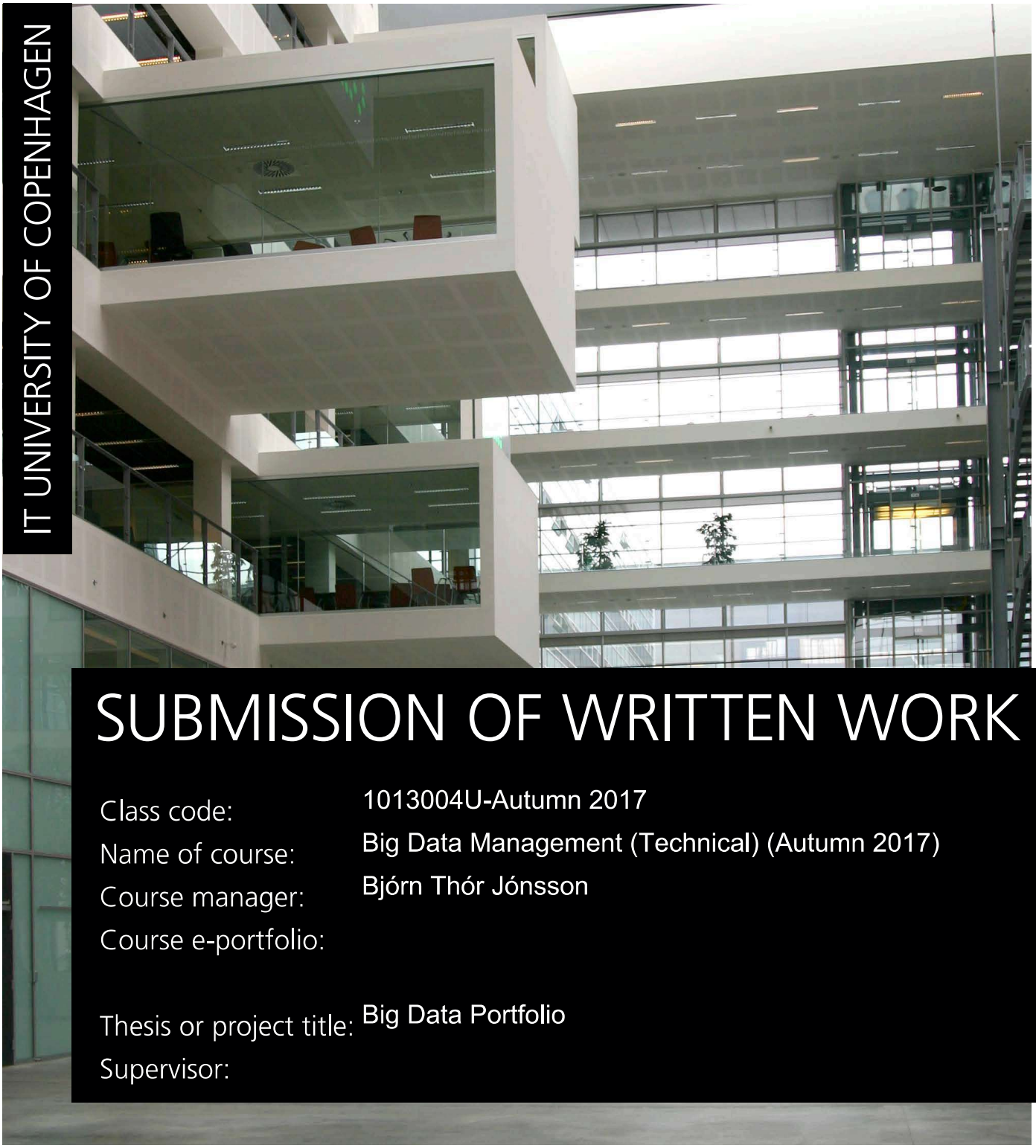
IT UNIVERSITY OF COPENHAGEN



SUBMISSION OF WRITTEN WORK

Class code:	1013004U-Autumn 2017
Name of course:	Big Data Management (Technical) (Autumn 2017)
Course manager:	Bjórn Thór Jónsson
Course e-portfolio:	
Thesis or project title:	Big Data Portfolio
Supervisor:	

	Full Name:	Birthdate (dd/mm-yyyy):	E-mail:
1.	Dennis Thinh Tan Nguyen	01/04-1993	dttn@itu.dk
2.	Jens Rimmen Bruus	14/01-1993	jerb@itu.dk
3.	Sofia Anni Sarauw	08/09-1993	ssar@itu.dk
4.	Daniel Nicklas Rosenberg Hansen	10/04-1994	daro@itu.dk
5.	Nicoline Scheel	22/02-1995	nisch@itu.dk
6.	Thor Valentin Aakjær Nielsen Olesen	14/02-1995	tvao@itu.dk
7.			@itu.dk

Big Data Management: Assignment 1

Jens Rimmen Bruus
jerb@itu.dk

Sofia Anni Sarauw
ssar@itu.dk

Nicoline Scheel
nisch@itu.dk

Daniel Hansen
daro@itu.dk

Thor Valentin Olesen
tvao@itu.dk

Dennis Thinh Tan Nguen
dttn@itu.dk

December 19, 2017

Contents

1 Introduction **1**

2 Conducting the "data detox" **2**
2.1 Commonalities after data detox 2
2.2 Meta-critical evaluation on the prospect of, and motivation to, 'detoxing' data 3

3 Databox **7**
3.1 Technical Architecture 8
3.1.1 Security 8
3.1.2 Availability 8
3.1.3 Trust 8
3.2 Nymote Databox Implementation 9
3.2.1 Mirage ("Security by Lean") 9
3.2.2 Signpost ("Availability by Cloud") 9
3.2.3 Irmin ("Trust by History") 10

4 Conclusion **11**

References **12**

5 Appendix **14**
5.1 Thor's Reflection 14
5.2 Dennis' Reflection 15
5.3 Sofia's Reflection 16
5.4 Nicoline's Reflection 17
5.5 Daniel's Reflection 18
5.6 Jens' Reflection 19

1 Introduction

From social media to IOT, we live in a world that is digitally connected, and with each connection, a vast amount of data is produced. This trail of data may be used as an advantage by the advertisers and online companies to generate business intel on the users. As such, this data driven trend is largely driven by the dominant online business model in which advertising is the primary source of revenue.[dat, 2017].

In order to understand the complications this data trail may lead to, the overall size of our data leakage must be addressed. This project will adddress the data production we generate every day and consider the effort it takes to produce data as well as stopping it.

Thus the members will take part in a "Digital data detox" experience over the course of eight days. The report will discuss the experiences and reflections produced by the members during the detox and the amount of effort it required to complete the data detox.

Furthermore, the proposed concept of a Databox which was introduced by Imperial College London will be discussed which includes whether the proposed solutions are technically feasible to address the issue of people not being able to control their data output.

Finally, the report will reflect and conclude upon whether the data leakage of the users may be fully managed by each individual. Nontheless, initiatives like the Databox strive to address the fundamental barriers, technical and social, that have yet to be addressed to make this realized in practice.

2 Conducting the "data detox"

This section will quickly address the different reflections, commonalities and unique experiences that was discovered by the members while analyzing the logs that each project member kept (referred to as 'data' from now on). During the analysis of the data, we discovered patterns of experiences, but also patterns of commentary that critically engaged and evaluated the Data Detox project which will be discussed.

2.1 Commonalities after data detox

Among the core commonalities, most of the members were already aware of the "cleaning" and privacy techniques presented such as deleting their browsing data or limiting their social media accounts. However, all members did not feel a great need to delete their data due to the benefits of leaving it be. In other words, all felt that deleting their identity data presented very little, if not unknown, rewards since the data are convenient and enhance their user experience with the services they use. For example, allowing their social media accounts to stay logged in when they use their personal computer or even allowing services such as youtube to personalize their experiences.

In regards to altering the members' online behavior, exposing them to the amount of data they produce and that they may be used for commercial use without explicit consent might not be enough to change their behavior.

The members agreed that they are aware of the vast amount of data they produce and the fact that one may not ever be fully aware of what data is produced and when and where they may be collected. However, this did not truly motivate them to alter their behaviors.

An interesting commonality in the experience of the project group is how much effort data fasting required for each member. Being exposed to this aspect of data detox truly showcased how challenging it is to avoid producing data than how easy it is to produce. The huge mental effort and preparations it requires to minimize data leakage do not increase the motivation of altering one behavior since it would be much easier just to adapt and accept their situations as is.

Consequently, if the outcome is just to accept how data is leaked uncontrollably, one may not be able to address the issues of this leakage. That is what the data detox tries to do, but it only showcases the amount of data we produce, how they are utilized by corporations and the considerable amount of effort it requires to limit one's data. Yet, it does not show the rewards for limiting our data or the rewards are not enough to convince people to limit it.

The group agreed that preventing corporations from using our data is not enough. It may require a more significant motivational factor to make people go through the effort of limiting their data. In this regard, using the safety of people as an argument may be a strong motivational factor. If the detox showcased how the data leakage might result in compromising peoples' own safety,

one might more likely change their behavior. This could be the consequences of cyber crimes such identity theft based on uncontrolled data leakage. The point being that the data detox did well in illustrating the data leakage and the effort of limiting it, but it did not do well enough to convince the project group to change their behavior.

2.2 Meta-critical evaluation on the prospect of, and motivation to, 'detoxing' data

To put a meta-critical spin on this project we will now attempt to question the epistemological underpinnings of data detoxing. We have decided to include a more in depth construction of arguments to highlight the importance and relevance of this due to previous feedback, and to do this we will briefly mobilize a few keenly connected concepts from philosophy, communication theory and educational psychology. The purpose of this section is to become more intimately acquainted with both sides of the table: Is corporate collection and processing of personal data as bad as the data detox premise presupposes, and, if not, how and when can it be considered acceptable? What drives motivation to hand over data, and how does this influence ethics?

From a social constructivist perspective, it is arguable that the prospect of 'detoxing' data consumption maintains some controversial features, as it is, to some degree, based on the underpinning idea that capitalizing on the collection and processing of data is problematic and bad; "toxic" even. By arming ourselves with a social constructivist viewpoint, we have to question the connotations that the usage of the verb "to detox" poses, and how it in turn positions how we are to perceive the noun 'data' in the context of the experiment. It is arguable, that if we accept the underlying construct that corporate capitalization on personal data is 'toxic', we are not challenging ourselves to consider our own relation to the data in question. In other words, we are reducing a complex construct to a simple dualist world view: toxic/nontoxic. Given the elicitation of this controversy, we have to more closely examine the epistemological foundation upon which the 'data detox' idea is grounded. To do this we have to look at the described premise of the experiment:

> *"The problem lies in what's happening with all of these pieces of data:* ***relentlessly collected*** *across our devices [...] and* ***eagerly analyzed, shared and sold.***"
>
> *"****Intimate digital patterns*** *emerge: our detailed habits, movements, relationships, preferences,* ***beliefs and secrets are laid bare*** *to* ***those*** *who collect and* ***capitalize on them***" (Data Detox, p. 1)

The premise quoted above can be said to appeal to an emotional response rather than a logical one, given the utilization of pathos-based rhetorical constructions (marked in **bold**). As such, the reader/experimentee/subject can be said to

be served with an emotionally conducive ideology rather than a logical argument. The idea of an unnamed "those" who "relentlessly collect[s]" and "eagerly analyze[s], share[s] and [sells]" our "secrets" illustrates this point splendidly. Emotionally imposing upon the reader an opinion on corporate action in big data processes as a premise for a data-flushing experiment leaves two outcomes for how the subject takes to the experiment. Piaget calls this 'assimilation' or 'accommodation'. Either the experimentee accepts the premise as a truth and dutifully understands the significance of the acts presented in the experiment itself because it fits with his/her existing schema, OR he/she is made to adapt to the premise, because the idea does not fit with existing schemas, which in turn results in cognitive dissonance. If accommodation induces dissonance, the experimentee will try to make sense of the learned by either elevating it as truth or meeting it with relative scrutiny [Piaget, 1954, Festinger, 1957]. In the case of our work with the experiment, most of the group members experienced the latter, as they found most of the tasks to be compromising data architectures of value to them. Whether this is due to the premise of the detox not being assimilated, or merely the experiments themselves being accommodated with dissonance (or both) is a question worth asking.

Thus, from a social constructivist perspective, we would argue the need for caution with the usage of a quirky buzzword such as 'detox', as it to some extent epitomizes and haphazardly simplifies the undertaking of a very complex ideological stance. Likewise, through our brief intervention in cognitive psychology and rhetorics, we find a problem in the way the premise for the experiments is formed, as it inflects importance on the subject's motivation to the task of flushing data. Now, while this could seem hypercritical, it is as important to be wary of the knowledge you consume as it is important to be cautious of the data you give away. And, while the takeaway from the above discussion is not to say that limiting data consumption does not have its benefits and place in the world, it does put in to question the mode with which the specific purpose of the 'detox' is presented and how this influences engagement.

Moving away from discussing semantics, we found, in our group, that that the actual end-effect of the data detox (on us) becomes not about the prospect of completely rid ourselves of corporate influence through delimiting data-processing. Rather, we found that it's effect on us was to highlight and gain insight into the massive (and, arguably, intrusive) datasets that are being collected on individual people. Mostly without our explicit knowledge. Likewise, *our* data detox became about realizing the impossibility of not producing and giving away data, as well as how difficult it is to cash-in on the right to be forgotten. And to this effect, the data detox is incredibly effective. It does this by having compiled a collection of websites that lets us see just how much data we are actually giving away. Interestingly, however, while we all agreed to be relatively intrusive, we found that we would rather continue to give our data than to miss out on the services we gain value from.

In this way, we were able to establish that the reason most group members

were unable to assimilate the experiments and act in complete accordance with the premise of the detox was that most of the experiments would result in the worsening of user experience with the data consuming services.

Two empirical examples in our logs included the problem of having forgotten details, but having it stored digitally (which, humorously enough also highlights the strength data has over the human cognition: reversibility and permanence). One group member remarked upon not remembering the exact name of a website visited months earlier, but finding value in being able to keyword-search in the browser bar to re-find them. The other example was the problem of having flushed data only to find that the access details to a certain social media website was unrecoverable, as she neither could remember the details nor recover access to the profile through a long-deleted email-address.

