SAP’s Strategies for Big Data Analytics

Business Analytics & Technology
April 10, 2012
Agenda

1. Big Data Analytics Overview
2. SAP’s Strategies for Big Data Analytics
3. Roadmap
4. Conclusion
1. Big Data Analytics Overview

1. Top 10 Strategic Technologies for 2012
2. 의사결정 지원시스템에 대한 현업사용자의 불만
3. BI(Business Intelligence)의 진화의 단계
4. Business Analytics가 사용 가능한 분야
5. 기존의 Business Analytics의 적용의 문제점은?
6. Big Data Analytics
1.1 Gartner Identifies the Top 10 Strategic Technologies for 2012

Gartner Symposium/ITxpo, October 16-20, 2011, in Orlando

- Big Data
- Next-Generation Analytics
- Contextual and Social User Experience
- In-Memory Computing
- Media Tablets and Beyond
- Mobile-Centric Applications and Interfaces
- Internet of Things
- App Stores and Marketplaces
- Extreme Low-Energy Servers
1.2 의사결정 지원시스템에 대한 현업사용자의 불만

- DW(BW) 및 EPM / BI 환경을 구축했는데, 활용이 제대로 되지 않는다.
- 계획데이터를 입력하는데 너무 많은 시간이 소요되서 업무가 힘들다.
- 의사결정에 활용할 정보가 없다.
- 텍스트 데이터 및 비정형 데이터에 의한 분석영역이 필요하다.
- 비즈니스에 대한 예측 및 최적화가 필요하지만 정보에 대한 확인도 없고, 어떤 방법을 적용해야 하는지 모른다.
1.3 BI(Business Intelligence)의 진화의 단계
1.4 Business Analytics가 사용 가능한 분야

<table>
<thead>
<tr>
<th>Application</th>
<th>Retail</th>
<th>MFG (MI)</th>
<th>Telco</th>
<th>Media and Entertainment</th>
<th>Banking</th>
<th>Insurance</th>
<th>Public Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>시장수요 및 판매 예측</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>비즈니스 시뮬레이션</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>고객이탈방지</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>품질분석 및 예측</td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>원자재가격 예측</td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNA(Social Network Analysis)</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>SMA(Social Media Analysis)</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>고객세분화 및 캠페인최적화</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>IT 장비최적화 및 정책예측</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>네트워크 트래픽 예측 및 최적화</td>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>운영 리스크 관리</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>부정 사기 적발 시스템</td>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>세금 징수예측</td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1.5 기존의 Business Analytics의 적용의 문제점은?
1.6 Big Data Analytics

- **Big data**
  - Structured
  - Semi Structured
  - Unstructured (Text, Social, VoC, Web)

- **Analytics**
  - Data Mining
  - Text Mining
  - Statistics
  - Forecasting
  - Optimization

- **Performance**
  - In DB Analytics
  - In Memory computing
  - Grid Computing

- **Real-Time**
  - Embedded Analytics
  - Trigger

**VARIETY**
**METHOD**
**VOLUME**
**VELOCITY**
2. SAP’s Strategies for Big Data Analytics

1. SAP’s Strategies for Big Data Analytics
2. In-database Analytics on HANA
3. SBOP Predictive Analysis
4. Hadoop Integration
5. Text & Social Analytics
6. Embedded Analytics
2.1 SAP’s Strategies for Big Data Analytics

**In-database/memory Analytics**
- R engine + PAL on HANA
- Predictive Analysis on HANA

**Integration of Hadoop & Text / Social Data**
- Hadoop Integration of BusinessObjects & Data Service
- Text Analytics using Netbase and Data Service (TDP)

**Industry / LOB apps adopt business analytics**
- ERP embedded BusinessObjects
- CRM with Netbase

Cloud, Collaboration ...
2.2 In-database Analytics on HANA

- **Project R Integration**
  - Project R is Open Source analytic SW., similar to commercial SAS and SPSS
  - Near 3000 external packages
  - Extend HANA analytic functions through integration with R
  - Use Shared Memory to reduce the extra cost

- **BFL (Business Function Library)**
  - New approach to build application of top of HANA
  - Performance Improvement by factors
  - Shorten Development Lifecycle
  - Deliver Customer’s real need Application
2.2 Project R – 3,000여 개의 통계 / 마이닝 패키지 제공

- Data miners report using an average of 4.6 software tools.
- R is used by the most data miners (43%).
- STATISTICA is the primary data mining tool chosen most often (18%).

