Question 1 (T): Master data set

A)

We store it in a folder on the HDFS. We are storing both the metadata about the access point, as well the readings from each individual data point. The schemas for these files looks as follows:

Readings: AP id, timestamp → connections, isdirty
Metadata: AP id → system, sourceName, timezone ...

Instances of readings schemas could look as follows:

```
21a60668-4e9a-5e9f-b3ca-70990dcd52d6;1475479740000          143;0
21a60668-4e9a-5e9f-b3ca-70990dcd52d6;1475479743600          130;0
…
```

An instance of the Metadata schema could look as follows:

```
21a60668-4e9a-5e9f-b3ca-70990dcd52d6   WiFi;Instrumentation;Europe/Copenhagen;…
E5c91e0d-6515-5b25-adc0-55743c5d371d   WiFi;Instrumentation;Europe/Copenhagen;…
...
```

For the readings, we have chosen this format in order to support a write-once read-many usage pattern. A possible alternative could have been AP id → readings, however this would run into issues when writing into a single file, since it's now necessary to find UUID in question whenever an insertion needs to be made. This performance issue could also have been solved by splitting files up by UUID, but wasn't done due to time constraints.
As for the metadata, simply using the AP id as key is sufficient here to allow us to simply append lines to the file.

B)

The sampling interval of the data is 60 seconds. Yes, the dataset does contain missing data. Sometimes the access points are unresponsive for several updates and sometimes the row is sent twice. This means that the dataset contains missing data, duplicates, or the sampling interval may be lower than 60 seconds – if there was a delay on the message queue.

C)

We store information about the missing records in the master data set because we want to be able to track which access point has the most errors. We chose to mark broken data with a “1” if the data is unusable rather than outright deleting it. We believe that this procedure would make it possible to diagnose bad components if the users want to.

D)

I. Yes. We used Hadoop for this for learning instances, and because there will be a point – if the project were to continue – where the amount of data would be large enough to warrant a HDFS. We have loaded the data onto our Hadoop file server and run a script there to clean every available dataset.
Close timestamps: First, in order to tag all readings that are very close to other readings, we check that they're at least 55 seconds older than the previous reading.
Outliers: Then we check for negative readings, which clearly doesn't make sense for a count of connections to an AP, making them outliers.
Missing data: In order to figure out whether we're missing any instances of data, we invoke a MapReduce job on the master dataset, which finds all the instances where there is a gap larger than two minutes between timespans for a given device, and then counts each missing minute as an missing instance.

II. We found 418.874 instances of missing data out of 40.814.170 reading, or roughly 1% of the total dataset. This seems like a rather reasonable failure rate.

Question 2 (C): Personal data

A.
The data is about users connected to the wireless networks (access points / routers) of ITU at a specific time, which is pushed continuously throughout the day. Data regarding location of the access points is also provided. The flow of information is about wireless chips of items (laptop, tablet or smartphone etc) and its behavioral pattern; location, transmission strength, time zone and user ID which is obtained through the MAC address of the device.

However, the true userID (in this case, MAC address) is anonymous so it should not be able to achieve personal information about the owners of the devices, rather gain meta data about patterns and how people interact with the logistics of ITU environment. Relevant information that is also missing is what kind of device is connecting to the network, we do not have any knowledge of the kind nor the user. Knowledge about the demographics of the users accessing the networks would be useful for further applications of this data, however this type of information is not provided in the metadata.

B.
The data is anonymous so it should be impossible to find any personal information about the individuals accessing the network. However, if the userID of the device stays the same in the logs of information ITU provides then it is possible to observe movement patterns of that user when he interacts within the ITU environment. The pragmatics of obtaining such information could enlighten ITU about where to strengthen their wireless connections, by observing transmission power of connections, then make sure the transmission is powerful enough in rooms that are often attended. They can also shuffle which wireless channels users connect to, so that the load is divided equally on the access point. ITU could also

prioritize internet traffic for different ports during peak hours, such as limiting known torrent ports and prioritizing on TCP port 80 so that people get smoother browsing experiences (which ITU probably does already).

It’s also possible to correlate students that are attending specific classes with specific rooms, such as devices that are connected to access points in the CRBDM room (knowing which class takes place there at a given time) that then also connect often to the cafeteria. There is no attendance requirement for courses when studying at ITU but the data could be used to see how many attend lectures of courses. This information could be used to make sure the room facilitates enough people when the lecture is scheduled. It could also be done to try to make courses better, to provide ITU information about which ones are poorly attended and try to change that.

Hackers have lately been using methods such as eavesdropping/wifi sniffing to obtain personal information from users connected to an open wireless hotspot. This information can be very sensitive to the user, such as credit card information and email passwords. Seeing ITU makes sure the userID are anonymous this is perhaps outside the scope of this assignment but important to keep in mind when connecting to wireless hotspots.

