

Gaining Insight from Big Data with SAP HANA: A Customer Case Story

Steve Lucas
May 16, 2012



What is big data?

Where is it going?

Volume

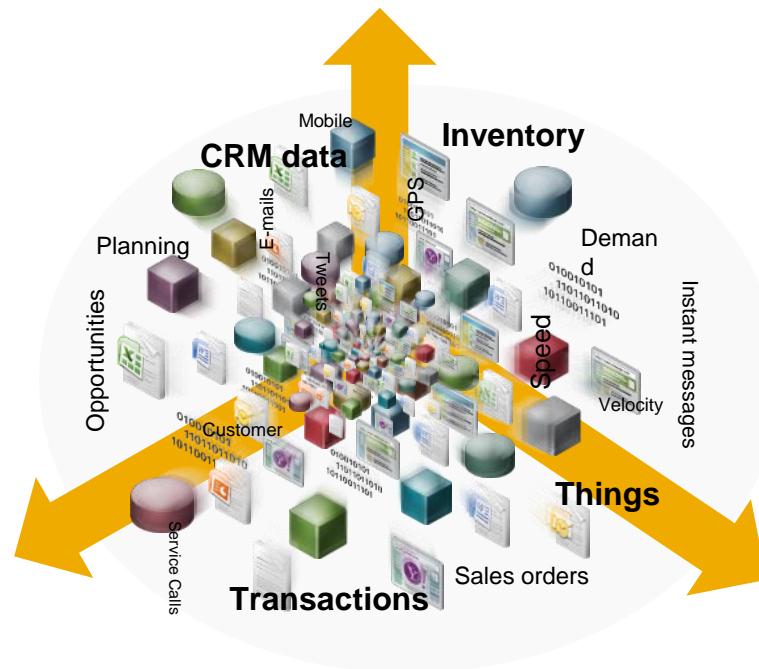
In 2005 humankind created 150 exabytes of information. In 2011 **1,200 exabytes will be created.**

The Economist

Velocity

Worldwide digital content will **double in 18 months, and every 18 months thereafter.**

IDC



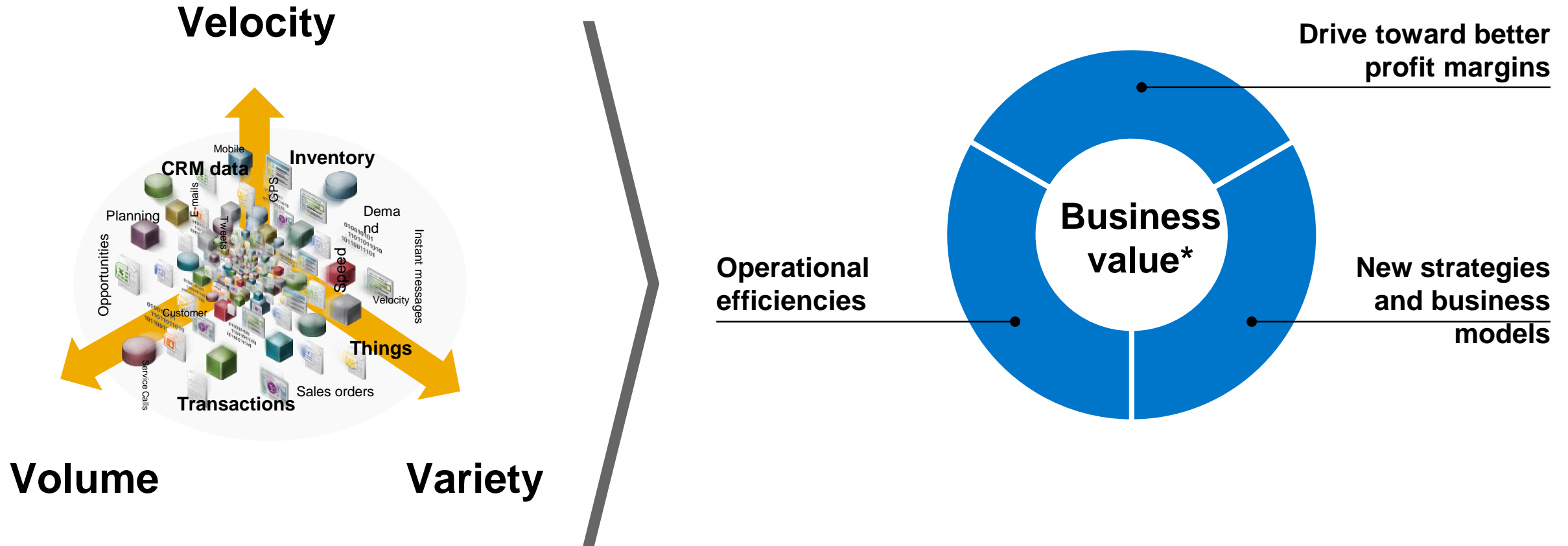
Variety

80% of enterprise data will be unstructured, spanning traditional and nontraditional sources.

