Yahoo! JAPANにおけるデータ利活用の実際

角田　直行

本稿では，Yahoo! JAPAN におけるデータ規模を示し，どのように利活用を行っているか事例を交えて紹介する．また，データを利活用するうえで必要となるシステムについて解説を行う．さらに，技術面だけでなく組織整備や文化の形成，個人情報の扱いやポリシーなど，円滑にデータの利活用を行ううえで技術的要素以外に必要となるポイントについて述べる．

キーワード：ビッグデータ，利活用事例，Hadoop

1. はじめに

Yahoo! JAPAN では「課題解決エンジン」というビジョンを掲げ，情報技術で人々や社会の課題を解決するべく日々サービスを開発，提供し続けている．また，Yahoo! JAPAN はビッグデータを最大限に利活用している企業であると自負しており，自身を「マルチビッグデータカンパニー」と定義づけている．本稿では，Yahoo! JAPAN がどのようなデータをどれほどの規模で扱っているのか，そのデータはどのように利活用されているのか，大規模なデータを処理するにあたりどのようなシステムを使っているのか，データを効果的に利活用にするにはどのような点に気をつけなければならないのか，について解説する．

2. Yahoo! JAPAN のデータ規模

現在，Yahoo! JAPAN では検索やメールをはじめ，ヤフオク，ニュース，知恵袋など 100 を超えるサービスを提供している．そしてこれらのサービスを運営していくうえで大規模のデータを扱い，そしてログをはじめとする大量のデータを出力している．規模の観点から例を挙げると，Yahoo! JAPAN の総ページビュー数は月間 580 億にもおよび，1 年間に検索されるキーワードの種類は 75 億にもなる．アクセスログや検索クエリのほかにも，広告ログやコマースの購買履歴，デモグラフィックと呼ばれるユーザー属性など多岐にわたるデータを扱っている．

Yahoo! JAPAN の強みの一つとして，検索やニュースなど日本市場において大きなシェアがあるサービスを持っており，そのサービスについてアクセスログをはじめとするユーザーの行動履歴に関するデータを大規模なレベルで保有していることが挙げられる．また，単一のジャンルに限らずさまざまな分野においてデータを持っているのも他企業と差別化する大きな強みとなっている．コマース企業であればコマースに関するデータ，ゲーム企業であればゲームのデータを持っているのが通常であるが，Yahoo! JAPAN は多くの領域に渡ってサービスを展開しており，幅広い分野を網羅している．そしてその一つひとつが巨大なデータ，いわゆるビッグデータを生み出しており，複数のジャンルにわたるビッグデータを扱っている企業という意味で，我々は自社を「マルチビッグデータカンパニー」と定義している．

3. データ利活用の事例

複数のジャンルにわたってビッグデータを扱っている Yahoo! JAPAN では，各サービスにおいてデータを利活用した事例が数多く存在しており，以下ではその中から一部について取り上げる．はじめに，データを使ったサービス改善事例として「A/B テスト」を紹介する．A/B テストとはウェブページなどにおいてデザインや機能の異なるパターンを複数用意し，実際にユーザーに利用してもらい効果を比較・分析することで最適化を図る手法のことである．Yahoo! JAPAN では各サービスにおいて，表示モジュールの配置や背景色，記事の本数などさまざまな単位で A/B テストを日々実施している．

以下に A/B テストの具体例を挙げる．スマートフォン向け Yahoo! JAPAN トップページを表示すると中央に検索フォームがある．

図 1 はトップページの二つのパターンを表しており，左と右を比較して何が異なるか一見しただけではわかりづらい．よく見ると，右のパターンのほうが検索フォー

かくだ　なおゆき
ヤフー株式会社システム統括本部データソリューション本部
〒 107–6211 東京都港区赤坂 9–7–1 ミッドタウン・タワー

図 1　スマートフォン向け Yahoo! JAPAN トップページの A/B テストパターン

ムの枠線が若干太いことがわかる．これは普通の感覚からするとささいな違いでしかなく，この変更を行ったところで何も影響がないように思うかもしれない．しかしながら，普段から大量にアクセスされるトップページにおいて，このような小さな変更でさえ多大な効果を与える．実際，この枠線の変更により検索フォームがより強調されたことでユーザーの検索利用が促進され，結果として年間 5 億円以上もの売上を増やすことにつながった．

