お客様事例

NTT DATA
Trusted Global Innovator

レセプト情報・特定健診等情報データベース（NDB）

全国民のレセプト情報を格納する次世代NDB構築において Hadoop/Sparkの活用により高い処理能力とスケーラビリティを実現

全国民のレセプト情報を解析し、医療の現場を可視化する研究は、日本の未来の医療に大きな変革をもたらす可能性がある。全国民のレセプト情報は約100億レコードと、非常に規模の大きいデータであるが、これを格納する現行のレセプト情報・特定健診等情報データベース（NDB）では、アーキテクチャの課題を含む様々な要因で利便性が低いといった課題があり、レセプト情報の分析、可視化の実施が困難な状況であった。
「新たなエビデンス創出のための次世代NDBデータ研究基盤構築に関する研究」プロジェクトでは、様々な分野のスペシャリストを集結させ、次世代NDBデータ研究基盤の構築を実施した。特にシステム基盤には、非構造データをスケーラブルに取り扱うことが可能なHadoop/Sparkを採用することで、高い処理能力とスケーラビリティを持つシステムが完成した。

お客様の課題	導入効果
リレーショナルデータベース（RDB）情報に特有な、硬直したシステム構造	Hadoop/Sparkの導入による高い処理能力とスケーラビリティを実現
データ利用の各段階で発生する各種手戻り	「レセプトの基礎知識学習コンテンツ」および「ダミーデータが利用可能な学習環境」の整備による学習効率向上
各種研究に適したデータ構造に再編成するための煩雑なデータ変換作業	分析用データマートによる分析作業の効率性／利便性の向上

導入の背景と課題

現行NDBにおける硬直したシステム構造やデータ利用時のハードルの高さなどが大きな課題

医療機関の受診により発生した医療費について、その一部または全部を保険者が被保険者に給付してくれる医療保険。特に、原則として全国民が加入する日本の公的医療保険制度は、世界でも最高水準といわれている。一方、公的な国民皆保険制度がない米国では、高齢者や障がい者、低所得者などを除いて民間保険に加入する必要があり、医療保険制度改革が進められてはいるものの、医療費が高額であったり、受診は登録医療機関に限定されたりと、さまざまな問題を含んでいる。
日本において保険診療が行われた際、医療機関は健康保険組合などに請求するための診療報酬明細書「レセプト」を提出する。このレセプトは医療機関が毎月、患者ごとに1人1枚作成しており、診療報酬も診療内容に応じて細かく規定されている。そして、この膨大なレセプト情報を集約しているのが、国家的データベース「レセプト情報・特定健診等情報データベース（以下、NDB）」だ。
NDBは、2008年4月から施行されている「高齢者の医療の確保に関する法律」に基づき、医療費適正化計画の作成・実施・評価に必要な調査や分析などを行う目的で構築された。その内部には全国民を対象としたレセプトおよび、特定健診・特定保健指導の情報が格納されている。しかし、この現行NDBのデータを活用するにあたっていくつか課題が生じていた。

リレーショナルデータベース（RDB）情報に特有な、硬直したシステム構造

問題となっていたのが、RDB（リレーショナルデータベース）特有の硬直したシステム構造だ。NDBは近年、利用者の増加が見込めることに加えて、今後は介護レセプトや健診データとの連携も期待されている。しかし現在のシステム構造ではデータ形式の変更に追従するのが難しいうえに、物理的なサーバのリプレースを伴う段階的な手順でしかスペック拡張が行えず、データ量の増加に対して柔軟な対応が難しかった。

データ利用の各段階で発生する各種手戻り

研究者、分析者などのユーザーが、NDBのデータ仕様や実際のシステムに触れられる機会が少ないことも問題となっていた。NDBを利用するためには、極めて高度なセキュリティ要件を満たし、なおかつ膨大なデータを取り込み・処理できる施設が必要になるからだ。このハードルを埋めるべく、京都大学と東京大学にNDBへアクセスできる「NDBオンサイトリサーチセンター」が設置されている。しかし、事前に学ぶための環境がないため、

センターに出向いてから手探りでデータ分析の方法を考えなければならず、望んだ分析結果が出るまでの試行錯誤による手戻りで多大な時間を要していた。データ量は100億レコード以上にもおよぶため、NDBのシステムに慣れていないユーザーの場合、検索・分析の仕方によっては数日待っても結果が得られないこともある。

