・コム社ウェブサイトのＨＴＭＬフォームに書き込みをして，目的とするＨＴＭＬのページをナビ

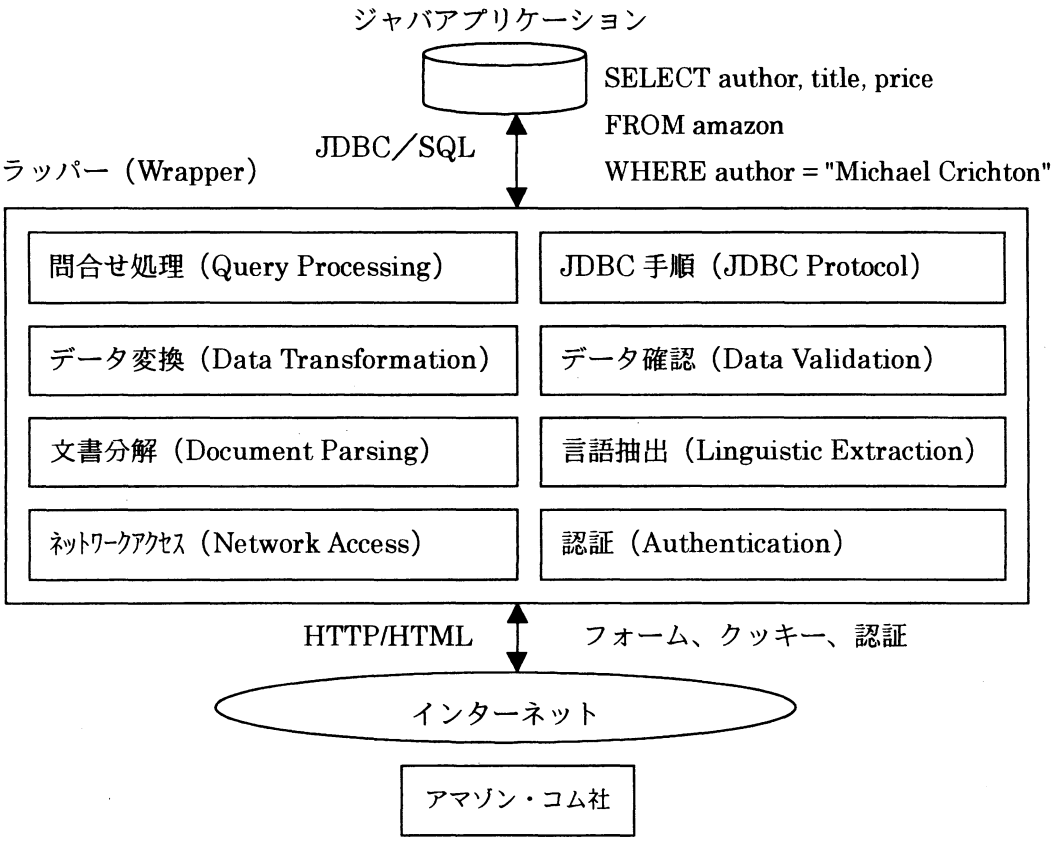

図２　動作時のラッパー

ゲートし，分解する。そして，そのデータをひとつのリレーショナルテーブル中の行に変換する。そのラッパーはウェブページから属性を抽出するために，複雑高度な言語処理（linguistic processing）を適用した“抽出ルール”を使い，スキーマにあわせてデータ変換をしたりデータを構造化するために“データ変換”ルールを使う。そしてデータの完全性を保証するために“データ確認”ルールを用いている。

図２はひとつ（または２，３の）データ資源と相互作用する簡単なジャバアプリケーションを表している。そのアプリケーションから見ると，おのおののデータ資源は独自のスキーマをもった全く別々のＪＤＢＣ資源に見える。そこでアプリケーションは個々の資源と個別に連携し，必要に応じてデータを結合しなければならない。

より多くのデータ資源を必要とする複雑高度なアプリケーションはＶＤＢサーバを利用する。図３はその一例である。ＶＤＢサーバは複数のデータ資源内のテーブルを一つのバーチャルデータベース（ＶＤＢ）上の“仮想テーブル（virtual tables）”として展開をする。ＶＤＢサーバは仮想テーブルに対して RDBMS のすべての機能をサポートしている。資源を横断して生成するビュー定義や問合せをも含めてである。図３の例においては，ＶＤＢは“アマゾン（amazon）”と“バーンズ&ノーブル（barnes & noble）”の仮想テーブルを結合したものをビュー“books”と定義している。ＶＤＢサーバが図に示されたような問合せを受け取ると，“問合せプ

