Manage Big Data in Real Real-Time with SAP HANA, Hadoop, R, …

Sanjay Patil, Kevin Wright
SAP Labs – Palo Alto
May, 2012

Agenda

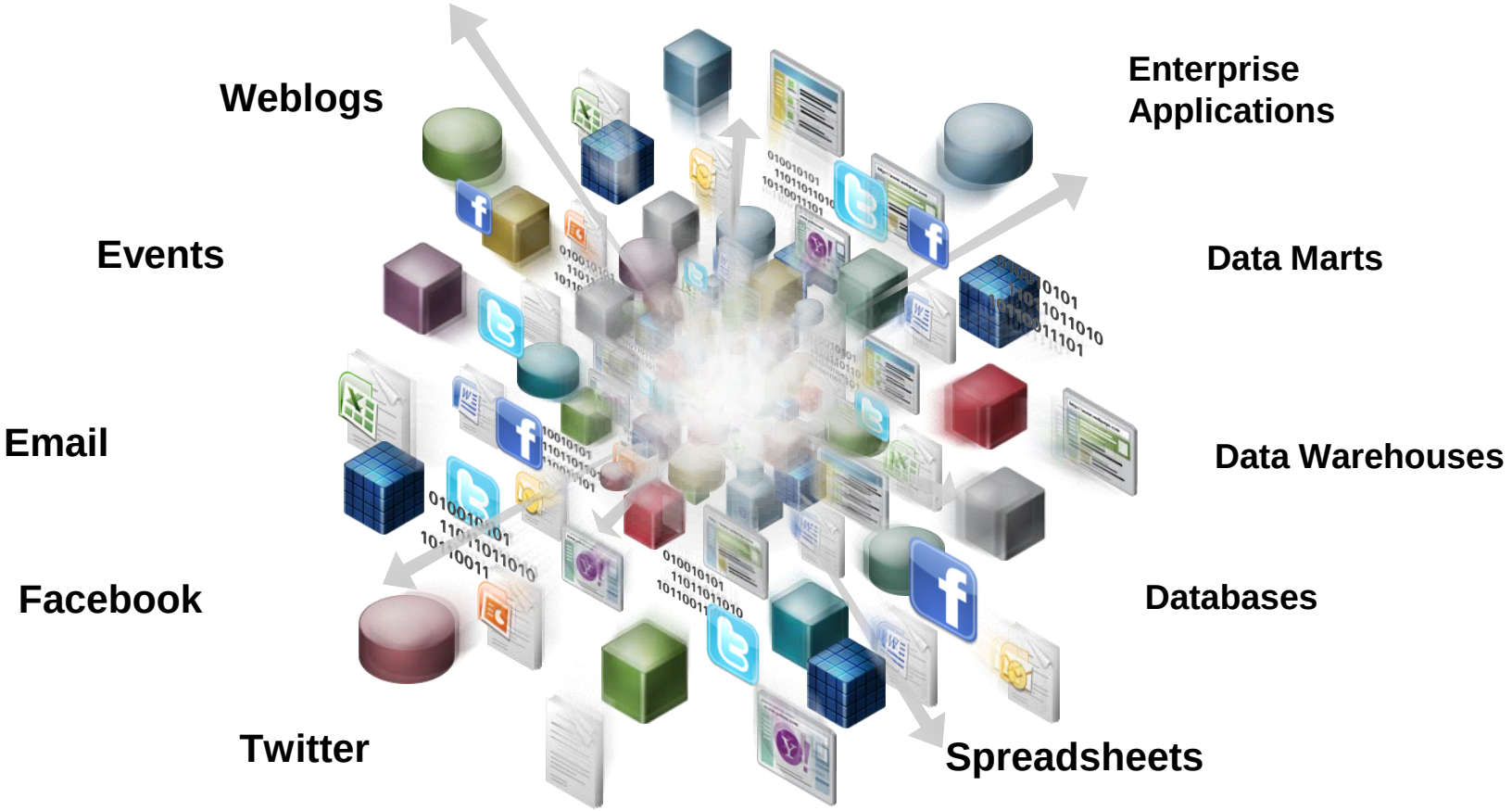
1. **Big Data Analytics – Key Challenges**
2. **Open Source and SAP HANA**
3. **Predictive Analytics with SAP HANA and R**
4. **Unstructured Data Processing in Hadoop using SAP technology**

Legal disclaimer

The information in this presentation is confidential and proprietary to SAP and may not be disclosed without the permission of SAP. This presentation is not subject to your license agreement or any other service or subscription agreement with SAP. SAP has no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation and SAP's strategy and possible future developments, products and or platforms directions and functionality are all subject to change and may be changed by SAP at any time for any reason without notice. The information in this document is not a commitment, promise or legal obligation to deliver any material, code or functionality. This document is provided without a warranty of any kind, either express or implied, including but not limited to, the implied warranties of merchantability, fitness for a particular purpose, or non-infringement. This document is for informational purposes and may not be incorporated into a contract. SAP assumes no responsibility for errors or omissions in this document, except if such damages were caused by SAP´s willful misconduct or gross negligence.