From these perspectives, it becomes apparent that there may be a need to weigh privacy needs against user experience needs in the question of limiting data consumption. Do we want to make it more difficult to use the Internet and be more private, or vice versa? Palen & Dourish (2003) discusses this at length through what Irwin Altman calls "boundary regulation processes":

> *"boundary regulation processes, where the person optimizes their accessibility along a spectrum of openness and closedness dependent on the context"*
> [Palen and Dourish, 2003, p.130-131]

In other words, context can incentivize the person to be more open (lessen focus on personal privacy) if there is something to be gained. In this sense, we can for example view the prospect of being able to interact with friends and family any place and any time through social media as potentially more valuable than any privacy sacrifices we make by quasi-invisible payments through data.

Ethically, the prospect of corporations collecting, processing and capitalizing data without the explicit consent and/or knowledge of the data subject adds urgency to amount of data we actually give away. It beckons us to reconsider how we pose ourselves in the realms of the digital, especially in cases where we are being tracked for no other reason than the fact that we did not read a mile long legalese text, which purports that access to all your data at all times is necessary for the app to function. This, however, in the case of the EU is what the forthcoming GDPR moves to address. As such, until this is the case, a small dose of paranoia may actually be healthy.

An important thing to consider is that data is neither objectively good nor bad, nor is it neutral. It all depends on how it is produced, collected, analyzed, and acted upon. Just like Latour's gun illustrates his translation theory, we can view the agency of consumer + data to afford certain goals, as well as the agency of corporation + data to afford certain goals. These goals and how they are reached determine ethical and moral viability, and "the responsibility for the outcome is shared by the various actants" [Latour, 1994, p. 34].

In this way, both parties need to be kept satisfied with their transaction: data for service – and neither should feel abused. However, with the current legislative landscape, this is perhaps not the case, as the consumer may not really know that he/she is being "abused" by in-transparent algorithms analyzing and predicting your behavior based on data trails that you unknowingly leave behind. From the perspective of Palen & Dourish we again need to reiterate the idea of boundary regulated processes and pose the question:

Even if it is the case that most people remain ignorant to just how much data they are handing over to corporations, would they really give up the services that in many ways are valuable to them as a member of society in a first world country?

This is tough question to answer, but the inherent abstractness of the value of data, fueled by the overwhelming intransparency of the processes that utilize it, would drive us to argue that it simply takes way too much engagement with knowledge to even begin to grasp the controversies lodged in this area. Hence, most people would arguably be more motivated by what they can gain in services than dissuaded by what they have to offer up in data to 'pay for' said service. Even us, as a group studying Big Data and having gained awareness of the complex problems of the field still find value in using services in exchange for data.

3 Databox

The personal Databox strives to help give back people control over their data by having your own "data box" of all personal information, through which an individual can control and mediate their data to third-party data processors. Arguably, this ideal conforms well with the current political debate of restricting access to personal data. Specifically, the EU GDPR regulation shares the same goal of giving back the control over personal data. In connection with this, it is natural to examine the Databox from a political perspective.

Currently, large companies like Google and Facebook store large amounts of personal data, without the data subjects being aware of what data is stored, how it is used and for what purpose. Consequently, these companies maintain a "data monopoly" and gets to classify people based on this (Shklovski, 2011). Due to this, the services that the companies provide are at a risk of becoming a "single source of truth".

An example, individuals are confronted with targeted advertisements and customized news feeds. In this way, data is processed with the purpose of manipulating individuals to spend money on certain products and vote for certain politicians. It can be argued that this is an attempt to manipulate what individuals think and believe to be true. Many data subjects are presumably unknowing of this attempt at manipulation, or at least of how it unconsciously affects their view of the world. As a result, preexisting biases in modern-day society may be amplified even further.

In this regard, the EU GDPR and the Databox project both strive to regain transparency on the Web by forcing companies to address how they use personal data and ensuring that any processing of personal data is done in compliance with the data subject.

From a security perspective, they also address the issue of companies collecting an abundance of personal data. Specifically, the initiatives require that companies only use the minimum data required for any processing to protect privacy online. Arguably, the idea of the data box and personally authorizing access to data processors gives back control and make it our choice whether we want to expose certain parts of our private lives or not. As opposed to today's current state, where many companies "freely" collect information about us and the terms of agreements on this process is somewhat "vague" and hard to understand because of the legal jargon.

Despite all of this, one may raise a concern regarding the Databox project, since all access to personal data is controlled and regulated by the individual. In this regard, the individual now determines what data is exposed, which leads to the final question of whether this data may even be considered "raw"? As Bowker and Star point out, claims to objectivity of big data are misleading, and the same applies when individuals chose only a subset of a larger set of data to expose to external clients (boyd and Crawford, 2011).

3.1 Technical Architecture

Databox is defined to be *a personal, networked service that collates personal data and can be used to make those data available*[Hamed Haddadi1, 2015]. Thus it is a piece of software that collects personal data and then manages how that information is made available to third parties. *In essence, it's "a networked service that collates personal information from all of your devices and can also make that data available to organisations that the owner allows"*[gua, 2015].

3.1.1 Security

First, the Databox must be trusted by the individual who uses it, since it gathers information about browsing habits, buying behavior, financial information, contact information etc. All together, storing this in a single online repository requires the Databox to be secure. At first hand, one would seem to think a single point of control over personal data is secure. Secondly, the Databox must allow controlled access to personal data from third parties, based on settings controlled by the user. Finally, there must exist incentives for people to use the service, e.g. by letting third parties pay for using the data and making it easy to use. In this regard, one may investigate how the databox actually addresses the concerns of trust, security and usability from a technical standpoint.

3.1.2 Availability

The Databox must be securely accessible, reliable and robust against loss of power, connectivity etc.[Hamed Haddadi1, 2015]. Thus, the service must be consistently available to help users manage their online interactions and their personal data proliferated from these actions. In order to maintain availability, all connectivity is pushed to the cloud. However, this introduces other challenges such as trust and cost.

3.1.3 Trust

The databox is a much more knowledgeable and intrusive system and trust is a critical requirement to address. Thus, it must provide means for the user to intervene in data collection and sharing. Further, logging mechanisms may be used so that users and third-parties may build trust in the system. To address all of these technical requirements, the people behind Databox are working on technologies, including **Nymote**, Mirage, Irmin and Signpost[nym, 2017].

3.2 Nymote Databox Implementation

The Nymote project and its constituent components including Mirage, Irmin, and Signpist is exploring the direction of a Databox. Thus, it is a concrete implementation of software used to control your networked personal data.

3.2.1 Mirage ("Security by Lean")

Mirage constitutes an alternative approach to deploying and managing services that are designed to be small, lean and secure [mir, 2017]. Namely, MirageOS is a programming framework using the OCaml language for building type-safe, modular systems. In this regard, it is used as a library to construct 'unikernels' running under the 'Xen' or 'KVM' hypervisor on Amazon's Elastic Compute Cloud and Google Compute Engine to secure network apps on the cloud.

Unikernels are lightweight application images with a small attack surface, as they are confined to the hypervisor. This is similar to a sandbox mechanism where programs are run in isolation to mitigate system failures and vulnerabilities from spreading. In pratice, code may be developed in a high level language (e..g OCaml) on a desktop OS (e.g. Linux or Mac OSX) and compiled into a fully-standalone unikernel that are run directly on Xen hypervisor machines. All together, MirageOS is one approach to controlling personal data securely in a databox application on the cloud [mir, 2017].

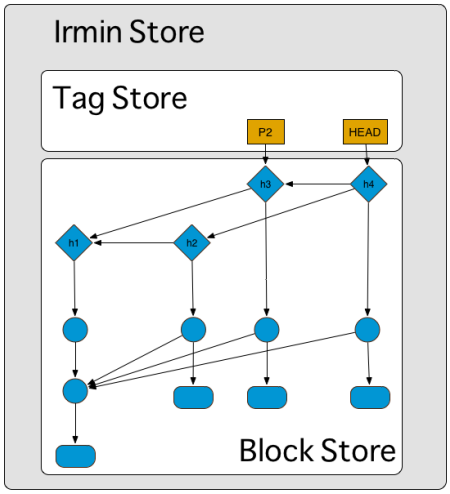
3.2.2 Signpost ("Availability by Cloud")

Signpost is used to build a "personal cloud infrastructure" to let all your devices see and reach each other without any complex configuration issues. The biggest challenge is to circumvent firewalls and NAT (i.e. Network Access Translation) servers that prevent devices from seeing each other. As a technical solution, cloud services need access to you personal infrastructure like e.g. Dropbox does. In this regard, Signpost addresses the problem by using probing connectivity and using DNS updates to constantly keep track of your devices [sig, 2017].

3.2.3 Irmin ("Trust by History")

Irmin is an OCaml library database similar to Git-like version control that is used to keep history between devices easily. Thus, Irmin uses the principles behind the version-tools like Git and applies them to solve the problem of storing and syncing data. Thus, it may be used to keep track of personal data as required in the databox. Instead of structuring the system like a conventional database, Irmin exposes the concepts of Git (clone, push, pull, branch, rebase) as a library to be used in any OCaml application. Under the hood, Irmin uses an append-only block key-value store and persistent mergeable queues, i.e. persistent queues with a fast merge operation.

Figure 1: The Irmin Store



The Irmin store is comprised of a block store and tag store that represent a global pool of data and a way to name values in that pool respectively. The Irmin block store connects blocks of data by using the hash of the block contents as an address to simplify the way distributed stores can be synchronized. The data sturctures are immutable, so once a block is created in the block store, its contents will never change. Instead, updating the immutable data structure returns a new copy. In terms of structure, Irmin uses branching to share information between processes. Namely, the Irmin tag store is the only mutable part of the system that maps global (branch) names to blocks with data in the block store. Ultimately, these tag names are used to pass block references between different processes. Finally, Irwin supports merging data structures consistently to provide an immutable history that help support computer forensics and rollback features. All together, Irwin rethinks how we persist data and provide history through version control to sync and coordinate devices, which may ultimately be used in a databox to store personal data[irm, 2017].

4 Conclusion

This project tried to enlighten the issues of our data leakage trough a practical data detox experience. It can be concluded that the detox project did well in showcasing the amount of data we leak, and the amount of effort one is required to limit such data. Especially if one was to prevent producing any data.

Also, the detox project tried to change people's behavior but the arguments such as how firms use our data without being transparent, was not enough to motivate people into changing their online behavior due to how much effort it required to limit such data. It would require additional motivational factors such as how data leakage may compromise one's safety, but it is arguable whether this point would be enough to make people go through the effort of managing their data.

Furthermore, by limiting the data we leak may affect many of the free services since the data we leak are their source of revenue for many online firms. Consequently, the services may have to use another business model to get their revenue such as setting up a paywall. This discussion regarding the balance between user experience and privacy is still an open question on how this should be solved.

On one hand, one may create regulations such as the GDPR that tries to put a limitation on the data gathering through legislation. On the other hand, the concept of the "databox" is to make the limitations through a technical implementation. The Databox helps showcase how the technical requirements of a data box may be implemented in practice to provide a secure, trustworthy and highly available service for managing personal data.

Regardless, attempts are being made to help people become aware of the proliferation of personal data online, and suggestions on the efficient management of personal data are being addressed in a way, such that companies can still retain online business.

References

[dec, 2011] (2011). A decade in internet time: Symposium on the dynamics of the internet and society. Available at SSRN: https://ssrn.com/abstract=1926431. [Online; accessed 26/9/17].

[gua, 2015] (2015). Fightback against internet giants strangehold on personal data starts here. https://www.theguardian.com/technology/2015/feb/01/control-personal-data-databox-end-user-agreement. [Online; accessed 24/11/17].

[nym, 2017] (2017). https://nymote.org. [Online; accessed 24/11/17].

[dat, 2017] (2017). How a box could solve the personal data conundrum. https://www.technologyreview.com/s/534526/. [Online; accessed 24/11/17].

[irm, 2017] (2017). Irmin introduction. https://mirage.io/blog/introducing-irmin. [Online; accessed 24/11/17].

[mir, 2017] (2017). Mirage official homepage. https://mirage.io/. [Online; accessed 24/11/17].

[sig, 2017] (2017). Signpost. http://nymote.org/software/signpost/. [Online; accessed 24/11/17].

[Boyd and Crawford, 2011] Boyd, D. and Crawford, K. (2011). Six provocations for big data. page 4.

[Boyd and Crawford, 2012] Boyd, D. and Crawford, K. (2012). Critical questions for big data, information, communication and society.

[Davis and Jurgenson, 2014] Davis, J. and Jurgenson, N. (2014). Theorizing the web [special issue]. interface 1.1. 1 - 14.

[Festinger, 1957] Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

[Hamed Haddadi1, 2015] Hamed Haddadi1, e. a. (2015). Personal data: Thinking inside the box. https://arxiv.org/pdf/1501.04737v1.pdf. [Online; accessed 24/11/17].

[Latour, 1994] Latour, B. (1994). On technical mediation - philosophy, sociology, genealogy.

[Palen and Dourish, 2003] Palen, L. and Dourish, P. (2003). Unpacking privacy for a networked world. new horizons. pages 129–136.

[Palen and Dourish, 2017] Palen, L. and Dourish, P. (2017). Databox project - about. https://www.Databoxproject.uk/about/. [Online; accessed 26/9/17].

[Piaget, 1954] Piaget, J. (1954). *The Construction of Reality in The Child*. The International Library of Psychology. Basic Books Inc.

[Shklovski, 2017] Shklovski, I. (2017). Thinking in terms of data [pdf slides]. retrieved from `https://learnit.itu.dk/pluginfile.php/194561/course/section/95785/week2.pdf`. page 9.

[Vertesi, 2015] Vertesi, J. (2015). How evasion matters: Implications from surfacing data tracking online.

5 Appendix

5.1 Thor's Reflection

Was this information you already knew?