また，検索フォームにはキーワード入力補助という機能がある．例えばフォームに「ビッグデータ」と入力すると，「ビッグデータとは」「ビッグデータ 活用」「ビッグデータ 問題点」…のように，入力したキーワードに関連するキーワードが並ぶ．これもデータを活用して実現した機能の一つで，ユーザーが入力する検索キーワードのログを集め，ノイズ除去など適切な処理を行った後，よく検索される順に表示している．データを使って検索をより使いやすくする機能としてこのキーワード入力補助のほかに関連検索ワードや，入力ミスを修正するスペラーなどがあるが，これらは全検索行動の 37%を占めており，データを活用した機能が非常にユーザー利用の効果を上げていることがわかる．

同じくスマートフォン向け Yahoo! JAPAN トップページを表示すると，下のほうに「あなたにおすすめの記事」というニュース記事が並んでいる枠がある．これはアクセスする人ごとにその人に合ったニュース記事が表示されるようになっており，いわゆるパーソナライゼーションを実現している．このパーソナライゼーション機能もニュースの閲覧履歴や Yahoo! JAPAN サービスの利用ログ，検索キーワードなどの行動履歴データを使って解析し，ニュース記事とのマッチングを行っている．マッチング対象となるニュースのコンテンツデータについても，単なるニューステキストをそのまま使うだけでなく，各キーワードに含まれるメタデータも利用しており，幅広くその人に合ったニュース記事が配信されるようになっている．

広告もデータを利活用している代表的な事例の一つである．Yahoo! JAPAN が配信している広告の一つに「行動ターゲティング広告」がある．これはユーザーが Yahoo! JAPAN のサービスにアクセスしたり，検索したり，広告をクリックしたりすることでユーザーがどのようなジャンルに興味関心を抱いているかを推定し，それに応じた広告を表示するものである．例えばあるユーザーが Yahoo!トラベルにアクセスしたり，ホテルや航空券に関する検索をしたりすると，旅行について関心があると推定し旅行に関する広告が出やすくなる．

Yahoo! JAPAN は自社で扱っているデータだけでなく他企業のデータもおおいに活用している．代表的なサービスとして「リアルタイム検索」が挙げられる．Yahoo! JAPAN では Twitter や Facebook などのソーシャルデータを保有しており，ユーザーが投稿した内容を即座に検索できるサービスを提供している．そのリアルタイム検索において，現在話題になっているキーワードをランキング形式にして並べた「注目のキーワード」機能や，特定のキーワードに関する投稿内容を分析して，全体的にポジティブな反応なのかネガティブに受け止められているのかを知る「感情分析」機能を提供している．これらの機能の裏側には自然言語処理の技術が使われており，Yahoo! JAPAN では形態素解析エンジンをはじめとする自然言語処理に関するプロ

ダクトを自社開発している．

Yahoo! JAPAN ではテキストデータだけでなく，バイナリデータも扱っている．Yahoo! JAPAN では音声を入力し，内容に応じて適切な内容を返す「音声アシスト」という Android アプリを提供している．このアプリにて音声を認識しテキスト化する音声認識エンジンや，「今日」というワードは現在日時を表し「8 時に起こして」という内容はアラーム登録の命令を意味するテキスト内容の意図を理解する意図解析エンジンなどが含まれており，日々入力された音声データを基に学習させ認識精度など品質向上に役立てている．また，コマース系サービスの画像データを学習し，物体認識や類似画像検索などの機能をスマートデバイスアプリの一機能として実現している．