All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.

The Information Explosion is Driving The Need for Better Information Management



Data Doubles Every 18 months

80% of Enterprise Data Is Unstructured

Information Is a Strategic Corporate Asset

Having Data is Not Enough!

Do You Have Real-Time Business Insights?

Customer Insights

- Which customers & channels are more profitable?
- Which customer profiles are suitable for loyalty rewards?
- How dynamic is your customer segmentation strategy?

Product / Service Insights

- How are products/services doing vs. their competition?
- Track complaints from call centers & social data in real-time?
- Where else is this part used in my company ?

Operations Insight

- How can you predict supply chain disruptions ahead?
- How do suppliers rank by cost, quality and timeliness?
- How is my “on-time / in full” delivery rate by customer?

Existing Approach is Slow and Limited

You Need an Agile Approach

Business Demands Agility

- Changing markets require frequent / fast changes
- Multiple and growing sources of internal and external data
- Growing business sophistication for self-service modeling and real-time analytics

Existing Approach is Slow

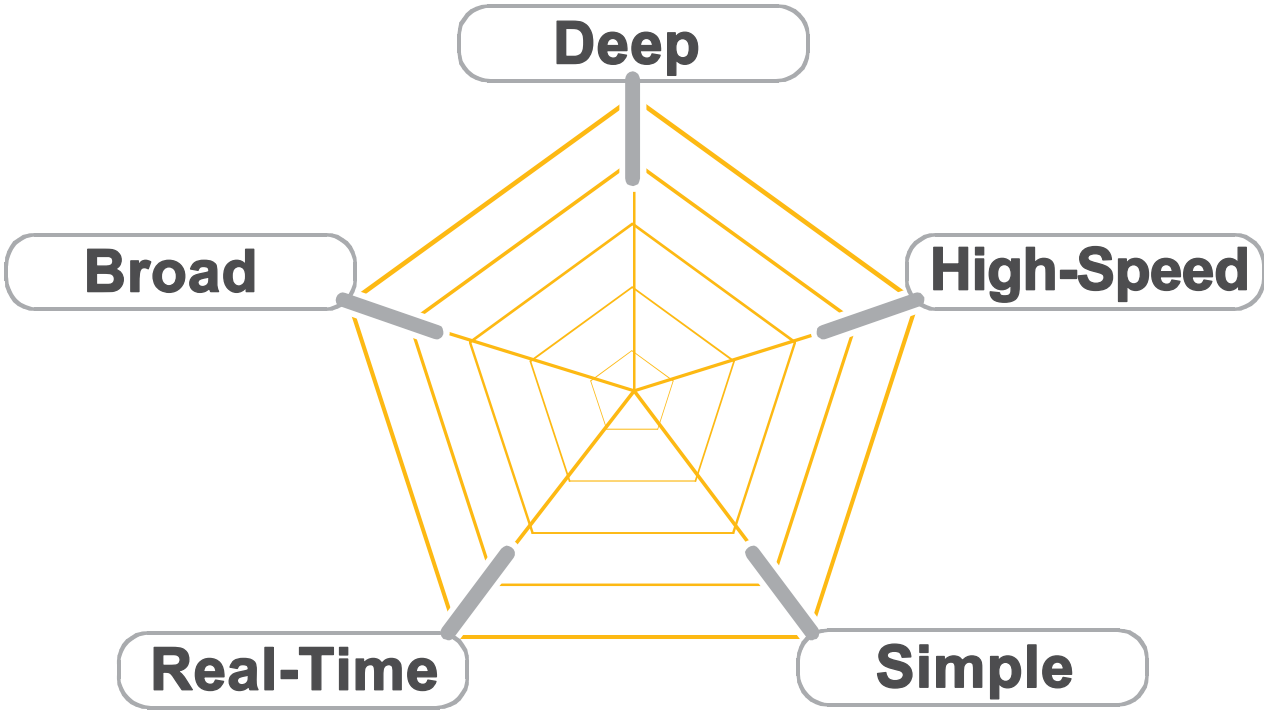
- Changes to EDW needs expertise modeling & months
- Database performance tuning leads to redundant data and latency from aggregates
- Poor data quality from out-dated information can impact business decisions

…& Not Future-roof

- Can you be ready without knowing the questions or the data in advance?
- Sensitive personal, financial or legal data may have to be isolated physically
- Multi-structured data (e.g., text, events, machine) makes current EDW technology obsolete

Need a breakthrough technology that delivers across the..