I have already used a lot of the detox techniques including confirming my Facebook settings, using adblock, private browsing, disabling data exposure on my phone (location, Bluetooth, wifi). I believe this to be a consequence of my rather technical background and my epistemology, being a software developer. However, I did not know about the anonymously collected data from Google that displayed my interests based in my search history. Also, I did not use few of the plug-ins that help block against any unwanted processing and collection of my data while searching. The biggest epiphany deduced from this is probably the fact that even I do not grasp all the channels of which my personal data is being collected. Especially, it frightens me that this might happen without my "understood" consent.

How much did you have to alter your routines and practices?

I already have been trying to minimize the exposure of my data through my computer and phone. Despite this, I realized that the detox is an ongoing process, which I think people tend to forget. Specifically, I had almost 100 applications on my phone and I was able to reduce it to less than half. Again, we sometimes seem to forget how the collection of data is aggregated by staying active in our daily lives and digital world. I liked the idea of leaving my phone at home. For some reason, it has become a tendency to check it regularly in the search or any potential notifications or reach outs from other people. I did not remove my cookies, since they are convenient and help make me more productive when using my laptop. However, I am aware of the data being collected. Despite this, it has a specific purpose in my life and is not just present for the "sake" of it as opposed to the many apps I removed because they do not have a purpose in my life.

Will this affect your own behavior?

I will probably employ some of the tools on a more regular basis. Especially removing apps and any means to data collection that do not deliberately serve a purpose in my life. Further, I would like to leave my phone at home occasionally.

5.2 Dennis' Reflection

Was this information you already knew? Most of the detoxing techniques were already known to me. Thus, before doing the detox process, I had already employed most of them, such as limiting my data exposure through VPN add-blockers and tracker-blocker, but also enabling many of the privacy settings on my social media accounts. In regards to the information that Google collects, I am already well aware that they collect a lot of data from me based on my internet behavior. What surprised me the most was how they had framed me as a completely wrong person. On the internet, I was a 44-54 age male based on Google. On one side, this allows me to stay anonymously, since they got the wrong notion about me, yet it is still somewhat disturbing to be framed and perhaps forever regarded as an entirely different person than what you are. How much did you have to alter your routines and practices? In regards to data fasting, not being connected and living a primitive life in a digital world is difficult, yet intriguing. A lot of preparations were required such as withdrawing money and deep reflection upon which channels may output data about you. What came into realization, was the fact that one may not be entirely disconnected from the digital grid. Even though I had to alter almost all my routines and practices, I was in some sense not able to avoid producing data. I had disconnected my phone and WI-Fi, used paper cash and avoided going outside. Yet, external factors such as surveillance cameras, my power and water usage at home were still producing data about me. To avoid this, I might need to move into the forest and live a stone-age life shunning all kinds of modern life aspects. One may reflect upon if this extreme change of lifestyle is worth

doing to avoid the big data sink or should we just adapt to this digital world we have created and appreciate its benefits. Will this affect your behavior? After trying out data-fasting and data-detoxing, I can conclude that the benefits of a digitized world have some great benefits that I can not live without after being exposed to it. Having a credit card and being connected to my friends and family does make life a bit easier. Although, after being aware that life with reduced exposure is still plausible I would try to limit the amount of data I produce on the internet by occasionally going through my privacy setup and see if they are up to date to what I consider is enough.

5.3 Sofia's Reflection

Did I already know the things presented by the Data Detox, and did I alter my behavior after the Data Detox? During the introduction to the Data Detox, the creators of the project fails to fully argue why the data build-up is toxic, and why we need to strive for less data. In this introduction, they state that the problem is how our habits, preferences, relationship etc. are exposed to companies, who capitalize on it (ref). They do not explain how exactly they capitalize on a person's data, but we can assume they mean personally targeted ads. I work with programmatic/algorithmic buying of ads that are targeted towards different segments of the population, so I have a lot of knowledge about this already. I did therefore not alter a lot of my. I enjoy being recommended articles, events, etc. that have to do with my hobbies. I also know that on an ad-exchange, I am a completely anonymous cookie. I did delete my cookies, and browsing history in order to follow the detox, only to find it really annoying and a lot of effort to keep my data output low. I also did not feel the Data Detox project provided me with any other argument for deleting my data other than to avoid ads. The call for action: "Confuse the companies" offers no explanation for why we want to ruin the business of ad and media companies that actually benefits the economy and makes up many jobs in the US.

Do I think the Databox project presents a viable solution? No, based on my experiences with the Data Detox, managing data is too much work for very little reward. The learning curve associated with using this app/platform seems steep, and even after education myself on the website, I am still confused in regards to how the Databox is supposed to help me. Wrap up I think the Data Detox project is somewhat problematic in the sense that it builds on the existing rhetoric about "big brother" (ref) and companies monitoring you. They do not provide any form of explanation of how this is done, to what extent one is being monitored, why it is bad and why we have to protect our privacy. They have also left the question of personalized data rather one-sided, since it also fails to emphasize the reformational things that collection, archiving and analyses of our data can do for society. In regards to my data output, I am primarily concerned with identity theft, and I did not feel that the Data Detox program provided me with much information about how to protect myself from it. All in all, I am really missing a concrete argument for the problematic aspect of companies capitalizing on personal data.

5.4 Nicoline's Reflection

Was this information you already knew? Many of the data detox tools were already known to me, however I only employ a fraction of them, including incognito browser, deleting browser history and cookies and keeping the privacy settings of my social media accounts very strict. I am aware that Google, Facebook and similar companies collect a lot of data about my online activities, and I was not surprised to find that what Facebook and Google think I like was mostly accurate. However, should any one of my Facebook friends dare venture into my profile and look at my interests, they could conclude the same as Facebook, and similarly should any one venture into my public YouTube profile, they could conclude the same as Google. Therefore, it did not phase me that Google and Facebook knew this about me, as I have chosen to make this information publically available. What did surprise me were the peculiar conclusions that Facebook drew about my interests that did not correspond to reality at all. In fact, Facebook seemed to connect me to a particular theme that I oppose. In addition, being able to view my search history sequentially was an eye-opener. Upon viewing my google search history alike this, I ended up deleting some search history elements. It made me think that if anyone gained access to my google account, they could get a look into my private life. They could get to know things about me that even my good friends do not know. How much did you have to alter your routines and practices? What did you NOT do? For the data fast, it was practically impossible not to produce any digital data: Could I go shopping groceries? No – I had no cash, and using my credit card would produce an entry on my transaction history. Had I had cash, could I actually have left the apartment without producing data? No - the premise around my building is surveilled, and surveillance tape surely qualifies as digital data as well - similarly to the measure of electricity I used, and the volume of water coming out of my tap. I could not contact my friends or family on my phone. In conclusion, living a life without producing digital data is practically impossible in modern day society. Having completed the seven days of detox, consider - will this affect your own behavior? It will not affect my behavior considerably. This is despite the fact that I recognize mass data collection to be problematic. As an example, data from Facebook has been used to psychologically profile entire populations of American states. This information has then been used when constructing targeted political campaigns. Basically, the politicians know what repertoire to use when addressing certain crowds. The idea of being able to influence populations in this manner, e.g. in connection with the election of the new U.S. president, is a scary notion to me. Even so, I am simply not willing to sacrifice the use of the services that collect the data.

5.5 Daniel's Reflection

Was this information you already knew? I knew most of this information already, for me I think the biggest surprise was that people go such lengths to avoid targeted marketing. I also get why people are upset that Google has blocked the extension that misleads the advertising tools, but Google primarily makes money from advertising, so I think it is understandable that they as a company will not allow this kind of software to be distributed through their own stores. I believe that the targeted advertising is the price you pay for using an extremely complex and otherwise expensive service for free. A good rule of thumb is that if something is free, you are the product. How much did you have to alter your routines and practices? I had to use different search engines, that were inferior to what I am used to, I also needed to use some extensions that I am not used to. Furthermore, I deleted some apps that I used sometimes, so I had to live without those. What did you NOT do? Why? I did not clear my history, simply because how inconvenient it would become for me if Google no longer knew who I was and what I wanted when I am using the service. If I wanted that kind of privacy I would have used inPrivate for every web session. I am used to deleting specific things from my history if I suddenly get a lot of annoying ads, but deleting the whole thing was too extreme for me. Did you learn anything new? I learned that much of the data that is stored about me is easy to extract from Google, which I think creates a great kind of transparency. I think Facebook could learn something here. Having completed the seven days of detox, consider - will this affect your own behavior? No not at all, I have always been picky in what to share and what not to. I did delete some apps, however, so this will change something I guess?

5.6 Jens' Reflection

An interesting discussion regarding this project is that of what privacy means to the individual. We can start this exploration by turning to Irwin Altman and his perception of privacy as a selective control of access to the self (Altman, 1975). Within the interplay of 'selective control' and 'self' we are able to perceive the concept of privacy as involving a degree of individual autonomy and free will. Merging this with the contemporary idea of data as the individual's existence on networked computers lets us perceive data as being the access third parties on these media have to the 'self' - to private individuals. In other words, ideally data should be only comprised of what an individual selectively decides to share with content providers on the internet. Taking this perspective, the concept of data as something "toxic"; something which one needs to rid oneself off becomes a frail one. Obviously, the vast majority of people do not read the TOS for every service they access on the internet, likewise they probably don't consider that some services collect significantly more data than they need to provide said service. Albeit, in order to understand data in the context of something which allows access to the individual, we must also identify data as something inherently personal. Something, which, by its very nature, cannot be toxic because it is merely a necessary and auxiliary part of existing in a networked society. "The problem" of data thus becomes about finding an equilibrium, not about ridding oneself of anything related to it. An equilibrium is necessary to control what can be known about oneself through data. However, through this we can also establish that no third party access to the self (apart from being nearly impossible) is as "toxic" as excessive third party access to the self. In fact, if we view data as something inherently personal; something that defines the self in a certain context, we can even argue that (a forced) 'data detox' is as much a breach of privacy as is excessive third party internet tracking.

Big Data Management: Assignment 2

Jens Rimmen Bruus
jerb@itu.dk

Sofia Anni Sarauw
ssar@itu.dk

Nicoline Scheel
nisch@itu.dk

Daniel Hansen
daro@itu.dk

Thor Valentin Olesen
tvao@itu.dk

Dennis Thinh Tan Nguyen
dttn@itu.dk

December 19, 2017

Contents

1 Introduction **1**

2 Architecture **2**

3 Data **4**
3.1 Data usages 4
3.2 Data cleaning 5

4 Defining the batch views **6**
4.1 Network traffic view 6
4.2 Room usage view 7
4.3 Courses and operating systems view 9
4.4 Implementation 12
4.5 Network traffic view 12
4.6 Room usage view 13
4.7 Courses and operating systems view 13

5 Legal issues **14**
5.1 Ethical considerations 16

6 Impact analysis **17**
6.1 The winners and losers of data interpretation 17
6.2 Other approaches to data based optimization 18

7 Conclusion **19**

References **20**

8 Log: Concluding Reflection on the Process **21**

1 Introduction

In this case study, IT-University of Copenhagen (ITU) has requested to implement a Big Data system providing insight on the wifi-usage at ITU. This insight will be utilized by ITU for beneficial purposes such as maintaining the existing network, improving facilities and so forth.

Thus ITU has provided a large set of data surrounding the WiFi infrastructure and facility usages at the IT University of Copenhagen, this project addresses how this data can be utilized to develop meaningful services of possible use to the University and what ethical considerations such services give rise to.

To address this, a prototype will be developed in the form of three batch views. The prototype is based on a chosen set of use cases based on the data. In developing the prototype, both the architectural considerations as well as ethical issues are considered and discussed.

The produced report will describe the data and its usages, how the data is cleaned and processed, the architectural considerations, an ethical discussion of the usage of the data and lastly an evaluation of the impact of the prototyped service.

2 Architecture

The Lambda Architecture is used to design and implement big data systems with robustness, scalability and fault-tolerance in mind. This fits well with the distributed nature of big data and its three data properties volume, variety and velocity. Namely, the architecture may help cope with the massive volumes of unstructured data and provide enough speed for big data processing and help process different types of data. The Lambda Architecture builds a big data system that fulfills the above goals by providing a 3-layer architecture, comprised of a batch layer, serving layer and speed layer (Marz and Warren, 2015, p. 19):

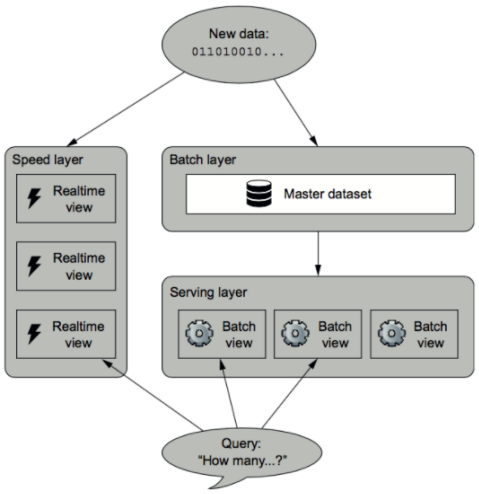


Figure 1.11 Lambda Architecture diagram

Figure 1: Lambda architecture

As seen on figure 2 , the Lambda Architecture uses both a batch layer and speed layer to process data. The batch layer holds the master data that represents an immutable, append-only set of raw data (Marz and Warren, 2015, p. 29) as a single "source of truth" in the system. This raw dataset is used to achieve accurate and consistent data processing results through recomputed views of batch data. These batch views are stored and indexed for performance improvements in a serving layer that decides how and where data is served to consumers of the system. Finally, the speed layer addresses the issue of latency at the cost of accuracy by providing real-time processing to complement the batch layer that strives to provide consistent and accurate data at the cost of being slow. As a result, batch views and speed views are combined in the serving layer to provide results to the client of the system. In this way, the Lambda Architecture becomes robust through the combination of two layers. Specifically, the real-time speed layer complements the batch layer by dealing with large amounts of data at a high velocity (speed), while accuracy errors

introduced by the speed layer in the data is corrected by the batch layer.

In this project, HDFS is used as the storage layer to store the master data, which is processed in Spark as the batch layer and finally served in the serving layer as csv files in Excel. As shown in figure 2, any new wifi data is stored in HDFS on the Spark cluster machines.

