

爆発的に増加するデータを 会社経営に活かす

ー 最近のBI、DWHの動向 ー

株式会社ニューシステムテクノロジー

野田信昭、高橋成明

株式会社 ニューシステムテクノロジー

- 代表取締役 野田信昭
- 港区新橋2丁目12番17号 新橋INビル7F
<http://www.kknst.com>
Tel 03-3597-0031 Fax 03-3597-0032
- 設立 平成18年4月
- 資本金 1億2,000万円
- 事業 ソフトウェアパッケージ販売、保守
- プロダクト
 - Chakra リアルタイムデータベースセキュリティ
 - Ingres VectorWise 分析・レポーティング用高速DB
 - CoSORT ソート／ETLソリューション
 - ClickIMRS ワーク・フォース・マネージメント・ソリューション

BI、DWHシステムの新しい動き

- Big Data & High Performance
- 新勢力の台頭
- 低価格化
- モバイル
- クラウド
- アプライアンス
- ソーシャル

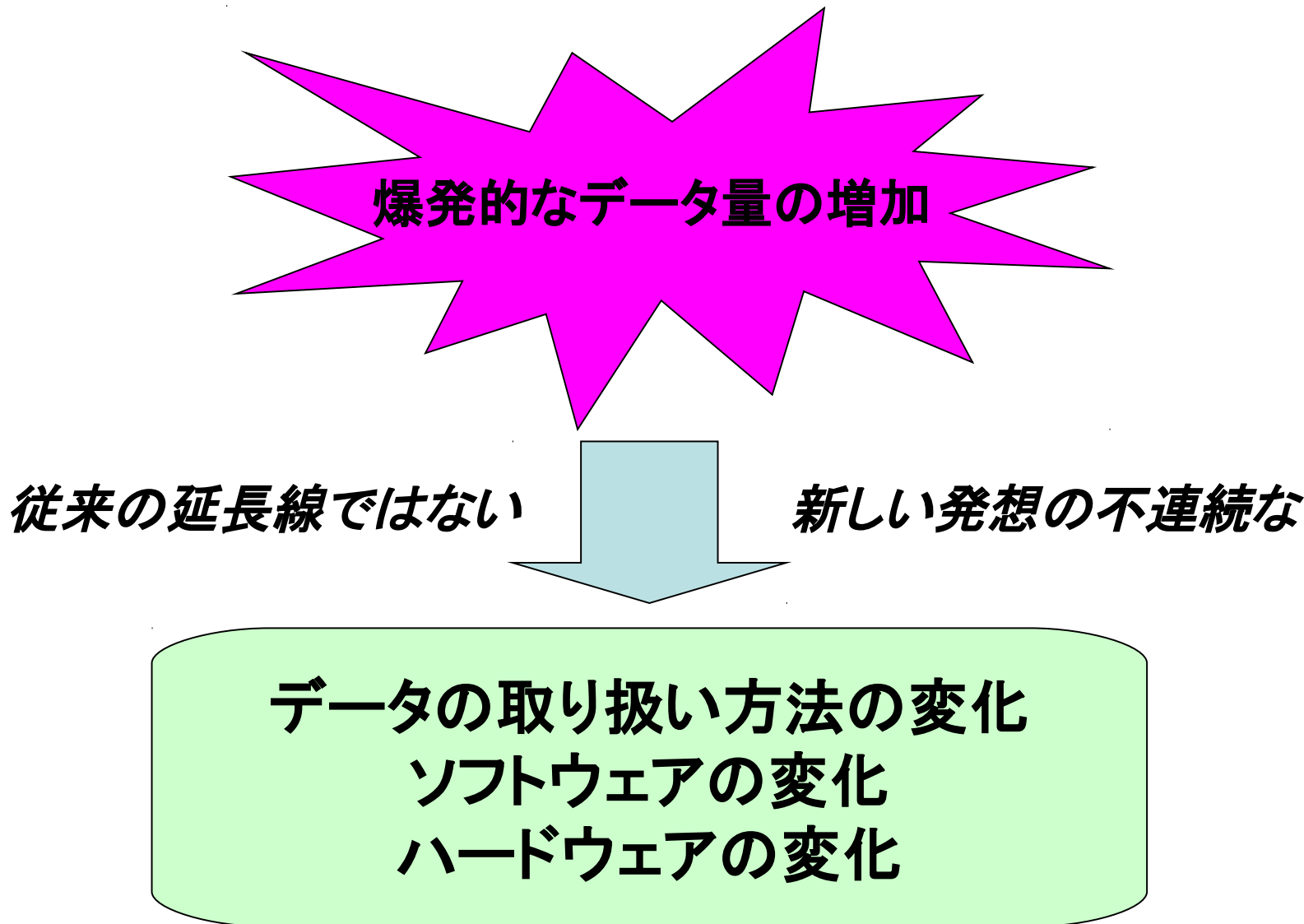
増え続けるデータ量

「2020年には、世界で作成されるデータ量は35ゼットバイトに達すると予測している。それらを全て保存したDVDを積み上げると、火星までの距離の半分になる。」

IDC White Paper: the Digital Universe Expands

1ゼットバイト【ZB】	10の21乗バイト	世界中の砂浜の砂の数
1エクサバイト【EB】	10の18乗バイト	5EB=これまでの人類の会話の全単語
1ペタバイト【PB】	10の15乗バイト	米国の全学術研究図書館のコンテンツの半分
1テラバイト【TB】	10の12乗バイト	大手病院の全レントゲン写真
1ギガバイト【GB】	10の9乗バイト	ベートーベンの交響曲第5番
1メガバイト【MB】	10の6乗バイト	薄い小説