5 dimensions of modern decision-processing



Deep
- Complex & interactive questions on granular data

Broad
- Big data, many data types

Real-Time
- recent data, preferably real-time

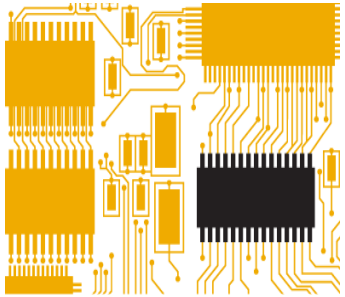
Simple
- No data preparation, no pre-aggregates, no tuning

High-Speed
- fast response-time, interactivity

SAP HANA

Ideal Platform for Real-Time Analytics

1. Revolutionary in-memory platform

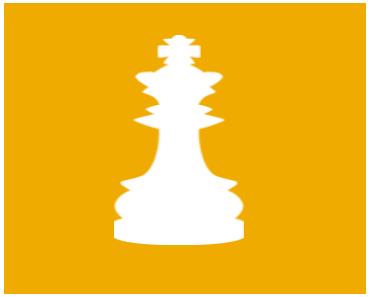


- Real-time analytics on detailed data on the fly
- In-memory calculations
- Real-time replication to eliminate data latency
- No aggregates, tuning of data for performance

2. Empowers you to interrogate data

- Wizard-driven data modeling for business
- Fast & easy creation of ad-hoc views
- Optimized for SAP BusinessObjects BI
- Open platform for other clients

3. Powerful predictive analytics



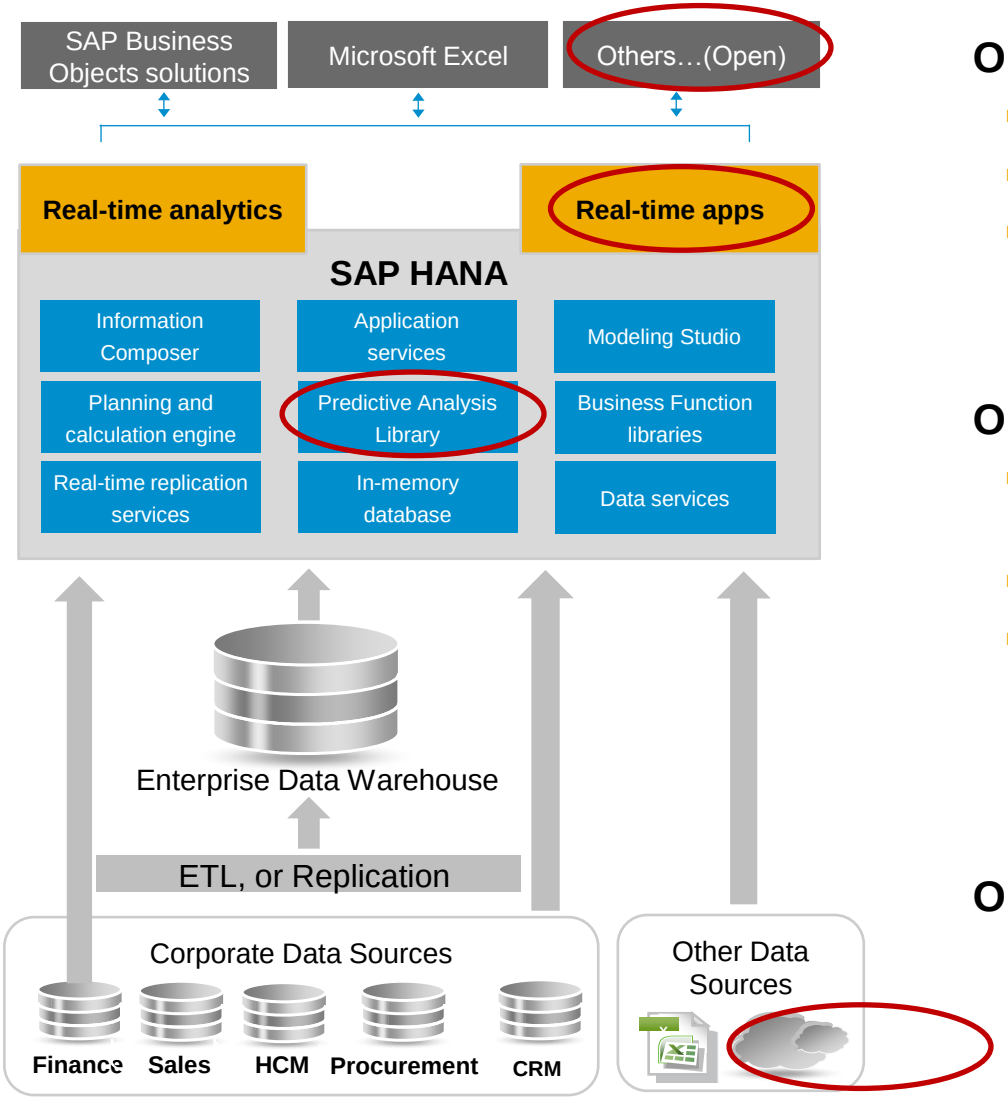
- Embedded data mining algorithms for predictive analytics
- Bring decision support capabilities to the business users through simplified experience and pre-built scenarios

4. High data quality

- Real-time replication and faster loading
- Tight integration with data quality capabilities

Revolutionary In-Memory Platform

Comprehensive Support for Open Standards and Open Source



Open Source in Consumption Layer

- jQuey
- jqPlot / D3.js
- …

Open Source in the Applications

- Spring, Apache CXF, Tomcat, Eclipse Link, …
- R libraries
- …

Open Source in Data Processing

- Hadoop
- …

Solving the Big Data Challenge with SAP HANA and Hadoop

What is Apache Hadoop/HIVE?

