第8回

Hadoopで実現する データ利活用のプラットフォーム

㈱NTTデータ　基盤システム事業本部
OSSプロフェッショナルサービス　シニアスペシャリスト
政谷 好伸

1 はじめに

NTTデータがオープンソースの大規模分散処理プラットフォーム「Hadoop」に着目し、取組みを始めてからおよそ5年になる。早くから先進的ユーザが関心を示す基盤技術ではあったが、ここ数年の間にビッグデータの主要技術として、さらに最近では主要な製品ベンダが各々の製品戦略の中でHadoopとの連携を謳うまでの注目を集めるようになった。NTTデータが手がけるシステム数も着実に増え続けており、数年に渡り稼働している実績や、クラスタの更改・増強を計画する案件も複数存在する。とはいえ、一般的な企業でHadoopを導入している事例はまだまだ少ないのも現実である。本稿では、現在Hadoopが着目されている背景・理由について振り返るとともに、データを利活用するためのプラットフォームの在り方をOSSやHadoopの視点から述べる。

2 データマネージメントの変化

あらゆるサービス活動のクラウド化・偏在化が同時進行する、さらにはInternet of Thingsと言われるような情報生成や集積が可能となる中で、ITの関心対象となる情報のスコープとサイズが各々飛躍的に拡大、増大すること（図1）を指して「ビッグデータ」と括ることが多い。しかしながらITの観点からすると、種々の取組みを包括した「ビッグデータ活用」がよりしっくりした表現である。広く捉えるならばビッグデータ活用とは、扱うデータがビッグ（大容量・大件数）であるために、従来のITアーキテクチャでは難しかった、もしくは甚だしく高コストにしか実現できなかったデータ活用に関する取組みである。あるいは、構想としては知られていたが現実には難しかったものが、分散処理技術や高速なネ

従来サービスとクラウドサービスの情報量の比較　経済産業省

サービス分野	従来サービス		クラウドサービス	
	システム	データ量	システム	データ量
交通／ITS	(例)VICS	2.5分毎に5万文字相当の情報を配信: 57.6MB／日	ITS	30秒ごとに120種類のセンサ情報(6KB)を1740万台から取得: 300TB／日 522万倍
ヘルスケア	(例)レセプト	2KB／人・レセプト×患者数(281万人／日): 5.6GB／日	ヘルスケア(健康管理、疾病予防)	3次元加速度センサによる行動解析:1秒に1回5KBのデータと一日1回50KBのバイタル情報を1000万人から取得: 4.32PB／日 80万倍
農業	(例)植物工場	1回/時間×ポット数(100万)×植物工場数(50箇所): 120GB／日	圃場管理(鳥獣被害対策、気候変動対策)	10MB/日×50万箇所のデータを過去10～30年分のデータと照合: 57PB／日 46万倍

出典: 2011.1.18 JISAセミナ: 情報サービス産業を取り巻く環境・国のIT戦略より (www.jisa.or.jp/news/771/download/102_1.pdf)

図1　関心となるデータセットのスコープとサイズが増大

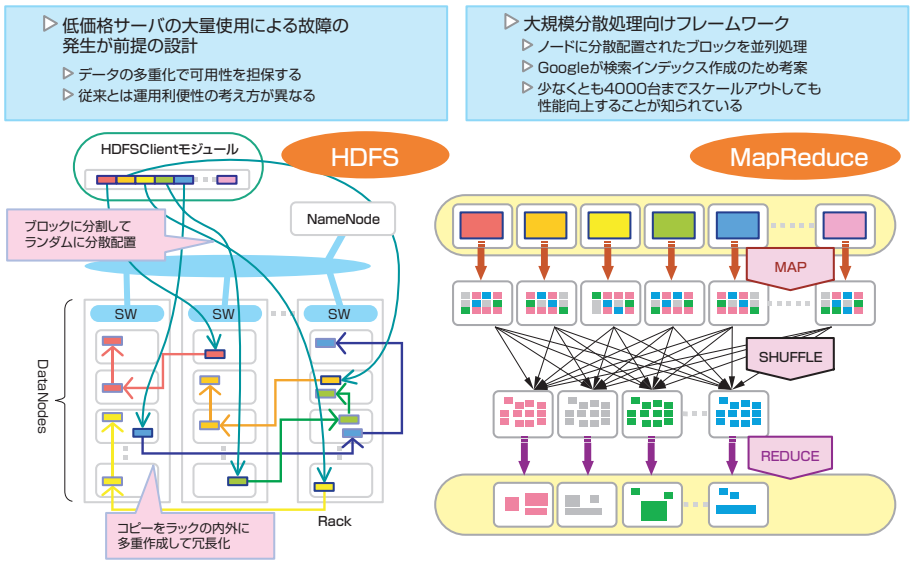

図2　分散ファイルシステムHDFSと並列処理フレームワークMapReduce

ットワークが手頃になってデータ活用が現実のものになった取組みである。例えば、ライフログを収集・格納・活用するシステム、オンラインゲームでの行動解析、決済や給付請求における不正利用検出、さらには、大規模なデータ収集に基づく様々なサービスの最適化などが挙げられる。

