## 

## Question 1 (T)

1. We store it in a folder on the HDFS. We use the schema X. We decided to store it in this way ***because...***
2. The sampling interval of the data is 60 seconds. Sometimes the access points are unresponsive for several updates and sometimes the row is sent twice. This means that the dataset contains both missing data, but also duplicates.
3. We store information about the missing records in the master data set, because we want to be able to track which access point has failed. We choose to ignore duplicate records. We also ignore records which are within 30 seconds of the last.
4. Clean data set
   1. Yes. We used hadoop for this ***because...***
   2. We found ***XX*** instances of missing data.

## Question 2 (C)

1. Basicly the data is about how many users have been using the wifi network available at ITU at a specific time. Data which tells us about where the access-points are located is also given. The access-points has unique id’s. There is not given any data that could identify a unique user and reveal that user's identity. There is not given any data about which kind of devices (mobile, pc, etc.) was connected. From the data, it is quickly available how many users had been connected to a specific access-point at a specific time.

- Devices accesing “Access Points” throughout itu. Behavioral patterns of wireless chips.

-known: timestamps in nano sec, location, transmission strength, time zone, track user using the user id through the devices’ MAC address

-unknown: what kind of device is connecting, information about the user,

1. The data seems to be anonymised, since none client MAC address has been revealed.Although, if someone compare the JSON file with credit card data from the canteen, and the MAC address also gets available, much more could be concluded, and almost everyone could be identified. If possible, then an algorithm/software could follow a person, without revealing the information to us as a readable file.

-moving patterns, distribution of the traffic throughout the university, throughout the day. Areas of no movement, correlate between transmission power and traffic in that area, used to strengthen transmission signal.

-it can reveal correlations of which areas have more people for longer periods of time and which don’t. It can reveal what classes are popular, amount of people in the room.

-

1. The data quickly reveals when there is peak hours at ITU. One could quickly find out which classes/area is being used, and when the usage peaks. If the data is streamed live, the canteen could easily get about 2-5 minutes preparation time. The climate control in the building could be regulated automatically based on the data available from the access-points. But generally, the data reveals the flow of people through the building, and how the students, teachers, etc. uses the data.

## Question 3 (C/T)

### A)

1. Visitors in timespans(minutes, hours, days, months, years) on a single access point.
2. Access points with most visitors in a specific time period, grouped by floor/room.
3. Access points which most often doesn’t respond in it’s heartbeat in a given time period. The batch contains total downtime in a given time period.

### B)

See code.

### C)

Yes. It seems as though the amounts of data we’ll be processing is rather immense, which means that having a significant throughput when analysing / cleaning the data will be important. ***Does this belong in Q1?***

Easy to clean incrementally?

## Question 5 (C/T)

1. Seeing what classes are most popular. Popular places at certain times. If there is a high concentration of users in a low transmission power area, this data could show that and it could be improved.
2. The first time a user connects, you would need consent from the user in order to track their information. After that consent would not be needed if the data is used for another purpose, or if the data needs to collect user descriptive information
3. Giving the user an incentive to provide their device information. For example giving them better access speeds or discount at the canteen or increasing print quota. From the unraveling effect the students will be more likely to see their information as university data and allow it to be taken. The university would have to promise not to make the data public/open.