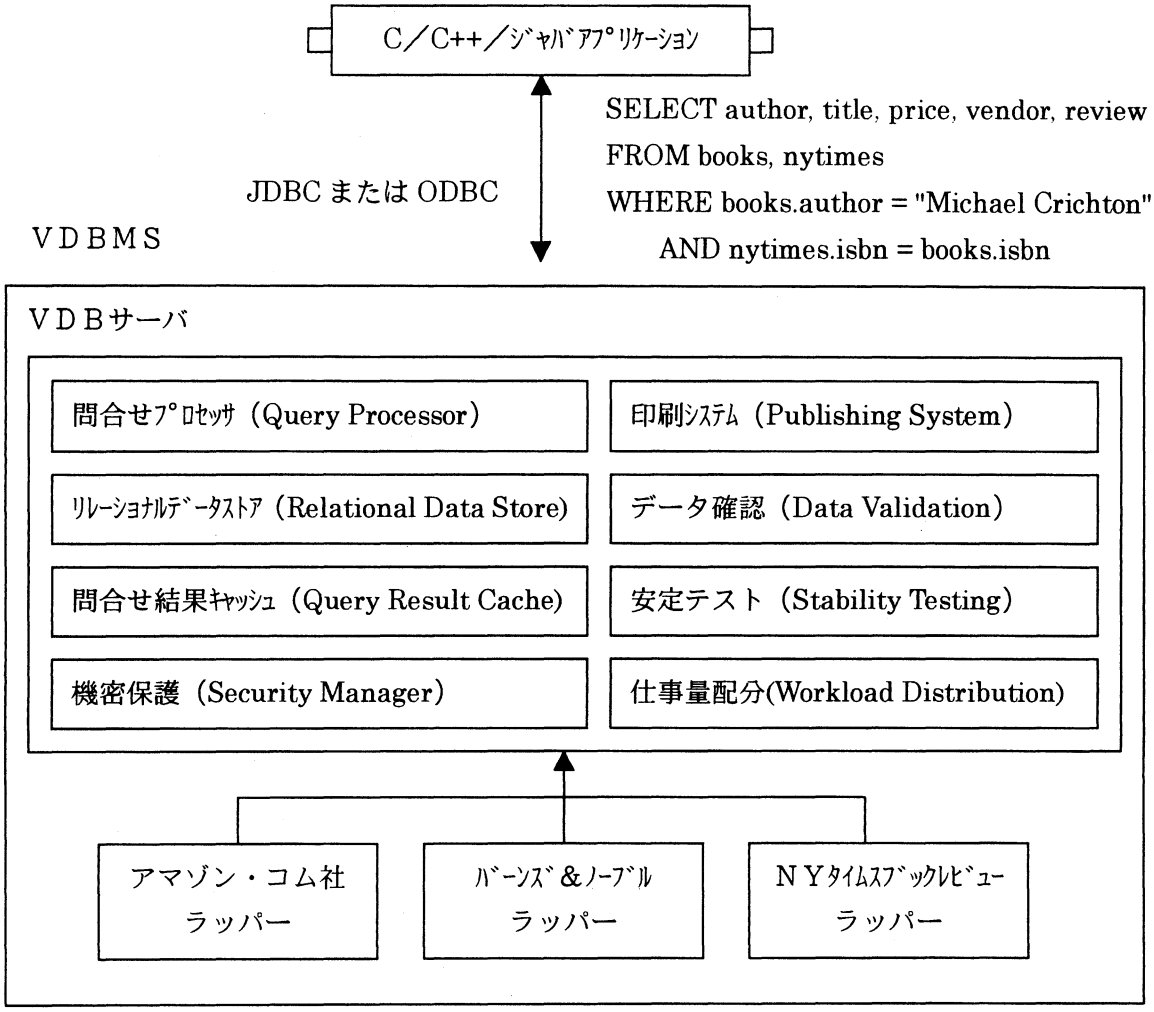

図3　ウェブ全体をひとつのデータベースに変換

ロセッサ”コンポーネントがその問合せを分解し，問合せのどの部分をどのデータ資源に送るかを決定して，そしてそれぞれの結果をまとめておく。“問合せ結果キャッシュ”はパフォーマンスを向上させるため，データ資源からの結果を一時的に保存しておく。加えて，“印刷システム（the publishing system）”は“リレーショナルデータストア”という場所に，仮想テーブルの物理的スナップショット（physical snapshots）を一時的に作り上げておく。ＶＤＢサーバはデータ確認テストをもすることができる。そのテストは個々のラッパーでのテストよりも高度である。一例として“安定テスト”をあげると，そのテストは，過去の統計的傾向とデータを比較して，もしその傾向から大きく逸脱していれば，警告を発する。