データの利活用先は自社サービスだけに限らない．世の中の課題解決を行っていくことを目的に「Yahoo! JAPAN ビッグデータレポート」を公開している．このレポートでは，Yahoo! JAPAN のデータを活用して世間の動向との相関性を分析したり未来予測にチャレンジしたりしている．その一つとして，データを使って日本の景気を把握し予測することにチャレンジした．景気を客観的に判断する値として内閣府が毎月発表している「景気動向指数」を基に，Yahoo! JAPAN のデータを使ってどれだけ景気動向指数に近い値や動きを再現できるか分析を行った．結果，内閣府は通常 2 カ月遅れで景気動向指数を発表するところを，それよりも早い時期にかなり近い値を算出できた．また，未来予測の別の事例として 2013 年夏に行われた参議院選挙では，比例区および選挙区での政党別獲得議席数を投票日前に予測し，結果として与党と野党の各議席数を完全に一致させることができた．

このように Yahoo! JAPAN では自社のサービスではもちろんのこと，世の中の課題解決などさまざまな方面でデータ利活用を行っている．

4. データ利活用を支える技術

これまでにいくつものデータ利活用事例を紹介したが，これらはどれもデータが適切に集められ，処理しやすい環境があってこそ成り立つものである．以降では，どうやってデータを集め処理しているのかという点について解説していく．

Yahoo! JAPAN はウェブブラウザーやスマートデバイスのアプリというインターフェースを通じてサービスを提供しており，ユーザーはこれらを操作してサービスを利用する．ユーザーがウェブブラウザーを介して Yahoo! JAPAN のサービスを使うことにより，Yahoo! JAPAN 側が管理するサーバーにアクセスログが出力される．ウェブサーバーのアクセスログはウェブの世界において一般的なものであるが，何らかの理由によりウェブブラウザー側で正しく表示されなくとも一つのアクセスとしてカウントされてしまうケースがある．広告事業を行っている Yahoo! JAPAN においてこの問題は非常に重要で，ユーザーがアクセスした回数と広告を含む画面が正常に表示された回数は一致しなければならない．このようにサーバーサイドのアクセスログによるカウントには問題があるため，クライアントであるブラウザーの挙動を利用してカウントを取る「CSC (Client Side Counting)」という手法を用いている．これは表示するウェブページ内に 1 ピクセル四方の透明画像，いわゆるビーコン画像をページ内の任意の場所に image タグとして埋め込んでおき，ページが正常に読み込まれた際にそのビーコン画像が読み込まれビーコン画像をホストするサーバー側に正常なアクセス回数がカウントされる，という仕組みになっている．この CSC という手法は正確なアクセスがカウントできるだけでなく，単純なクローラによる不要なカウント回避や JavaScript を利用した動的なページの動きに合わせたカウントなど，柔軟な対応が行えるメリットもある．

また，「ユーザーが何をクリックしたのか」も非常に重要で取得したい情報の一つである．通常，アクセスログ内に含まれる Referer を参照すると遷移の経路をつかめる．しかし，Yahoo! JAPAN 外のページにリンクした場合，リンク先のサービスのサーバーにはアクセスログが残るものの，Yahoo! JAPAN 側には何も情報が残らない．この問題を解決すべく，Yahoo! JAPAN では「リダイレクタ」と呼ばれる仕組みを使っている．これは，リンク URL に対して Yahoo! JAPAN 側で用意したリダイレクタサーバーを介した URL に書き換え，ユーザーがそのリンクをクリックするといったんリダイレクタサーバーに飛んだ後，HTTP ステータスコード「302 Found」を返して本来の遷移先の URL にリダイレクトする，という仕組みである．これにより，Yahoo! JAPAN 側で管理するリダイレクタサーバーに「ユーザーが（Yahoo! JAPAN 以外の）どのページに遷移したか」を表すログが出力される．このリダイレクタには，遷移先のサイトに何かしら問題があった際に遷移自体をコントロールできるなどの副次的なメリットもある．