Gartner Group Inc.

Big data matters

From jargon to transformational business value*



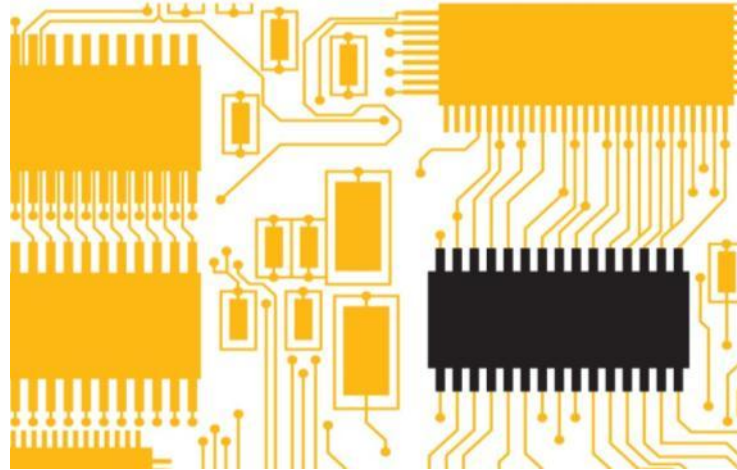
*A McKinsey study has found huge potential for big data analytics with metrics as impressive as 60% improvement in retail operating margins, 8% reduction in (U.S.) national healthcare expenditures, and \$150 million savings in operational efficiencies in European economies. Source: *"Big Data: Next frontier for innovation, competition, and productivity,"* by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. May 2011.

Big data opportunity

As a business manager you want to . . .



Drive better profit margins



Improve operational efficiencies



Uncover new strategies and business models



Big data

Open-source solutions

Present

Big data applications?