IT業界に起きている変化



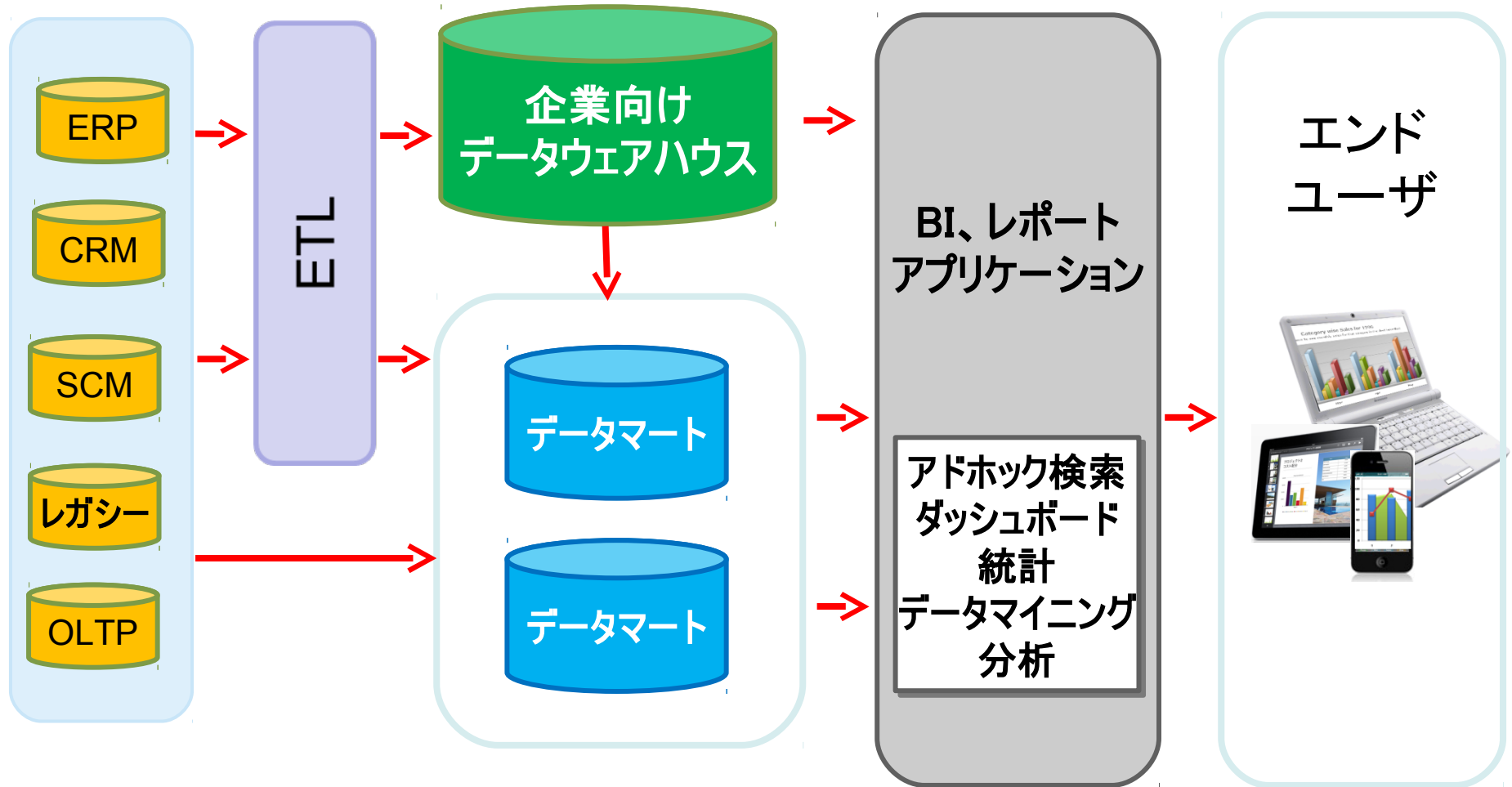
爆発的なデータ量の増加

従来の延長線ではない

新しい発想の不連続な

データの取り扱い方法の変化
ソフトウェアの変化
ハードウェアの変化

BI、DWHシステムの構成



これまでのBIベンダーの動き

- BO ⇒ SAP
- Cognos ⇒ IBM
- Brio ⇒ Hyperion ⇒ Oracle
- Hyperion ⇒ Oracle
- Microstrategy
- SAS
- (Dr.Sum)

新世代のBIベンダー

- Jaspersoft
 - OSS、帳票系、ETLはtalend
 - 販売会社: ワークブレイン ⇒ 日本法人設立?
- Pentaho
 - OSS、OLAP系、独自ETL、モバイル
 - 正規代理店: KSKソリューションズ
- Yellowfin
 - ベンダーソース、バランス良い、低価格、モバイル
 - 国内総代理店: 京セラ丸善システムインテグレーション
- Tableau
 - ベンダーソース、ダッシュボード、低価格、C/S⇒Web、
 - 国内代理店: エス・エス・エス
- QlikView
 - ベンダーソース、インメモリーBI
 - 日本法人: クリックテック・ジャパン

データ活用系データベース

ずっと**Teradata**メインできていたのが、
Netezzaが2007年くらいに殴り込みをかけて、
Oracle ExadataのV1が出たのが2008年ですか
ね。今は、
MicrosoftもDWHの**アプライアンス**製品を出してきましたし、
IBMは**ネティーズ**を買収したし、
EMCは**Greenplum**を買収し、
HPは**Vertica**を買収し、
SAPまでが**Sybase**を買収してHANAを作り．．．、

データ活用系データベース

- Teradata
 - アプライアンス、大規模向け
- Netezza ⇒ IBM
 - アプライアンス、大規模向け
- Oracle Exadata
 - アプライアンス、大規模向け
- Greenplum ⇒ EMC
 - カラム指向DB、オープンソース、アプライアンス化
- Vertica ⇒ HP
 - アプライアンス化

データ活用系データベース

- Sybase IQ ⇒ SAP
 - カラム指向DB
- SAP HANA
 - インメモリーDB
- AsterData ⇒ Teradata
 - カラム指向DB
- InfoBright
 - OSS、カラム指向
- Ingres VectorWise
 - カラム指向、ベクトル処理、CPUキャッシュ