2.2　**バーチャルデータベース管理システム**（The Virtual Database Management System; VDBMS)

ジャングリー社のバーチャルデータベース管理システム（VDBMS）は書籍ＶＤＢのようなバーチャルデータベースの生成と管理をするものである。図4はVDBMS2.0，つまり第2世代

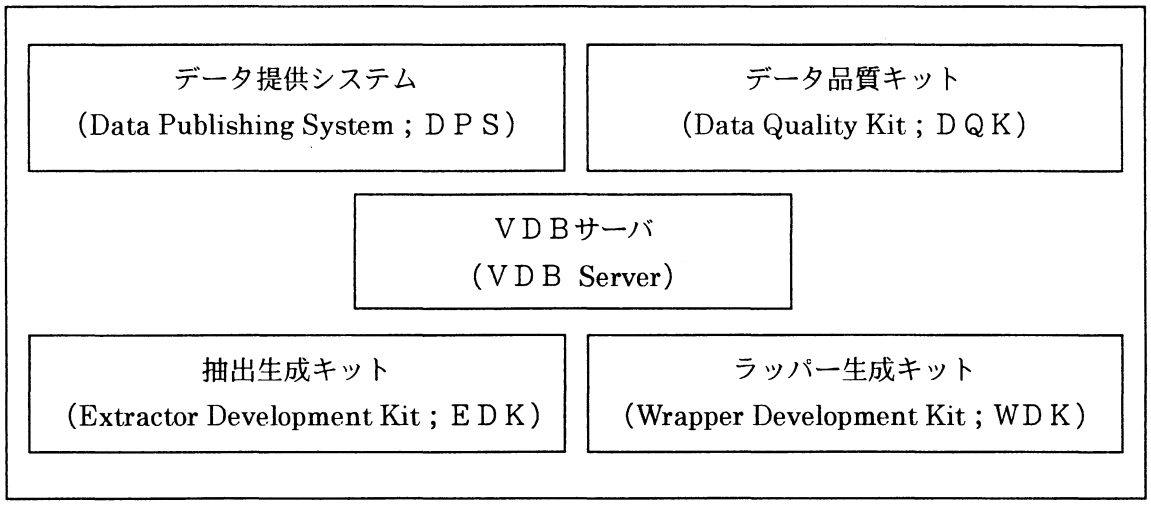

図４　VDBM 2.0 の構成

ＶＤＢ技術の構成要素を表している。

ラッパー生成キット（Wrapper Development Kit; WDK）

ラッパー生成キットは，ウェブサイト，ファイルシステムやその他のネットワークデータ資源のラッパーを迅速に生成することを可能とさせる。ＷＤＫは“ラッパーフレームワーク”の概念にもとづいてつくられている。ラッパーフレームワークはどのような分類があるかの集合体（a collection of classes）とプログラミング部品とから成っており，それらはともに一連のデータ資源に対するラッパー生成を容易にしている。たとえば，“ナビゲータ”フレームワークはウェブサイトに対するラッパー生成を簡単にする。つまり，数行の簡単なジャバコードで，最も複雑なウェブサイトの構造さえをもとらえることができる。それには，フォームへの書き込みに応じて動的にページを生成するようなサイトも含まれる。このフレームワークはウェブサイト内でのハイパーリンクページ間の関係をとらえ，そのサイトに対応するバーチャルデータベースのテーブルの内容にその関係を反映している。