このように Yahoo! JAPAN ではデータを生成する

時点でいろいろな仕組みが用意されている．次に生成されたデータをどのように集めているかについて述べる．世間ではログデータを収集するソフトウエアとして fluentd や Apache Flume などのオープンソースがあるが，Yahoo! JAPAN でも同様の機能を持つシステムとして「Data Highway」と呼ばれるインハウスのシステムを利用している．Yahoo! JAPAN では大量のデータを扱っており，1 日に流れるデータ量は約 13 TB にも呼ぶ．また約 8,500 台ものウェブサーバーからデータを収集しており，トラフィックの多いサーバーや少ないサーバーなどが混在する中，安定的かつ効率的にデータを転送する仕組みが用意されている．また，アクセスログは大事なデータであり欠損があってはならないため，Data Highway ではデータの完全性を保つ仕組みも用意されており限りなく 100%に近い回収率を実現している．

最後にデータ処理を行うシステムについて概説する．Yahoo! JAPAN では主に Hadoop，Teradata，Storm の三つの処理システムを活用している．以下，それぞれの特徴について述べる．

Hadoop はオープンソースソフトウエアとして公開されている大規模データ処理システムである．HDFS (Hadoop Distributed File System) と呼ばれる分散ファイルシステム上に MapReduce をはじめとする分散処理フレームワークが動くようになっており，数十台から数千台のマシンの規模までスケールするよう設計されている．Hadoop 上で動かすアプリケーションについて，以前は MapReduce と呼ばれるプログラミングモデルに基づいてアプリケーションを書く必要があったが，最近では Pig や Hive のような MapReduce を隠ぺいした SQL に似たなじみやすい DSL (Domain Specific Language) を用いたり，Spark などの MapReduce とは異なるプログラミングモデルによって分散処理を行う仕組みができたりなど，オープンソースの利点を活かしさまざまなデータ処理を効率的に行うためのエコシステムが発展している．

Yahoo! JAPAN では複数の Hadoop クラスタを運用しており，最も大きなクラスタは 4,000 台規模になる．サーバー 1 台のハードウエアとしての寿命を 3〜4 年ととらえた場合，4,000 台となるとほぼ毎日故障している計算になる．実際，4,000 台規模の Hadoop クラスタでは 1 日に平均 1.5 台の間隔でハードディスク障害などの故障が発生している．Hadoop は，一時的に少数台のサーバーが故障しても大きな問題にならずリカバリしやすいように設計されているため，実際には通常のシステムと比較して復旧に関する運用コストはかかっていない．Yahoo! JAPAN では各サービスのアクセスログや広告ログ，購買履歴，検索クエリなどのログの加工や集計をはじめ，レコメンデーション，機械学習，音声解析など主なデータ処理の大半を Hadoop 上で行っており，数百人のユーザーが一つのクラスタを使うマルチテナンシー運用を行っている．

Teradata は，Teradata 社による商用のデータウエアハウス製品である．並列分散処理に非常に優れており高速にデータ処理を行えることが特徴で，SQL が扱えるためエンジニアでない人でもデータ処理を行えるのも魅力の一つである．Yahoo! JAPAN では日本最大規模の Teradata を導入しており，広告のレポーティングや広告モデルの効果測定やコマースの購買分析などに使われている．Teradata も Hadoop と同じくマルチテナントによる運用を行っており，数十〜数百人のユーザーが並行してクエリを投げている．Teradata はリソース管理に優れており，ユーザーまたは用途の単位で CPU やディスク IO の割り当てができるため，全体的に効率の良いデータ処理が行える．

Storm はオープンソースソフトウエアの分散ストリーム処理プラットフォームである．ストリーム処理とはログや Twitter のつぶやきのような逐次流れるように出力されるデータに対して，ためることなくそのままリアルタイムに処理することを意味する．Storm はそのストリーム処理を大規模な環境で実現するシステムであり，Twitter 社を中心に Apache プロジェクトとして開発が進められている．Storm では一つの処理単位を Topology と表現し，Topology は Spout と呼ばれるデータを発生させるコンポーネントと，Bolt と呼ばれるストリームデータの小さな単位を処理するコンポーネントで構成される．Yahoo! JAPAN ではページ内の各リンククリック速報や広告改善向けのデータ一次加工，スマートデバイスアプリのクラッシュやエラー情報の速報などに使われている．