データ活用

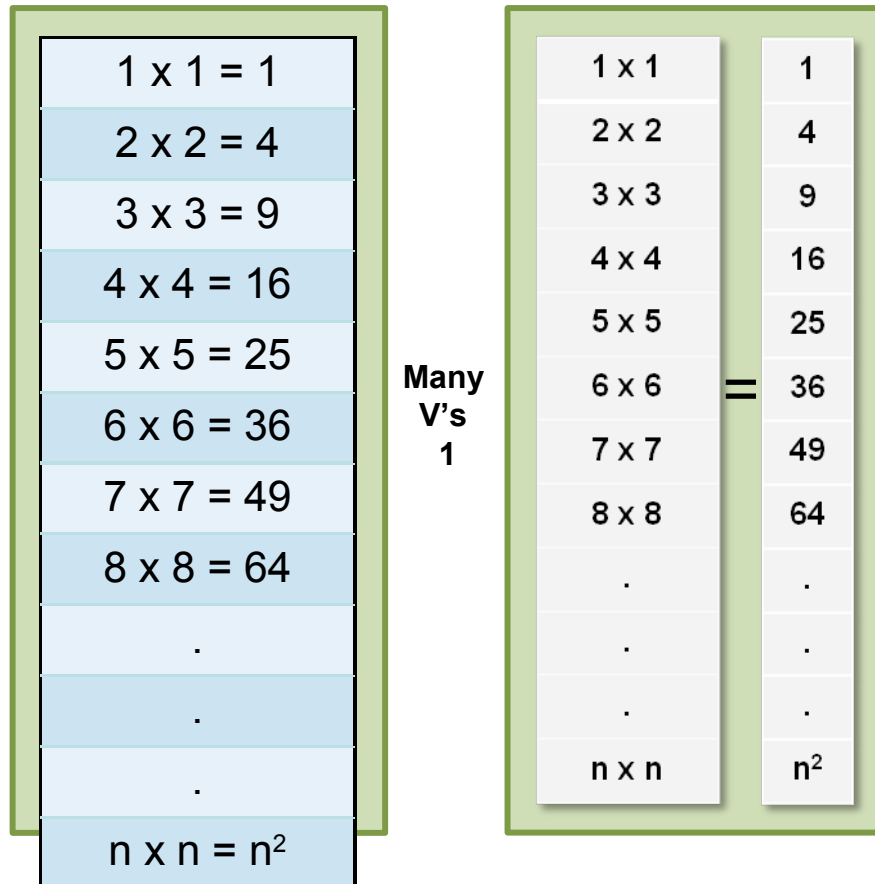
分析・レポートニング

データウェアハウス・データマート

データベースが **速い** 理由

ベクトル処理

- 単一の命令で、複数のデータ処理



SSE (ストリーミング SIMD 拡張命令)

16個の128bitのレジスタ

(Intel Sandy Bridgeは256bit)

- * 32bit float * 4

- * 16bit integer * 8

- * 8bit byte/char * 16

etc.

- * 加減算、積除算、比較、最大最小など

- * 文字列の処理でSSE4.2が効果大
(GROUP BYやLIKEなど)