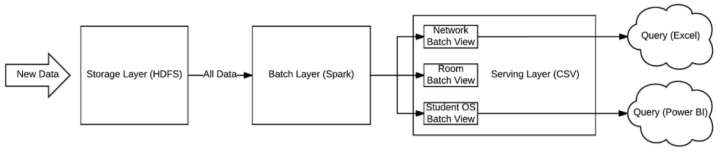


Figure 2: Lambda architecture

This data is passed to Spark as the batch layer, which processes it and produces a set of batch views. In this project, three batch views are produced that each provide data for potential insight about the network traffic, room usage and student operating systems at ITU respectively. All together, these batch views are served as CSV data to be exported in business intelligence tools like Excel and Power BI for visualization. These parts of the lambda architecture have been implemented to help provide ITU with analytics of their wifi data. As mentioned already, the data is stored in HDFS that distributes the data set across the provided cluster machines in multiple file partition nodes. The distributed file system helps provide high-performance access to the wifi data across the cluster, as opposed to storing all of the data as one big file csv file in the built-in file system. Further, HDFS is designed to be highly fault-tolerant and enables the system to continue running even if a node fails by having copies of data distributed as well. Finally, the data may be processed in parallel in the Spark batch layer, because HDFS takes in data and breaks it into separate partitions, distributed to different nodes on the cluster. It uses a master/slave architecture, with each cluster consisting of a single NameNode that manages file system operations and DataNodes that manage data storage on individual compute nodes.

In terms of the Lambda Architecture, the speed layer has not been implemented and the serving layer does not index of batch views for performance improvements. However, the Lambda Architecture has been implemented with Apache Spark and HDFS.

Finally, to further address the requirement of robustness and the issue of latency, one may implement a speed layer that uses Apache Spark Streaming. Such an implementation complements the consistent precomputed batch views with real-time incremental views, however this sacrifices accuracy to minimize latency by providing real-time views of the most recent data. Essentially, this helps fill the "gaps" caused by the batch layer not being able to serve views based on recent data.

3 Data

3.1 Data usages

Before one may formulate suggestions to how the provided data may be utilized, it is critical to get an understanding of how the raw data is defined, what information they provide and which limitations that may arise from such data.

Firstly, the data provider is a web server at ITU, which provides daily reporting on three types of data. This includes a meta-dataset, which list all data regarding WIFI-routers and their location at ITU, time-stamp datasets reporting the connected devices to the WIFI endpoints at some given point in time, and finally, datasets reporting all the room bookings for different courses enlisted at ITU.

By observing what data the three datasets provide, a union of the meta-dataset and the time-stamp dataset may produce a WIFI-usage dataset, that offers insight on multiple aspects regarding the WIFI use at ITU, such as the general traffic on the WIFI-infrastructure. Namely, the data includes simple information about how many devices that are connected to a router or how many routers there are placed at a giveb location. More complex information may also be derived from this data, such as the overall load on given endpoints and the general signal to noise ratio at some given area. All together, this information can be used by the IT-department to tune and optimize the WIFI-infrastructure at ITU. For example it may help them identify areas with a need to install additional WIFI-endpoints due to weak reception or overloads during spike periods.

Further, the room dataset may be used to provide insight about which rooms are currently occupied within a given time frame. This information is, reasonably, the most apparent use case but given a union on the WIFI-dataset and meta-dataset the general WIFI-activity during lectures may be tracked. For example, one may derive how many clients are active during a lecture, as well as the network load between different lectures. Again, this may help the IT-department in optimizing and tuning the WIFI. For example, they may check whether a router can handle a large number of clients in Auditorium 1 by analyzing the number of clients that were dropped and needed to continuously reconnect. If there are many drop-offs, then they may consider replacing the current router or check if there are any external factors that may disturb the reception.

What may not be obvious is the fact that the provided datasets do not expose the exact traffic or network activity of a given client, so it may not be possible to monitor each client directly and reveal their network activity. Therefore, the datasets may not be used to verify whether the network activity is related to the current context of a given lecture, or even if it may involve some security tampering. Also, the datasets do not explicitly state the identity of a client, so it may not be immediately possible to reveal the person behind a given client.

Although, each client does have a unique anonymous identifier or "CID" in the time-stamp dataset, which cannot be linked to a given person directly using the provided dataset. The CID is mapped to one unique client, and thus one may be able to monitor the behavior of an anonymous client. For example, this may bring insight to how often a person is at ITU, where this person usually is located and also which courses the person may take based on a union of the

timestamp dataset and the rooms dataset.

By utilizing this knowledge, one may be able to extract information on individuals such as which year/semester they currently are enrolled on, as well as which education they are enlisted to. Still, this does require some additional data regarding how ITU educations are structured and which courses are bound to which ITU educations. The provided datasets do not supply any raw data regarding the courses or their relation to a given education line, so one must extract the data from a different data source.

In this project, it has been decided that the data can be used as a means of bettering network services (institutional focus) and ultimately improve the WIFI user experience. In particular, the rooms data can be used to improve facilities, supplies, and logistics. As an example, if a room is determined to be used very little, a survey might be conducted as to why this is, which may result in optimizing the use of facilities provided at ITU.

3.2 Data cleaning

Data surrounding client accesses to devices is sampled approximately every minute with some seconds deviation. This data is saved into a JSON file containing attributes such as did, RSSI, snRatio, cid, clientOS and ssid for every day. The number of attributes vary for some instances of the daily JSON files. As an example, some instances of the data do not contain attributes clientOS and ssid. These missing attributes are observed on the files from the 30th of September to the 4th of October.

Thus, data cleaning is performed to cater for these missing attributes in the client access data. Namely, a schema for the data is defined in the Spark pipeline that allows for null values in every attribute. An alternative approach would be to disregard the JSON files that contain data with missing attributes. However, that would result in a substantial loss of data that could otherwise be utilized in views where some null values are acceptable.

To make use of the data on client accesses, the JSON files are first converted to a dataset in Spark. This dataset is then flattened by exploding the arrays readings. Next, the clients are defined as in the JSON, meaning there exists a row containing the attributes did, RSSI, snRatio, cid, clientOS and ssid for each client access. Furthermore, if the clientOS attribute is null or empty string, it is replaced by 'unknown', so that all clients rows with an unspecified operative system can still be grouped, when processing the views. This signifies that the data is missing; not that the router has been unable to identify the OS of the client. It is only a small part of the data that is missing information on clientOS, e.g., approximately 7% client rows were found to be missing data in the clientOS attribute on the file 25-10-2017.json.

The data contained in the rooms JSON files are cleaned by splitting the string contained in the room attribute, as it contains several rooms in a single string. Instead, a row is added to each room contained in the string, as seen:

```
val roomsCleaned = rooms.flatMap(r => r.room.split(",").
    map(room => (room, r.startDate)))
```

This allows for views, which may contain data for each specific room rather than data for a combination of rooms. All cleaning and processing is done during precomputing of batch views.

4 Defining the batch views

For this project, the following three batch views have been computed. The batch views can be utilized in the query layer, using various Business Intelligence (BI) tools. This section will showcase different data reports based on the computed views. We utilize the Microsoft PowerBI tool to analyze the data.

4.1 Network traffic view

The network traffic view provides data that the IT department may use to monitor the use of specific devices in the network infrastructure at ITU. This data provides the means for the IT department to identify overburdened devices or specific time periods where a specific device is burdened. The view is precomputed in a way that allows support for both general queries and more specific queries. As a result, both of these types of queries may be answered within a reasonable time. This comes at the cost of a slightly larger precomputation time. In a real-world use case, the large computation time would be unacceptable in the absence of the speed layer. In this project, the generality and speed of the batch view are prioritized more so than the precomputation time to be able to cater for more interesting queries in the query layer.

The view achieves this generality through time granularities as inspired by (Marz and Warren, 2015, p. 201). That is, the view is precomputed with two different time granularities, namely hour and day. Therefore, the schema of the view is the following:

$$NetworkView(location : String, deviceName : String, \\ granularity : String, value : Int, cnt : Int)$$

In this schema, the granularity is a single letter such as 'h' for hour and 'd' for day, where the value signifies the value for the given time granularity. As an example, a granularity of 'd' with a value of '200' would mean day 200 since the Unix epoch. The attribute 'cnt' is the count of clients.

As an example, using the time granularities, the view can answer the more general question, "How many clients connected to the devices on the fourth floor in October?", without having to examine every hour of every day, but rather every day of the month. The more specific question, "How many clients were connected to the device(s) in 4A07 between 10.00 - 12.00 on October the 5th?", can be answered using the same view.

The following figure is based on the network traffic view and provides an example showing how the view might be used to answer a query in the query layer. The view enables selection of which devices in which room to inspect and which granularity to use. This functionality is implemented using 'slicers' that create checkboxes for the user to select from.

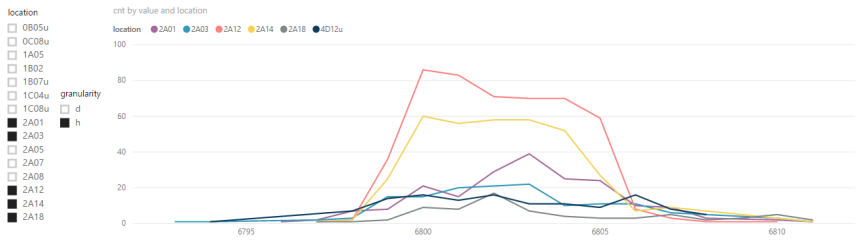


Figure 3: Network traffic per rooms

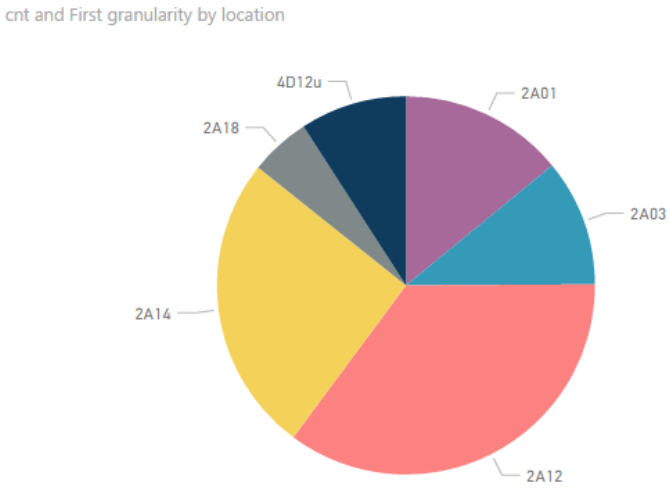


Figure 4: Count of connections per rooms

The figure has two different representations of the data. The representations are shown as a line chart with connections over time and a pie chart that shows the distribution of the total amount of connections in the dataset.

4.2 Room usage view

The room view shows the usage frequency of the rooms at ITU, based on how many devices being connected. Below is a report that compares the different usage frequencies between the auditoriums on a given date. On the 25/10-2017 Aud3 was used the most. For example, this view may be useful for facility managers to optimize rooms that are not used as intended, and further gain insight as to why some rooms are used the most. It might even help determine which

rooms require the most frequent cleaning and which may potentially be skipped.

The schema of the view is as follows:

$$RoomView(room : String, date : java.sql.Date, usage : Integer)$$

The usage attribute in the schema corresponds to the number of times a given room has been used for a given date.

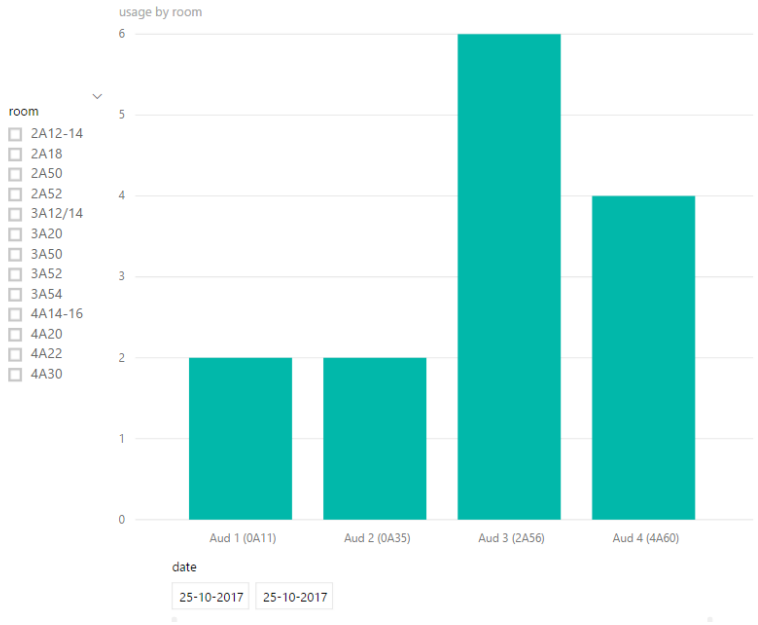


Figure 5: Usage by rooms - bar chart

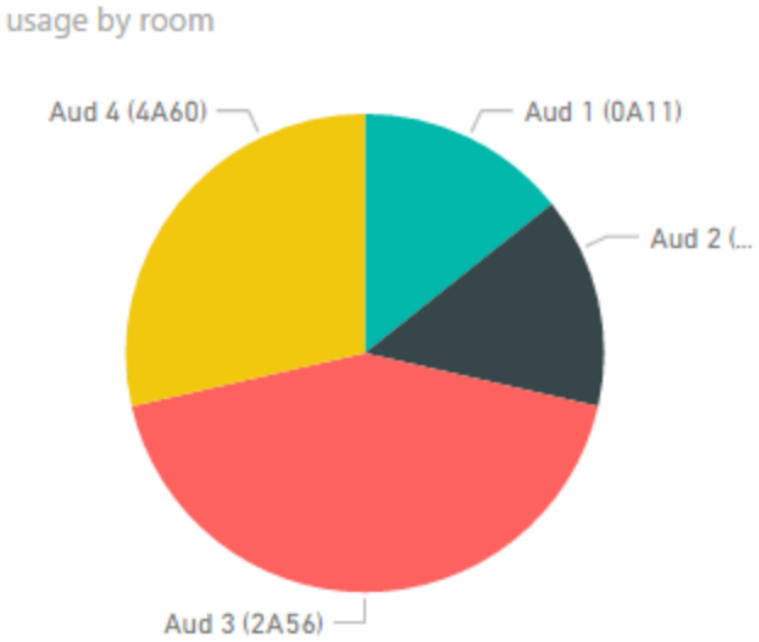


Figure 6: Usage by rooms - pie chart

The above illustrations showcases a comparison between the usage frequency between auditorium AUD1 to AUD5 on the 25th of October. Interestingly, the view yields the possibility to select and compare different rooms, based on a time interval or a specific date. For example, one can compare rooms, floors and see which floors are mostly used. Further, one can select all rooms on e.g. the fourth floor and the third floor, sum their frequencies, and ultimately determine which of the two floor has the most room bookings. Again, this may help facility managers to decide which floors may be optimized and help distribute people evenly around the building. Thus, this may avoid some floors from being overcrowded.

