厚生労働科学研究費補助金（政策科学総合研究事業（政策科学推進研究事業））
分担研究報告書

レセプト NDB（ナショナルデータベース）特別抽出データ利活用推進のための課題

研究分担者：
今中雄一　(京都大学大学院医学研究科医療経済学分野　教授)
研究協力者：
國澤　進　(京都大学大学院医学研究科医療経済学分野　講師)
大坪　徹也　(京都大学大学院医学研究科医療経済学分野　助教)

要旨
目的： NDB の特別抽出データを有効に利用するための環境を提言していく
方法： NDB の特別抽出データを受け取り、データベースとして解析を行った。この間に生じた問題点と改善点を検討する。また抽出データの受入れの内部環境要件を検討するにあたり、研究室データベース環境のセキュリティー診断を実施した
結果・考察：　１）NDB の特別抽出データの解析について課題が存在する。
○ 受取りデータ形式が特殊でかつテキストを RDB に格納するまで膨大な時間(1 か月以上)がかかりうる　→　Microsoft SQL の RDB 形式でのデータ渡しにより、利活用の効率が大幅に向上する。
○ 受け取りデータが膨大であり、その解析のためのデータサーバーが膨大に必要、しかしデータ整理後はそれらが不要になる　→　必要なセキュリティを確保したデータ部分のクラウドサーバーの活用により、必要なセキュリティを確保した利活用の効率が大幅に向上する。
○ 内部環境のみで解析まですべて終了させることが不可能、GIS システムやスーパーコンピューターでの解析が必要　→　万全のセキュリティを確保できるように単純なデータに落とし込んだものを解析環境に持ち出せることにより、必要なセキュリティを確保した利利活用の効率が大幅に向上する。
２）当研究室のデータ管理は、情報セキュリティーマネジメントシステム適合性の第三者審査登録機関による認証を取得して維持されている（国際規格 ISO/IEC 27001:2013, 国内規格 JIS Q 27001:2006）。この上で、ソフトウェア管理によるセキュリティーの脆弱性の存在を、専門機関に委託し診断したところ、重大な脆弱性はなく、継続的なセキュリティーの確保を行うことで、物理的にもソフトウェア的にも高いセキュリティーの維持が可能とみなされた。
結論： NDB データを有効に活用するためには、データの受け渡し方法から解析環境まで柔軟な対応が求められる。また、必要なセキュリティー対策を講じていることで、大学内研究室レベルで安全性の高いシステムが維持できる。

A．目的

NDB の特別抽出データを有効に利用するための環境を改善するための条件を検討し、改善のための施策案を提示することを目的とした。

B．対象・方法

1）NDB の特別抽出データの、データベースとしての有効活用に関する問題点と改善点の検討

当研究室では、NDB より特別抽出として、2016 年 3 月まで NDB データの提供を受けた。このデータを用いて解析を行い、報告を行っている。このデータ解析実施期間中に生じた問題点のうち、データベースを効率的に利用するにあたっての問題点を記録し、その解決策を考察した。

2）研究室サーバーという環境のセキュリティーの評価

当研究室では、情報セキュリティー管理方針は、情報セキュリティーマネジメントシステム適合性の第三者審査登録機関による認証を取得して維持されている（国際規格 ISO/IEC 27001:2013, 国内規格 JIS Q 27001:2006 認証登録番号 IS75998）。NDB 特別抽出に関しては別途セキュリティーを確保した運営を行っているが、今回、そのほかのサーバーシステムについて、自営（オンプレミス）運用でのセキュリティーの脆弱性について、専門機関に委託し診断を行った。

脆弱性診断を行う期間は、複数の候補を挙げ、最終的に株式会社ラックへ発注した。

診断は 2016 年 2 月 22 日、23 日にかけて行い、大学のファイアーウォールのさらに内部での脆弱性を診断するため、研究室内で直接ネットワークに接続し診断を行った。