SSE2はPentium 4, AMD64以降

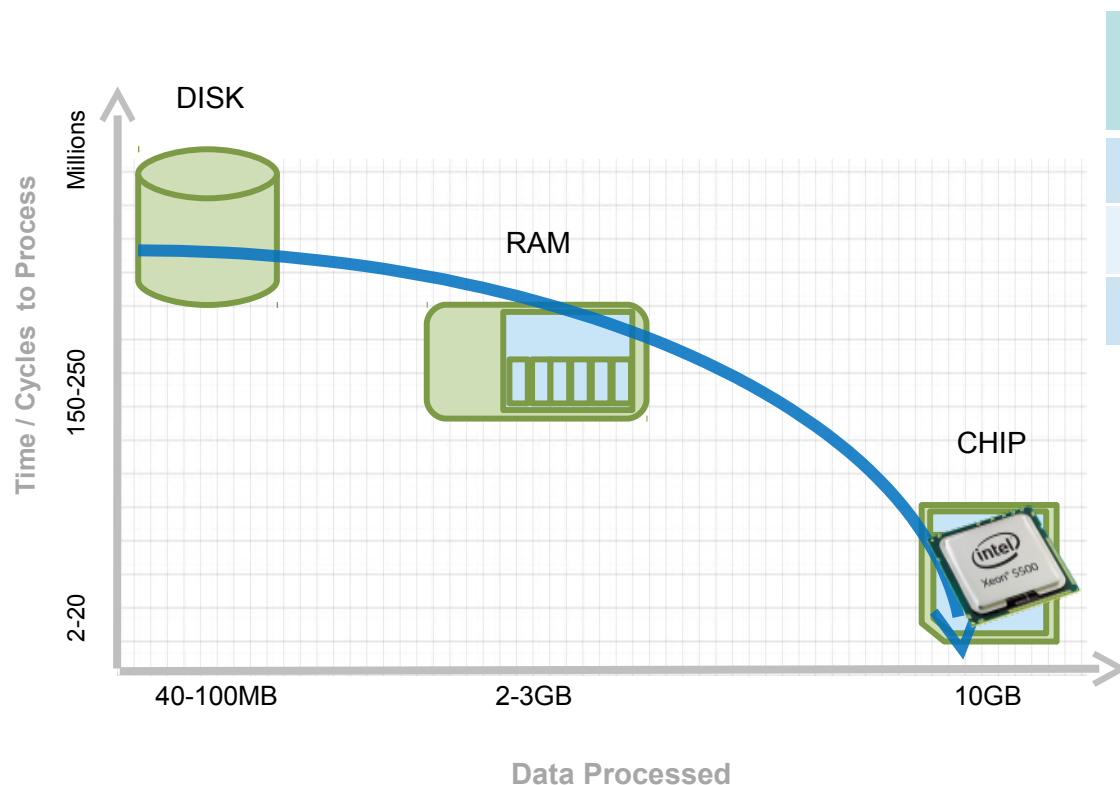
SSE3は後期Pentium 4,

後期Athlon64以降

SSE4は後期Core2以降

CPUキャッシュ内で処理

- CPUキャッシュのアクセスは、RAMより非常に高速
- すべてのベクトルがCPUキャッシュ内に収まるように最適化



	アクセス タイム	転送スピード (毎秒)
DISK	10-15ms	40-100MB
RAM	150-200ns	2-3GB
Cache	2-20ns	10GB

CPU	キャッシュ サイズ
Xeon	512KB/1-2MB
Xeon 50X0	4MB
Xeon 33X0	8 – 12MB
Xeon 75XX	12 – 24MB

