Big Data exercise 1

Torbjørn Bakke

Jens Mjønes Loe

Torbjørn Øverås

Question a

Describe the three V's of the data sources used in the IT service requests management system at Florida State University.

Volume: The size of the data. In context with Big Data, the data sets are so large that traditional data processing software is inadequate to deal with them. It is not unusual for companies to have petanytes of data in storage devies and on servers.

To quote the article:

> Information Technology Services at Florida State University provides IT support to 16 colleges and more than 110 centers, facilities, labs and institutes, serving 11,000 faculty and staff along with more than 40,000 students. It offers more than 100 IT services, including email, desktop support, file storage, software licensing, classroom and learning management system support and online training.

As you can see, the data is coming from many sources and services.

Variety: The type of the data. Whether the data sets are structured, semi-structured or unstructured. The data can also be a mix of the three structure types. On top of that, unstructured data comes in a near infinite different forms, not just plain text, but also video, sound recordings, graphics, differnt document types, social media feeds and so on.

Quoting the article:

> Today, ITS's system captures, among other things, textual descriptions and manual classifications of IT problems (e.g., desktop, email, network), when they were reported, who reported them and which channels were used to report them (e.g., email, phone).

The actual content of the support ticket is unstructured text, but they also collect additional data witch is semi-structured data.

Velocity: The speed and timing with which the data is coming in and being analyzed. The data could be analyzed in real time, near real time or periodic batches with delays.

From the article:

> As part of an enterprise-wide CRM installation in 2011, ITS decided to add a component for capturing and managing IT service requests across business units.

and:

> At present, more than 100,000 service requests and their resolutions have been logged by the system since its inception in 2011.

They are using the data they collected from year 2011 up to the present year (2016). The data stream is analyzed near real time.

Question b

In the lecture notes, we explained that the benefits of data analytics are 1) to draw insight from data, 2) to make better decision based on the insight, and 3) to automate the decision and bake it into a business process, hence process automation. Can you relate these to the cases presented in the paper and give one example for each of the three benefits? Your examples can related to any of the three cases in the article and the three examples can related to different cases in the articles.

1. *to draw insight from data*

   From Hilti customer service:

   > The objective of this analysis is to unearth the root causes of recurring product failures, monitor persistent complaints about customer service and identify potential business process improvements.

   The data consist of who, when, where and how an event occures. The insight is the what and why the event happened. In the ITS example they got a detailed history of service requests from the analysis, showing when requests rose in frequency and gaining insight into why there where more requests in a specific time period.

2. *to make better decision based on the insight*

   If you understand why an event is happening, you can improve / prevent it from happening.

   For example, in Hilti they aim to monitor the volume of different types of service requests to quickly identify changes that could imply certain systematic problems; if the volume of a topic varies by more than two standard deviations, alerts for that topic might be triggered.

3. *to automate the decision and bake it into a business process, hence process automation*

   In the Hilti example, more automated categorization of service requests are desired also valuable because they significantly save time and money. Output from their tests gave categorizations that agreed with expert judgment 77% of the time, which Hilti deemes satisfactory.

   From the Inventx example:

   > Based on real-time analysis of its textual data streams, Inventx implemented a set of automated processes, such as the automated routing of requests to the most suitable agent or the datadriven recommendation engine to aid agents in providing the best answers possible.

   A new support ticket that matches a previous ticket from the knowledge base that resolved the issue could be automatically answered, eliminating the need of manual handling of the new ticket.

Question c

Discuss whether each of the following activities is a database query or a data mining/machine learning task?

1. **Dividing all the service tickets into two groups: open and closed requests.**
    - Database query. Simple select statement based on if the request is open or closed. When developing a support ticket system handling reported issues, one would assume the implementation has a database flag indicating whether the ticket is open or closed.
2. **Computing the total number of requests issued in the past week.**
    - This is a simple database query, returning the number of entries in a table with the time frame as the specification.
3. **Group more than 100,000 service tickets into topics (printer, email forwarding, etc...)**
    - Data mining/machine learning task. This requires analyzing the service tickets to determine which topic it belongs to (text analyzis etc). The machine learning model can classify tickets into clusters and then determine whether tickets have some sort of correlation.
4. **A “recommender” system that suggests the most appropriate expert for solving a problem and the possible solutions based on how requests were handled in the past.**
    - Data mining/machine learning task. This requires the system learning based on the historical data. The machine learning model can match the incoming ticket with a resolved ticket from the knowledge base. If the tickets match by a certain percentage, the resolved ticket can be applied as an answer.