Process

Azkaban Oozie Pig Hive

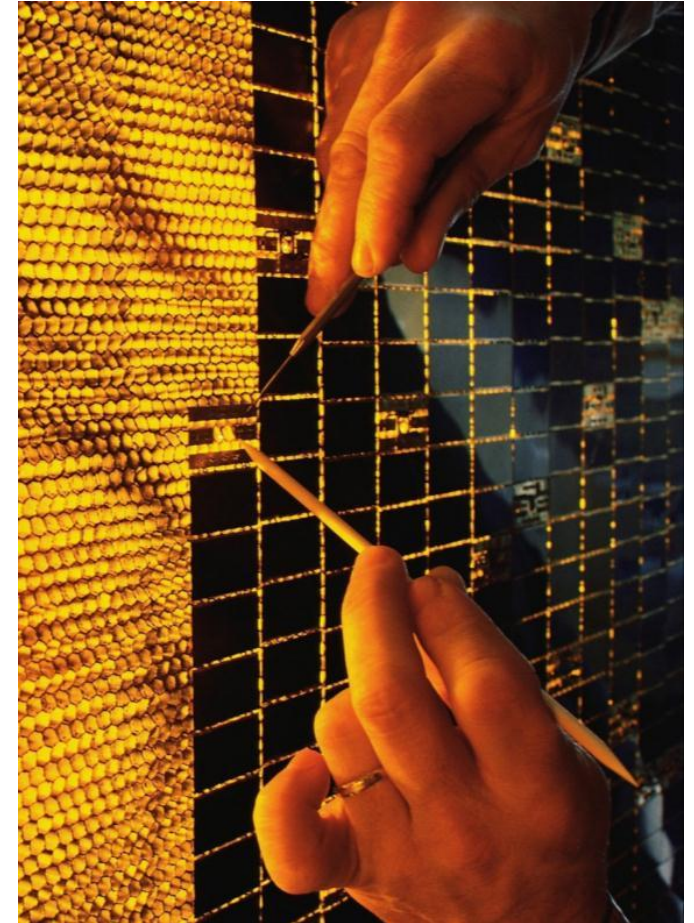
Hadoop MapReduce S4 Storm

Store

Voldemort Cassandra Hbase

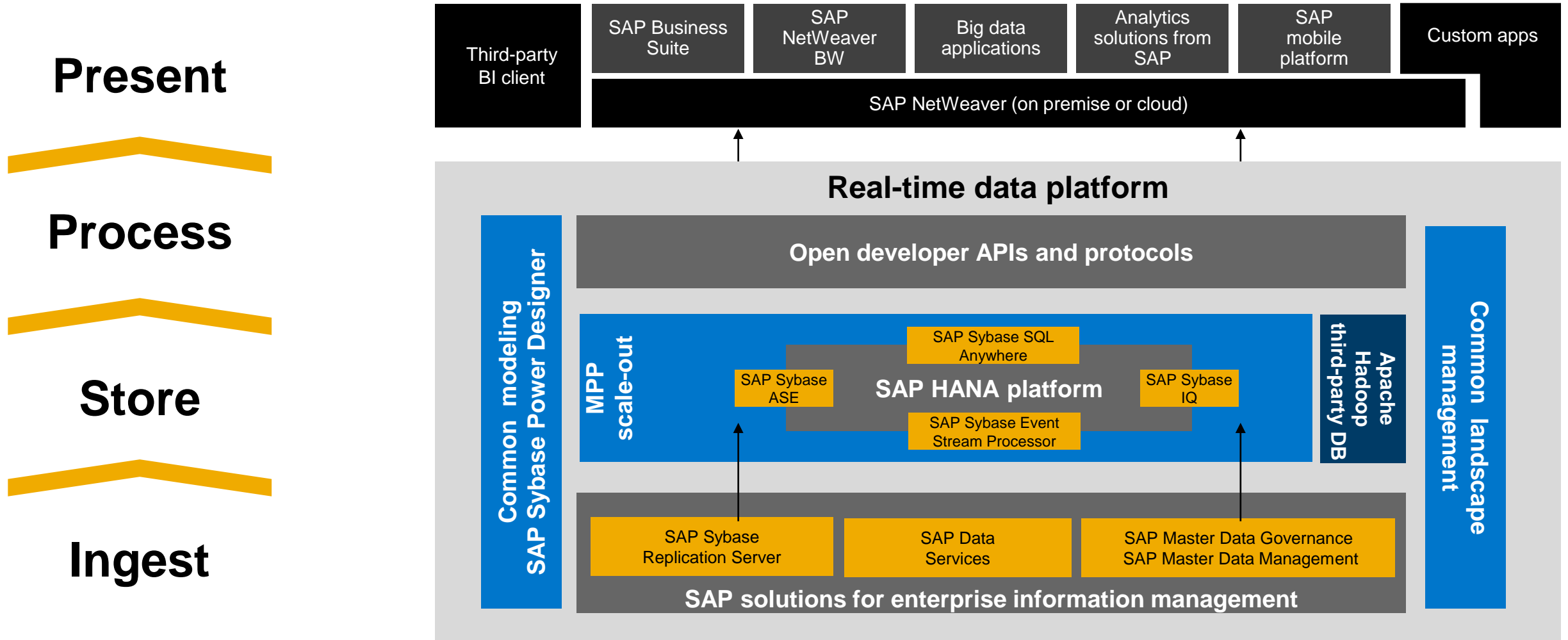
Ingest

Kafka Flume Scribe



Real-time data platform

Real-time insight and foresight, comprehensive integration, and packaged business scenarios



SAP HANA platform

Accelerated advanced analytics on big data with in-memory computing

The power of SAP HANA:

- Gain in real time – data insights from any data source
- Run faster – analyze big data at the speed of thought
- Get flexibility – eliminate prefabrication requirements
- Act broadly – manage large volumes of data
- Go deeper – predictive analytics via R on SAP HANA and Apache Hadoop



SAP HANA in action

Comprehensive and real-time big data solution to deliver new business opportunities



Real-time big data analysis to improve profit margin

- Increased reporting time by 1131x and ability to manage over 2100x more data
- Increased data compression by over 600%



Smooth and comprehensive big data process to improve operations

- Able to handle over 100,000,000 records and runs 900x faster than before
- Gain insight from large and complicated data scenarios



New business opportunities to expand business models

- Identify driver mutation for new drug target
- Reduced genome analysis from several days to 20 minutes

Cancer genome analysis by leveraging SAP HANA

May 16, 2012

Mitsui Knowledge Industry

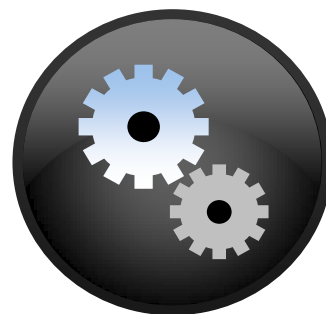


We offer the following services to pharmaceutical companies, universities and research institutes



