Project 2

Big Data Management / Fall 2017
Irina Shklovski / Björn Þór Jónsson

Project description

The university wants to implement a big data system but it would like to start small and leverage some of the more easily available data first.

The dataset they have identified consists of building-related data that has been collected for a short while, but is continually added to. 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). A more detailed description is found below.

You are part of a diverse consultancy team that has been asked to develop new services and potential applications for data management and processing for the university based on the particular dataset that is made available to you.

Your goals are to:

1. Determine the most appropriate architecture for such a big data repository, and implement the storage, batch and service layers.
2. Produce suggestions for how these data may be used and to estimate potential benefits given these services.
3. Create a mock-up for the services, identify potential legal and data protection issues that would need to be further investigated and/or addressed.
4. Conduct an ethical analysis—critically consider the winners and losers in this scenario—if your proposals were to be implemented.
5. Suggest potential technical approaches/alternatives to address potential ethical challenges.

Each group should prepare a single document that answers the project questions below. This document should be submitted on Piazza by November 3rd at 23:55, in order to get feedback on it. Document length is up to you, but remember: “brevity is the soul of wit.”

Questions

1: Determine the most appropriate architecture for such a big data repository, and implement the storage, batch and service layers.

A. How do you store this master data set? Explain your answer.
B. Which layers of the Lambda Architecture are required to manage this dataset? Explain your answer.
C. What is the sampling interval of the data? Are there missing data in the dataset? If so, how many instances of missing data did you find?
D. Define a procedure to clean the data set and handle the missing data. Give arguments for your approach.
E. Generate a clean data set. (Do you use Spark for this process? Explain your answer.)

2: Produce suggestions for how these data may be used and to estimate potential benefits given these services.

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?

3: Create a mock-up for the services, identify potential legal and data protection issues that would need to be further investigated and/or addressed.

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. (Do you use Spark for this process? Explain your answer.)
C. What would you implement as your consent procedure? Do you need consent here?
D. What are some of the ethical issues you would need to think about?

4: Conduct an impact analysis:

A. Who are the potential winners and losers given your selected scenario?
B. Are there other approaches that might change the balance?
C. What might be some unintended or secondary effects of your implementation? Why?

5: Address any ethical issues you have uncovered:

A. Are there technical approaches that might mitigate any ethical issues you have identified?

B. Are there policy or design approaches that might help?

6: Log

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

Proposed Work Process

What you can do

Interact with students, faculty and staff in a lightweight fashion: observe, ask questions, but not conduct interviews.

Review university information to understand how the organization functions and potential inefficiencies that could be addressed.

What you can not do!

Do NOT contact the IT department for more details about the dataset. All of the available details are provided here. You can not have any more information.

Meetings

You are encouraged to have a minimum of four meetings about the project:

Meeting 1: Project kick-off. Björn and Irina will explain the project to all groups and you will have a chance to ask questions. You then need to have a kick-off meeting to decide on your approach and set expectations.

Meeting 2: Brainstorming about solutions. You have now developed the beginning of the system and conducted some research about the organization; bring these together.

Meeting 3: Create a mock-up of services, conduct an ethical analysis of these services, generate a report detailing the necessary technical systems that would need to be implemented.

Meeting 4: Propose some solutions to deal with ethical challenges, and finalize the report.

Data Files

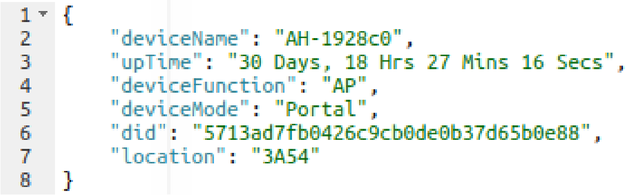
The main data source is from the WiFi infrastructure at ITU. Each access point reports different status data (e.g., transmission power, used WiFi channels, no. of connected clients, Mac addresses of connected clients etc...). Besides the WiFi dataset, the data contains calendar bookings and course base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

The collection consists of three types of data: (i) time-series data, (ii) metadata and (iii) room booking data. The time-series data contains time-indexed values. Each series has a unique device ID (“did”: a 128-bit key). The metadata contains descriptions for each time-series. You can retrieve the time series data, metadata and booking information from: http://130.226.142.195/bigdata/project2/

This folder contains a file that contains the metadata (meta.json), a file for each day of time series data (e.g., 4-10-2017.json) and room bookings (e.g. rooms-2017-10-04.json). We will provide you incrementally with new time series data for each new day. The new files will be automatically added to the above link. The metadata file is just updated each day.

You can download the files using wget or curl directly to the Spark server or simply download them with your web browser.

