## Question 1 (T)

1. We store it in a folder on the HDFS. We use the event schema where each device (uuid) has a record. We decided to store it this waybecauseit makes the batch processing more elegant and faster.
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 missing data, duplicates, or the sampling interval may be lower than 60 seconds – if there was a delay on the message queue.
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 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.
4. Clean data set
   1. 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.
   2. We found ***XX*** instances of missing data.

## Question 2 (C)

1. The data is about how many users have been using the wi-fi network available at ITU at a specific time throughout the day. Furthermore, the location of the access points is given in a separate file, making it possible to determine the number of logons for a period of time in a specific room. The access-points have unique IDs, but there is no data that could help identify a unique user and reveal that user's identity, and it is also not possible to determine the kind of device that the users used to log on to the wi-fi.
2. The data seems to be anonymised, since none client MAC address has been revealed. Although, if someone compares 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)

We have decided to consider the following three views:

1. Visitors in timespans (minutes, hours, days, months, years) on a single access point.
2. Access point with most visitors in a specific time period.
3. Access point which most often does not respond in its heartbeat in a given time period. The batch contains total downtime in a given time period.

### B)

Our batch view has been set up via Map-Reduce. First we map the data to the following structure

uuid\_1 [reading\_11]  
 uuid\_1 [reading\_12]

… …

uuid\_1 [reading\_1n]

uuid\_2 [reading\_21]

… …

uuid\_n [reading\_nn]

Then we run a reduce-job to ***XXX***

If the data is not pure (i.e. is a duplicate, timestamps within a few seconds of each other, etc.) it is flagged with a 1. We count the number of impure data-rows.

Otherwise, see code.

### C)

We use Hadoop for the batch run as well since the data is there. It seems as though the amounts of data we will be processing is rather immense in the long run, which means that having a significant throughput when analysing / cleaning the data will be important.

## Question 4 (C/T)

We were surprised that Hadoop did not automatically split the final dataset into smaller pieces since the dataset is more than one gigabyte in storage. We chose not to investigate this further but had the project continued, we would have needed to define a partition for Hadoop.

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 for follow-up purposes it would be better to flag them.

## 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.