Apache Hadoop addresses some of the key challenges mentioned, but leaves some wishes unanswered

- Open-source project administered by the Apache Software Foundation
- Allows for scalable and accessible storage of massive data amounts (structured and unstructured) on commodity hardware clusters
- Designed for **non-real time analysis** of both structured data and complex data

Key Hadoop/HIVE Services:

- Reliable data storage using the Hadoop Distributed File System (HDFS) – structured and unstructured
- HIVE is a data warehousing solution on top of Hadoop – direct access to HDFS and Hbase
- Parallel data processing and query execution using MapReduce

Companies starting to adopt Apache Hadoop

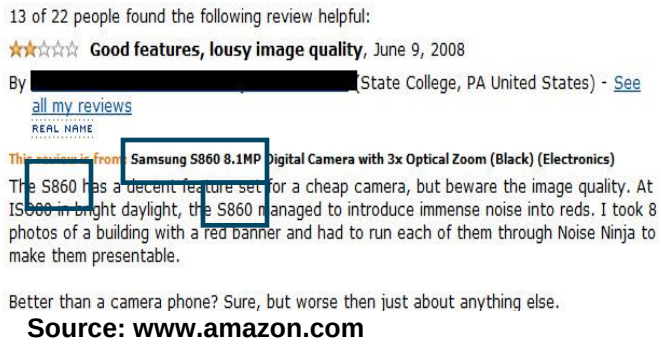
- Originally developed and employed by dominant Web companies like Yahoo and Facebook
- Today used in finance, technology, telecom, media and entertainment, government, research institutions and other markets with significant data

Understanding the Business Value of Unstructured Data

Some key challenges

- Understanding data and identifying the relationship between the embedded information
- Bringing meaningful structure into unstructured data
- Relating unstructured, complex data to structured information to get 360 degree insights
- Extract meaningful information

Example – Public Product Reviews



13 of 22 people found the following review helpful:

☆☆☆☆☆ **Good features, lousy image quality**, June 9, 2008

By [redacted] (State College, PA United States) - See all my reviews

REAL NAME

This review is from: **Samsung S860 8.1MP Digital Camera with 3x Optical Zoom (Black) (Electronics)**

The S860 has a decent feature set for a cheap camera, but beware the image quality. At ISO80 in bright daylight, the S860 managed to introduce immense noise into reds. I took 8 photos of a building with a red banner and had to run each of them through Noise Ninja to make them presentable.

Better than a camera phone? Sure, but worse then just about anything else.