これまで三つのシステムをそれぞれ紹介してきた．各システムにはそれぞれ特性があり，その特性に応じてシステムを使い分けている．例えば，Hadoop には大規模なデータをためられる HDFS という仕組みがあり汎用的なマシンを追加することでスケールすることが可能なため，リーズナブルにデータ容量を増やすことが可能である．それにより，Hadoop にデータのほぼすべてを格納するようないわゆるストレージのような役割を果たしている．そしてほぼすべてのデータがあることを活かし，その大規模なデータセットを一括で処理

するようなバッチ処理がメインになる．Teradata でもサイズ拡張などのスケールを行うことは可能であるが，商用製品であるため Hadoop に比べて価格コストが高く，Hadoop ほど気軽にデータ容量を増やすことはできない．しかしながら，Teradata は分析用途に特化した RDBMS でありリレーショナルなデータ操作に強いため，大規模なデータ同士を結合したり複雑な SQL を効率よく実行するのは得意である．Yahoo! JAPAN では，JOIN 処理を多用したりアドホックな分析を行ったりする必要がある場合には Teradata を選択することが多い．Storm は Hadoop や Teradata のようにデータをためることは苦手である一方，入力として入ってきたデータをそのまま処理するのに適しているため，リアルタイムにデータを処理したいニーズに適している．

5. データ利活用を行うための組織整備

これまで，Yahoo! JAPAN におけるたくさんのデータおよび，そのデータを処理するシステムについて紹介してきた．一見，豊富なデータとそれを活用する技術および人材さえ用意すればデータ利活用が十分にうまくいくように思えてしまうが，実際はそれだけでは十分とはいえない．以降では，利活用を進めるうえで技術的要素以外において重要となるポイントの一つである，組織面について解説する．

企業において効果的にデータを利活用するためには，その企業に在籍する全員が何かしらデータに触れている環境を目指さなければならない．昨今，ビッグデータという言葉がウェブや雑誌，テレビなどいくつものメディアに登場しており，データサイエンティストという職業が脚光を浴びている．そして各企業でもデータに関する取り組みが活発になりデータ分析に関する専門部隊や部署が立ち上がった事例も多く目にするようになった．Yahoo! JAPAN でもデータソリューション本部というデータ利活用を専門的に行う部署があるが，けっしてその部署のみがデータを扱うようにはなっていない．前述したようにデータの多くはセントラライズされた Hadoop に格納してありデータソリューション本部が管理しているが，そのほとんどすべてのデータがフレームワークや HiveQL などの DSL，ウェブインターフェースのツールを通して技術に精通していない職種の方を含め，社員全員がアクセスできるようになっている．普段からダッシュボードを見たり生のデータに触れる機会を作ることにより，人の感覚でなく数値やデータで意思判断を行う習慣が浸透していき，「データにアクセスすること自体は何も特別でない」という環境を作り出せる．もちろん，自然にこのような環境を作り出すことはできないため，各サービスで KPI を設定したりイベントを開催したりなどさまざまな働きかけが必要である．Yahoo! JAPAN ではサービスの KPI の一つとして DUB (Daily Unique Browser) という指標を設定しており，これは 1 日にどれだけのブラウザーによってサービスが利用されているかを表すものである．同じ人が PC とスマートフォンでそれぞれアクセスすると，それは異なる DUB としてカウントされる．この指標は，まさに現在訪れようとしているスマートデバイス中心の世界を想定し，「PC だけでなく，いろいろなデバイスを通じて，常日頃から使っていただけているサービスになっているか」を表すために考えられたものである．各サービス担当者はどうすればユーザーに満足してもらえ DUB につながるかについて考え，アクセス解析ツールを使って分析したり A/B テストを行ったりする．また，Yahoo! JAPAN では社内でデータに関するコンテストのようなイベントを開催してきた．このイベントは，データを軸にしたユニークなアイデアを発表したり新しくアプリケーションを作ったりして，投票や審査を通して表彰するものである．このイベントを通じて，別サービスのログや検索クエリ，Twitter のような自身が担当するサービスのアクセスデータとは異なるデータに触れる機会を与えたり，社内にデータを利活用する文化を作り活性化させる効果がある．

