Big Data infrastructure and skills from a consulting point of view

„The Bottom up View“

Kevin Berwind, M.Sc.
University of Hagen
EGI Community Forum
November 12th 2015, Bari, Italy

EDISON – Education for Data Intensive Science to Open New science frontiers
Grant 675419 (INFRASUPP-4-2015: CSA)
Background and CV

Education:
• University of Ludwigshafen am Rhein
  – Business Informatics (B.Sc.)
  – Business Informatics with the specialization “Information Management & Consulting” (M.Sc.)

Working Experience:
• Accenture Technology Solutions GmbH, Kronberg im Taunus
  – Consulting: Relationship Management and CRM Analytics

• BTEXX GmbH, Mainz
  – Consulting: Big Data, Predictive Analytics and Predictive Maintenance
Agenda

• Big Data in Enterprises

• Big Data Reference Infrastructures

• Needed Big Data Skills For Consulting Projects
Big Data in Enterprises I
Industrial Application Domains

- Business Transaction Data (ERP, CRM etc)
- Business Documents (Text Data)
- Industrial Data (R & D, Monitoring, Simulation, Analysis, Sensor)
- HealthCare Data (Multimedia Data, Security!)
- Logistical Enterprise Data (RFID Data)
- Other Data . .
Big Data in Enterprises II
Industrial Big Data Characteristics

3 Vs of Big Data
- Volume
  - Terabyte
  - Records
  - Transactions
  - Tables
- Velocity
  - Batch
  - Near-Time
  - Real-Time
  - Streams
- Variety
  - Structured
  - Unstructured
  - Semi-Structured

5 Vs of Big Data
- Volume
  - Terabyte
  - Records
  - Transactions
  - Tables
- Velocity
  - Batch
  - Realtime
  - NearTime
  - Processes
  - Streams
- Variety
  - Structured
  - Unstructured
  - Probabilistic
- Veracity
  - Authenticity
  - Origin, Reputation
  - Accountability
- Value
  - Statistical
  - Events
  - Correlations
  - Hypothetical
Big Data in Enterprises II
Big Data as immeterial Assetmanagement

**Data Value:**
Detect risks, information about customers, competitors, supplier, processes, new opportunities

**Data Volume:**
„Increasing of data volume up to 65% each year“

**Data Types:**
Structured, unstructured, semi-structured

**Various Data Sources:**
Data bases, In-memory data bases, geographical distributed

**High Velocity:**
Near-time, real-time, batch
Big Data Reference Infrastructures II
Architecture and Processes

Decide

Analysis

Transform

Extract Collect

Applications, Services, Visualisation, Reporting

Parallel Processing Platform

DB/IMDB

DWH

Streaming Data

Extract

Collect

Trans- form

Ana- lysis

Decide
Big Data Reference Infrastructures II
Architecture and Processes

Data sources:
Transaction systems, data warehouses, protocol data, streaming data, social media

Parallel Processing Platform:
Hadoop Distributed File System (HDFS) and other software products like SAP HANA

Analysis and Visualization:
Data analysis within the PPP, data forwarding to other systems, data analysis and reporting via applications or services
Big Data Reference Infrastructures III
Big Data Distribution Models

SaaS – analysis- and visualization tools

Paas – middleware (example Hadoop system)

IaaS – server, storage, network components
<table>
<thead>
<tr>
<th>Enterprise Big Data infrastructure</th>
<th>External Big Data Cloud infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Well-matched hard- and software.</td>
<td>• Cloud provider is taking care of well-matched hard- and software.</td>
</tr>
<tr>
<td>• Special staffing to operate, develop and manage the infrastructure.</td>
<td>• Staffing is done by the cloud provider.</td>
</tr>
<tr>
<td>• Cost reduction concerning missing staff and infrastructure, investments in new technology and infrastructure.</td>
<td>• Various price models like pay-per-use, staffing and investments are done by cloud provider.</td>
</tr>
<tr>
<td>• Useful if you are using high volume or sensible data or data streams for example from machines, sensors, processes.</td>
<td>• You may use storage from the cloud provider for non-sensible data or data from the internet to keep your data and information within the company and to prevent high network workload.</td>
</tr>
</tbody>
</table>
Needed Big Data Skills For Projects

Successful Big Data Projects

Technical Data Competences
- Statistics
- Operations research
- Data mining
- Machine learning
- Visualization
- Computer science

Technical Software Competences
- Software and product overview
- Software architecture
- Service architecture

Management Competences
- Coordination and planning
- Reporting and budgeting
- Leadership and staffing
- Quality management
EGI Community Forum 2015
November 12th, 2015 - Bari, Italy

Presentation: „Big Data infrastructure and skills from a consulting point of view“

Kevin Berwind, M.Sc.
Backup
Big Data Methods And Processes

- SEMMA
- CRISP_DM
- OWN
- Organisations
- KDD Process