![Data Mining Software](image)

- Survey Questions:
  - What Data mining/analytic tools did you use in 2009? (rate each as "never",
2.2 R on HANA

- R Client in CalcEngine of HANA
- Data Sharing between HANA and R using Shared Memory Manager
- R command in HANA Script

```sql
1. DROP TABLE "spamClassified";
2. CREATE COLUMN TABLE "spamClassified" LIKE "spamEval" WITH NO DATA;
3. ALTER TABLE "spamClassified" ADD ("classified" VARCHAR(5000));

4. DROP FUNCTION USE SVM;
5. CREATE FUNCTION USE SVM( IN train "spamTraining", IN eval "spamEval", OUT result "spamClassified")
6. LANGUAGE RLANG AS
7. BEGIN
8. library(kernlab)
9. model <- ksvm(type = data.train, kernel = rbf(epsilon = 0.1))
10. classified <- predict(model, eval)(which(names(eval) %in% "type"))
11. result <- as.data.frame(cbind(eval, classified))
12. END;

13. DROP FUNCTION R_PARTOFMORE;
14. CREATE FUNCTION R_PARTOFMORE(OUT result "spamClassified")
15. LANGUAGE SQLSCRIPT AS
16. BEGIN
17. subset1 = select * from "spamEval" where "capitalLong" > 14;
18. subset2 = select * from "spamEval" where "capitalLong" <= 14;
19. train = select * from "spamTraining";
20. newtrain = CE_UNION_ALL(subset1, train);
21. CALL USE_SVM(newtrain, subset2, result);
22. END;

23. CALL R_PARTOFMORE("spamClassified") WITH OVERVIEW;
24. SELECT * FROM "spamClassified";
```
2.2 BFL (Business Function Library)

- the calculation library for the Applications which is built on top of HANA
- It resides in HANA CalcEngine and It is written in C++
- PAL (Predictive Analysis Library) is subset of BFL
2.3 Data Mining Process in HANA

**Data Visualization and Sharing**
1. Visualize the model for better understanding
2. Store the model and result back to HANA
3. Share results via PMML and with other BI client tools

**Solution**
- Frontend: R console, SBOP PA
- Backend: PAL

**Data Loading**
1. Understand the business and identify issues
2. Load the SAP and non-SAP data into HANA

**Solution**
- SLT / Data Service/
  Sybase Replication Server
- JDBC/ODBC / CSV Loading

**Data Preprocessing**
1. Visualize and examine the data
2. Sample, filter, merge, append, apply formulas

**Solution**
- Frontend: R console, SBOP PA
- Backend: PAL

**Data Processing**
1. Define the model via clustering, classification, association, time series, etc.
2. Run the model

**Solution**
- Frontend: R console, SBOP PA
  other potential product
- Backend: PAL

**Data Mining**
1. Define the model via clustering, classification, association, time series, etc.
2. Run the model

**Solution**
- Frontend: R console, SBOP PA
  other potential product
- Backend: PAL

**Data Visualization and Sharing**
1. Visualize the model for better understanding
2. Store the model and result back to HANA
3. Share results via PMML and with other BI client tools
2.3 SBOP Predictive Analysis

- Predictive analytic Process design Tool
- HANA PAL and R Integration
- Data Visualization with new chart type for predictive
- Support large data volume
2.3 SBOP Predictive Analysis
- Data Loading 및 분석 프로세스 설계

- HANA, Sybase, 일반 RDBS, CSV, BusinessObjects Data ...
2.3 SBOP Predictive Analysis
- Data Preparation: 기본 통계량 및 분포조회
2.3 SBOP Predictive Analysis
- Data Preparation: Parallel Coordinates
2.3 SBOP Predictive Analysis
- Data Preparation : OLAP 분석
2.3 SBOP Predictive Analysis
- Data Processing : K-Means Clustering 결과
2.3 SBOP Predictive Analysis
- Data Processing: Decision Tree 분석결과 (C4.5)
2.3 SBOP Predictive Analysis
- Data Export: BI Platform과의 통합
2.3 SAP Business Analytics Framework 적용 모델 (예시)

- BI Platform (SAP BusinessObjects)
  - 다차원 분석
  - 정형 리포팅
  - 대쉬보드
  - Office기반 고급분석
  - 인메모리 Ad-hoc
  - Mobile

- 분석 엔진 (SBOP Predictive Analysis)
  - Data Mining
  - Data Visualization

- HANA Appliance (included BW)
  - Data Mart
  - R Engine
  - PAL (Predictive analysis Library)
  - Hybrid Model

- ETL
  - Others
  - ERP
  - 기간계

- 기간계 연동모듈
  - Parameter Alert

- No ETL
- Integration with BI
- Better Performance
- Rapid development of Analytics using R & PAL

© 2012 SAP AG. All rights reserved.
2.4 SAP BusinessObject Semantic Layer for Hadoop Hive

- **Data Service**를 통한 HANA로의 통합
- **BusinessObjects**의 Universe를 통한 BI로의 직접 통합

*Empower business users with the autonomy they need to access, analyze, enrich, and share information freely and securely using familiar business terms*

Empower all people, enable all workflows

High performance, feature rich and secured access

The new information design tool is your point of entry to business intelligence solutions

**Common user experience for all front-ends**
- Web Intelligence
- Crystal Reports
- Dashboards
- Explorer

**Best access method for each specific data source**
- Direct Access
- Universe Access

**All data sources**
- SAP BW
- Sybase
- SAP HANA
- Any Relational Database
- Files
- Web Service
2.4 Hadoop Integration using Information Design Tool

Database Middleware Driver Selection

Select the driver for your database middleware.