実施された診断には下記が含まれた。

1．ポートスキャン：稼働サービス特定を実施し

2．市販脆弱性ツールによる診断：

(サービス、ミドルウェア、OS 等に存在する脆弱性を調査)

3．独自ツールによる診断および 2 の結果確認：弊社独自ツールを使用し②の診断では検出できない HTTP や SMTP 等の問題を診断、および市販ツールで検出された問題が存在するか確認

C．結果

1）NDB の特別抽出データの、データベースとしての有効活用に関する問題点と改善点の検討

I

受け取りデータ格納、元データからの抽出

現在特別抽出申出に際して、CSV ファイルを EXE（特殊なプログラム）で圧縮された、1,000 個以上にわたるファイルを受け取っている。これらを個別に解凍し、読み込み、RDB（最も解析に利用しやすいと考えられるデータベース形式）に格納するのに、かなりスペックの高いサーバーでも 1 か月以上かかる。全国データになるとさらに多くなる見通しである。このように RDB 格納に要する時間が膨大であり、研究実施スケジュールの遂行に多大な支障をきたす。

そこで、データ利用期間内に確実に上げるため、以下の二つを提案する

改善案１　　RDB のデータベースファイルとして提供を受ける

具体的には、研究室の指定する RDB のデータベースファイル形式で提供を受ける。例えば SQL Server 2014 Enterprise Edition に付随するクラスター化カラムストアインデックスによるデータベースの圧縮を施したデータベースを、データベースファイルとして提供受けることができれば、研究室でのデータ運用は、アタッチするだけで開始でき、データベース容量は非常に小さく受け渡しも簡便になり、その検索能力も高くなる。

※　データ提供としては、SQLDB と CSV 両方をもらうのが最良

改善案２　　受領データの格納形式や圧縮方法について、申請者が指定できるようにする。

現在の仕様､圧縮 EXE ファイル以外での提供が無理な場合､圧縮 EXE ファイルの一覧表と、それぞれに圧縮されているファイル一覧（内容を含む）を提供いただく

II　セキュリティーを確保しながらより合理的な解析環境を構築するための提案

特別抽出データは、申出者が管理権限をもつ物理ディスクにデータを格納し、外部ネットワークとは断絶した環境内で運用することでセキュリティーを確保するとされている。

一方、受領データの運用に必要なハードウェアのスペックはデータ受領前では未知であるため、複数年度で全国にわたる大規模データを研究に要する場合、事前に必要な器材を選定・調達することは極めて困難となる。特に、受領データ後にスペック不足が明らかとなった場合、追加の器材を調達するための予算は直ちに確保することは極めて困難となる。

受け取りデータを加工し、解析用に抽出するためのサーバーとして、全国規模の解析を行う際データが大量なり、研究室で用意できる環境では、能力の不足が起こり得る。従来、独立したネットワークにつながらないサーバーのみでの運用をセキュリティーの高い状態として考えられているが、この状態を維持するだけでは能力を増強するのが物理的にも予算的にも困難となってくる。

このことを改善するため、以下を提案する。

改善案３　クラウドコンピューティングの活用

解析用のデータベースを構築するまでに、大量のデータを展開する場所と、そのための高スペックなサーバーが必要となる。いったんデータベースを構築した後にも別途高性能のサーバーが必要となるが、必要なスペックが異なり、用途ごとに高性能なサーバーを購入し、設置・運営するのが非常に困難になる。クラウドコンピューティングでは、データ整理や、多変量解析など、その用途に応じてサーバースペックを可変できるため、柔軟で効率的な運用が可能になる。さらに、クラウド内でのみの解析運用を行う場合、データを一切「外」へ出す必要がなくなり、かえってセキュリティーの高い運用が可能になるとも考えられる。ただし、この信頼性はクラウドコンピューティングを提供する会社の信頼性に依存する。