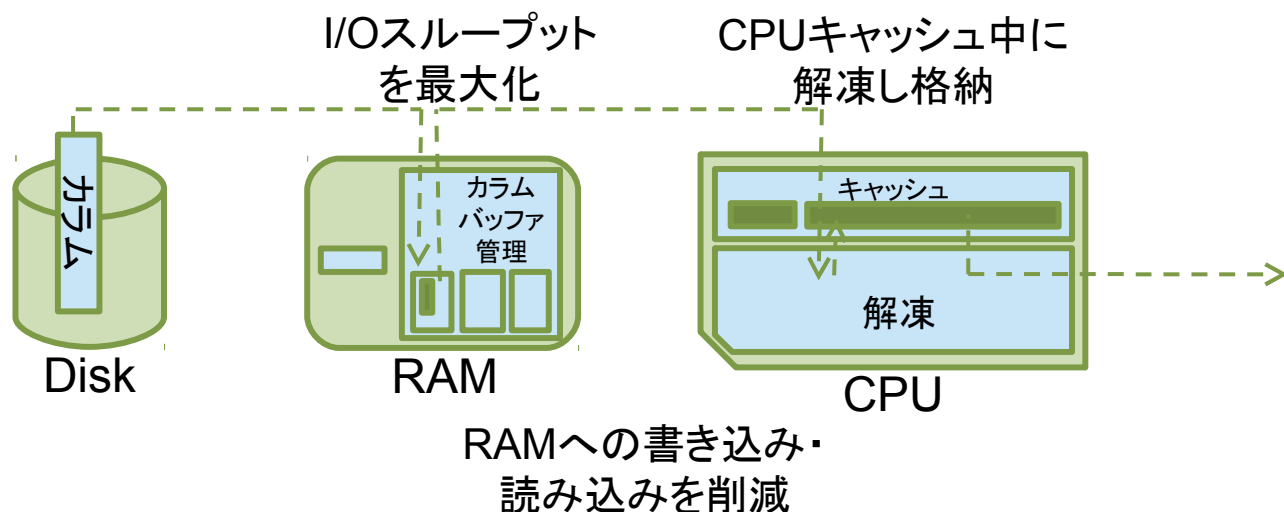
更新可能なカラム毎の格納方式

- 必要なデータだけにアクセス
- 効率的な”増分更新”が可能
 - 以前のカラム毎の格納方式では、弱点だった

Cust_Num	Cust_sur_name	Cust_first_name	Cust_mid_name	Cust_DOB	Cust_Sex	Cust_Add_1	Cust_Addr_2	Cust_City	Cust_State
46328927956	Jones	Steven	Sean	17-JAN-1971	M	333 StKilda Rd		Melbourne	Vic
98679975745	Smith	Leonard	Patrick	04-APR-1964	M	Unit 12, 147 Trafalgar Sqr		Birmingham	London
52634346735	Rogers	Cindy	Carmine	11-MAR-1980	F	Belmont Rail Service	421 Station St	Belmont	CA
346737347347	Andrews	Jenny		14-SEP-1977	F	Apt1, 117 West 42 nd St		New York	NY
88673477347	Cooper	Sheldon	Michael	30-JUN-1980	M	Ingres Corporation	Level 2, 426 Argello St	Redwood City	CA
34673447568	Kollwitz	Rolf		22-DEC-1975	M	IBM Headquarters	123 Mount View Crs	Atlantic City	PN
99554443044	Wong	Penny	Lee	13-NOV-1981	F	Ming On Tea Shop	1777 Moa Tien Rd	Ming Now Restaurant	Shanghi

自動的な圧縮と解凍

- 複数のアルゴリズムを使ってカラム毎に圧縮
 - 最適なものをIngres VectorWiseが自動的に採用
- 解凍はベクトル処理
 - CPUキャッシュ中でデータ処理



ブロック単位の自動インデックス

- データブロックごとに、最小値/最大値を保持
 - 小さく、読み込みも速い(カラムサイズの0.1%以下)
- 候補となるデータブロックを効率よく見つけることが可能になる
- 自動作成、自動メンテナンス

データベース設計、管理

- 手間のかからないデータベース設計
 - 特別なエキスパートによるスキーマ設計が不要
 - インデックス設計やマテリアライズドビュー、投影などが不要
- 継続的なチューニングが不要

ハードウェアコストの低減

- 安価なx86ベースの汎用サーバー・PCで動作
 - 1 CPUで20CPU以上の作業をこなす
 - 単一のサーバで複雑な複数ノードのクラスターを超える
 - 運用や空調のエネルギー使用・コストを低減
 - メンテナンスコストを低減し、故障も少なくなる