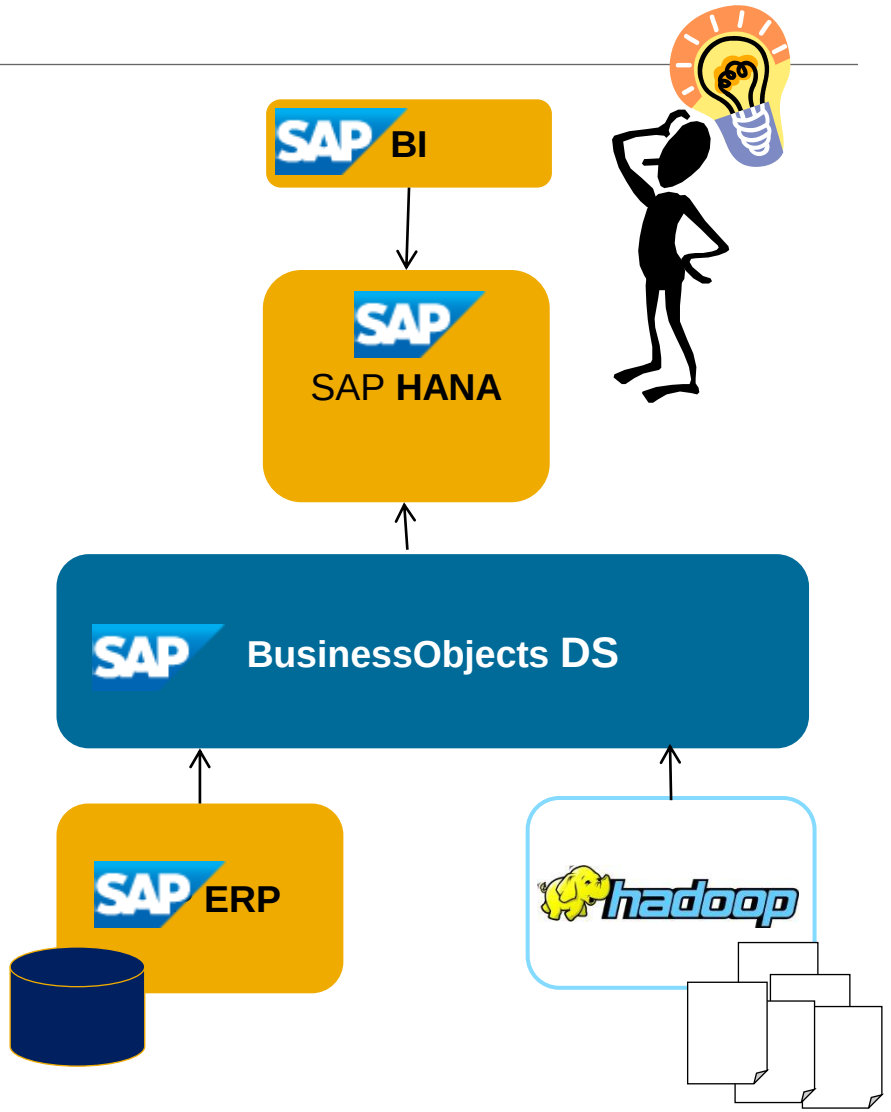
Source: www.amazon.com

STANDARD_FORM	TYPE	CONVERTED_TEXT
Polaroid	PRODUCT	Very disappointed. Because of other duties, I dic
Quicktime	PRODUCT	June 12, 2008 I bought this camera yesterday a
Samsung S860	PRODUCT	Samsung S860 8.1MP Digital Camera with 3x Opt
Samsung S860	PRODUCT	Samsung S860 8.1MP Digital Camera with 3x Opt
Samsung S860	PRODUCT	I bought one of the Kodak C813's and returned i
Samsung S860	PRODUCT	I bought a pink Samsung S860. I used to use car
Samsung S860	PRODUCT	I own a Canon Power Shot that works perfectly
Samsung S860	PRODUCT	The Samsung S860 is a low priced camera that is
Samsung S860	PRODUCT	Admittedly, I am a novice. But this camera is per
Samsung S860	PRODUCT	Admittedly, I am a novice. But this camera is per
Samsung S860	PRODUCT	I purchased a pink Samsung S860 for my mother
Samsung S860	PRODUCT	I purchased a pink Samsung S860 for my mother
Samsung S860	PRODUCT	Let me start off my saying, I am not a camera ex
Samsung S860	PRODUCT	Let me start off my saying, I am not a camera ex
Samsung S860	PRODUCT	The Samsung S860 has features I wanted. It do
Samsung S860	PRODUCT	I needed a digital camera to take pictures of iten
Sony A700	PRODUCT	I also own a Sony A700 but it is a big camera to l