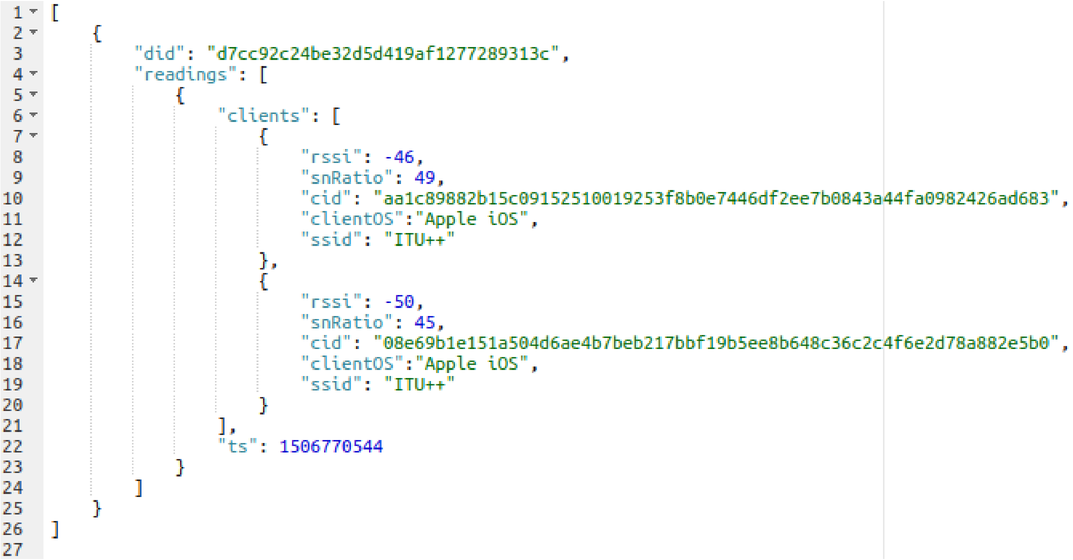
For example, the following contains the metadata for the stream of clients associated to an access point at different points in time:



```
{
    "deviceName": "AH-1928c0",
    "upTime": "30 Days, 18 Hrs 27 Mins 16 Secs",
    "deviceFunction": "AP",
    "deviceMode": "Portal",
    "did": "5713ad7fb0426c9cb0de0b37d65b0e88",
    "location": "3A54"
}
```

did	Device Identifier

The time series data looks as follows:

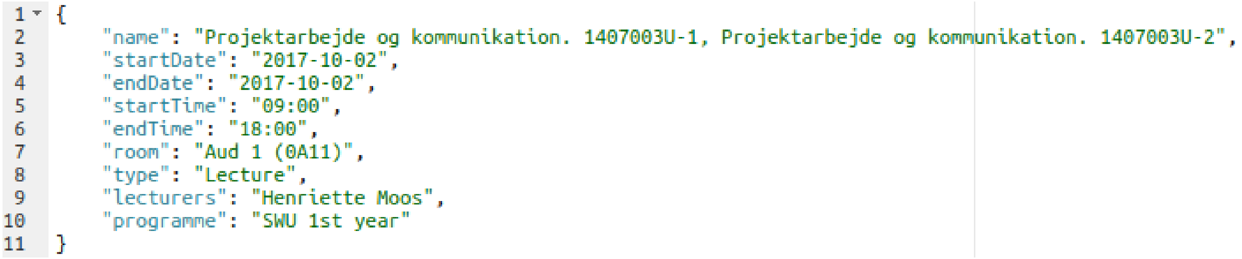


```
[
    {
        "did": "d7cc92c24be32d5d419af1277289313c",
        "readings": [
            {
                "clients": [
                    {
                        "rssi": -46,
                        "snRatio": 49,
                        "cid": "aa1c89882b15c09152510019253f8b0e7446df2ee7b0843a44fa0982426ad683",
                        "clientOS":"Apple iOS",
                        "ssid": "ITU++"
                    },
                    {
                        "rssi": -50,
                        "snRatio": 45,
                        "cid": "08e69b1e151a504d6ae4b7beb217bbf19b5ee8b648c36c2c4f6e2d78a882e5b0",
                        "clientOS":"Apple iOS",
                        "ssid": "ITU++"
                    }
                ],
                "ts": 1506770544
            }
        ]
    }
]
```

did	Device Identifier
rssi	Received signal strength indicator
snRatio	Signal-to-noise ratio
cid	Client Identifier
ts	Timestamp

Each time serie is assigned a unix timestamp in seconds.

The room booking data looks as follows:



```
{
    "name": "Projektarbejde og kommunikation. 1407003U-1, Projektarbejde og kommunikation. 1407003U-2",
    "startDate": "2017-10-02",
    "endDate": "2017-10-02",
    "startTime": "09:00",
    "endTime": "18:00",
    "room": "Aud 1 (0A11)",
    "type": "Lecture",
    "lecturers": "Henriette Moos",
    "programme": "SWU 1st year"
}
```