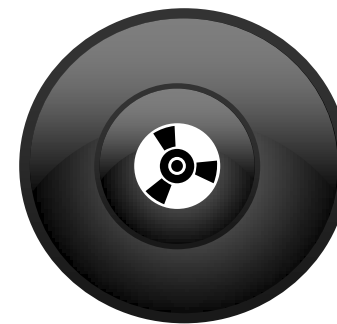
Consulting

- Providing our bioinformatics consultants' expertise to enable client's research project to accelerate



Systems development

databases for genome, proteome and metabolome
custom system for in-silico drug discovery, pathway analysis and biomarker discovery



Products sales

- Develop and sell software for bioinformatics analysis



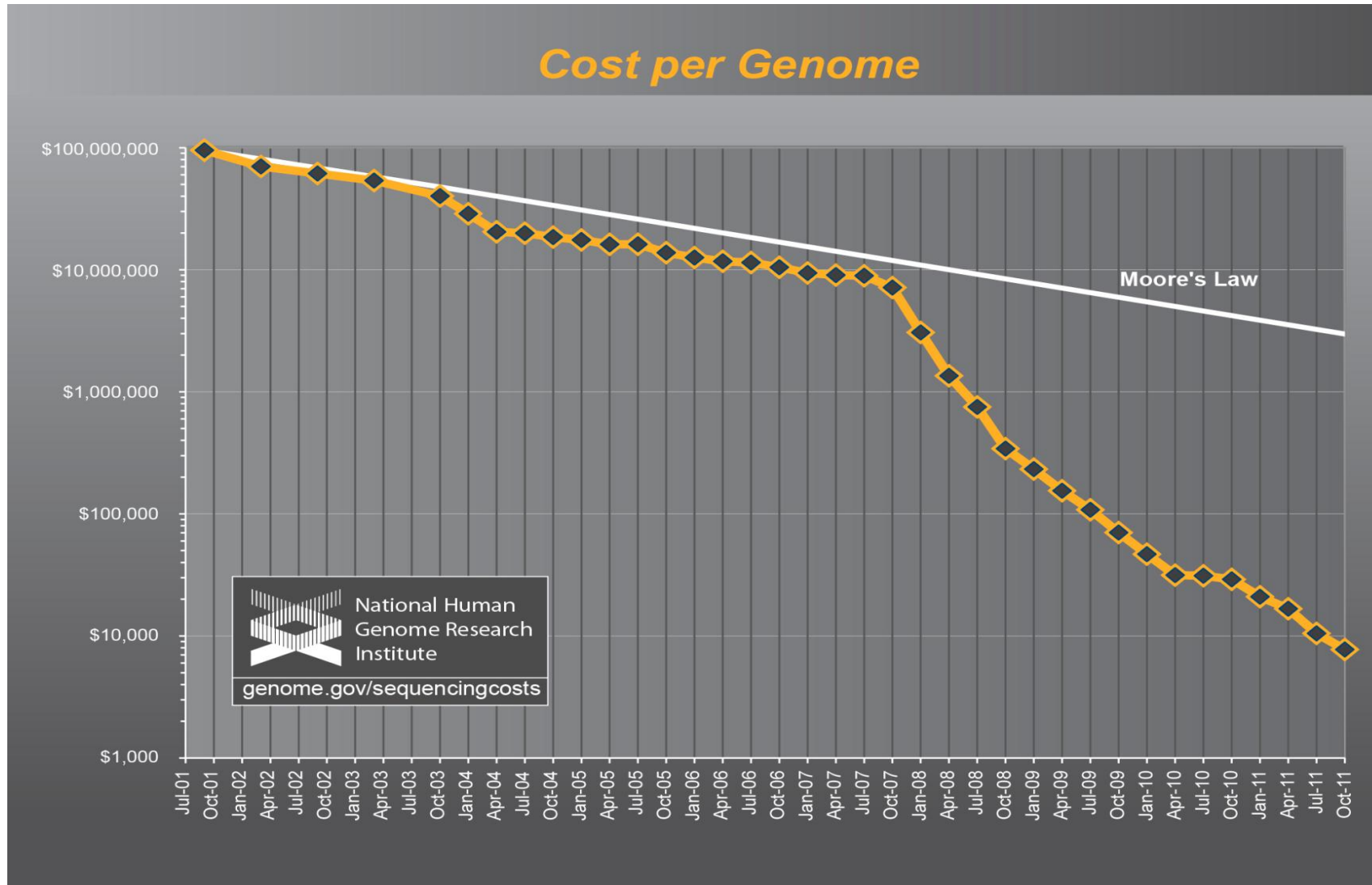
Research

- Think tank

Our goal: analytics for personalized medicine



Sequencing cost per genome



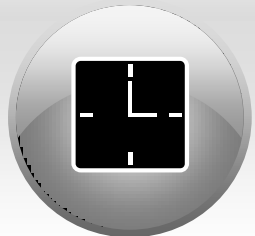
Toward personalized medicine



Time and cost for sequencing are reducing



Data analysis remains time consuming task



Process time:
a few days