ＷＤＫには，ＨＴＴＰ，ＦＴＰや HTTPS のようなプロトコルを取り扱う“ネットワークサービスライブラリ”と，ＨＴＭＬ関係のサービス（書式分解のようなサービス）とフォーマット無関連のサービス（パターンマッチングのようなサービス）を提供する“ドキュメントサービスライブラリ”が含まれる。ＷＤＫは“スキーママネジャ”に対して，対象のスキーマと抽出ルール，データ確認，データ変換操作を結びつける機能と，多くの類似したデータ資源間のスキーマ情報を共有する機能とを提供する。たとえば，企業の求人ウェブサイトの集まりは，同じスキーマを共有しており，“職種”，“給料”，“採用条件（requisition text）”という属性のテーブルを持っている。そのスキーマには，給料はある範囲内の数値であるべきだと明記されたデータ確認ルールや，職種と給料の情報がたとえなくても“採用条件”に対してある抽出ルールを適用することによって，そうしたデータが得られることを記述したルールが含まれる。

抽出生成キット（Extractor Development Kit; EDK）

データ統合とは，しばしば非構造的なテキストデータから構造を引き出すことを意味する。たとえば，アパートの賃貸を載せている新聞ウェブサイトを考えてみよう。そのアプリケーションは寝室の数や浴室の数，場所，家賃のような特徴の欄を持ったテーブルを必要とする。しかしながら，それぞれのアパート分類リストは，一般的に区別のないテキストのかたまりである。抽出ルールには，そのようなテキストからいかに必要な特徴を引き出すかを記述している。

“抽出エンジン（Extractor Engine）”は，複雑高度なテキスト処理ルールを表すために設計された言語であるＪｅｌ（Junglee Extraction Language；ジャングリー抽出言語）のためのインタプリタである。テキストの一部分から特定の特徴を抽出するためのＪｅｌプログラムは“抽出ルール（Extraction Rule）”とよばれている。抽出エンジンは，それぞれの抽出ルールを“直列的で有限の文章自動変換装置（cascaded finite state automaton）”に変換し，その自動装置がその場で抽出を実行する。ＥＤＫは，立地場所，電話番号，価格，その他多くの共通属性のための抽出ルール標準ライブラリを持っている。ＷＤＫを使って生成されたラッパーは，抽出ルールを広く利用している。

抽出ルールは属性をあらわす用語やフレーズの辞書に依存している。例えば，立地場所抽出ルールは合衆国の都市や州の名称とともに，その名称の一般的な略字もとりあげた辞書を用いている。ＥＤＫはそのような辞書を生成したり管理するための“辞書管理ユーティリティ”である。

データ品質キット（Data Quality Kit; DQK）

ＤＱＫは“データ変換”（“マッピング”とも呼ばれている）や“データ確認”の役割を果たしている。

ラッパーは任意のデータ資源をリレーショナルデータテーブルのように取り扱うのだが，これらのテーブルは一貫性のないスキーマや用語を持つ傾向がある。簡単な例として，あるラッパーは給与を“ドル／月”で表わし，別のものは“ドル／週”で表わし，アプリケーションは“ドル／年”を必要としているようなシナリオを考えてみよう。要求される属性名と単位変換は“項目レベルのマッピング（field-level mapping）”によっておこなわれる。“行レベルのマッピング（row-level mappings）”では他のカラムの値に基づいて新しいテーブルのカラムをつくる。

バーチャルデータベースはしばしば，ＶＤＢ管理者の制御外の資源から，かなり不規則なデータを取り扱うために，何も気づかずに大規模な変化を受けたりすることがある。それゆえ，データの完全性を保証することはＶＤＢにとって最も重要な問題である。そこでＤＱＫは３種類のデータ確認チェックルールを用意している。

1. **項目確認ルール**。これはテーブルの各行内にある特定カラム（または項目）の条件であ

る。たとえば日付は論理的に正しい（well-formed）日付か？

2. **行確認ルール**。これはテーブルの各行内にある2つ以上のカラムに関係する条件である。例えば，カラムAとカラムBの少なくともひとつはゼロでない。
3. **データベース確認ルール**。これはテーブルのすべての行に関係する条件である。たとえば，テーブル中のすべての行はカラムAについてそれぞれ一意の値（“主キー制約（primary key constraint）”とも呼ばれる）をとるべきである。データベース確認ルールの中でも特殊なものとして，“総計制約（aggregate constraints）”（Aカラムにおいてゼロ値を持つ行の数はそのテーブルの行数の10%以下であるべきであるというような制約）や安定テスト（過去データの総計制約（historical aggregate constraints））がある。

“安定テスト”はラッパーのデータごとに固有なものであるから，特別に述べておく必要がある。ラッパーはしばしば，変化にさらされたウェブサイトからのデータを取り出しているから，これらのテストは無効なデータに対する最初の防御線となる。これらのテストはあるテーブルのデータを同一テーブルの過去の統計値と比較し，過去の傾向から大きく逸脱していれば報告する。

