Trip Report – Big Data im Public Sector

Where: Fraunhofer FOKUS, Berlin
When: 30/01/2013

Author : Martin Strohbach, AGT

Scope and Audience:
The workshop was organized by Fraunhofer FOKUS. My impression was that his was a community building event for identifying potential collaboration between FOKUS and the industry.

As this was only a half day workshop and I had to travel back shortly after the workshop, there was extremely little time to talk to people. So it is hard to give a good impression on the audience. But it certainly was directed to industrial practitioners rather than researchers. The people I talked to were all from IT companies active in the public sector.

The workshop was opened by Prof. Dr. Radu Popescu-Zeletin, director of Fraunhofer FOKUS. He mainly gave an overview of the upcoming talks.

Big Data in the Public Sector: of DigiNats, Swarm Intelligence, and Loosing Power

Speaker: Andreas Reichel, Dataport

Talk Summary: Inspiring speaker giving a basic introduction about Big Data with many examples and provocative statements. He emphasized the importance of data protection. For instance although there are about 600 municipalities (?, to beverified) in Germany on Facebook, this is a legal grey area corresponding to the German data protection laws.

He pointed out that the right technologies are not yet available in the public sector. In a response to a question when Big Data would arrive at municipal authorities in Germany, he replied that it is not there yet, but that there are a few cities such as Bremen (?) that are working on Open Data initiatives. It was not clear what his company is actually doing on the Big Data topics.

Company Information: IT provider with municipal customers looking at modernizing administration

Other Highlights

The era where knowledge is power is over for two reasons (1) data increases too fast for a single entity to own the knowledge, and (2) an ever increasing amount of data is publicly available.

Internet of Things contributes to increasing number of data.

Use Cases

- Prediction of divorces
- Mobility traces help to locate origin of Malaria

- If you call from Reeperbahn at unusual hours, an alarm will be triggered as there is an increased likelihood of an a high bill that could potentially not be paid→**data protection**
- Algotrading
- With Cybercrime makes more turnover than the drug "industry"
- In municipal administration
    - Economy projections, climate change, demographics, …
    - Urban planning
    - energy
    - MIT research: use twitter to predict traffic congestion

Big Data and the Opportunties for Public Administration

Speaker: Dr. Eckert, Fraunhofer

Talk Summary: Similar to the first talk, Dr. Eckert provided a high level and very basic introduction into the Big Data topics. This time from the perspective of Fraunhofer FOKUS. For me a key takeaway message was the identification that a **roadmap is needed** in order to follow a strategy for Big Data technologies. In this context he also pointed out that are open questions with respect to **standardization**: e.g. definitions, type of frameworks[1], understanding of strength and weaknesses of existing Frameworks for certain problem domains, and ways for quantifying the quality of the solutions. He also pointed out that Big Data is **combining a lot of data sources** together. For the public sector this means that the thinking in data silos, i.e. the administrative departments, must stop.

Other Highlights

- Relevance of the topic
    - Gartner Hypecycle
    - Survey about ICT topics that will be important in Germany
    - **Gartner's Big Data Opportunity Heatmap is hottest for eGovernment**
- Mentions FH FOKUS project such simTD (with T-Systems) and the Trusted Cloud Project MIA

Big Insights in Health Care. The Scientific Institute of the AOK analyses with IBM Smart Analytic Systems in the speed of seconds

Speaker: Alexander Schmidt, IBM Deutschland GmbH

Talk Summary: This talk was the first one that talked about actual solutions. The speaker emphasized the **importance of the capability to be able to deal with uncertainty** illustrating that based on the recommendations that Watson can give to doctors. He also pointed out that good systems in the security domain would require a crime inspector to wait a day for query, and that the fact of processing this information quickly could deliver a whole new quality for using this systems.

Other Highlights

- Unstructured data has the greater share on the data explosion
- Emphasizes the importance of having a Big Data roadmap

[1] As examples he pointed out NIST and the Telemanagement Forum. I am not sure what activities he was referring to if any. Maybe this was just used as examples for standardization bodies.

Use cases

- Patent research
  - Platform to support the patentability of an idea
  - Sentiment analysis for the idea
  - Full text search on databases
  - Nearest neighbour search of similar patents and publications
- Assessment of pharmaceutical companies and their market potential
- US hospital visits: second hospital visit is free. Thus treat the first time in a way the patient does not come back. Use analytics to find the best treatment
- Watson: recommendation for doctors
- AOK – a large German health insurer
  - Collaboration with the **AOK Scientific Institute** www.wido.de
  - Pointed out that there is a lot of cost saving potential without requiring sacrifices to the patients
  - Provide examples of the amount of data they have
  - AOK insures a quarter of all non-privately patients in Germany. This allows to reach (disputable) statistical relevance for Germany when analysing their data. For instance they found indications that threre is a trend that ADHS appears at an earlier age

Big Data Supporting Funding Proposals: „How to cope with the increasing amount of data in the public sector“

Speaker: Martin Lange, Talend GmbH

Talk Summary: This talk was the most technological including a live demo of a hadoop cluster with a couple of notebooks visible by the audience to demonstrate the Talend tool. Talend demonstrated an Eclipse Plugin that helps in importing, exporting and executing jobs on hadoop and its various extensions (e.g. Hive and Pig). It works with all major Hadoop distributions (HortonWorks, MapR, Cloudera). As pointed out by a member of the audience, the speaker fell short to show the relevance of the technology for writing proposals. It seemed to me that the talk was almost to technical for the audience, and at the same time not solid enough for people that know about hadoop and related technologies.