In addition, we have discussed in class the fact that people tend not to care about privacy when they are getting a convenience or gift, however, people still want to have the choice of limiting the access that others have about them. This data can tell us something about the people at ITU, about how in order to get wireless connection they are willing to give up a certain level of privacy (willing to be tracked but are anonymous).

C.

As mentioned in the section above it can be used to correlate usage of rooms and also gather descriptive statistics of ITU facilities. It is possible to estimate how many use the cafeteria, bathrooms etc. The framing of the question is also strange since we cannot see any concrete information about the devices connecting to the access points. Just the pattern of logistics for the userID. If the data is aggregated it might be possible to find out which of the userIDs are teachers, students and guests by seeing which facilities are being used by each individual userID. A teacher for example has access to rooms that students and guests don’t have, and guests probably use a very limited part of the building.

Then there is also the discussion of whether or not this type of use of the data can be considered surveillance. Even when the data is anonymous, it would still be used to infer who is a student, a teacher, staff and guests and there would be knowledge of who accesses which facility. From Westin’s four states of privacy, anonymity serves the purpose of making the user feel like they have control and that they are protecting themselves, however, in lecture it was discussed that even when individuals are willing to give up their privacy to an extent, most people do not agree to the idea of being tracked. These are problems that we might encounter, if we try to use this data to reveal more information about ITU occupants.

Question 3 (C/T): Batch layer

A)

We have decided to consider the following three batch views:

1. Visitors in timespan (hourly) on a specific access point.
2. Network connections established on a given floor in a specific timespan (hourly).
3. Number of invalid readings in a specific timespan (hourly) by access point.

We'll from now on refer to these as the first, second, and third batch view.
For the first batch view we structure the data by having a composed key of the AP identifier and the beginning of the hour that the record describes. The value of the record are gonna be the sum of the connections made for that hour. A slice of the view would look like:

```
E5c91e0d-6515-5b25-adc0-55743c5d371d;1475479740000          1776
E5c91e0d-6515-5b25-adc0-55743c5d371d;1475479796000          1653
E5c91e0d-6515-5b25-adc0-55743c5d371d;1475479799600          1643
…
```

Running the first batch process we ignore records which has been marked as dirty during preprocessing.

The second batch view is very much like the first, but unlike the first we use the floor to group the number of connection by. The second batch view is more directed towards students, whereas the first more is towards system administrators. An example of the second batch view created by the process looks like:

```
1;1475479740000        52032
1;1475479796000        48291
1;1475479799600        47281
…
```

The third batch view can be used for e.g. locating defects in the ITU network. Unlike the other two, the third batch view will ignore records that are not flagged as dirty. The records are stored with a composed key of the AP id and hour. The value is a single number matching the number of failed readings in the hourly timespan. A snippet of the third batch view is as follow:

```
E5c91e0d-6515-5b25-adc0-55743c5d371d;1475479740000          3
E5c91e0d-6515-5b25-adc0-55743c5d371d;1475479796000          0
E5c91e0d-6515-5b25-adc0-55743c5d371d;1475479799600          4
…
```

B)

To generate the three batch views we create three corresponding MapReduce jobs. Like our data cleaning, these are implemented in Java. To generate the batch views we have to run the jobs on the full dataset, which is not to be prefered, but since the data set is of a manageable size this works in the current situation. We wrote bash scripts to automate the process of retrieving the new data set and executing the batch processes to update our batch views. These did however have to be runned manually.

All the batch processes did as much work as possible in the mappers to limit the amount of work the reducer would have combine.

C)

We use Hadoop for the batch processes as well as storing the data. Hadoop allows for an easy way to process data and generate new views as needed. It works well with semi-structured data and is really flexible when it comes to reusing views from earlier computations. E.g. could the second batch view have been calculated using the first batch view and the meta data from the master data set. Even though we didn't go down this road, it surely was an option for quickly extending the data set. With Hadoop It's easy to distribute the workload across multiple nodes. We acknowledge that we in this case don't actually work with large amount of data and we could easily have handled this in memory. This is more of a toy example, but the incremental nature of the data stream speaks towards having an architecture that support it. We could set on a cron job for retrieving the data daily using our bash script and put it in a folder '.temp' that we clean before doing so.

Question 4 (C/T): Log

Development log
Data cleaning:
We argued how we should clean the data – if it was better to delete the corrupted readings or flag them someone. In the end, we agreed that it would be better to flag the data, rather than to delete it, which also helps towards making the data immutable. This could prove to be especially relevant in relation to data cleaning, as this makes it possible to still retain data if we redecide upon our definition of unclean data, avoiding scenarios where we'd have to recover deleted data.