各種研究に適したデータ構造に再編成するための煩雑なデータ変換作業

ユーザーが各種分析作業を実施する際、研究・分析内容に合わせて、NDBデータの再編成やデータ変換等の煩雑な分析前処理を、ユーザー自身が実施しなければならない、ということも、NDBを利用する上での利便性を大きく下げる要因となっている。

特に大きな課題は、同一患者のIDが途中で変わった際に分析時の追跡が途切れてしまう「ID問題」であった。NDBはその特性上、匿名でデータを集積しているが、結婚や退職をはじめとしたライフイベント、さらに誤記などにより同一患者のIDが変わり、分析時にデータを追跡する障害となってしまう。

こうした現行NDBの課題を払拭するための次世代NDBの研究に向け、国立研究開発法人日本医療研究開発機構（以下、AMED）では2016年度の研究事業として「新たなエビデンス創出のための次世代NDBデータ研究基盤構築に関する研究」プロジェクトを公募・採択した。

選定ポイント

柔軟性の高いスケーラビリティに加えベンダーロックイン回避も必須要件に

本プロジェクトでは、京都大学医学部附属病院、奈良県立医科大学、東京大学大学院医学系研究科から、それぞれ医学と情報分野に精通したプロフェッショナルたちが集結。そして基盤構築に関しては、この分野に関して数多くの実績を持つNTTデータが参加した。

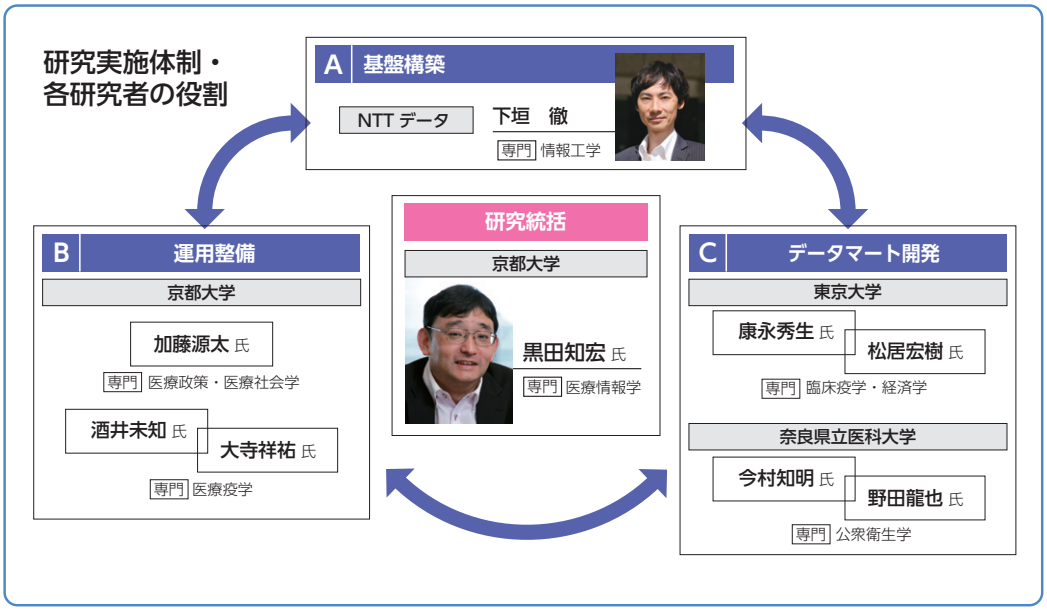
プロジェクトの統括を務めた、京都大学教授の黒田知宏氏は「公募要件としては、十分なスケーラビリティを確保することが挙げられます。従来のようにリプレースを伴う段階的な手順でしかスペック拡張が行えないのでは、今後見込まれるデータ量の増加や、介護をはじめとした新たな情報の追加に対応できないため、将来を見据えたスケーラビリティが必要不可欠です」と語る。

そしてもうひとつ、公共特有の調達条件として重要だったのが、ベンダーロックインを避ける仕組み作りだ。「本プロジェクトで次世代NDB実現の目処が立った後、実際に本番環境を構築する際は入札案件となります。しかし、特定ベンダーの技術に依存していると公平性が保てなくなるばかりか、価格競争ができずにコストの増加を招いたり、将来的なシステム改変にも影響が出てきたりと、国家プロジェクトとして数多くのデメリットが生まれてしまうため、ベンダーロックインを避ける必要がありました」と続ける。

