04/06 initial data analysis

data question

1. data files have more fields than schema:

    *RNC*BA* has 42 fields, schema specifies 40

    *RNC*BS* has 84 fields, schema specifies 77

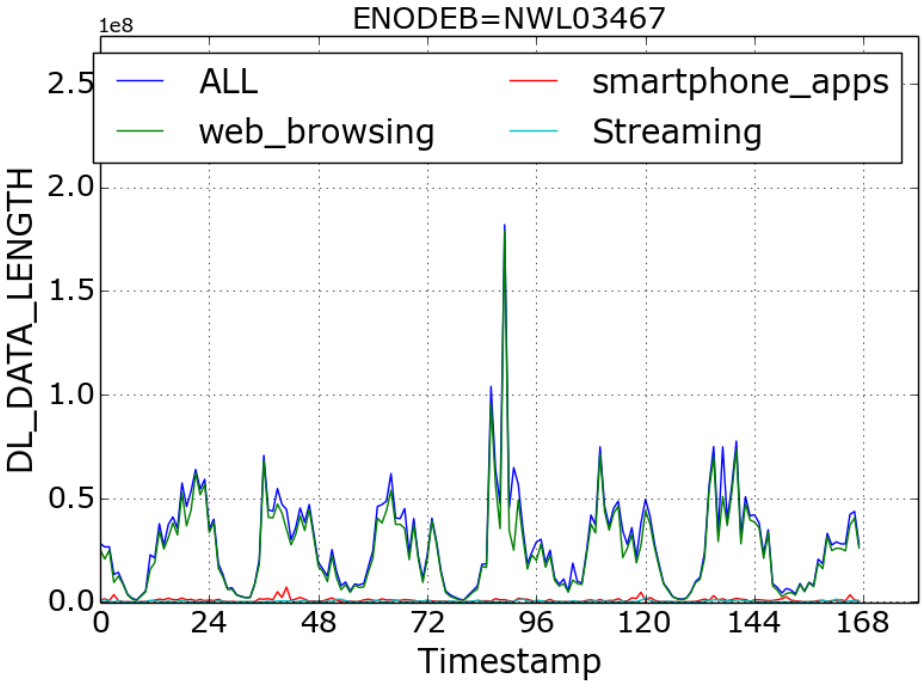
    *RNC*PA* has 41 fields, schema specifies 39

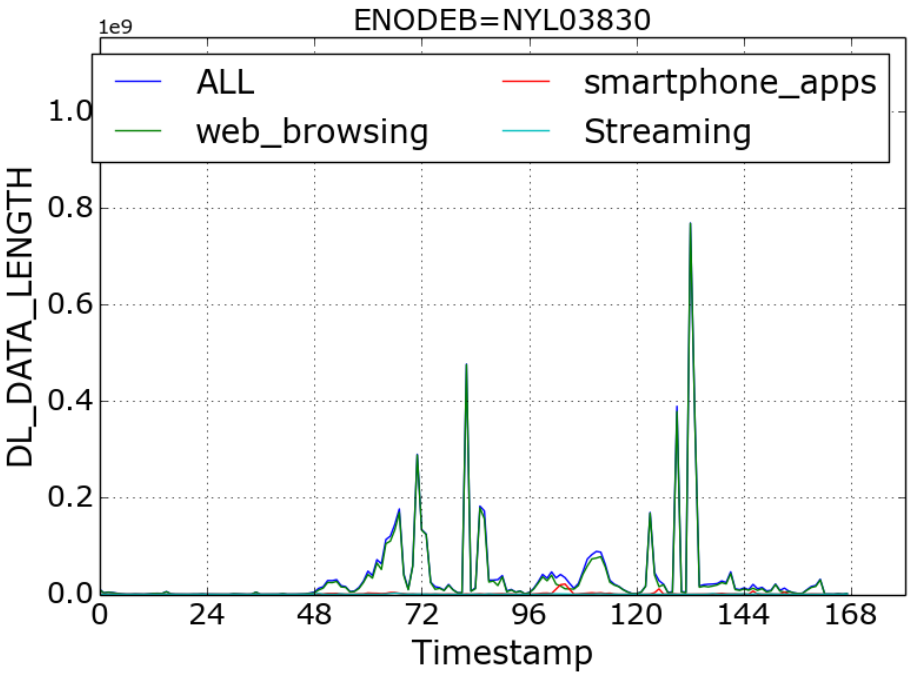
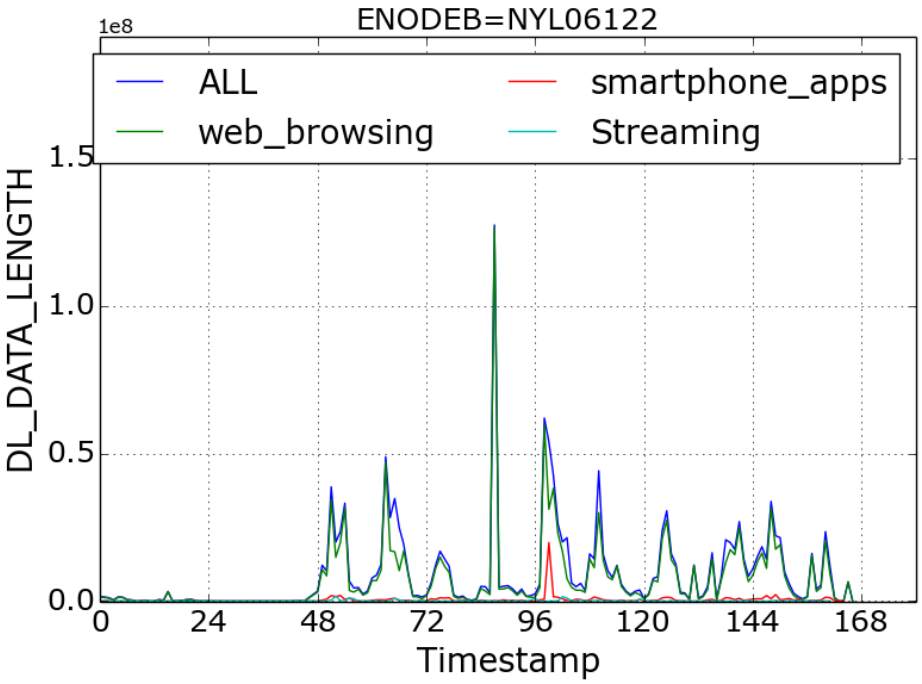
    *LTE*BA* has 42 fields, schema specifies 44