Deployment:
During the development of the batch processes, we quickly realised just how cumbersome our deployment pipeline was. It took several minutes to get from building the project to actually running the processes on the hadoop cluster. In retrospect, a better approach would be to set up an infrastructure that supports faster testing. We ended up using deploy scripts to quickly run the batch processes on , which sped up the deployment phase, but it still had to be run manually. In future projects it's recommended to use unit tests that can be executed on the developer's local machine even for simple projects.

Input types:
We experienced some issues with matching keys correctly for the mappers, as well as input and output for both mappers and reducers. Creating custom key values for our batch jobs proved to be somewhat troublesome, along with the task of matching output-input values between mappers and reducers, as well as between batch jobs, since there isn't any type checking available.

Privacy thoughts:
When discussing the missing parts, we discussed what actually could be concluded if the MAC id or IMEI number was attached to the log file. We discussed that no matter what, the data which could identify an user actually was available, but was deleted, before the logfiles was made publicly available. What we further discussed was that if that's possible here, then it's also possible elsewhere. We tried to relate the case to how Copenhagen Airport uses triangulation to track and measure guests. In our case, we decided to look at what we could see and avoid the missing data in our considerations.

Critical groups thought on developing:
A part of the critical group has programming skills and could therefore easily understand the technical group. It was easy to grasp for everyone in critical group that the technology has some benefits and limitations. The critical group learned the difference between Hadoop and SQL. The technical gap was not that big, it was easy for the technical group to explain stuff to critical group, the problem was not that big.

Working together:
Working together as a big group including 2 subgroups, technical & critical, had some managerial issues regarding finding a timeslot where everyone could be available and join. But we managed to fix that issue by sharing notes and updating those who could not be available as soon as possible, so they were involved and could add their thoughts to the accumulated knowledge.

Question 5 (C/T): Ethics/consent

A)

Trying to infer which classes are popular and which ones are poorly attended. The system would look at the ratio of total number of people signed up for the course and number of connections in the room during lecture. If there are very few connections that would be an incentive for ITU to try to make the course better so that it will be attended by most of the students throughout the semester. In addition you could see which courses are popular and offer bigger classrooms in the following semester.

One idea would also be to look at popular places at specific times. For example, if there is a concentration of users in an area where there is low transmission power for the users, the data could be used to try to prevent that and optimize the internet experience for all users.

As mentioned in 2c. a variety of inferences and applications can be made using this data, however, there is a need to put in place a consent/privacy policy to make sure that users do not feel that they are under surveillance and being watched (even while remaining anonymous).

B)

According to Westin's privacy theory, anonymity allows the user to feel in control of their private information. However, the first time a user connects to the network he should be notified that ITU keeps logs of internet activity of the device and what the purpose of doing so is. If that purpose changes at any time in the future, that consent / disclaimer needs to be updated so that it's users are aware of what is being logged and why.
One possible problem is that users hardly ever read passages about consent and disclaimers, they just blindly accept it. So an idea would be to inform students during the welcoming day at ITU or in some other manner which delivers the message in a better way than a huge text which is prompted when connecting to the network for the first time – which hardly anyone reads at all. Firms such as Fruit Ninja and Fitbit provide data to third party companies. This quite often slips through the eyes of users of those apps because they simply are not informed enough.

It is commonly known that users express concerns about leakage of their data but their actions when reading disclaimers and accepting consent suggest otherwise, this problem is known as the "privacy paradox". The problem of users not giving disclaimers any thought could also stem from "learned helplessness", where users experience a loss of control where they experience the consent of terms to be a scenario which the user can't escape from. Users silently agree although they might be able to reflect later that the terms they accepted might feel like a violation of their privacy.

C)

People value their privacy much, but as the saying goes – "everything has a price". By giving the users an incentive to provide more information than the logs already provide it could be possible for ITU to obtain data which then would be even more valuable than otherwise. However, this could be a slippery slope seeing this could possibly harm individuals if any information were to become public by any chance.
Ideas on how to facilitate the unraveling effect would be to give the students print quota, faster internet speeds when connected to the ITU network and even free cup(s) of coffee at the cafeteria.

As Peppet et al (2011) presented, employees have been fired due to personal data that was being tracked which for example revealed how often they were not present at their work desk at specific times. It can be argued also that some traits are hard to monitor with tracking information which sometimes can lead to inconclusive results. There have also been pragmatic approaches for users such as offering better insurance rate if firms are allowed to track the everyday habits of users, such as number of walking steps and driving behavior.

If the unraveling effect were to pushed at ITU then there would have to be an additional disclaimer for the students to accept these terms. ITU would also have to make sure no information is leaked and possibly invest more in cyber security and promise that no data will be public and open. This is extremely important because a lot of information can be gathered through wireless connection information, such as what browser the user is using, what websites he visits, what internet dependent apps he’s using, how much data the websites are using, time spent online etc. The logs could in theory also hold information of all transmitted text which is not encrypted.