Complex process:
several analysis softwares,
Apache Hadoop and R

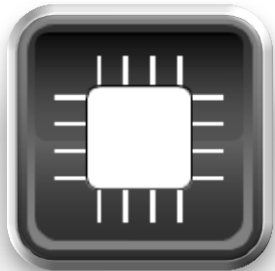


Data size:
a few gigabytes



Preprocess

- Alignment of DNA sequence from cancer to normal



Data analysis

- Variant calling from preprocessed sequencing



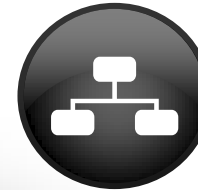
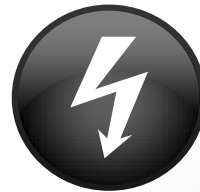
Annotation

- List of actionable mutated genes and related medicines
- Create predictive model (prognosis, driver mutation, etc.)

Why SAP HANA?

High performance

Powerful real-time computation capability



R + Apache Hadoop

Apache Hadoop connector and R integration



Reliability

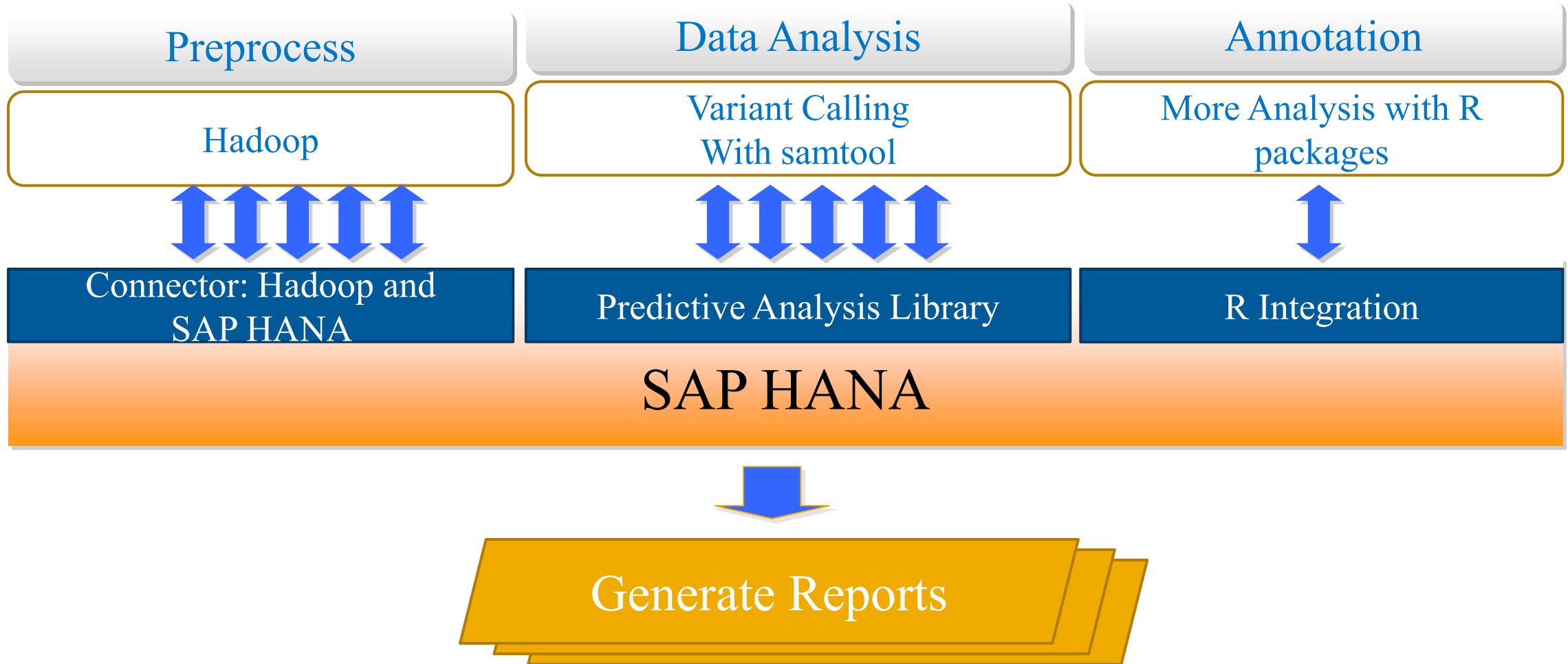
A platform for mission critical application