このように、現行NDBの課題を払拭しつつ、「スケーラビリティ」と「ベンダーロックイン回避」という公募要件を満たす必要があった次世代NDB。その中核となるソフトウェアについては、分散処理技術によって大規模データの蓄積・処理を実現できる「Apache Hadoop」と、大量データの繰り返し処理を高速に行える「Apache Spark」の2つのオープンソースソフトウエアが採用された。高コスト・高性能なサーバを使わず、汎用的なIAサーバの追加だけで容易に性能を拡張できるシステムが構築可能である。

「課題解決と2つの公募要件を満たすソフトウェアフレームワークとしては、Hadoop以外に選択肢はありませんでしたね。そこでHadoopやSparkに関する技術力とノウハウを持つ企業を探したところ、コミュニティの中心にいたのがNTTデータでした。Hadoopに関して世界でも有数の開発実績を有するのはもちろん、コミュニティの中心でもっとも情報を発信できているというのは、技術の中核をしっかりと把握している証でもあります。また、国家プロジェクトという特性上、仕事の細やかさに加えて、企業体力があるかも重要なファクターとなるため、そうした点でNTTデータは最適な企業だと判断しました」と、黒田氏はHadoopおよびNTTデータを選んだ理由について語る。

京都大学 大学院医学研究科 医療情報学 教授
京都大学 医学部附属病院 医療情報企画部 部長
黒田 知宏 氏



プロジェクトの内容

各分野のスペシャリストが集結した高い技術力を誇るドリームチーム

本プロジェクトは2017年1月に発足し、以下のような役割分担にて遂行した。

京都大学
NDBの研究にかかる学習・試行・運用のすべてを実現できる環境の整備

東京大学
分析用データマートの作成。特にAIを視野に入れた次世代NDBの高度利活用に注力

奈良医科大学
分析用データマートの作成。特に患者の追跡性（名寄せ）の確立に注力

NTTデータ
Hadoop / Spark 等を用いた高い処理能力と拡張性を有する基盤の構築

こうした各分野のスペシャリストたちが集結した本プロジェクトを統括したのが黒田氏だ。黒田氏は医療情報学を専門としており、普段の業務では電子カルテのマネジメントおよび、そのデータを使った病院の経営マネジメントなどで手腕を発揮している。レセプトなど医療関連のデータ解析にも精通し、その実績と経験がプロジェクト全体の統括にも活かされている。

黒田氏は、「各分野におけるスペシャリストの方々が、それぞれのゴール設定や進捗状況を共有しながら進めてくれたので、マネジメントする立場としては非常に楽でした。まさに“エース級が集まったドリームチーム”という印象でしたね。その中でもNTTデータのHadoopや Spark に関する高い技術力に加えて、医療分野・レセプト業務に関する知見、数多くの実績に裏付けられた詳細なプロジェクトマネジメントには驚かされました」と語る。

導入効果

高い処理能力と柔軟な分析を実施可能とするデータ研究基盤の実現

Hadoop/Sparkの導入による高い処理能力とスケーラビリティを実現 処理時間は“時間単位”から“分単位”へ

2017年1月に発足した本プロジェクトは、要件定義の後、2017年4月からHadoop/Sparkを活用した次世代NDBの構築に着手、同年11月より第三者提供データを用いた試験をスタート、2018年1月にデータマート作成、分析、検証を実施した。例えば、これまでオンサイトでの処理に約4時間を要していたデータ抽出が、試行環境の整備と事前のデータの準備を行った次世代NDBではわずか数分にまで短縮されるという高い処理性能が発揮された。また、各大学の研究者が実際に次世代NDBシステムで分析を実施し、本システムの使いやすさとスピードを実感した。