ビッグデータ活用へとITアーキテクチャへのニーズが変化する中、注目されているのがHDFSと呼ばれる分散ファイルシステムとMapReduceと呼ばれる並列処理フレームワークを組み合わせた大規模分散処理プラットフォームHadoopである（図2）。

3 Hadoopの特徴

Hadoopは「入れ物」として優れていると供に、その入れ物の中に蓄積したデータを「並列にバッチ処理」できる。つまりデータを複数のデータソースから集約しつつ蓄積する、さらに蓄積したデータに対して集計・分類・分析するという、2つの役割が1つに統合されている形にデータプラットフォームとしての魅力がある。

そもそも大量のデータセットは単一のディスクや装置に納めることが適わない。それ故、システム統合や大規模アーカイブ、データの可用性を目的とした分散ストレージや分散ファイルシステムは様々なものが存在している。大量のデータを言わば結果として蓄積するシステムは存在していたものの、蓄積された大規模なデータを一括して処理する発想・用途は重視されていなかった。ビッグデータ活用において重要なのは、大きなデータに対する処理をどのように実現するかである。処理を複数の演算ノードで並列して行う、ノードに分散配置されたデータのローカリティを活かす。そのために処理ノードとデータ蓄積ノードが同居した構成とし、チャネルスループットを最大化するというアイデアがHadoopの本質といえる。大規模なデータを経済的に蓄積・処理できるだけではない。試行錯誤を繰り返しながらデータを柔軟に変換し情報間の関係性を検討したり、広く深くデータを分析しモデルを検証したりといったアドホックな処理プロセスを加速できるに足る性能を有することもビッグデータ活用では重要である。

従来の業務中心・標準化といったプロセス指向のデータマネージメントではなく、互いに関連するサービスとそれらの外延で生成する膨大かつ多様な生のデータを蓄積し、その中から有益な素材を見出し、サービスや業務の価値へと還元するデータ指向のデータマネージメントが重視されるようになった。蓄積されたデータを様々な観点で再整理し、複数のシステムからのデータを集約し掛け合わせるといった、いわゆるマッシュアップに対してHadoopが採用している「準・非構造データに強い」Key-Value型のデータモデルは適合性が高い。

OSS活用の観点からすると、スケーラビリティに優れるデータプラットフォームを、コモディティー製品を活用して構築できることもHadoopの大きな魅

力である。従来、大規模なデータを扱うシステムを構築しようとした場合ファイルシステムやDBMSにOSSを選択したとしても、ストレージは高性能なベンダ製品を選択することがほとんどであった。ミドルウェアで可用性や耐故障性を配慮したHadoopを活用することで、特定の機器やベンダに依存せず、最適なハードウェアをバランスよく選択しながら「Free(自由)」にデータプラットフォームを構築することができる。

4 データ利活用のプラットフォーム

現時点でのHadoopのユーザ層は、おおまかに2つのタイプに分類される。前述のビッグデータ活用に加え、既存処理の高速化に取り組むユーザが多い。両者はそれぞれ別の動機に端を発する取組みであるが、ビッグデータを支える技術の多くが既存処理の高速化、より正確には並列・分散処理化に役立つのは事実である。既存のDBMSやDWHに比べてHadoopは「何に使えるのか」「どう使えばよいのか」という用途をイメージして活用するというアプローチ、従来の手法では不可能なこと、より大件数が求められる対象をHadoopにオフロードするという考え方が分かり易く受け容れられている。しかしながらそのような端緒であったとしても、ITアーキテクチャとして鑑みるならば、よりデータ利活用へとシフトしていく。つまり、Hadoopとそれに連携するシステムの主客は次第に逆転していくと予想される。

準・非構造データの入れ物として、且つまた、大容量・大件数データのバッチ処理に優れるHadoopこそが、最もデータソースに近い場所に配置されると供に、データの流通を一義的に担うインフラ的な存在、データ利活用のプラットフォームの中心になると考えられる。複数のシステム、さらには複数の組織を横断して、企業外からのデータも含めすべてのソースデータを一義的に集約する「ハブ」となる。一次的な分析や横断的なマッシュアップ、長期アーカイブのための再構成など大規模なデータ処理を一手に司る。蓄積した内容を活用目的に応じた一定の情報の固まりとして相手側のフォーマットに適した形式に整形して引き継ぐ、あらゆるドレインに対しても「ハブ」となる。DBMSやDWHなどが、それらが得意としていたトランザクション処理や低レイテンシ処理へと特化していく。そのような形でHadoopの利用は進むと考えられる。

5 ビッグデータ活用のポイント

最後に、これまでの導入経験、お客さま利用実績を踏まえHadoopを効果的に導入・活用するためのポイントについて述べ、本稿のまとめとする。

①アイデアを持つ

「どのデータを」「どのように処理して」「お客さまにどう効果をもたらすか」「改善のサイクルをどう回すか」という具体的なビジネス上のアイデア、イメージを持つことが肝要。捨てていたデータを漠然と溜めるだけでは価値は見いだせない。データを分析したりマッシュアップしたりするための仮説、アイデアを検討するのがスタートラインである。