4.3 Courses and operating systems view

The courses and operating systems view provides data describing how many clients in a specific course that use a specific operating system. This view could be utilized by the course administration of ITU to identify what software tools to use in the lectures. A possible scenario could be that the course administrator of a user interface design course wishes to identify which design software tools to use in his course. Namely, should the students learn a tool like 'Sketch' for MacOS or 'Gravity' that enables browser-support? In this scenario, if the data revealed that the majority of students for the given course are Windows users, 'Gravit' might be the better option. The schema of the view is as follows:

$$(clientOS : String, count : BigInteger, course : String)$$

The following figure shows how the data might be used in the query layer:

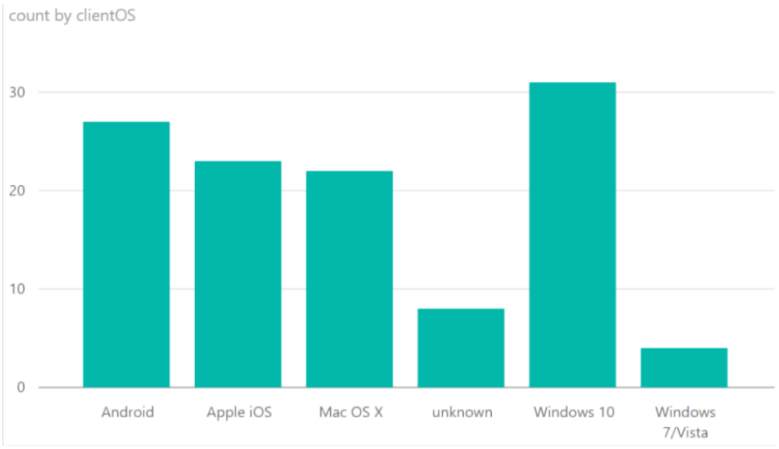


Figure 7: OS usage based on every courses

The above figures showcase the number of connected devices, based on their OS and a given course. In this case, it illustrates the spread in choice of operating systems between students in the "Critical Big Data Management(...)" course. This allows a course administrator to decide which tools are most suitable to use to maximise device compatibility and help accommodate the students, based on their choice of hardware and OS. Also, another interesting insight is revealed in the choice of operating systems compared between different courses. By way of example, which operating system do students in a creative course use, as compared to students in a computer science course. Below are two illustrations on such report. The illustration 4.3 shows the distribution between students using Windows and Apple in the "Applied Algorithms" course. As shown, more than two-thirds are windows users.

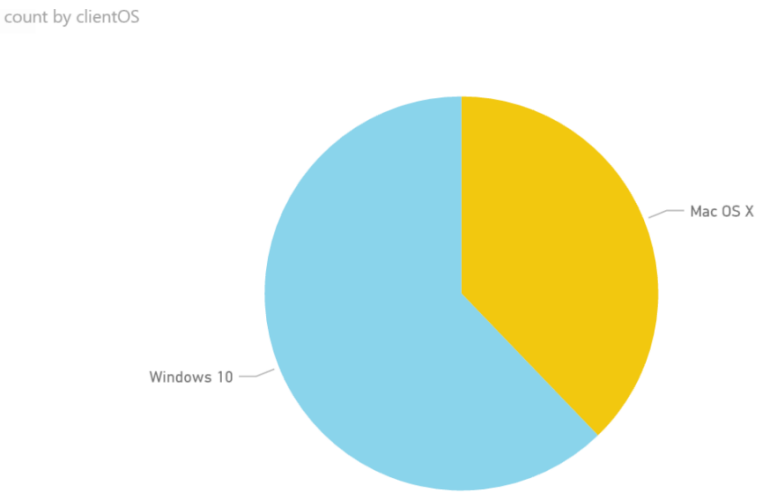


Figure 8: Os usage in Applied Algorithms course

Further, the illustration 4.3 below shows the distribution between students using Windows and Apple in the "Social Media Usage" course. Again, more than two thirds are windows users.

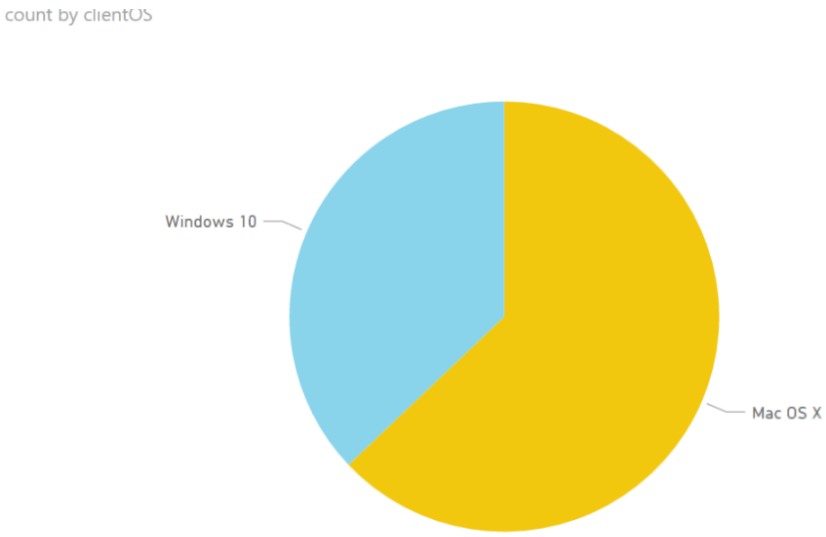


Figure 9: OS usage in Social Media Usage course

Interestingly, one might argue that the computer science students prefer

to use Windows, while the business/creative students prefer to use Apple OS. Based on this information, one might decide which software tools to use across the courses within the technical and creative/business area respectively.

4.4 Implementation

The views are implemented in Spark using Scala. The following figure provides an overview of which data is used within each view.

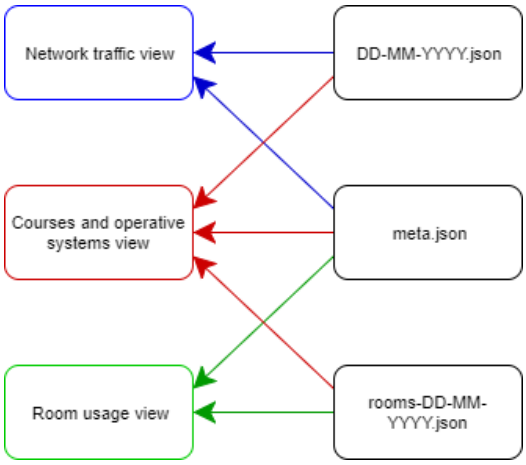


Figure 10: Overview of the views

4.5 Network traffic view

For the network traffic view, the data is grouped by the hour for each device. The count of the number of connections to a device during that hour is then aggregated using an approximation in the form of the function approxCount-Distinct. An approximation is used, which sacrifices some of the precision, but instead causes a large speed-up of the precomputation of the view. The hour of the given data is calculated by taking the day of the year that the given traffic data is recorded, multiplying it by 24 hours and adding the hour of the day to this result. Both the hourly data and the daily data is stored in the same file identified by a granularity, as described earlier.

Furthermore, the result of the above is used to compute the number of connections during a whole day by taking the hour and dividing it by 24 - thereby finding the given day that the data was recorded. This produces a day granularity, upon which the data is grouped and the sum per day is then aggregated. Lastly the result of this is unioned with the rows that resulted from calculating the hour granularity to form a complete view with both hour - and day granularities.

Lastly, this resulting union is joined on the meta data. The join on the meta data is done with the purpose of enriching the data with the location for a given network device. The resulting schema of the result includes device

location, deviceName, granularity, value (for the given granularity) and count of connections.

4.6 Room usage view

The room usage view utilizes the daily rooms dataset. The query groups the dataset by the rooms and their usage date and counts the number of time it has been booked for each grouping. The query produces a new dataset containing the room name, a date and the total bookings by selecting the attributes from the result of the grouping.

4.7 Courses and operating systems view

The courses and operating system view employs all available types of data files provided and creates a final dataset showing the client OS, course name and count of devices. The query takes the flattened WIFI readings dataset and rooms dataset, and performs an inner join with the meta-dataset on the "did" column, between meta and wifi, and "location"/"room" between meta and rooms dataset. Subsequently, the produced dataset is grouped by OS and Course followed by a distinctive counting of the total devices. All together, this will create a column on the new dataset showing the total number of devices based on their OS and Course.

5 Legal issues

Change in the EU Data Protection legal landscape and the question of consent in our system

In considering how a consent procedure is necessary, we need to consider how the concept of consent is reflected in the current legislative landscape, and discuss the specifics of our case against the same. Being that the geographical location of our system is based in Copenhagen, Denmark, and our general interest is the collection and processing of personal data, we need to refer to EU Data Protection legislation.

At the current time (Nov 2017), the EU is preparing to introduce new legislation to the area. Hence, it may be essential to make a note of some of the key differences between current data protection legislation and the forthcoming General Data Protection Regulation (GDPR), which will be enacted into law as of May 2018. Making this brief exploration, we may carefully comprehend the necessary measures that need to be taken into consideration in the case of a data-driven facilities management system at ITU and thus ensure that one comply with the impending regulatory changes.

Under current legislation (according to Directive 95/46/EC on the protection of natural persons in the processing of personal data), it is withheld that the collection and processing of data must have a clear and expressed purpose as decided by the 'data controller'. Also, the 'data subject' should be made aware and have the ability to unambiguously consent to said purpose. However, until now, conditions for said consent has been based mainly on whether or not it was given unambiguously (Directive 95/46/EC, Art. 7a), and relatively few regulations have been enacted in regards to the format upon which the contract is presented to the data subject.

GDPR moves to address the problematics of consent given by involuntary data protection ignorance due to strategically unintelligible Terms and Conditions consisting of long legalese texts (EUGDPR.org, 2017). Recital 32 of the GDPR withholds explicitly that consent needs to be given by a "clear and affirmative act", such that the act of consenting is a "freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data relating to him or her". Hence, the specifics of the data collection and processing should be made transparent, clear and specific (Rec. 39; Art.5(1)(a)).

This, along with the move towards "privacy by design" and "data minimization" as core legal requirements of the GDPR, means that not only should the Terms and Conditions be clear, but the access to personal data should also be limited only to what is absolutely necessary for the provision of a service (Rec.39; Art.5(1)(c);(EUGDPR.org, 2017)).

Explicit freely given consent from the data subject is necessary under the GDPR when certain conditions are met. If data is collected and processed for a specific Provision of a Service (henceforth POS) of which the data subject is engaged, explicit consent remains unnecessary. For example, Google Maps

can legally use your location data if you ask it to locate you. However, if data is collected (and processed) for anything beyond the provision of the primary service, the level of compatibility with the primary service needs to be evaluated to determine the need for consent[1]. An example of a case where an explicit consent procedure would be necessary could be that of previous versions of the Google Photos IOS app, which would collect your location data all the time rather than just when needed to document photo geolocation.

Given these few extrapolations on the current regulatory landscape, we now need to determine how legally compatible our data is with how we intend to use it in our system. Given that our system incorporates network traffic data and Wi-Fi router history as a primary model for the solution, legally we have to perceive it as the foundation of our system. In other words, like the router, the fundamental POS of our system can be established quite simply as being the provision of Internet access. Without Wi-Fi data, we would not be able to solve any of the proposed problems. Thus, all data considered absolutely critical to the provision of Internet access requires no explicit consent procedure. The consent is given by the data subject's mere engagement with the Wi-Fi. Put differently; our proposed system will have to remain directly compatible with the POS of providing Internet access, lest we will need to implement an explicit consent procedure.

One of the functions of our system is to use the data to view network traffic in order to reduce Wi-Fi noise and thereby improve internet connectivity around the university. This can be considered directly compatible with the POS, in the sense that the improving of Wi-Fi speed and connectivity is general maintenance of the same - and thus in line with the interest of the data subjects. However, while this requires no consent, the other functions of our proposed system are less directly compatible with the POS. Processing the Wi-Fi data to help account for room usage and course vs. operating system is incompatible with the primary POS, because it is no longer projecting relevance to the maintenance of Wi-Fi access. Rather, it is using the data picked up by routers to solve tertiary problems, using separate datasets. In this case, implementing our system would require explicit consent for these processing purposes.

> GDPR Art. 7(4): Where the performance of a contract, including the provision of a service, is made conditional on consent to the processing of data that is not necessary for the performance of that contract, this is likely to call into question the extent to which consent can be considered to be freely given.
>
> GDPR Recital 43: "the performance of a contract, including the provision of a service, is dependent on the consent, despite such consent not being necessary for such performance."

Furthermore, qua GDPR Art. 7(4) and Recital 43 which withholds that POS should not be contingent on consenting to process of data which is unessential to the POS. Also, we would be required to implement a way that Wi-Fi users can access and view their data, and opt out without being sanctioned (i.e., lose

[1] According to 00569/13/EN "Opinion on purpose limitation" the purpose limitation principle states that data collected for one purpose should not be used for another(Party, 2013).

Wi-Fi access or speed). Doing so would establish a system in which the consent to collect and process data for our communicated purposes would be freely given and legal under the GDPR.