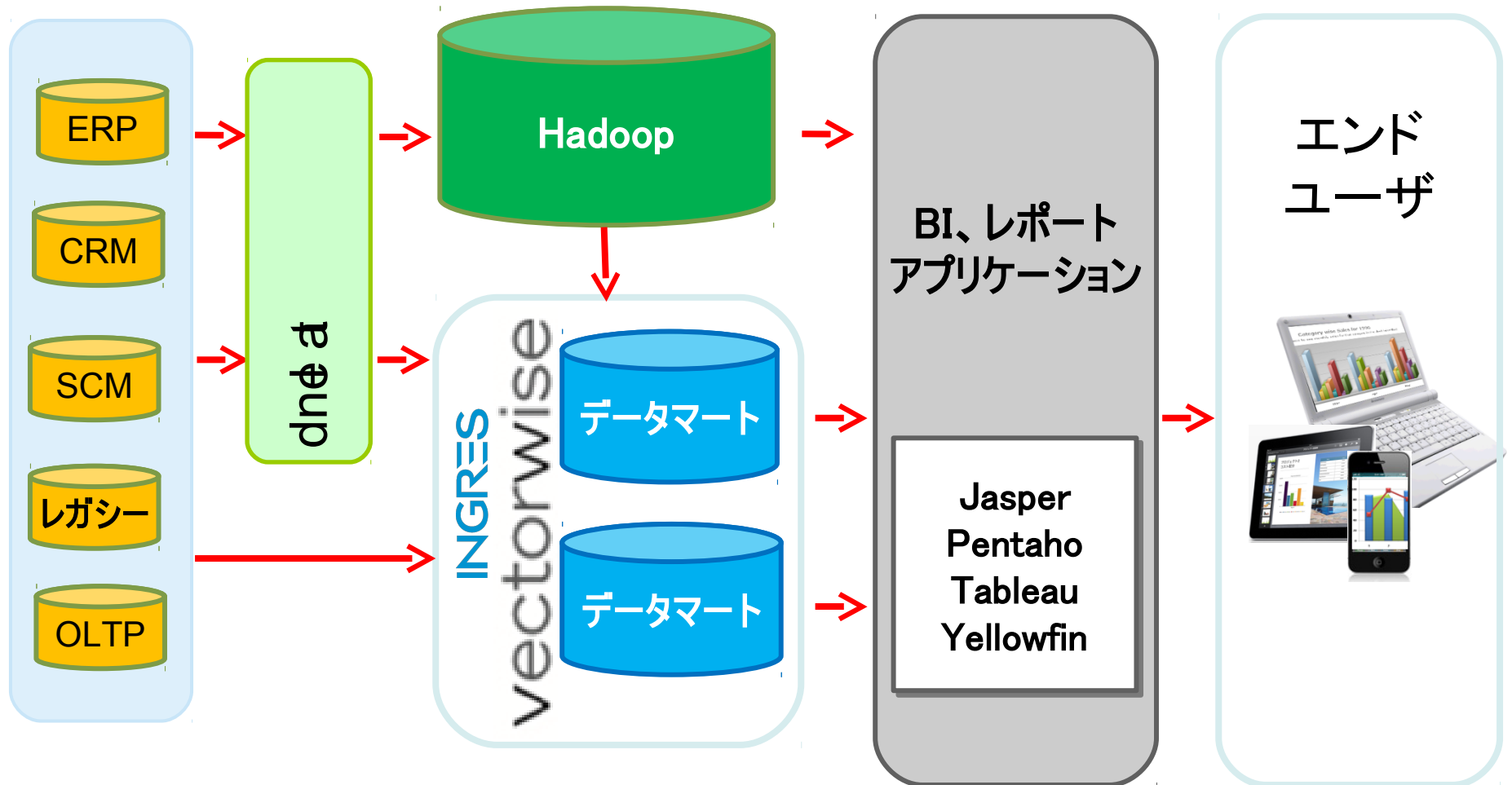
NoSQLデータベース

- キーバリュース型
 - 米グーグル「MapReduce」
 - 同オープンソース版「Hadoop」
 - アマゾン「SimpleDB」
- ドキュメント指向型
 - CouchDB (OSS)、MongoDB (OSS)
- グラフ指向型
 - Neo4j (OSS)、Sones (OSS)
- 列指向型
 - Hbase (OSS)、BigTable (Google)

ETL、EAI

- Talend
 - OSS、Talendジャパン
- Infomatica
- DataStage ⇒ IBM
- DataSpider
 - アプレッソ
- Asteria
 - インフォテリア
- jBolt
 - マジックソフトウェア
- Waha !
 - ビーコンIT

今後のBIシステム



爆発的に増加するデータを 会社経営に活かす

ー 最近のBI、DWHの動向 ー

株式会社ニューシステムテクノロジー

野田信昭、高橋成明