②データを蓄積する

「準・非構造データの蓄積に強い」というHadoopの特徴を活かして、データを取り込むタイミングでは極力生データに近い形で蓄積する。分析の検討・検証には一定期間に渡る適当量のデータ蓄積が不可欠であり、収集自体に手間と時間を要することからまずは生データの蓄積を開始する。パイロットプロジェクトに着手し収集可能なデータの特徴や特性の把握に早期から努めることが望ましい。必要なデータをどこで入手するか／どの程度のデータ量が必要かをまず考える必要がある。

データの構造・意味を考える必要がないという訳ではない。重要なのはデータに構造・意味を与えるタイミングである。データ管理や性能管理のためにデータを予め構造化・正規化しなくても構わない、

意味付けの処理を分析時にデータを読み出す時点まで持ち越すことができるということである。もちろん、収集から蓄積（廃棄）までのプロセスをどう管理するか、どうすればデータの質を高められるのかなどは、なるべく早い時期から検討するのが望ましい。

③価値の発見

多様なデータを集積し、それまで試みていなかった複数のデータをマッシュアップした分析をどれだけ広範囲に実施できるかが、新たな素材を見い出し新たな価値と出会うポイントになる。分析に基づくデータ活用という考え方は新しいものではない。企業活動のデータを分析し経営に活かす実践はBIとして従前より知られている。ビッグデータ活用において異なるのは、分析対象となるデータセットのスコープが企業や特定組織内に留まらない点にある。むしろ外部から取得するデータセットや分析に際して新たに生成するデータセットの方が大量であり、かつ対象として「全件データ」を扱うことで価値が生まれている。従来のBIでは分析結果のフィードバックが主に経営層を対象とした限定的なループであったのに対し、ビッグデータ活用では全件が対象、つまりフィードバックがサービスを利用する個々の全ユーザにパーソナライズされることや、システムの全構成要素毎に最適化されることで価値が生まれるのである（図３）。

一般的なシステムでは、目的別にサブシステムを用意し各々にデータを蓄積していることが多い。利用目的や形式の異なるデータを貪欲に受け容れHadoopの処理対象にする。さらには、テープデバイスやNAS上に死蔵されているアーカイブデータも選択的に取り込み、活用可能な状態にすることが望ましい。

④スケーラビリティ

Hadoopはサーバを増設することで処理能力をリニアに伸ばす「スケールアウト構成」を採ることができる。そのため構成設計やサイジングは容易と考えられ安易な製品選定が行われがちだが、MapReduceの処理特性、運用要件を考慮してコモディティー製品の幅広い仕様の中からバランスの良いコンポーネント（コア数、メモリ量、I／O帯域など）の組合せを見極めるべきである。

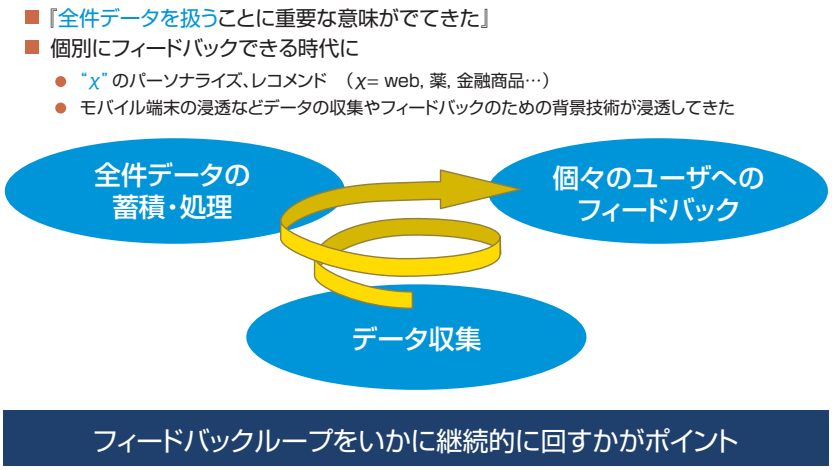

図３　ビッグデータ活用の背景と本質

同様に、アプリケーション開発において並列処理最適化のためのスキルも不可欠である。データの流れと処理順序をリファクタリングする(例えば、大きな走査単位を見つける、データを用途に合わせ非正規化し最適配置するなど)など分散並列処理特有の性能改善ノウハウがある。

⑤セキュリティ・プライバシー

ビッグデータ活用では、個人情報に接点を持つことが多い、高い意識と十分な注意にもとづいて活用する必要がある。また、どのようにデータ活用をしているのかきちんと説明した上で、世の中に受け容れてもらう流れを作っていくべきであり、いくつか法律上の手当が必要なケースもあるので、状況をきちんと見極めながら対応を進めることが重要である。

お問い合わせ先

株式会社NTTデータ
基盤システム事業本部
OSSプロフェッショナルサービス
TEL：050-5546-2496
Email：hadoop@kits.nttdata.co.jp
URL：http://oss.nttdata.co.jp/hadoop/