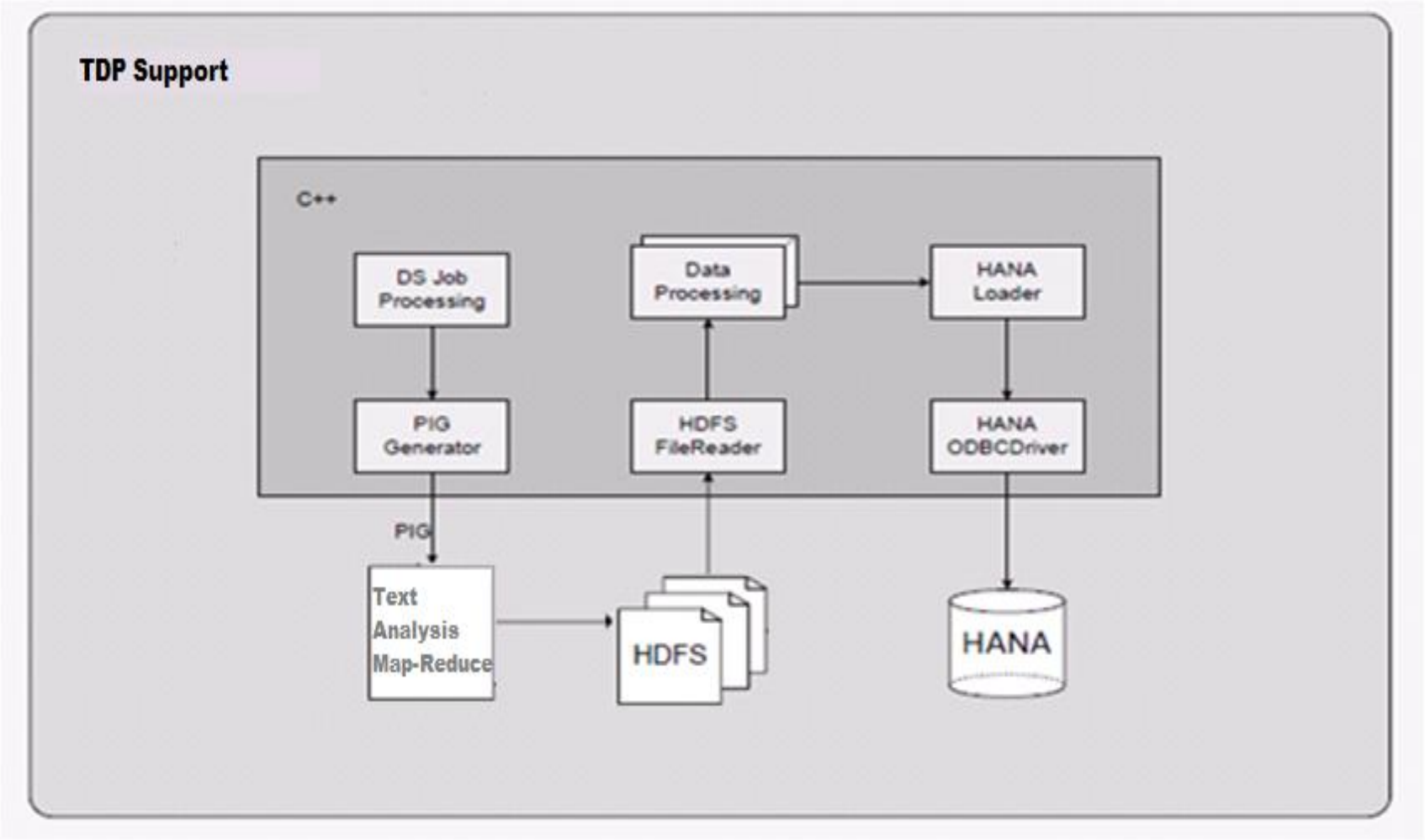
The information shown here is for demonstration purposes only. Organizations wishing to access Amazon public reviews should contact Amazon Web Services (https://aws.amazon.com/) for information licensing their Hadoop service & public content.

Demo Scenario: Getting Rapid Insight with EIM power and HANA speed

1. Automatically create TDP dictionary based on ECC product entities
2. Extract product master data from ECC into HANA
3. Analyze Hadoop data files against the generated dictionary via TDP pushdown
4. Load result into HANA
5. Relate structured ECC data with analyzed Hadoop data based on product entities
6. Get advanced insight by reporting on the related data



Deep Integration with Hadoop for Text Data Processing



Demo

- Data & TDP Entity extraction from Hadoop HDFS
- - Correlation with data from an SAP ERP system
- - Loading data into HANA for analysis and reporting