また「異なる組織間で連携し共通の課題解決に取り組む」ことも重要になる．データを扱う専門部署を作ってしまうと陥ってしまう問題の一つに，他の部署の人たちが「データに関する課題について当事者でなくなってしまう」ことが挙げられる．「データ分析については専門の部署に任せてしまえばよい」「あの部署に相談すれば何かデータ分析してすごいことをしてくれるだろう」といった，自身の課題についても任せてしまい当事者意識が薄れてしまう体質になる危険性がある．Yahoo! JAPAN では 100 以上ものサービスを抱えており，すべてのデータが集約されデータソリューション本部にて管理していることは事実であるが，一部署が全サービスの問題一つひとつを深くまで把握することはあまり現実的ではない．やはりサービスのドメイン知識について熟知し，抱えている課題について常に考え，最もサービスを今より良くしようと思っているのはそのサービス担当者自身である．Yahoo! JAPAN ではサービスを抱える事業部門とデータソリューション本部それぞれ少数名からなるミニプロジェクトを結成

し課題解決に取り組んでいる．両組織から人的リソースを出し合うことにより，互いが共通の課題に対してきちんとコストをかけて取り組んでいるという意志表明につながり，組織間の偏りから発生する政治的問題を減らすことができる．また，あまりに人数を多くすると調整や管理のコストが増え全体の動きが鈍くなってしまうため，一つのプロジェクトにつき 4〜5 名程度に抑えるのもポイントの一つである．

6. データの扱いとプライバシーポリシー

ビッグデータの利活用にはいろいろなメリットがある一方，ネガティブな側面として企業が個人情報を不当に扱うことへの懸念が挙げられる．節操なくデータを取って利活用を進めることでユーザーに不安な思いをさせてしまい，結果的にユーザーが離れては本末転倒である．日本には個人情報保護法や通信の秘密などの法律が存在し，データをうまく利活用するにあたってこれらの法律についてよく理解し，企業で掲げたプライバシーポリシーに関してうまく対応していくことも重要なポイントの一つである．

Yahoo! JAPAN では継続的にプライバシーポリシーを見直しており，必要に応じて改定を行っている．また，「社外秘」や「極秘」などデータの情報区分に応じてシステムを別に分けたり，特定の社員以外がアクセスできないようにするなど，システム面においても対応を行っている．事例に挙げた行動ターゲティング広告について不要と感じるユーザーに対しては，その機能を無効化するための「オプトアウト」機能を用意している．

データを取れば取るほど，また，利活用を進めていけばいくほど，ユーザーの満足度を向上させることにつながる一方で，少しやり方を間違えるだけで大きく信頼を損失してしまう．Yahoo! JAPAN では「ユーザーファースト」というキーワードを掲げており，これは「ユーザーが本当に求めているものは何かを追求して提供できるようにしよう」という，いわゆる顧客至上主義を意味する言葉である．本来サービスにあてはめる言葉であるが，データの扱いについても同じことが言える．データを取得するときはユーザーが不安にならないように配慮し，利活用を行う際はユーザーの満足につながるような形で提供しなければならない．ともすればデータは企業の持ち物のように思えるが，ユーザーファーストの観点からいうとデータはユーザーのものでありユーザーのために利活用されなければならない．

7. まとめ

Yahoo! JAPAN は数多くのサービスやアプリを提供しており，その裏では多分野にわたるビッグデータを扱っている．そして各サービスの至るところでデータを利活用しており，継続的に改善を続けている．大規模なデータを収集し処理するにあたり，Hadoop をはじめとする複数のシステムを特性に応じて使い分けている．また，データをうまく利活用するには組織や社内文化での整備，プライバシーポリシーなどデータの取り扱いなどもあわせて検討しながら進めることも重要である．