改善案４　通常運用のサーバーによる取り込み、処理を許可してもらう、処理後は、データベースを独立した運用に移行する

従来の申請では、データ自体を特別に指定した独立した機体に格納することとしている。この方法では、目的とする能力が場面場面で異なり、いずれにおいても十分な性能で作業ができない。つまり、データの格納に必要なサーバー

（主に作業スペースなど）と、解析に必要なサーバー（主に計算能力）など、解析全体を通じて必要なサーバーが異なっている。

そこで、特にデータ整理を行う期間、通常ほかで運用している研究室内のサーバーを一時的に NDB 用に割り当て、研究用データベースを確立し、その後、独立した状態に戻す、という柔軟な運用を提案する。

具体的には、データ自体を NDB 専用のサーバー（NDB ストレージと呼びます）に置き、データアクセス、処理の命令を行うサーバーと分けます。この「頭」となるサーバーを、通常運用のサーバーから一時的に流用し、NDB ストレージ内のデータを整理し、整理後、頭を切り離す、という流れができる。

この利点は、有限な資源を効率的に利用できるほかに、「頭」は NDB ストレージを切り離した状態では通常運用ができるため、セキュリティパッチのアップデートなど、重要なメンテナンスがスムーズに行えることもある。

つまり、物理的なつながりを遮断する旧来の方法にこだわることなく、システムとしての接続可能性に依った運用を柔軟に行うことで、資源を有効利用できることが利点になる。

この前提としては、セキュリティーの高いサーバーの運用が確保されている必要がある。

改善案５　サーバーの概念の見直し

現在の運用では、物理的な違いを持って、サーバーが違うのでアクセス権が違うことを明確にしてる。具体的には、甲サーバー（元データ）と乙サーバー（解析データ）を物理的に分けて管理するなどを示している。しかし、いずれも高性能の処理が求められ、本来であれば能力を補完して運用するものになる。これは前述の改訂案２と同様の内容である。

このことを解決するために、サーバーに対する管理というものを、機能単位で明示し、そのアクセス権を制御することでセキュリティーを保つようにできる。

具体的には、1 つの物理的サーバー内（DB1）に、甲サーバーと、管理の異なる乙サーバーを構築し、それぞれ適切なアクセス権を付与します。例えば Microsoft では、Domain Controller と呼ばれる機能により、アクセス権の集中管理を行い、解析端末でのアクセス権との整合性を保つことができる。

このシステムでは、従来甲・乙、独立した 2 台のサーバーを、上記 DB1 と同等の機能を持つ DB2 を並列に処理させることで、どの処理についても約 2 倍の処理能力を得ることが可能になる。

III

解析環境の問題点

独立系の PC での解析の限界

最近では、解析対象のデータ量が膨大になってきている。また、解析手法自体も洗練されてきており、それに伴い必要な処理が高度化してきている。

データ量：単純に地域やデータ内容が増えればその分増加する

解析方法：単純な回帰モデル　→　マルチレベル

高機能 PC でも 1 昼夜必要

また、ネットワーク分析、GIS 分析など、専用の解析ソフトやサーバーが必要になってきている。

このため、独立系の解析環境に、大量データを扱える機材と、そのソフトウェアを準備することが、現実の予算的に困難となる。例えば GIS 解析を行うために、その解析サーバー自体が必要になり、一つ一つのソフトウェアが非常に高額となる。

また、実際には、マルチレベル分析は、スーパーコンピューター（大型計算機）による分析

が必要になるなど、いわゆる外部（通常運用場所）での解析が必須となる

このため、現行では独立したネットワーク、空間での解析のみではなく、次のように改定を行うことで、より迅速で有用な解析を実施できると考えられる。

改善案６　データを個票レベルであっても、単独では意味のないデータに落とし込み、通常環境での解析を行う

具体的には、データの加工はすべてサーバー室内（独立した環境）で行い、解析直前のフラグとなったデータのみを、通常環境へ移行し、解析を行う。

イメージとして、図表 1 を参照されたい。このイメージは、実データにまったく関係なく作成したデータであるが、実データで作成しても同様に、単独ではまったく意味を持たないデータとして作成が可能である。このイメージ図では説明をわかりやすくするため変数に意味が若干推測しやすくしているが、変数名を 2 文字以内にするなどの制限により、さらに単独ではわかりにくいデータとして規定することができる。