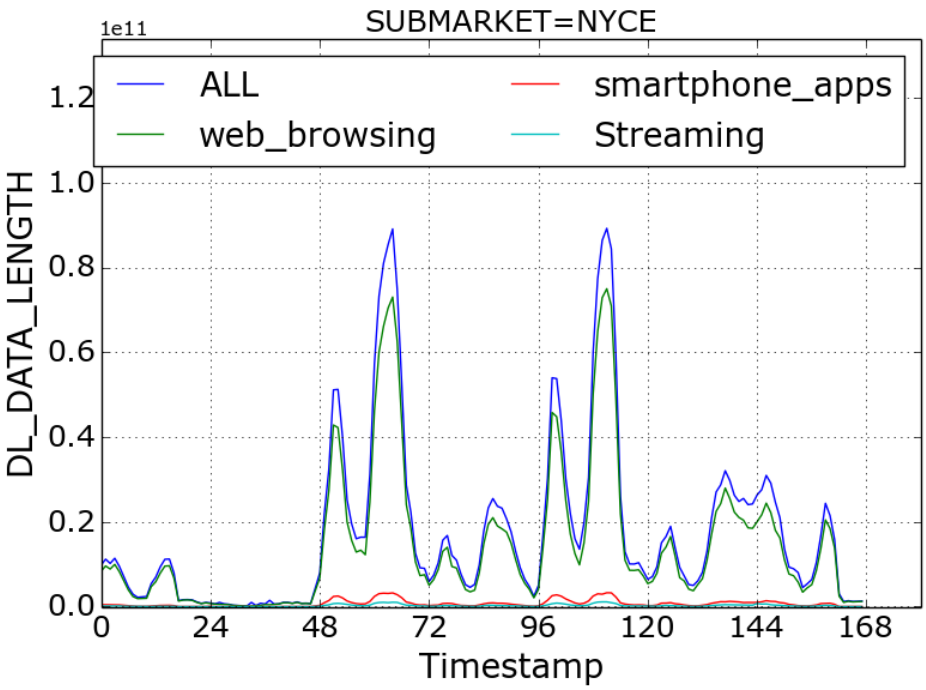
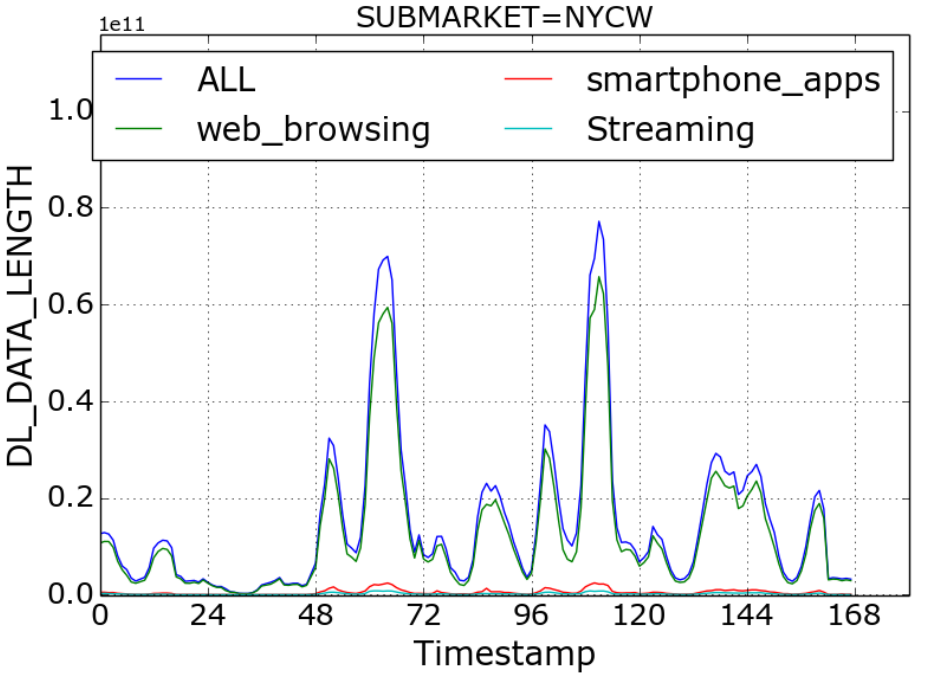
2. is there a robust way to know if the data is ready? LTE data for 2015 is not ready, although there are lots of LTE files there (they are empty). Also, RNC files are not ready for 2015/03 too, although files are there and they are not empty but they are much smaller than 2015/02 RNC data... e.g., 4652 bytes vs. 1002977 bytes)

