Project 02

The project contains both technical and critical questions to be addressed. Technical questions are identified with a T, and critical questions are identified with a C. Questions that require both types of expertise are identified with T/C. This means that the relevant part of your group is responsible for different parts of the assignment. However, you will be judged on the assignment as a whole and not only on your parts so it is a good idea to discuss all parts of the project together. Please note that the whole group will need to understand the dataset well in order to answer the questions below so make sure to schedule a time to discuss what's in the dataset and what kind of information these data represent.

Each group should prepare a single document that answers the project questions. This document should be submitted on LearnIT both in the technical class and in the critical class by **October 28th at 23:55.** Document length is up to you but remember "brevity is the soul of wit".

The dataset for this project consists of building related data that has been collected for the IT University of Copenhagen. The data is JSON formatted and consists of two types of data: (i) time-series data and (ii) metadata. The time-series data contains time-indexed values. Each series has a unique ID (a 128-bit UUID). The metadata contains descriptions for each time-series.

The main data source is from the WiFi infrastructure at ITU. Each access point reports different status data (e.g., transmission power, used WiFi channels, no. of connected clients, Mac addresses of connected clients etc...). Besides the WiFi dataset, the data contains calendar bookings and course base information from ITU (e.g., is a room booked/available).

The data is structured hierarchically: /Instrumentation/ITU contains all ITU related data grouped into floors and rooms. Data on individual WiFi-clients can be found in /Instrumentation/ITU/WiFi-Clients.

For example, the following contains the metadata for the stream of clients associated to an access point at different points in time:


```
{
  "Path": "/ITU/2/AUD32-3A56/AH-D41080/no_clients_device",
  "uuid": "e5c91e0d-6515-5b25-adc0-55743c5d371d",
  "Properties": {
    "StreamType": "numeric",
    "UnitofMeasure": "Clients",
    "UnitofTime": "ms"
  },
```

```
  "Metadata": {
    "Extra": {
      "NumericalID": 85
    },
    "Instrument": {
      "Model": "Access Point"
    },
    "Location": {
      "Building": "IT University of Copenhagen",
      "City": "Copenhagen",
      "Floor": "2",
      "Room": "AUD32-3A56",
      "Street": "Rued Langgaards Vej 7"
    },
    "SourceName": "Instrumentation",
    "System": "WiFi",
    "Timezone": "Europe/Copenhagen"
  }
}
```


The time series data looks as follows:


```
[{"uuid":"e5c91e0d-6515-5b25-adc0-55743c5d371d","Readings":[[14754797
48000,1],[1475479748000,1],[1475479808000,1],[1475479808000,1],[14754
79868000,0],[1475480335000,1],[1475480395000,1],[1475480455000,1],[14
75480515000,2],[1475480575000,2],[1475480635000,2],[1475480695000,2],
[1475480755000,2],[1475480816000,1],[1475480875000,1],[1475480935000,
1],[1475480995000,1],[1475481055000,3],[1475481115000,2],[14754811750
00,3],[1475481235000,4],[1475481295000,4],[1475481355000,4],[14754814
15000,7],[1475481475000,14],[1475482953000,37],[1475482954000,38],[14
75483421000,38],[1475483519000,40],[1475484505000,42],[1475484579000,
44],[1475484639000,45],[1475484699000,45],[1475484759000,45],[1475484
819000,44],[1475484879000,44],[1475484939000,33],[1475484999000,31],[
1475485059000,24],[1475485119000,23],[1475485179000,23],[147548523900
0,23],[1475485299000,22],[1475485359000,23],[1475485359000,23],[14754
88022000,7],[1475488023000,7],[1475488259000,11],[1475488260000,11],[
1475488775000,36],[1475489001000,55],[1475489116000,64],[147548917700
0,66],[1475489237000,66],[1475489297000,67],[1475489357000,65],[14754
89417000,67],[1475489477000,66],[1475489537000,63],[1475489597000,63]
,[1475493357000,61]]}]
```


Each value is assigned a [unix timestamp] in nanoseconds.

Data Files

You can retrieve the time series data and metadata from
http://130.226.142.195/bigdata/project2/

This folder contains a file that contains the metadata (meta.json) and a file for each day of time series data (e.g., 2016-10-04.json). We will provide you incrementally with new time series data for each new day. The new files will be automatically added to above link.
The metadata file is just updated each day.
You can download the files using wget or curl directly to the hadoop server or simply download them with your web browser.

Questions

Question 1 (T): Master Data set

A. How do you store this master data set? Explain your answer.
B. What is the sampling interval of the data? Are there missing data in the dataset?
C. Define a procedure to clean the data set and handle the missing data. Give arguments for your approach.
D. Generate a clean data set.
i. Do you use hadoop for this batch process? Explain your answer.
ii. How many instances of missing data did you find?

Question 2 (C): Personal data

A. What are these data about? What is known/not known about WiFi use in this data set? What is made obvious/visible? What is overlooked?
B. What kinds of stories can these data tell about people at the ITU (what can these data reveal about individuals if anything)?
C. What can these data reveal about all occupants at ITU in general? Can you say anything about things other than the devices connecting to WiFi access points and the locations of these access points in the building?

Question 3 (C/T): Batch layer

A. Define three views that can be used to get insights about this data set.
B. Implement the corresponding batch processes that take the clean data as input.
C. Do you use hadoop to answer Q3B? Explain your answer.

Question 4 (C/T): Log

List the problems/challenges you faced during this project and explain how you tackled them.

Question 5 (C/T): Ethics/consent

A. If you were charged with a problem of coming up with ways to make these data useful to ITU in new ways what might be some options for doing so? Select and describe one potential way you might implement a way to use these data - what kind of system would this be?
B. What would you implement as your consent procedure? Do you even need consent here?
C. What are some of the ethical issues you would need to think about? (Critical students, think about the "unraveling effect" - week 7 lecture).