集計され、許可されたデータのみを持ち出すのではなく、個人情報を全く含まず、かつ単独で意味をなさない加工されたデータを、持ち出し、通常環境で解析を行い、その解析結果の解釈を含めて初めて、意味のあるデータとしての持ち出し許可を判断してもらうことでもセキュリティーを確保しつつ運用が可能と考えられる。

２）研究室サーバーという環境のセキュリティーの評価

研究室で運用されるサーバーおよび PC をサンプリングし、診断を行った。サーバーは、実サーバーのほか、仮想サーバー、およびサーバー管理ポートも診断に含めた。

Windows サーバーの Windows Update の遅滞に起因する脆弱性、サーバーの BIOS レベルでの管理ポートに起因する脆弱性が、緊急性の高い事項として指摘をされた。いずれも速やかに対応を行った。そのほか中等度以下の脆弱の可能性としての指摘を受けたが、必要に応じて随時対応を行っている。重大な脆弱性は見つからなかった。

D．考察

１）NDB の特別抽出データの、データベースとしての有効活用に関する問題点と改善点の検討

NDB データの性質上、どれだけセキュリティーを確保できるかに焦点がおかれ、非常に制約の高い状況での解析が求められている。今後、ますます高度な解析が求められるようになる中、セキュリティーが確保できる中での、柔軟な運用が期待される。