some initial analysis/plot

Not sure which field I should focus on, I plot some results for DL_DATA_LENGTH, and assume this means the hourly "sent bytes" for each base station (I'm not sure some other fields contain correct data: DL_BYTES are mostly 0, UP_BYTES are either 0 or 1 digit number, mostly 8; DL_PKTS, UL_PKTS are mostly 0, for non-zero values, they have lots of trailing 0s, e.g., 100000, 200000, 500000. **I just figure out probably the out-of-date schema issue can explains this strange UP_BYTES/DL_BYTES problem: 0/8 are some category ID, but not bytes count. We need to get the latest schema to confirm this, and if this is true, then below plots are for UL_DATA_LENGTH instead of DL_DATA_LENGTH. But still, I don't know what's the reason for trailing 0s problem, even if they mean UP_BYTES/DL_BYTES in conjectured new schema**).

Time: 12/01/2014 - 12/07/2014 (0-24 are 24 hours for 12/01, 24-48 are 24 hours for 12/02, etc.) **Base station:** One basestation NWL03467, in NYCNJ market. **Takeaway:** 1. We see time-of-day effects, but not sure why it peaks at Thursday (12/04/2014). 2. Most of data are web-browsing, steaming takes a really small portion (we can't even see its line).	

Setting Same setting, but on a different base station, NYL03830. **Takeaway:** not sure if this plot makes sense...	
Setting Same setting, but a different base stations. **Takeaway:** not sure if this plot makes sense...	

Time: 12/01/2014 - 12/07/2014 **Base station:** Aggregated result for basestations of submarket NYCE (NYCNJ/NYCE). **Takeaway:** 1. We see time-of-day effects, but not sure why it peaks at Wednesday and Friday and doesn't show traffic on Tuesday. 2. Most of data are web-browsing, steaming takes a really small portion	
Time: 12/01/2014 - 12/07/2014 **Base station:** Aggregated result for basestations of submarket NYCW (NYCNJ/NYCW). **Takeaway:** 1. We see time-of-day effects, but not sure why it peaks at Wednesday and Friday and doesn't show traffic on Tuesday. 2. Most of data are web-browsing, steaming takes a really small portion	

discussion

1. How do we quantify "capacity" (What's the definition of "capacity")? This is a radio network question. For example, sometimes the sending rate is 100MB/h, sometimes it's 1MB/h, but this metric doesn't quantify capacity. Even 1MB/h can be fully loaded for a base station, if it's trying *its best* to reach a distant device; similarly, 100MB/h doesn't mean it's fully loaded neither. Actually, in Magus, the model assume base stations always use its maximum capacity: assume if a base station serves device A only, the device gets rate Rmax, then if there are N devices, A gets rate Rmax/N, meaning that each device gets a fair share, but the basestation is always 100% busy. The basestation is idle when it serves 0 devices, otherwise even 1 device takes its maximum "capacity".

2. More queries/experiments? I assume running/plotting queries are simple, it takes some time though. Currently my script takes 3 input parameters, time range, a field name to sum up, and a filter which can specify multiple field names and their values to only compute on qualified rows:

```
python .py [TIME RANGE] [FIELD NAME] [FILTER]
        e.g.: python .py 01-07 DL_BYTES ENODEB=abcde,SERVICE_CATEGORY_ID=30
```

3. How to update? My first preference is Google Doc, but I guess for privacy reason you probably don't like this idea. My second preference then is to write update in Word and generate PDF attachment. What do you think?