加えて、Hadoop / Spark基盤のスケーラビリティに関する検証も実施し、サーバを増強することで性能が向上することも確認された。

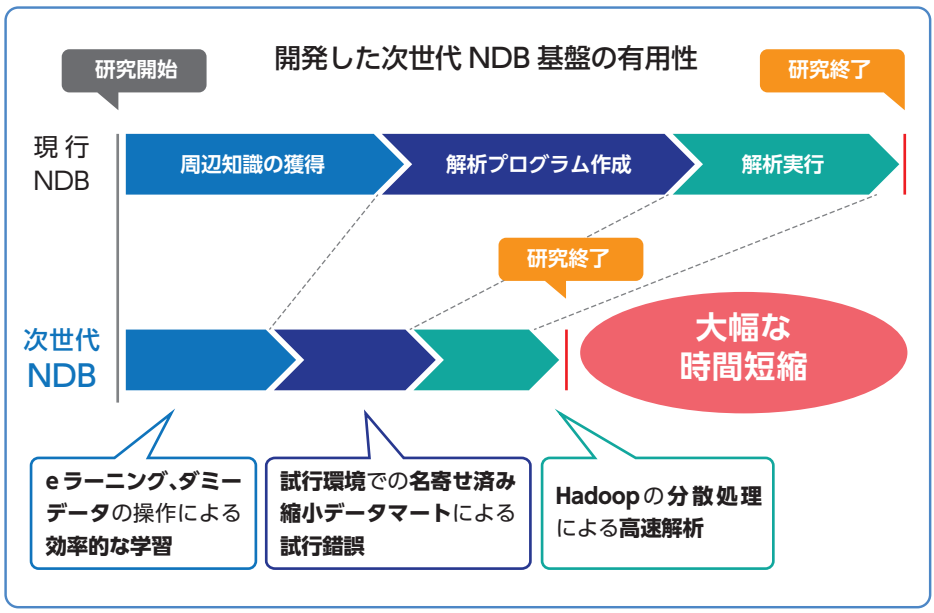
「レセプトの基礎知識学習コンテンツ」および「ダミーデータが利用可能な学習環境」の整備による学習効率向上

京都大学は、NDBのレセプトデータの知識を持っていないユーザー向けに、レセプトデータの基礎知識をオンラインで学習可能とするe-learningのコンテンツ及び、NDBのダミーデータに対して検索・分析を試行可能なNDB解析ハンズオン環境を整備した。これにより、ユーザーは、e-learningコンテンツでレセプトデータの基礎を学習・理解し、NDB解析ハンズオン環境にて、分析を事前に試行することで、ユーザーの学習効率を向上させ、新規にNDB分析研究を行いたいユーザーが、実際に分析を開始するまでのリードタイムの縮減を可能とした。

分析用データマート作成による分析作業の効率性／利便性の向上

NDBで分析作業を実施する上での非常に大きな障壁であったデータの再編成、変換作業のなかでも、特に大きな課題であった「ID問題」については、奈良県立医科大学より提供された名寄せアルゴリズムをNTTデータが実装した。これにより、次世代NDBでは、分析時におけるデータを追跡する際の障害が払拭され、患者単位での集計の精度向上を可能とした。

また、特によく利用されると見込まれる分析については、東京大学より提供された分析アルゴリズムをNTTデータが実装し、事前に分析用データマートを作成した。これにより、次世代NDBでは、各種研究にあたり、新規に分析用データマートを作成せずとも事前に作成した分析用データマートを用いて研究ができるようになり、分析作業の効率性や利便性向上を可能とした。



今後の展望

次世代NDBの更なる活用・展開を目指して

利用を通じて次につながる可能性を見出す一方で、すでに次の課題があることも分かった。

「日本の公的医療保険制度は全国民が加入していることに加えて、細かく価格が設定されているため、レセプトを見れば診療内容がわかります。これだけの数量と精緻さを持つデータは、世界中でも類を見ません。しかし残念ながら、従来はこのデータ価値を活用できる仕組みがありませんでした。次世代NDBは、この“宝”を活かすためのシステムです。また、本プロジェクトではHadoop / Spark基盤を用いたシステムというベースラインが完成しましたが、そこから新たに見えてきたこともあります。そのひとつが、次世代NDBの内容をより多くの人に利用してもらうためにどこにどのようなデータがあるのかを理解するための“カタログ作り”です。

今回の成果で、さまざまな切り口でデータマートを作成することが可能となったので、研究に活かしたいと想像が膨らみますが、たくさんありすぎても利用しづらいため、カタログ化して需要の判断も実施していきたいところです。すでにカタログ作りに必要な仕組み作りも提案済みですので、新たなプロジェクトに反映されると思います」と、今後の展開について語る黒田氏。