Analysis library

Embedded predictive analysis libraries

Genome analysis process on SAP HANA



Design optimized process for genome analysis



Original data analysis process :
2~3 days
 -Containing many manual tasks
 -Execute on low spec machine

Think Centre M91P
 CPU: i7 3.4GHz, 4 cores
 Memory: 16G



Optimized process:
2~3 hours
 on high spec machine

Sequence alignment	80.50 min (single node)
Variant calling	65.2 min (single node)

Apache Hadoop: 1 Z600 (2x6 cores namenode) + 9 Think Centre M91P
SAP HANA: CPU:4x2.6GHzx6 cores
 Memory: 512G
Network: Inter-connected with 1G Switch.



Accelerated process:
20~40 minutes
 with SAP HANA and Apache Hadoop

Sequence alignment	15.2 min (SAP HANA and Apache Hadoop)
Variant calling	19.5 min (SAP HANA)



Performance of recommend environment

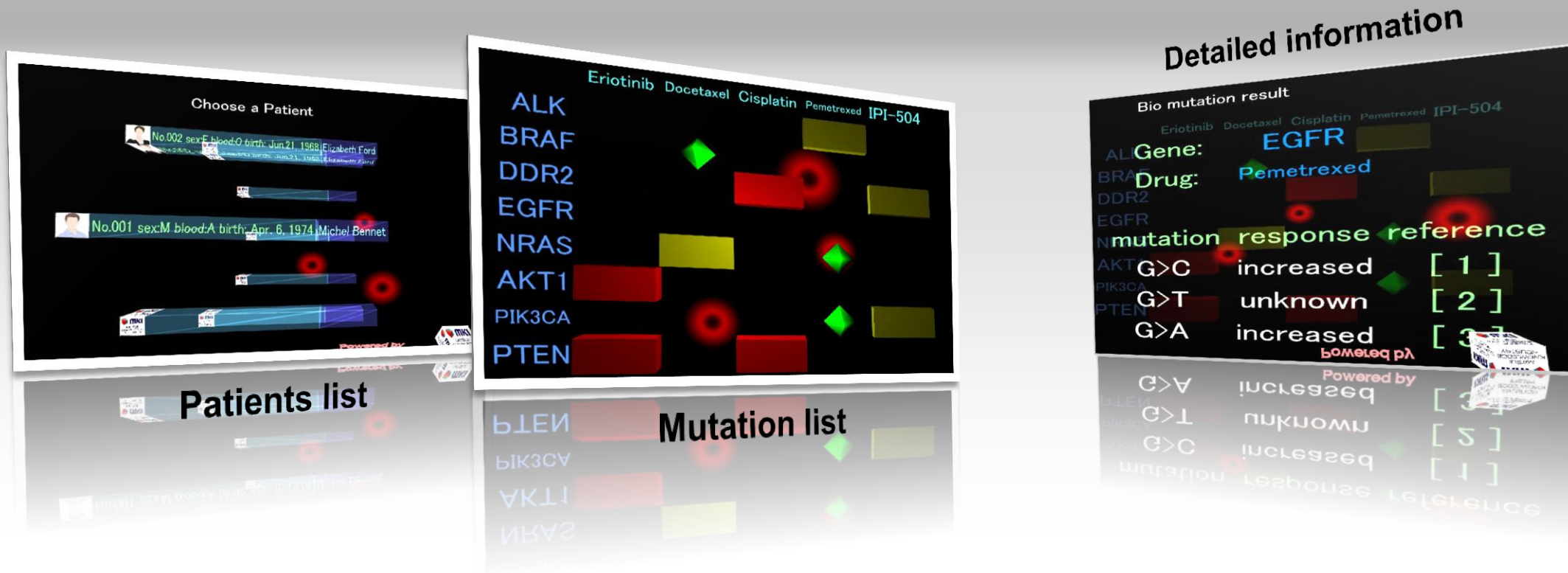
Recommended environment and estimated performance

* Input data: one 4.6 G PRQ file.

	Estimated time	Best environment
Sequence alignment	2.1	Apache Hadoop: 64 nodes (each with 8 cores, 2.43GHz)
Variant calling	6.5	SAP HANA: CPU:40 cores (80 threads)*2 Memory: 512G*2 Network: Inter-connected with 10G Switch.