ＶＤＢサーバ（VDB Server）

ＶＤＢサーバは複数のデータ資源上のテーブルを一つのバーチャルデータベース（ＶＤＢ）内の“仮想テーブル”として表現する。ＶＤＢサーバは，資源を横断してのビュー定義や問合せ処理を含む，仮想テーブル上での完全な RDBMS の機能を提供している。ＶＤＢサーバはＪＤＢＣと ODBC APIs を支援し，ブラウザベースの“ＶＤＢコンソール”によって管理される。ＶＤＢサーバにはリレーショナルデータストアも含まれる。それは流れを強化した（commercial-strength）RDBMS である。データストアはアクセスを迅速にするために，仮想テーブルのスナップショット（snapshots）を貯蔵するのに使われる。そして，データベースアプリケーションによって使われる他の物理的なデータテーブルをも蓄積している。

データ提供システム（Data Publishing System; DPS）

ＤＰＳは VDBMS のデータ貯蔵コンポーネントである。それは仮想テーブルと物理テーブルの間で変化する“データインテグレータ（data integrators）”と，時間順（chronological）あるいは依存度順（dependency-based）のどちらのジョブスケジューリングをもすることのできる弾力的な“ジョブスケジューラ（job scheduler）”をもっている。

自律的な外部データ資源が関係しているので，故障時の取り扱いと思いがけない事故は非常に重要になってくる。ウェブサイトが思いがけなく落ちたり，その構造が変わってしまうこともある。ネットワーク関係がくずれることもある。コンピュータシステムが壊れるかもしれない。データ提供システムはそのような故障が適切に処理されることを保証する。例えば，デー

タ資源にアクセスできなくなった時，そのラッパーは資源が使用できるようになるまで定期的に問合せをする。そして，システムが壊れた場合などすべての状況下で，データ提供システムは獲得したデータの完全性を守る。

3. 動作時のＶＤＢ：アプリケーションの実際

ジャングリー社はいくつかの重要な分野にＶＤＢ技術を提供している。例えば，求人広告，消費者ショッピング，不動産業，アパート（賃貸マンション）リストなどである。ジャングリーはこれらすべてのアプリケーションを顧客に対して作り上げ，インストールをおこなった。

3.1 オンライン求人

ジャングリー社の“ジョブキャノピー（JobCanopy）”ＶＤＢアプリケーションは，求人をおこなっているジョブサイトやフラットファイル（flat file），過去のデータ資源をも含めた700を超えるデータ資源からジョブリストをつくりあげている。このＶＤＢのスキーマは雇用者や求職者にとって関心のある，職名，職種，勤務場所，そして問合せ先を含めた31の属性を含んでいる。ユーザインターフェース（front-end）は高度な配列（configurable）となっており，ブラウザベースのデータベースアプリケーションを迅速に展開することができる。ジョブキャノピーは，“ウォール・ストリート・ジャーナル・インタラクティブ・エディション（The Wall Street Journal Interactive Edition）”，“ワシントン・ポスト（The Washington Post）”，“サン・ノゼ・マーキュリー・ニュース（The San Jose Mercury News）”を含む，いくつかのメディア会社のウェブサイトで展開されている。

3.2 ウェブコマース

“ショップキャノピー（ShopCanopy）”ＶＤＢアプリケーションは，書籍，音楽，コンピュータハードウェア，消費者向け電子機器（Consumer Electronics）を含む８つのカテゴリーで４０を超える販売店を比較検討して購入できるようになっている。ショップキャノピーは“ヤフー・ビザ・ショッピングガイド（Yahoo! Visa Shopping Guide）”と“ＺＤネット・コンピュータ・ショッパー（ZDNet Computer Shopper）”ウェブサイトで展開されている。

4. むすび

ＶＤＢ技術は，以下の特徴のうち少なくともひとつの特徴を持った資源において，アプリケーションを迅速に展開することができる。

・非常に多くのデータ資源
・自律的であり，集中制御を伴わないデータ資源
・構造的，非構造的なデータの混合であるデータ資源

ワールドワイドウェブとほとんどのイントラネットは，これらの特徴をすべて備えている。ＶＤＢ技術はインターネットをひとつのデータベースに変換する技術なのである。