Predictive Analytics with SAP HANA and R

What if you could …

. . . **Identify hidden revenue opportunities within your customer base?**

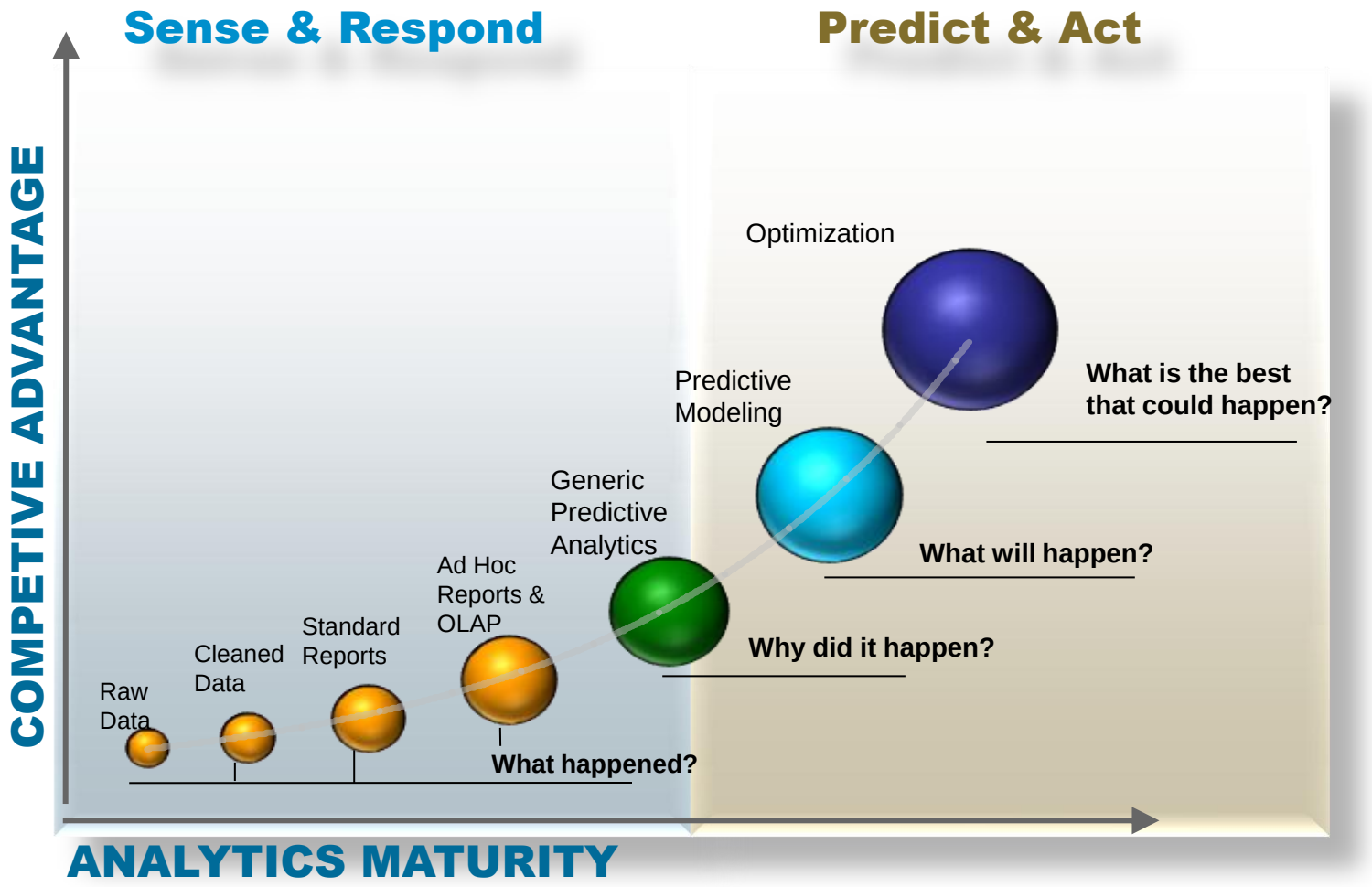
. . . **Retain your high-value customers, employees, vendors and partners with the right retention offers?**

. . . **Delight customers with accurate next-step recommendations for product usage?**

. . . **Increase cross-sell and up-sell effectiveness through cross-channel coordination?**

. . . **Build long-term relationships with customers, employees, vendors and partners via intelligent interactions?**

Extend Your Analytics Capabilities



The key is unlocking data to move decision making from sense & respond to predict & act

R Integration for SAP HANA

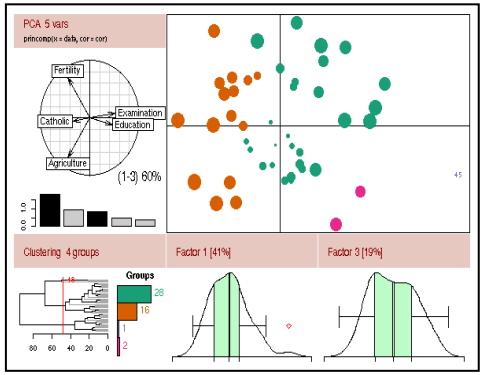
What is R?

R is a software environment for statistical computing and graphics

- Open Source statistical programming language
- Over 3,500 add-on packages; ability to write your own functions
- Widely used for a variety of statistical methods
- More algorithms and packages than SAS + SPSS + Statistica

Who's using it?

- Growing number of data analysts in industry, government, consulting, and academia
- Cross-industry use: high-tech, retail, manufacturing, CPG, financial services , banking, telecom, etc.