To take a challenge our idea of primary and secondary purpose in evaluation the provision of service, we can introduce the concept of data-services as a public good. (Nissenbaum and Solon, 2014) argues that privacy for public good can be viewed as contextual integrity. In this way, the institutions we are participating in can (in much the same way Google assumes consent for the use of geolocation in Google Maps), assume consent for the development of services that objectively improves working conditions for the people involved. The question then becomes, how and when do we limit what can be considered within the provision of service, and how we ensure that the data being utilized does not compromise or go against personal interest and/or privacy procedures outlined in the GDPR.

5.1 Ethical considerations

A general question that may be asked in terms of ethics is whether a system such as this serves the general good of all involved parties, or whether the benefit is hierarchically skewed. In the case of Wi-Fi improvement, the benefit is arguably evenly distributed between the actors. However, in the case of the management of the physical facilities and knowledge of student OS preferences, one could consider it as a form of surveillance, which primarily benefits the organization. The data is indirectly able to identify a person, as it maintains the ability to cross-database reference OS, cid, did, location and time. Furthermore, should a dataset be added that maps the aforementioned data to something like their usernames? Thus, a more direct surveillance of individual users could be established, which could reflect negatively on the system. In this way, establishing a way to further pseudonymize or anonymize the data (and keeping it historical rather than real-time) may be key to reduce ethical problems and dissuade potential attempts at identifying individuals. However, as Lane et al. also points out: "Even when individuals are not 'identifiable', they may still be 'reachable' (Lane et al., 2014, p. 45). Thus, doing so may still not reflect the desire of the (consenting) data subjects, as they are still potentially being reached through the way in which the findings of the processed data is acted upon.

Furthermore, on a grander scale, the controversy still remains whether the legitimate purpose of the system will be abused to go against the will of the data subjects. i.e., will night-guards be sent to constantly check if people are sleeping at ITU during nights where Wi-Fi activity are registered? Or, will lecturers willfully use incompatible software such as Microsoft software in classes filled with Mac users? Arguably these scenarios are unlikely, but the possibility of being unfair to the data subjects is certainly still enabled as they are made reachable through their data, which proves a critical problem to take into consideration.

6 Impact analysis

6.1 The winners and losers of data interpretation

In this section, we will cover some of the possible implications that the implementation of our suggested three solutions can have, both positively and negatively, for both ITU as an institution but also the users whose data and privacy might be at stake. Preceding our actual impact analysis, we will briefly cover the aspects of data interpretation and the way we, as individuals, interpret, read and use data since the sense-making process of third-party data is significant to the discovery of patterns, errors and capturing value from it. This will be done mainly in accordance with Peter Tolmie's study and Emma Garnett's research on legibility in personal data and making data accountable.

Accounting for how we make sense of data is important because we might be looking at data at face-value and think that things work in certain ways. This is also the case with the gathered WiFi data from ITU, and we thus need to acknowledge the limitations of our knowledge in regards to exactly what we can tell from this data, how we make sense of it, and how we turn it into the core empirical foundation for our three solutions. When we interpret data, we use a lot of social understanding to attempt to account for deviations in data (Tolmie et al., 2016), which in our case is one of the biggest hurdles since we are not familiar with the students' own accounts of their WiFi-data. What classifies as good data is according to Garnett a question of being able to "understand and materialize error as a form of 'other' to that being studied" (Garnett, 2016, p. 3). We have discovered null values and empty strings in for example operating systems, but since we have not extracted the data ourselves, errors and/or deviances have been difficult to identify and classify.

The analysis of the business value for ITU and the ethical implications for students of ITU from our selected use cases should however be somewhat valid enough, since we have developed some local knowledge, or 'domain knowledge', of the institution and what it means to be a student there. The knowledge is however situated to the student aspect of our proposed solutions, and we can only try to predict how our suggestions can benefit ITU as an institution.

Network traffic load: The most obvious benefit of optimization of the network, is that of making it a better experience for the users, i.e., mainly the students. It would especially benefit groups working together where the load is heavy on web-based group work platforms, such as having several and/or large files and data sets open in Google Drive. For ITU the benefits also seem clear, since optimizing the use of the network would perhaps only have to be done every once in awhile and not as a continuous process, thus eliminating the number of man hours needed to provide this service to the students.

Use of rooms: Getting insight into how much and often the rooms are used would create a mutual benefit, seeing as ITU administration would be able to see which rooms would need to be cleaned and optimized (maybe in terms of tech), and which would not if they had not been used. This would also benefit students during exam periods or other holiday seasons where the school is closed

and unstaffed. This way ITU could schedule cleaning in certain rooms/areas of the school for the many students who are still using the school as a place to write exams.

Operating systems: Our goal by using information about operating systems is to potentially be able to structure courses and plan tool-and-software use based on what would benefit most students. We can see exactly which operating systems are used during which courses, and that would be efficient for both students and professors, seeing as time spent on troubleshooting and finding multiple alternatives potentially could be minimized. However, it could also lead to a discussion of exclusion of certain students, who might not feel they are being catered to by their professor.

Ethics-wise, we can ask whether these personal data based solutions mainly benefit the user or the corporation, or if it is an equal split, and also whether using this data potentially can end up harming the user. Under GDPR, personal data has a much more broad definition, and now specifically includes operating system/platform-level, browser-level and application-level identifiers, and the gathered WiFi data is therefore considered personal data. Mishandling or unnecessary retention of this personal data could lead to very expensive fines for ITU, and it would therefore be important to make sure that there is a continuous upkeep in regards to what data is being gathered and what is actually being used purposefully. In that case, the use cases we have presented might not be of sufficient business value for ITU as an institution, seeing as they are more or few optimizations of pre-existing services. The resources for upkeep and personal data handling might therefore not present a trade-off that is beneficial enough.

6.2 Other approaches to data based optimization

Instead of using the WiFi data as a basis for these suggestions for optimization, an alternative solution could be to initiate a much greater ethnographic optimization project within ITU. Here, it could be to gather fully anonymous survey data and/or conduct interviews with students and staff, and through this try to identify emerging patterns in regards to improvements and changes within the institution. This would also be a way to tap into Tolmie's idea of how to best make sense of data in order to develop services through domain knowledge: "It will be necessary to actively involve the people in making the data accountable to third parties to deliver services that work in the local context" (Tolmie et al., 2016, p. 11). This way ITU could allow students to account for their data, allowing the making of correct interpretations, all while not compromising the students' data.

7 Conclusion

The focus of this project was to develop services to provide possible value to the IT University of Copenhagen, and more specifically the IT department and administration. Three views were developed as part of the definition of such services, in particular a view containing information on the network traffic on devices within ITU, a view complete with data on room usage and a view presenting data on the distribution of operative systems for clients within a given course.

The definition of the above-mentioned views gave rise to the discussion of the legal issues related to the intended use of the views. In particular, it was discussed how the use of data represented by the network traffic view is compatible with the Provision of a Service, and therefore no consent procedure is needed. However, the room usage and OS view is not related to the primary Provision of a Service, and the services relying on these views therefore need to collect explicit consent from the data subjects.

Lastly, it was discussed that the prototyped services present ITU with a opportunity for optimizing already offered services. However, the question remains whether the resources necessitated by the upkeep and handling of personal data such as the development of a consent procedure outweighs this benefit.

References

EUGDPR.org (2017). Key changes with the general data protection regulation. http://www.eugdpr.org/the-regulation.html. (Accessed on 11/03/2017).

Garnett, E. (2016). Developing a feeling for error: Practices of monitoring and modelling air pollution data. *Big Data & Society*, 3(2):2053951716658061.

Lane, J., Stodden, V., Bender, S., and Nissenbaum, H. (2014). *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press.

Marz, N. and Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co.

Nissenbaum, H. and Solon, B. (2014). Big data's end run around anonymity and consent. In *Privacy, big data, and the public good: frameworks for engagement*. Cambridge University Press, Cambridge.

Party (2013). Article 29 data protection working party, opinion 03/2013 on purpose limitation, 00569/13/en, wp 203. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf. (Accessed on 11/03/2017).

Strandburg, K. J., Barocas, S., Nissenbaum, H., Acquisti, A., Ohm, P., Stodden, V., Koonin, S. E., Holland, M. J., Goerge, R. M., Elias, P., et al. (2014). Privacy, big data, and the public good: frameworks for engagement.

Tolmie, P., Crabtree, A., Rodden, T., Colley, J., and Luger, E. (2016). "this has to be the cats": Personal data legibility in networked sensing systems. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 491–502. ACM.

8 Log: Concluding Reflection on the Process

Starting this project, one of our main challenges was determining exactly which services or optimizations we could provide for ITU. Since we are not allowed to do speak to the IT department or do requirements requisition, we can only speculate on which solutions would actually benefit both ITU and its students. Furthermore, the technical aspect of the assignment presented a very steep learning curve due to the introduction of new technologies such as Spark, HDFS and Scala. Consequently, the main focus of the project was learning to use the presented technologies, troubleshooting and creating work-arounds in order to have a viable working solution.

An example is how the large amount of data limited what could be computed for the batch views. Due to insufficient memory, the operations performed on in-memory data had to be simplified and limited. The technical process was thus stalled numerous times. Likewise, establishing a mutual language between critical and technical students was essential to achieve a common understanding of how certain technical data processes may be construed legally and ethically. This was not an easy task, seeing as how our respective skill sets differ starkly, and the fundamental knowledge required to understand certain concepts and ideas had to be simplified and communicated to establish a common knowledge base.

Big Data Management: Assignment 3

Jens Rimmen Bruus
jerb@itu.dk

Sofia Anni Sarauw
ssar@itu.dk

Nicoline Scheel
nisch@itu.dk

Daniel Hansen
daro@itu.dk

Thor Valentin Olesen
tvao@itu.dk

Dennis Thinh Tan Nguyen
dttn@itu.dk

December 19, 2017

Contents

1 Introduction and Goals **1**

2 Architectural Alternatives **3**
2.1 Data Storage Layer Options 3
2.2 Data Processing Layer Options (Batch) 5

3 Dataset **7**
3.1 The Data 7
3.2 Data cleaning 7

4 Batch layer **9**
4.1 Traffic view 9
4.2 Speed view 10
4.3 People view 10

5 Query layer **11**
5.1 People Density 11
5.2 Traffic Density 12
5.3 Actual Speed 13

6 Sales pitch **14**
6.1 Definition of Stakeholders 14
6.2 Pitch: Our Proposals 14
6.3 Critical Reflection on our Proposals 15

References **16**

7 Log: Concluding Reflection on the Process **17**

1 Introduction and Goals

This application takes into account the movement of pedestrians and cars. It aids in giving the user the ability to efficiently navigate traffic, by creating a map-overview of the speed of cars, as well as car/pedestrian density in any given area.

Thus, the goal for this project will be to find a way to use traffic data to improve resident mobility, while at the same time providing value to the city's stakeholders (Scenario 2). We will be utilizing pedestrian and vehicle movement data to establish a platform, which visualizes historical traffic data, by producing interactive maps from which users can learn about traffic patterns and modulate their activities accordingly.

Making this data accessible and readable to residents should ideally improve traffic flow by alleviating tailback tendencies in trouble-zones through heightened resident-awareness of the same. Likewise, the utilization of pedestrian movement data will be used to visualize the movement of clusters of people over time. This will be marketed as a "ZenWalk" function, where pedestrian users can gain knowledge of where they can walk in peace, without being disturbed.

While these goals, in and of themselves, maintain a potential direct influence on traffic flow, it will also have secondary benefits: Widespread adaption and modulation of personal movement based on the data, will ideally allow a significant decrease in pollution (noise, radiation, environmental, etc.), due to the increased awareness of traffic flow patterns.

Our research question will thus be:

How can we build an application that enhances citizen mobility through traffic awareness, and how will this benefit the stakeholders?

Technically-speaking, the application will include the creation of three map-based visualizations **1) pedestrian density**, **2) vehicle density** and **3) speed overview**. These will be further explicated later.

A critical problem that needs to be addressed in terms of the data on vehicle density and speed, is how it enables an overview of states of traffic in practice. For one thing, in order to create a functional unbiased overview and avert rerouting traffic unnecessarily, we would need to know how different speed limits and impediments (such as lights) are dispersed throughout the city.

This would allow us to dynamically average the speed data according to a specific area viewed by the user. For example, data on traffic movement on main arteries could potentially be shown as constantly dense and slow if it is averaged solely according to a cross-city average: It is not slow to drive 35km/h if that's the historical average speed on the road. It is, however, if the historical average is 70km/h.

Without the ability to normalize the data according to average traffic density and more local speed-data, this could potentially pose a challenge in creating an effective traffic overview. Given access to data such as this could perhaps even help evaluate the efficiency of speed limits in certain areas, and improve city infrastructure. In short, to do this it would be valuable to have more complex data, describing speed differences on disparate roads, in order to create a dataset that defines speed averages according to local speed limits.

If these tweaks are not implemented in the final version of the application, our visualizations will presuppose a relatively intimate level of knowledge of local traffic, i.e. that the user knows what speed is normal in certain areas. Implementation of an algorithm to guide the user from A to B, would also increase the user experience and make sure that residents are given the most efficient journey possible, based on the data available.

From a privacy perspective, it would be important to take into consideration how this data on speed could be used to target drivers. For example, if the average speed of a road in a 50km/h zone is 55km/h, would this be used to patrol the street for speed offenders, or would it be grounds for re-evaluating the speed limit?

In the case of "ZenWalk", we also need to take into consideration that the app may potentially be used for illicit activities, such as helping criminals marking areas for mugging/drug dealing and the likes. This should be evaluated on a trial basis. These points will be further explicated later.

2 Architectural Alternatives

The vehicle and pedestrian data is made available in both XML and CSV, of which CSV has been chosen to be stored into HDFS (Hadoop Distributed File System) and processed with Apache Spark as the computation framework. In this regard, one may consider the pros and cons of using two different systems to store and manipulate the data. Namely, the following sections consider the trade-offs between choosing different systems for the storage layer, where master data (i.e. CSV/XML) is stored and the batch layer where data is processed.