- Apache
  - Apache Hadoop HIVE
    - JDBC Drivers
  - Derby 10 Embedded

The selected driver is: Apache Hadoop HIVE (JDBC Drivers)

Query Panel

Universe: DataMart

Result Objects for Query #1

- Year
- Quarter
- Revenue

Query Filters for Query #1

- Region Equal to North America

Result set (8 rows: 39187 ms)

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>1</td>
<td>745.17</td>
</tr>
<tr>
<td>2</td>
<td>1301.33</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>870.64</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1275.05</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>1628.91</td>
</tr>
<tr>
<td>2</td>
<td>1521.72</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1887.72</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1404.80</td>
<td></td>
</tr>
</tbody>
</table>
2.4 BusinessObjects Explorer on Hadoop Hive
2.5 Text Analytics

SAP BusinessObjects Data Services 4.0

- Business UI (Information Steward)
- Technical UI (Data Services)

Unified Metadata

One Runtime Architecture & Services

ETL
Data Quality
Profiling
Text Analytics

One Administration Environment
(Scheduling, Security, User Management)
One Set of Source/Target Connectors
2.5 Text Analytics

Speech: Martin_Luther_King (1963-08-28)

The most occurring words found in the speech.
2.5 Netbase

Social Media

- Blog
- Facebook
- Twitter
- Chat

Natural Language Processing Engine
- Deep parsing
- Grammatical analysis
- Industry lexicons and filters
- Cloud-based
- 100+ Servers

NetBase Web Services API
- 20B documents
- 12 Months history
- 70+ servers

Social Intelligence Warehouse
- Key social metrics
- Conversation drivers
- Comparisons & trending
- Hosted on SMA servers

Scorecard
- Topic definition
- Disambiguation
- Excel Export
- Analytics
- Comparisons
- Scorecard creation

Workbench
- Language
- Textual analysis
- Sentiment analysis
- Natural Language Processing Engine
2.5 Netbase

**Workbench**
- Create Topics
- Setup Scorecards
- Deep dive analysis

**Scorecard**
- Setup & Provision
- Deep dive analysis
- Publish analysis

**Analyst Users**
- Key social metrics
- Across competitors
- Over time
- With conversation drivers

**Brand Managers**
- Analyze & Track:
2.6 SAP BusinessObjects Embedded Analytics within the SAP Business Suite

- Legacy reporting technology
- Analytics for sophisticated users
- Limited report formatting capabilities

- Modern user experience
- Compelling analytics for all users
- Best-of-breed BI solutions

*SAP BusinessObjects BI solutions integrated Within SAP Business Suite*
3. Roadmap & Conclusion

1. Roadmap
2. Conclusion
3.1 PAL (Predictive Analysis Library) Algorithm Roadmap

SP3
Cover Classical predictive Analysis algorithms in each category
- Clustering
  - K-Means
- Classification
  - KNN
  - C4.5 Clustering Tree
  - Linear Regression
- Association
  - Apriori
- Classification
  - ABC Classification
  - Weighted Score Tables

SP4
Extend algorithms in each category Cover Times Series and some Preprocessing algorithms
- Clustering
  - Apply it to Anomaly detection
- Classification
  - Bi-Variate Exponential / Geometric / Logistic / Natural Logarithmic Regression
  - Naïve Bayes
  - CHAID Decision Tree
- Time Series
  - Moving Average
  - Single / Double / Triple exponential smoothing
- Preprocessing
  - Outlier Detection (Inter-Quartile Range)
  - Correlation calculations
  - Sampling

SP5
Cover more complicated algorithms
- Clustering
  - Hierarchical Agglomeration
  - Self Organized Map
- Classification
  - Polynomial / Stepwise / Cox Regression
  - C & RT Decision Tree
  - Neural Networks
  - Support Vector Machine (SVM)
- Social Network Analysis
  - Sequential / Link Analysis
- Optimization
  - Linear Programming
  - Monte Carlo Simulation
  - Generic Algorithm

Nov 2011 May 2012 Sep 2012
3.1 Roadmap of SBOP Predictive Analysis

**SBOP PA 1.0 Beta**
- Modern UX/UI
- Univ, JDBC, CSV
- Outlier detection, regression, exp smoothing.
- R integration of 16 algorithms
- Predictive Visualizations
- PMML support

**SBOP PA 1.0 RTC**
- ...plus updates based on Beta customer feedback
- HANA PAL Support
- Sybase IQ as a data source

**SBOP PA 1.1 RTC**
- Universe UNIX as a data source
- HANA R Integration
- SBOP PA Wizard
- Sybase IQ, IN-DB algorithm support
- SAP BW support
- Visual Numerics Integration
- Large data volume data visualization

- Q4’2011
- Q1’2012
- Q3’2012
이제는 기업들은 더 이상 좋은 제품을 생산하고, 좋은 서비스만을 제공하겠단데 집중해서는 안됩니다. 기업들이 더 성장하기 위해서는...
감사합니다.