さらに黒田氏はNTTデータに対して次のような期待を語ってくれた。「NTTデータは情報革命の観点から、その技術力・信頼・実績ともに『この指とまれ』といえる数少ない重要な企業です。ぜひ今後も積極的に、官学に対して社会全体を変えるような提案を発信していってほしいですね」

実用化に向けてのベースラインを突破した次世代NDB。今後の日本医療界を支える重要な基盤として、NTTデータも全力でサポートしていく予定だ。

お客様 Profile

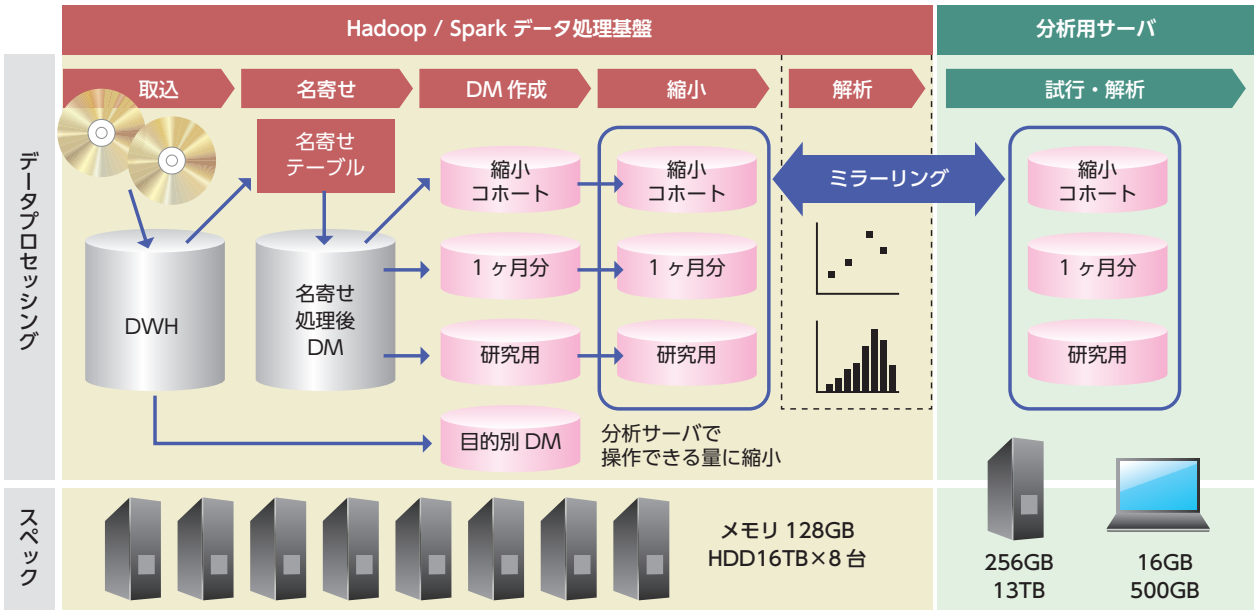
AMED

国立研究開発法人 日本医療研究開発機構（AMED）

所在地	東京都千代田区大手町1-7-1 読売新聞ビル20階
設　立	平成27年4月1日
概　要	医療分野における基礎から実用化のための一貫した研究開発を、大学などの研究機関の能力を活用しながら行うことを目的として設立。文部科学省、厚生労働省、経済産業省とともに内閣府が所管している。

URL　https://www.amed.go.jp/index.html

本稿に登場したサービス　Hadoop / Sparkを活用した医療ビッグデータ分析基盤



スケーラブルでコストパフォーマスの高い並列分散処理基盤「Hadoop / Spark」をプラットフォームとする、次世代NDBデータ研究基盤です。
約130億件という膨大な診療記録の利活用を促進するシステムとして世界的に注目される試みでもあります。

株式会社NTTデータ
第二公共事業本部 社会保障事業部　技術革新統括本部 システム技術本部 方式技術部
TEL 050-5546-9048　hadoop@kits.nttdata.co.jp　https://oss.nttdata.com/hadoop/

サービス自体に関するお問い合わせは、直接各サービス担当者へお願いいたします。

2018.10