Cancer genome analytics platform

One stop service for cancer genomic data analysis supporting personalized therapeutics



The interface is divided into three main sections:

- Choose a Patient:** A list of patients with their IDs, sex, blood type, and birth date. Two patients are visible: No.002 (Elizabeth Ford) and No.001 (Michel Bennet).
- Mutation list:** A heatmap showing mutations across various genes (ALK, BRAF, DDR2, EGFR, NRAS, AKT1, PIK3CA, PTEN) for different drugs (Eriotinib, Docetaxel, Cisplatin, Pemetrexed, IPI-504). Mutations are indicated by colored diamonds and rectangles.
- Detailed information:** A table showing the bio mutation result for a specific gene (EGFR) and drug (Pemetrexed). The table includes columns for mutation, response, and reference.

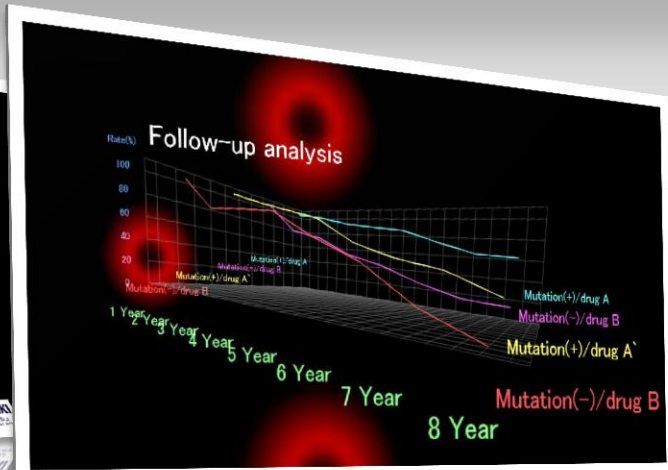
Gene:	Drug:	mutation	response	reference
EGFR	Pemetrexed	G>C	increased	[1]
EGFR	Pemetrexed	G>T	unknown	[2]
EGFR	Pemetrexed	G>A	increased	[3]

Cancer genome analytics platform

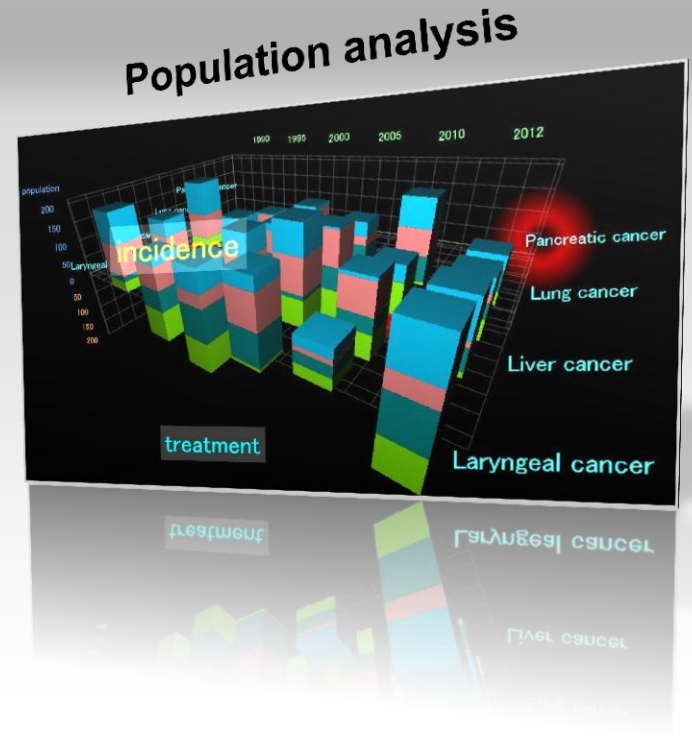
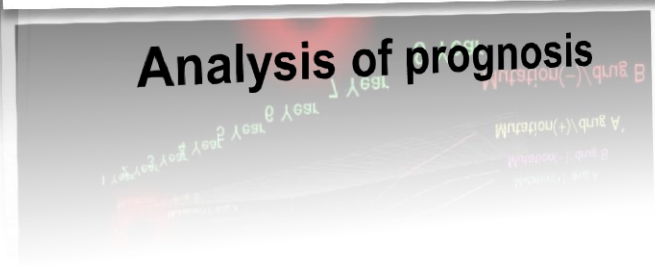
Choose a Patient

No.002 sex:F blood:O birth:Jun.21, 1968 Elizabeth Ford

No.001 sex:M blood:A birth:Apr. 6, 1974 Michel Bennet



Patients list



Acknowledgment

Cheney Sun
Xiaowei Xu
Jianhuang Liang
Wang Peng
Caro Ge
Rick Liu

Technology Innovation and Platform,
Design and New Applications
SAP China

Manabu Matsudate
SAP Japan

Hideo Shirota
Motohiro Kikkawa
Akira Kobayashi
Tomohiro Sakuma
Kenichi Aoki
Kumiko Kawasaki