2.1 Data Storage Layer Options

The storage layer is where the data resides once it has been gathered from its sources (i.e. CSV/XML). For this purpose, HDFS has been used to store the data. On top of that, a range of alternatives may be considered to store, and categorize the data instead of HDFS. As opposed to the traditional relational database approach, only non-relational strategies are considered, as they fit well with big data that is characterized by its variety, unstructuredness and rapid growth (see (Gerasimou, 2016)). For this purpose, a NoSQL (non-relational) solution may be considered, where data may either be stored in a key-value format, document format, column format or graph format as indicated below:

NoSQL database types (see (Kumar, 2016))

- Key-Value Store (e.g. HBase): Data is stored as keys and values in a hash table
- Document Store (e.g. MongoDB): Data is stored in a document made up of tagged elements
- Column Store (e.g. Cassandra/HBase): Data is stored in blocks that each contain data from only one column
- Graph-based (e.g. Neo4j): Data is represented as a graph composed of edges and nodes

For the sake of brevity, the chosen solution (HDFS) will be compared with one NoSQL solution only, namely a column-based store. It should be noted that NoSQL solutions often work on top of the Hadoop Distributed File System. This is the case of the column-based store "HBase" that works well in conjunction with the Hadoop framework. HBase is a non-relational distributed column-store, in which data is stored in rows and columns as shown below:

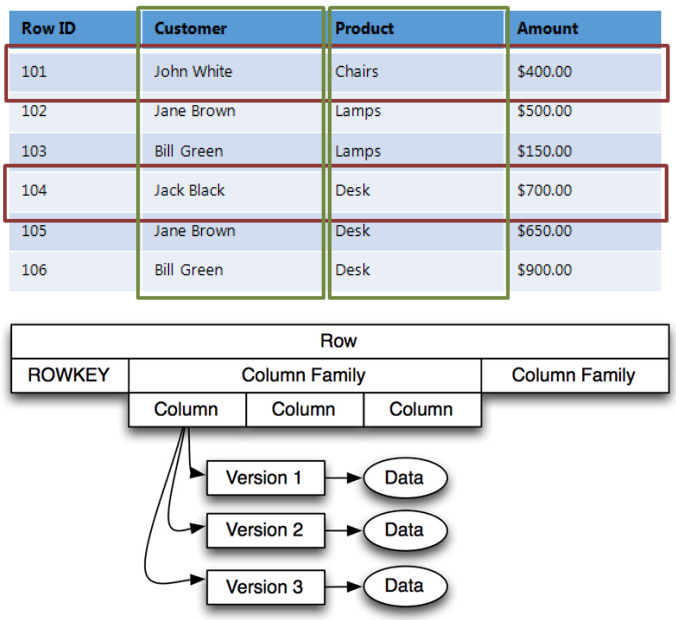


Figure 1: `https://www.netwoven.com/wp-content/uploads/2013/10/hbase-12.png`

As a benefit, HBase enables random access and allows for fast random reads and writes that are otherwise not handled by on HDFS alone. Otherwise, one may potentially risk brute-force reading the entire file system (with e.g. Hadoop MapReduce or Spark). By comparison, both HDFS and HBase are great for managing large and rapidly growing data sets. Namely because they both handle a variety of data, scale well horizontally and shard (i.e. partition) data. In other words, they fit well with big data that is unstructured and the scalable distributed storage help leverage parallel computing. However, NoSQL is primarily intended for real-time access.

Thus, tools like HBase and Cassandra both use a DFS (i.e. HDFS) to make the data storage layer scalable. In addition, they enforce a set of possibly desired properties like random read-write access. Importantly, by doing this they allow for fast read-write, which is important for real time data handling. However, for a batch processing layer, it may be sufficient to use the Vanilla Hadoop framework with MapReduce on HDFS. In our case, no speed layer with real-time processing is used and Hadoop MapReduce has been replaced with Spark because it accommodates fast in-memory processing in the batch layer. For further investigation, one could also compare the features of other NoSQL types, e.g. a key-value store database like 'Cassandra' or 'BigQuery', where data is stored as key-values in a hash table. Most importantly, all these storage options each impose trade offs and merits to be considered in the process of selecting a big data ecosystem.

2.2 Data Processing Layer Options (Batch)

Most famously, the Hadoop framework has commonly been used with MapReduce for data processing. Arguably, it was the first processing framework for batch processing. Specifically, it does this by reading in data from the HDFS filesystem, partitioning the data into chunks, distributing it among nodes, applying computations on the nodes with the MapReduce paradigm and writing the result back to HDFS. Consequently, Hadoop can be used for scalable distributed storage and parallel processing of big data using MapReduce. Nontheless, it is not a suitable choice for real-time processing.

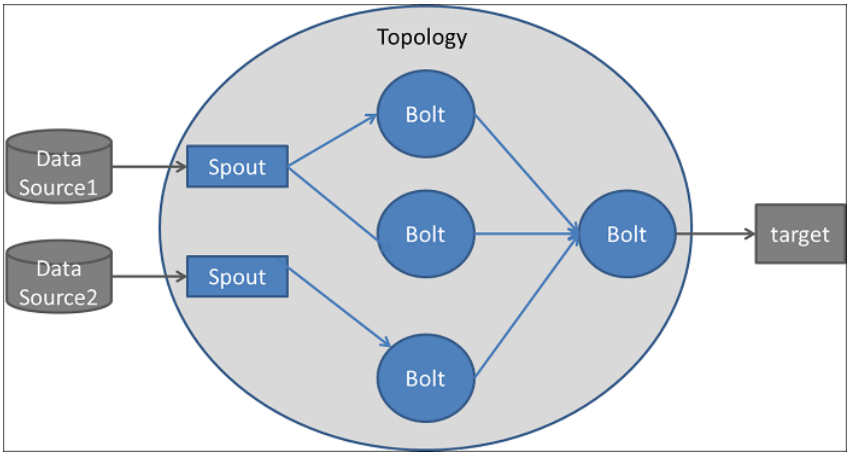
Alternatively, one may use stream processing (e.g. Spark Streaming, Flink or Kafka), where data is continuously fed into the processing engine in real time. Either case, batch processing is often used when dealing with large amounts of data, where you do not need real-time analytics, and when it is more important to process large volumes of information than it is to get fast analytics results. On the other hand, stream processing is key to leverage analytics in real time, where data streams are used to feed data into analytics tools for near-instant results using e.g. Spark Streaming. Thus, one may wish to complement the batch processing option with streaming to obtain faster results and react more quickly to problems or opportunities from a business perspective.

As mentioned already, a stream processing approach has not been considered in this project. However, Spark was chosen for data processing, as it leverages in-memory computing. In general, when choosing a tool for data processing, one must consider different systems and their merits. Namely, one may consider the proficiency of the system user, the performance requirements and the data format to be processed. Firstly, it is important to determine how technical the person trying to analyze the data is. For example, if the person only knows SQL, "Hive" may be a good data processing tool, as it allows you to use a SQL-like query interface. On the other hand, the person may be proficient in a programming language like Java or Scala, which makes Spark a good choice.

Secondly, one has to consider how fast data needs to be processed. Tools like Hive, Pig and Spark all use batch oriented processing, which means data processing may take a long time. As already mentioned, this may not be sufficient if real-time data insight is crucial.

Thirdly, the format of the data and its structure also have an impact on the tools used to both store it and process it. Specifically, data may be structured like in a CSV file or come unstructured in the form of a log. If the data is well structured already, it may not require any cleaning and can be analyzed in e.g. a Hive table directly. However, it may not be structured and require some data cleaning, which leaves options like Pig and Spark to be considered. Finally, one may consider some of these alternatives to Hadoop and Spark in more detail. For this purpose, a brief comparison with a streaming system (e.g. Storm, Flink or Kafka) is made for batch processing. Such tools help move from the Lambda to the Kappa architecture, where batch processing is completely removed and data is fed through a streaming system quickly (see (Repository, 2017)).

In general, a stream processing tool compute over data as it enters the system, as opposed to the batch paradigm where operations are applied on an entire dataset. Further, it is optimized for functional processing on immutable data without side effects to avoid race conditions and enable parallel processing. In this regard, Apache Storm is a stream processing tool that strives to achieve low-latency and is ideal for real-time processing. It does so by arranging DAGs (i.e. Directed Acyclic Graphs) in a framework of so-called "topologies" that describe transformations to be applied on data coming into the system. Each topology is composed of **streams**, **spouts** and **bolts** (Tiwari (2017)):



As shown above, streams represent data arriving at the system. Further, spouts represent sources of data streams (e.g. API or queue). Finally, bolts represent a processing action (i.e. transformation) that consumes a stream of data, applies an operation to them and outputs the result as another stream.

Compared to other tools, both Flink and Spark support batch and stream processing, whereas Storm only supports stream processing. However, Storm offers at-least-once processing guarantees, meaning each data message is processed at least once. For this purpose, a **Trident** abstraction is used to guarantee items are processed exactly once. Further, Storm with Trident enables you to use micro-batches instead of pure stream processing. In other words, micro batching is when the data stream is processed by dividing it into small batches to achieve some of the performance advantage of batch processing, without increasing latency for each task completion too much.

This may be beneficial in systems where the amount of incoming tasks vary. Further, "batching" may help increase the throughput of the system by not executing each task separately but grouping them into bigger batches. In this regard, micro batching is a variant that attempts to strike a better compromise between latency and throughput. Specifically, it does this through shorter "batch cycles", meaning it waits a shorter time interval to batch tasks before processing them. All together, the tools employ different techniques to processing, fault tolerancy, processing guarantees, state (i.e. stateful operations or stateless streams) and impose different trade offs between latency and throughput.

3 Dataset

3.1 The Data

The data provided for this project is limited to the following attributes:

- id for unique identification
- x and y longitude and latitude of the current position
- angle describing the current angle of the vehicle
- speed as the current speed
- edge/lane as a unique id of the current road

The information is limited historical position data, similar to the data stored by companies like Google in the real world. The value of this information alone is limited if one cannot contextualize it. Namely, e.g. Google gathers similar data but they have many different ways to join the data, to provide multiple purposes with the same kind of data. This increases the value of the data significantly by having multiple alternatives in terms of data analysis. One tool that Google utilizes for consumers is to warn them about traffic tailbacks (e.g. to and from work), thereby creating value for them when gathering the data. For this, Google uses their own map to plot in the data and create more value from the data. This data is also used for marketing by looking at pedestrian locations and gathering information about e.g. which stores you visit and thereby your interests. This is very valuable for companies to target their advertisement, which provides a completely different value to another target group. In our project, to visualize the given data in a meaningful matter, a service was needed to plot the data onto a map. As we are not provided with much information about the city, which we are creating batch views for, the usefulness of the data is very limited. Especially in comparison to what a company like Google could accomplish with the same data, if they existed in the world we are simulating.

3.2 Data cleaning

The vehicle data from Sumo includes data, which is 'teleporting', meaning that the vehicles are taken off the road and at a later time placed back on the road again further down, where there is room for the vehicle again.

Various methods to clean this data has been considered. One solution could be to find the distance between the start of the teleportation and the end of the teleportation, and thereby create the missing datapoints in the timesteps that were skipped to convert the teleportation into a normal movement.

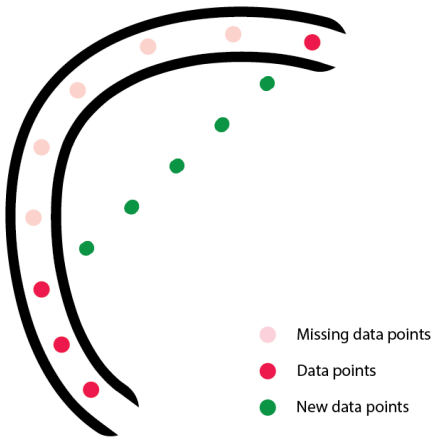


Figure 2: Road Illustration

According to the Sumo documentation (Sumo, 2017), the position of the vehicle could be incorrect if the road isn't a straight line as shown in the Figure 2 above. To solve this issue, using OSM (i.e. Open street maps) data could be a solution. The data from OSM could be matched on the lanes so we could ensure that the generated data is on the road and not as shown above. This would however take processing power, given that the OSM file for Copenhagen is almost 200MB of memory. Thus, this approach only works depending on the usage of the data and the size of the map. Further, one might employ a testing approach to find the probability of cases like the one displayed in Figure 2 and see if this actually constitutes a problem in reality.

For the views that have been created in this project, cleaning of the data does not serve any real purpose. The data is joined on the lanes to count distinct vehicles (ids) over time. Given that we only have information from before and after the teleportation, it wouldn't make a difference in the result (count), if we fabricated these missing data points. Finally, if we didn't use the advanced version using OSM, the cleaning could cause errors in the X and Y coordinates used for creating the visualizations.

4 Batch layer

This section describes three batch views that have been defined in the project.

4.1 Traffic view

The traffic view utilizes the mobility data to present processed data on the usage of lanes in particular time intervals. This data may be used to navigate users around high density traffic, as it can be used to visualize high density traffic for a given time interval.

The schema of the view is the following:

$$TrafficView(lane : String, tenminutetimestep : Integer,$$
$$midX : Double, midY : Double, vehicleCount : Integer)$$

Within the schema, the attribute 'tenminutetimestep' signifies which ten-minute gap the number of vehicles (vehicleCount) is summed for. This means that a 'tenminutetimestep' of 0 corresponds to the interval from timestep 0 to timestep 600, as this is equivalent to the number of timesteps over 10 minutes. This design means that the volume of the traffic view becomes smaller, than if a row existed for every low level timestep.

In this way, the query layer may query the traffic view for larger time intervals, e.g. it may query for traffic data for a whole hour, without having to aggregate every row contained in the unprocessed dataset. Assuming the data would later be provided by the city in realtime, this view could answer questions such as 'How many cars have been driving on my way from A to B within the last 10 minutes?'. Answers to questions such as this would provide an idea of which ways to preferably navigate through traffic, in terms of minimizing waiting time.

The attributes 'midX' and 'midY' correspond to the middle x and y coordinates respectively for the given lane within the given ten-minute timestep. These coordinates can be used to visualize the data in e.g. a heatmap or similiar within the query layer.
The view is implemented by adding a column with the ten-minute timestep, which is calculated by dividing the value of the time attribute by 600. The resulting dataset is then grouped by lane and ten-minute step, which is then aggregated by the maximum x and y and minimum x and y and the distinct count of vehicles for such particular lane.