２）研究室サーバーという環境のセキュリティーの評価

継続的なセキュリティーの確保を行うことで、物理的にもソフトウェア的にも、高いセキュリティーの維持が可能であった。

E．結論

大学研究室内の運用にてセキュリティーの高い環境を維持できることが第三者的にも示され

た。
　NDB のデータの効率的な運用のためには、柔軟な対応が期待される。

Ｆ．健康危険情報

特になし

Ｇ．研究発表

特になし

図表１　イメージとして、何らかの解析データにまったく関なく創作したデータ

staindex	sex	agec	outcome	sev	com1	com2	com3	com4	area1	area2
1	2	4	1	1	1	1	1	1	232	232
2	1	4	0	4	0	0	1	0	176	176
3	1	1	1	3	0	1	1	0	52	51
4	1	1	1	3	1	0	1	1	319	321
5	1	2	0	2	1	0	1	1	158	157
6	1	5	0	4	0	1	1	0	165	166
7	2	2	0	4	0	0	0	1	200	201
8	1	5	0	1	0	1	1	1	393	394
9	1	4	1	3	0	0	1	0	419	419
10	2	3	0	1	1	0	0	1	352	353
11	1	2	0	1	1	0	1	1	131	133
12	1	3	0	4	1	1	1	0	180	180
13	1	3	1	4	0	1	1	1	364	363
14	1	5	0	3	1	0	1	0	246	247
15	2	3	1	4	1	0	0	0	267	267
16	2	3	1	1	1	1	0	1	171	171
17	2	1	1	3	0	0	0	1	304	305
18	1	4	0	3	0	1	1	0	130	131
19	2	4	0	3	1	1	1	0	278	279
20	2	1	1	2	1	0	0	0	256	257
21	1	4	0	4	0	0	0	0	382	383
22	2	2	0	2	0	0	1	0	267	266
23	2	4	1	3	1	0	0	1	322	322
24	2	3	0	2	0	1	0	0	88	90
25	2	3	0	1	0	0	1	0	393	394
26	1	2	0	2	1	0	1	0	311	313
27	2	5	1	4	1	1	0	0	361	360
28	2	2	0	2	1	1	0	0	187	189
29	2	3	1	1	1	1	0	1	405	407
30	1	2	1	3	1	1	0	1	221	220
31	2	4	0	2	1	0	0	0	190	189
32	2	1	0	2	1	1	1	1	338	340
33	2	4	0	4	1	1	0	0	156	156
34	1	5	0	2	1	0	1	0	118	120
35	2	5	0	2	1	1	0	0	115	116
36	2	2	1	3	1	0	1	0	293	293
37	1	5	1	3	0	0	1	0	43	44
38	2	5	1	3	1	0	0	1	313	314
39	1	4	1	4	0	1	1	1	216	215
40	2	1	0	2	1	0	0	0	224	225
41	2	1	1	1	1	1	1	0	367	367
42	2	5	0	4	1	0	1	0	420	421
43	2	1	1	3	0	1	0	0	49	51
44	1	1	1	2	0	0	0	1	154	153
45	1	2	1	3	0	0	1	1	127	128
46	2	2	0	3	1	1	0	1	447	448
47	2	4	1	1	0	0	0	0	404	406
48	2	3	0	1	1	1	1	1	64	66
49	1	2	0	4	1	1	1	1	209	210
50	1	3	1	1	0	0	1	1	304	304
51	2	2	1	3	1	1	1	0	301	300
52	2	2	1	1	1	0	1	0	141	141
53	2	1	0	1	0	1	1	1	103	103
54	2	4	1	2	1	1	1	0	166	167
55	2	3	0	2	0	1	0	1	349	350
56	2	1	0	2	1	1	0	0	122	124
57	1	5	0	1	0	1	0	1	113	115
58	2	3	0	4	0	1	0	1	373	373
59	1	4	1	1	1	1	0	1	446	445
60	1	2	0	1	1	0	1	1	131	131
61	2	4	0	1	0	0	1	0	201	203
62	1	1	1	2	1	0	1	0	231	230
63	2	5	0	4	1	1	1	1	358	359
64	1	4	0	3	0	0	0	1	162	162
65	1	1	0	4	0	0	0	0	218	220
66	2	2	1	2	1	0	1	1	450	452
67	1	2	1	1	1	0	1	1	237	239
68	1	4	1	2	1	1	1	1	210	211
69	2	4	0	2	0	0	1	0	295	296
70	1	2	1	2	0	0	1	1	282	282
71	1	5	0	1	0	0	1	0	144	146
72	1	4	1	2	1	1	0	1	263	265
73	2	1	0	4	0	1	0	0	9	9
74	2	2	1	4	0	0	0	0	19	18
75	2	3	0	4	1	1	0	1	173	175
76	1	3	0	4	1	1	0	1	181	182
77	2	2	0	2	1	0	0	1	346	346
78	1	3	1	2	1	1	1	0	116	118
79	2	5	0	2	0	0	0	0	393	393
80	1	1	1	1	0	1	1	1	303	303
81	1	1	0	1	0	0	0	1	268	268
82	2	2	1	4	1	0	1	1	444	446
83	2	2	0	4	0	0	1	0	125	124
84	2	5	0	4	0	0	0	1	66	65
85	1	4	1	3	1	1	1	1	130	132
86	2	2	1	3	1	0	0	1	67	66
87	2	2	0	2	0	0	0	1	289	288
88	2	2	1	2	0	1	0	1	297	298
89	2	5	0	1	0	0	0	0	56	58
90	1	5	0	1	0	1	0	1	321	320
91	1	5	1	4	1	0	0	1	247	248
92	2	3	0	4	1	0	0	0	128	129
93	2	1	1	1	1	1	1	0	437	437
94	2	4	1	1	1	1	1	1	70	70
95	2	1	0	4	1	0	1	0	68	67
96	2	3	0	4	0	0	1	0	275	277
97	2	1	1	4	0	0	1	0	149	150
98	1	5	0	4	0	1	1	0	307	307
99	2	4	1	4	1	1	1	1	39	38
100	1	2	0	2	1	1	0	0	73	75