Other Highglights

- Nice illustration how a statement in a private context can have a different meaning, impact and interpretation than when put in public without further comment: “It’s mummy’s fault that the birds don’t sing anymore”
- There is no unstructured data, although the computer may not understand it. This was really an odd statement, especially as for his demonstrator he used information about US data because they were structured as opposed to the unstructured EU information that supposedly does not exist.
- **Big Data is not a Replacement for analytical systems**
- **Not an alternative for business critical storage systems (CRM, ERP)**
- Q: **why not use typical BI system ?**
  - A: inconclusive, pseudo-technical

Better Control By Using Early Indicators – Big Data Opportunities for Homeland Security and Citizen participation

Speaker: Georg Rau, SAS

Talk Summary: This talk emphasized on using analytics for **predictions and correlation analysis** and also included a demo somewhat similar to typical OLAP scenarios, but including **a visual analytics solutions** that has the capability to automatically choose the right visualization for a given set of dimensions to explore. The speaker contrasted the **reactive** analysis (BI and Big Data BI) vs. **proactive analysis** (predictions) of Big Analytics and Big Data Analytics. The difference between Big Analytics and Big Data Analytics is the speed rather: there are cases where there is an improvement from executing a query from hours to minutes. This changes decision making as one can make more queries and simulate whether to make an investment decision or not. He also pointed out that in order to fully benefit from Big Data, a **mind change is required to recognize that data itself is an asset.** The speaker pointed out that the **real value is in combining data between different departments in administration,** but data protection is an issue. He sees almost a contradiction in the fact that more transparency on governmental data is asked for and at the same time we insist on data protection. **SAS are discussing with Fraunhofer how a self-service analytics solution for Open Data could look like.**

Company Information: according to Forrester, **SAS is leader in predictive analytics solution**.

Other Highlights

- Tool alone is not sufficient: a full platform and experts are required

Use cases

- EU Web Portal in order to analyze proper design
- UN Global pulse: prediction of job unemployment based on sentiment analysis. Results: **unemployment could be predicted 5 months before actual figures showed the increase**
- Oklahoma Gas & Electric: **increasing amount of data from smart meters**. Solution enables **faster reactions for suppply-demand control**
- Demo: SAS Visual Analytics Tool
  - In memory DB connection to DB in Heidelberg
  - They advertise **Self-Service Analytics** rather than Self-Service BI vs (probably following the idea of making predictions)
  - Demosntrates typical OLAP queries
  - Improvement: self-service w/o IT, user cannot harm the data
  - Visual analytics: auto charting to show correlations between age and other parameters
- Question: which part of public sector would this solution serve
  - A: where there are already a lot information, e.g. ministries or their department, e.g. agricultural. But **real value is in combining data between different departments in adminstration**, for instance in combining open job offers with existing job profiles, but data protection is a real issue.
  - A lot requiremenets regarding transparency but on the other hand data protection: is almost contradicting
  - Discussion w Fraunhofer: self-service analytics for Open Data

In-memory Databases

Fabian Kaspereit, Software AG

Talk Summary: The speaker positioned in-memory databases as an important enabler for Big Data, but it was somewhat obvious that it's not the key enabler. He made provocative statement claiming that **memory is the new disk and the disk is the new tape**. He envisions that all data is processed in memory and that **feeding back data in operative processes are the low hanging fruits** where in memory databases are the right technology.

Other Highlights

- Durability problem with in memory DBs? Of course not. They have an extremely high availability.

Use Cases

- Customer: paypal
    - Billing must happen within 1 s
        - Includes risk checks, requires an increasing amount of data to be processed
        - Feedback transaction payments in sub-second analysis
- Police: Suspicious person meets another person. Should the forces act or wait?
    - Integrates various governmental sources
    - 2010: 10GB/day big, but manageable
    - 2013: **96 GB Data/day, becomes unmanageable**
- Use case: web portal for doctors informing about various drugs
    - High demand, increasing amount of data made website collapse
    - **Increasing number of DB does not scale: the more I add the less I gain**
- Real-time trend analytics during election campaigns of president Obama
    - Social media analysis
- Real-time alanysis of tactical information in fight scenarios (rockets ?)
    - Video and geo-data in real-time
    - Rocket decides what data is trasnmitted
    - Solution: **im memory DB for real-time analysis in combination with CEP**

Summary

The workshop was very valuable for understanding the current landscape of Big Data in the public sector in Germany. Personally I found that the format leaves to many room for discussions and networking. Nevertheless, I positioned BIG flyers next to Fraunhofer brochures and handed out some of them personally in particular to the presenter Alexander Schmidt from IBM, who I asked to be available for an interview for the BIG Health Sectorial Forum. A link to Sonja Zillner has been established.

Apart from the activities between AOK and IBM and SAS' involvement in the EU portal there **all other examples or success stories were non-European**. This clearly shows the necessity for creating a European community as envisioned by the BIG project. All the **presenters were extremely positive about the potentials of Big Data** as they have illustrated in their success stories. As many presenters and talks with participants during the break showed, **data protection laws are a major road blocker for Big Data** in the public sector that needs to be addressed. However this workshop did not discuss how these concerns could be addressed.

Another important take away message of the workshop was the fact that **Big Data problems occur when combining many data sources**. The talk from SAS was particular interesting as they were advertising **value of large data sets for making predictions**.

IMHO, for BIG it is important to perceive the workshops as a potential community of Big Data stakeholders and we should think about how we can collaborate with them as a whole.