4.2 Speed view

The speed view provides data on vehicles with particular regard to speed. That is, the view includes the same parameters, as in the traffic view, except for vehicleCount. Furthermore, the speed view includes attributes describing the max speed, the minimum speed and the average speed recorded on the lane within the given ten-minute timestep. The schema of the view is as follows:

$$SpeedView(lane : String, tenminutestep : Integer, midX : Double, midY : Double, maxSpeed : Double, minSpeed : Double, avgSpeed : Double)$$

Ultimately, having this view allows us to query the data to investigate average, minimum and maximum speeds recorded in different traffic areas. As a result, this may be used to provide insight and predict traffic speed, by selecting the optimal historical dataset from the big data, using distributed parallel processing. This caters for efficient traffic speed forecasting, based on the assumption of having massive heterogeneous historical data at our disposal. Again, we are unfortunately constrained by the limited access to data and the somewhat small size of our data set.
The speed view is generated by adding a column of ten-minute step in the same manner as for the traffic view. A column of the speed in km/h is added to this dataset by multiplying the speed by 3.6. This is then grouped by lane and ten-minute timestep and aggregated to find the maximum x and y, the middle x and y of a given lane, the maximum-, minimum- and average speed.

4.3 People view

The last view defined in the project, namely the People view, provide data on how many people are present on a given edge within a given ten-minute timestep. Alike the Traffic view, the People view holds middle x- and y-values for a given ten-minute timestep. For a given ten-minute timestep, the number of people (peopleCount) is included, which provides an idea of the human activity for the given edge. The schema of the view follows below:

$$PeopleView(edge : String, tenminutetimestep : Integer, maxX : Double, midX : Double, midY : Double, peopleCount : Integer)$$

This view is used to provide insight to the pedestrian activity in different locations. In this way, the batch view may leverage the use of big historical data to help city pedestrian planning. For example, it may help track the movement of people through busy urban spaces and greatly help transport planners. Further, it may even help city planning when arranging mega events or emergency planning, provided the data is sufficiently real-time. Nonetheless, a speed layer approach should probably be adopted to fully benefit from this.
The people view is also computed by first adding a column of ten-minute timestep in the same manner as the above views. The result is then grouped by edge and the ten-minute timestep respectively, which is then aggregated to find the maximum x and y coordinates, and minimum x and y coordinates as well as the distinct count of people per edge for the given timestep.

5 Query layer

Within the query layer, the three precomputed batch views have been used to create data visualizations. The query layer utilizes the Business Intelligence (BI) tool Microsoft PowerBI to produce visualizations, which can be analyzed to address the research questions of the project.

The following visualizations illustrate the usage of the three views as previously described.

5.1 People Density

The people density visualization provides insight on the density of people in a given area. This visualization addresses several use-cases, such as a need for insight in places where the people density is minimal, or "Zen" spots where one may not become disturbed by crowds of people. Alternatively, one can also identify areas where the people density is high, if one wish to traverse the city with people around them (or use it for city planning). Figure 5.1 illustrates the people density in a park. The size and color denotes how dense a given area is. Thus, a big and dark red circle denotes a large density, whereas a small light gray circle denotes a low density. The PeopleCount range denotes the upper and lower bounds count of people to consider and the per 10/min range denotes the time frame from 0 to N*10 minutes in Figure 5.1 a range from 0 to 60 minutes is chosen.

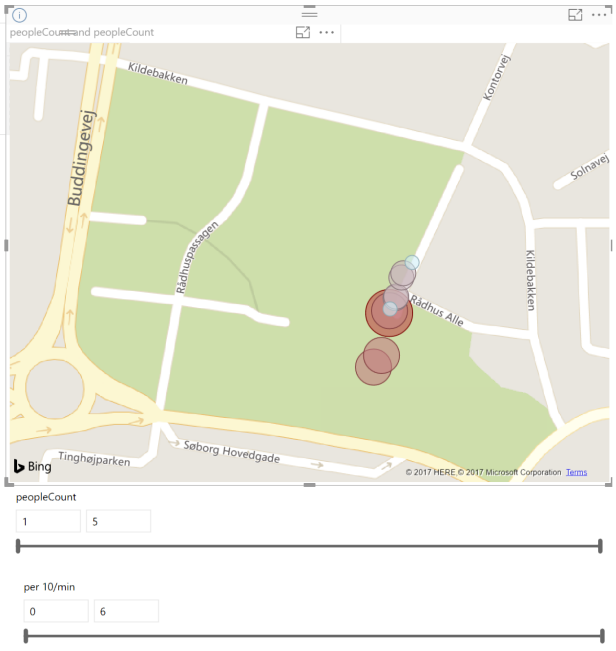


Figure 3: People density in a park

5.2 Traffic Density

The traffic visualization provides insight into how densely trafficked roads are, and at which road positions, traffic bottlenecks and large traffic densities may occur. This visualization addresses the case, where a user wishes to know the fastest way from A to B. Using historical data, e.g. the last 10 minutes of traffic data; the user can gain insight into which roads to avoid in terms of traffic. This, along with the speed, as will be described later, can provide a notion of what will be the fastest route to go from A to B. Figure 5.2 shows the density on different roads. The size and color denotes how trafficked a road is. Thus, a big and dark red circle denotes large traffic density, whereas small red circles denote low density. If there is no circle, then the density is below the desired minimum range. The vehicleCount range denotes the upper and lower bounds count of people to consider and the per 10/min range denotes the time frame from 0 to N*10 minutes in Figure 5.2 a range from 0 to 1000 minutes is chosen.

Figure 4: Traffic density in various roads

5.3 Actual Speed

The Actual Speed visualization represents how fast the traffic is currently flowing. That is, it provides insight into how fluent traffic is currently going. Combined with the traffic visualization, the speed visualization can provide insight about traffic speed and predict the occurrence of traffic tailbacks. Further backed up by knowledge about whether you can expect to drive slow through a given road. Figure 5.3 shows an example of this type of visualization for a given ten-minute timestep. In the example, the range between 50 km/h and 80 km/h is chosen, and every spot, where the average speed lies in this range is marked with a green circle. The avg_speed range denotes the upper and lower bounds for the preferred speed on any given roads.

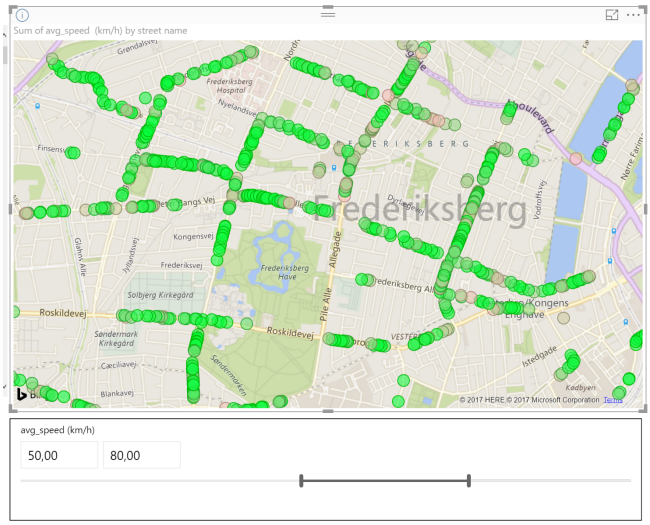


Figure 5: Optimal speed based on given speed range

6 Sales pitch

6.1 Definition of Stakeholders

For this pitch we want to propose our ideas to the government or city council of this city, as they might be interested in being in charge of it and getting insight into the data as well. Seeing as the city has no bikes or public transportation, we imagine that traffic is very congested, and both the city and its residents could benefit from a system that could 1. Help the residents not spend too much unnecessary time in traffic, and 2. Give way to a better environment overall by possibly minimizing pollution from cars.

6.2 Pitch: Our Proposals

In this pitch we will present and cover the following aspects of our three proposals: How our app can help the city move towards smart city status and utilize big data to help solve the fundamental urban problem of too much traffic, how it can improve the lives of the residents and at last, the costs and implications of the app considering both economic and also social factors in a critical reflection of the final product.

What this app is essentially going to do, is utilize crowd mechanisms to not only provide the residents a safer and more pleasurable way to navigate through the city, but also to push the city towards a status of 'smart city' by empowering the citizens to become a part of the urban governance. Smart cities seek to "anticipate and resolve problems proactively" (Zook, 2017, p. 2-3), which is exactly what this app aims to do, since it paves the way for problem solving for more than one actor, and does so by a using a continuous inflow of data. Our app will not only provide the citizens of the city a way to travel faster, but also safer. Seeing as how driving and walking are the only means of transportation, the app could provide a way for drivers to disperse themselves across the city, thus minimizing the risk for accidents. As mentioned earlier, the app can show the least dense areas and how fast they can get to their destination by going using that route instead (see fig. 4). If drivers began to take faster routes instead of being stuck in traffic, this app could essentially help decrease the pollution coming from cars and also provide cleaner air for the pedestrians who are walking in close proximity to the more congested areas. Furthermore, the pedestrians can also use the app to check where the density of pedestrians is minimal (see fig. 3). This way, pedestrians can choose how they want to travel. Using this app, we see a benefit for the residents who might prefer the less crowded places, which could be people who want to go for a stroll with a baby, people who are walking their dog or going for a run outside – but the app is also for people walking home alone at night, who might prefer a route with a bit more people around.

Regarding the cost of the app from a financial perspective, the resources required to do maintenance and upkeep of data are relatively small compared to the value it will create for the residents and for the potential development of the city in the long run. By tracking citizen mobility data and the way the residents use the app, the city can also get a feel for how they will develop

infrastructure in the future. Analyzing the data from the app could provide insight into places where it could be useful to establish bike lanes, which could be economically beneficial for the city. It would essentially require a couple of developers and a compliance officer of some kind. A data analyst will also be necessary if the city was to utilize the data from the app to pursue infrastructural development and city planning.

6.3 Critical Reflection on our Proposals

Ashton et al. (2017) argue: "The first key step in analyzing urban big data and smart cities, is the recognition of the different kinds of crowds implicated in the use of these platforms" (Ashton et al., 2017, p. 3). This means that we, as developers of the app, need to be able to account for its use and by whom it is used. The reason for that is that we are making very broad assumptions and speculations about the patterns of the mobility of drivers and pedestrians, who essentially make up the data foundation for this app, while at the same time being the users of this app. We have no domain knowledge of this city, so pulling on such crowd-heavy data could present a problem in not being able to account for our sense making. Tolmie et al. (2016) argue that social understanding is a key component for reasoning about data, which we as third parties lack in this scenario. This is where the question of citizen empowerment comes back, since there is juxtaposition in the existing 'smart city' rhetoric by always referring to the citizens as being empowered when they are seldom allowed or asked to account for their own data (Tolmie et al., 2016). Zook (2017) discusses how the users of such an app might be a younger and wealthier crowd, which might contribute to a bias in the conclusions we are making about that data and the residents of the city overall, thus giving data that is not representable of the residents' mobility patterns.

Furthermore, assessing the usefulness of this application is also crucial, since we are not providing the residents with visualizations of the traffic based on real time data. The proposed solutions are based off of historical data, which is an issue since the app could never give a fully accurate representation of the current traffic. Also, we cannot as developers actually predict how the solutions will be used. We imagine that residents will use the app for taking a less crowded route to their destination, but when people can see where there are no other people, we are essentially enabling the use of said service to be strategically utilized for criminal activity, such as for example drug dealing. In regards to the use of personal data and user experience, it can also become troublesome that you can scale the range all the way down to show only one pedestrian or driver (see fig. 3). This would have to be changed, so that the minimum amount of pedestrians visible to the user is at a higher number, so that it is not possible to locate a single person from the map.

References

Ashton, P., Weber, R., and Zook, M. (2017). The cloud, the crowd, and the city: How new data practices reconfigure urban governance? Big Data and Society, [online] 1-5. Available at: http://journals.sagepub.com/home/bds.

Gerasimou, V. (2016). Big data and the 3vs: What is the fourth 'v' and what are the implications for not embracing it? https://www.thinkbiganalytics.com/2016/03/29/big-data-3vs-fourth-v-implications-not-embracing/. (Accessed on 28/11/2017).

Kumar, G. (2016). Exploring the different types of nosql databases. https://www.3pillarglobal.com/insights/exploring-the-different-types-of-nosql-databases. (Accessed on 28/11/2017).

Repository, K. (2017). kappa-architecture.com. http://milinda.pathirage.org/kappa-architecture.com/. (Accessed on 28/11/2017).

Sumo (2017). Why vehicles are teleporting. http://sumo.dlr.de/daily/userdoc/Simulation/Why_Vehicles_are_teleporting.html. (Accessed on 19/11/2017).

Tiwari, A. (2017). Apache storm: Architecture. https://dzone.com/articles/apache-storm-architecture. (Accessed on 19/12/2017).

Tolmie, P., Crabtree, A., Rodden, T., Colley, J., and Luger, E. (2016). "this has to be the cats": Personal data legibility in networked sensing systems. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 491–502. ACM.

Zook, M. (2017). Crowd-sourcing the smart city: Using big geosocial media metrics in urban governance. Big Data and Society, [online] 1-13. Available at: http://journals.sagepub.com/home/bds.

7 Log: Concluding Reflection on the Process

One of our main takeaways creating this project was a disconnect between the data and the chosen scenario. Creating a service for the residents to improve their journey is a difficult task when the data is historical data. All of our proposed solution and improvements would be drastically better if the data was real time. We have not really accounted for this in our pitch seeing as it was easier to write (and sell) if we assume that the app would operate with real time data. The data being limited to only driver and pedestrian data, along with the city being described in the project as not being "a very progressive city," provided us with questions of ambiguity. Because we do not know what is meant by a city that is not "very progressive" (as it is described in the project description), we can only speculate on what this city is like.

As explained in our critical reflection of the proposal, we have no domain knowledge of this city. The whole scenario, and in turn, our proposals, become very hypothetical and based on our own interpretation of the description of the city. Therefore we also found that it was difficult to imagine what the residents of such a city would see as improvements to their journey, other than being able to get to their destinations faster. We even had discussions about whether or not Google or Google Maps existed in this city. Google Maps is able to provide real-time traffic information, so in order for our solutions to make sense and provide value for both the city and its residents, we had to pretend that we did not have competition from other actors in the market.