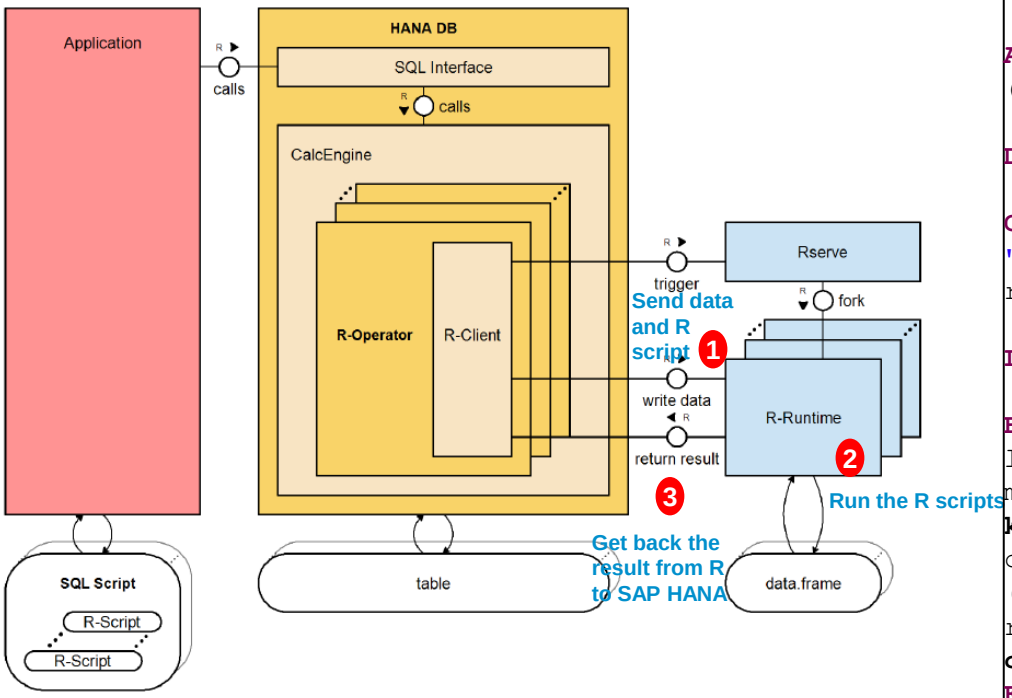


Why do they use it?

- Free, comprehensive, and many learn it at college/university
- Offers rich library of statistical and graphical packages

R Integration for SAP HANA

Functionality Overview



Sample Code in SAP HANA SQLScript

```
DROP TABLE "spamClassified";

CREATE COLUMN TABLE "spamClassified" LIKE
"spamEval" WITH NO DATA;

ALTER TABLE "spamClassified" ADD
("classified" VARCHAR(5000));

DROP PROCEDURE USE_SVM;

CREATE PROCEDURE USE_SVM( IN train
"spamTraining", IN eval "spamEval", OUT
result "spamClassified")

LANGUAGE RLANG AS

BEGIN
library(kernlab)
model <- ksvm(type~. , data=train,
kernel=rbfdot(sigma=0.1))
classified <- predict(model, eval [,-
(which(names(eval) %in% "type"))])
result <- as.data.frame(cbind(eval,
classified))
END;

CALL USE_SVM("spamTraining", "spamEval",
"spamClassified") WITH OVERVIEW;

SELECT * FROM "spamClassified";
```

MKI Uses SAP HANA to Speed Cancer Research and Improve Patient Support

Company
MITSUI KNOWLEDGE INDUSTRY

Headquarters
Tokyo

Industry
IT services

Products and Services
Services to pharmaceutical companies, universities and research institutes

Employees
1,990

Web Site
www.mki.co.jp

Objectives

- Reduce delays and minimize the costs associated with new drug discovery by optimizing the process for genome analysis
- Improve and speed decision making for hospitals which conduct cancer detection based on DNA sequence matching

Why SAP

- High-performance real-time computational capabilities of SAP HANA
- Ability to leverage the combination of SAP HANA, R, and Hadoop to store, pre-process, compute, and analyze huge amounts of data
- Breadth of predictive analytics libraries

Benefits

- Reduced time of genome analysis from several days to 20 minutes making real-time cancer/drug screening possible
- For pharmaceutical companies, ability to provide required new drugs on time and aid identification of “driver mutation” for new drug targets
- Able to provide a one stop service including genomic data analysis of cancer patients to support personalized patient therapeutics

“Our solution is to incorporate SAP HANA along with Hadoop and R to create a single real-time big data platform. Data mining will be handled by R and assisted by HANA. Data pre-processing prior to data analysis and high-speed storage will be managed by Hadoop. With this we have found a way to shorten the genome analysis time from several days down to only 20 minutes.”

Yukihisa Kato, CTO and Director of MITSUI KNOWLEDGE INDUSTRY

Faster

Genome analysis

Better

Insight to support the needs of cancer patients in real-time

Greater

Personalization to individual patient needs

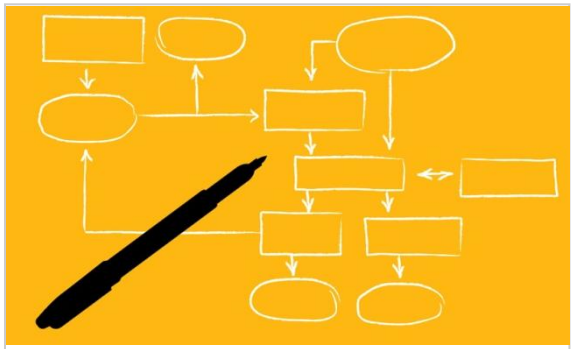
Example of Big Data in Real Time Customer Energy Management

Customer Energy Management (CEM) – B2B Powered by SAP HANA

Customer Energy Analytics

Easy to use end customer web application which helps understanding & managing the energy consumption of different sites

Energy Services



Energy Services which brings added values around Energy (CO2 Reduction Service, Alerting Service, Industry Benchmarking etc.)

End to End Communication

End to End processes which allows a efficient communication with the customer, including mobile devices

Demo

Customer Energy Management

Thank You!

Sanjay Patil sanjay.patil@sap.com
Kevin Wright kevin.wright@sap.com
SAP Labs – Palo Alto