Mitsui Knowledge Industry





Appendix

SAP HANA: Big data features and benefits

FEATURE	BENEFIT
In-memory architecture	Subsecond analysis of detailed data records
SAP HANA: grid architecture	Store well into the terabytes of raw data
Unstructured data	Analyze documents, Web content, and freeform text
R language support	Predictive analysis in-database using all data
Hadoop integration	Combine real-time analysis of high-value data with batch analysis of all data
Integration with SAP Data Services	Load data into SAP HANA in real time from all data sources

Deliver value from big data

Accelerate advanced analytics on big data with SAP HANA platform



Precision

- **Plan accurately** – SAP Planning and Consolidation and SAP NetWeaver BW on SAP HANA
- **Go deeper** – Predictive analytics via R on SAP HANA and Apache Hadoop



Acceleration

- **Answer faster** – immediate results
- **Move quicker** – Increase frequency of analytics, plan, forecast, and scenarios evaluation (HILO)



Efficiency

- **Manage simply** – eliminate unnecessary aggregation, caching (in-DB OLAP)
- **Reduce complexity** – One solution for data warehouse, dimension analysis, planning, and query acceleration

© 2012 SAP AG. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

Microsoft, Windows, Excel, Outlook, PowerPoint, Silverlight, and Visual Studio are registered trademarks of Microsoft Corporation.

IBM, DB2, DB2 Universal Database, System i, System i5, System p, System p5, System x, System z, System z10, z10, z/VM, z/OS, OS/390, zEnterprise, PowerVM, Power Architecture, Power Systems, POWER7, POWER6+, POWER6, POWER, PowerHA, pureScale, PowerPC, BladeCenter, System Storage, Storwize, XIV, GPFS, HACMP, RETAIN, DB2 Connect, RACF, Redbooks, OS/2, AIX, Intelligent Miner, WebSphere, Tivoli, Informix, and Smarter Planet are trademarks or registered trademarks of IBM Corporation.

Linux is the registered trademark of Linus Torvalds in the United States and other countries.

Adobe, the Adobe logo, Acrobat, PostScript, and Reader are trademarks or registered trademarks of Adobe Systems Incorporated in the United States and other countries.

Oracle and Java are registered trademarks of Oracle and its affiliates.

UNIX, X/Open, OSF/1, and Motif are registered trademarks of the Open Group.

Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame, and MultiWin are trademarks or registered trademarks of Citrix Systems Inc.

HTML, XML, XHTML, and W3C are trademarks or registered trademarks of W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.

Apple, App Store, iBooks, iPad, iPhone, iPhoto, iPod, iTunes, Multi-Touch, Objective-C, Retina, Safari, Siri, and Xcode are trademarks or registered trademarks of Apple Inc.

IOS is a registered trademark of Cisco Systems Inc.

RIM, BlackBerry, BBM, BlackBerry Curve, BlackBerry Bold, BlackBerry Pearl, BlackBerry Torch, BlackBerry Storm, BlackBerry Storm2, BlackBerry PlayBook, and BlackBerry App World are trademarks or registered trademarks of Research in Motion Limited.

Google App Engine, Google Apps, Google Checkout, Google Data API, Google Maps, Google Mobile Ads, Google Mobile Updater, Google Mobile, Google Store, Google Sync, Google Updater, Google Voice, Google Mail, Gmail, YouTube, Dalvik and Android are trademarks or registered trademarks of Google Inc.

INTERMEC is a registered trademark of Intermec Technologies Corporation.

Wi-Fi is a registered trademark of Wi-Fi Alliance.

Bluetooth is a registered trademark of Bluetooth SIG Inc.

Motorola is a registered trademark of Motorola Trademark Holdings LLC.

Computop is a registered trademark of Computop Wirtschaftsinformatik GmbH.

SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, SAP HANA, and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.

Business Objects and the Business Objects logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius, and other Business Objects products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Business Objects Software Ltd. Business Objects is an SAP company.

Sybase and Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere, and other Sybase products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Sybase Inc. Sybase is an SAP company.

Crossgate, m@gic EDDY, B2B 360° , and B2B 360° Services are registered trademarks of Crossgate AG in Germany and other countries. Crossgate is an SAP company.

All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

The information in this document is proprietary to SAP. No part of this document may be reproduced, copied, or transmitted in any form or for any purpose without the express prior written permission of SAP AG.