Report 1

LI, Ruizhe | 1155076990

Github: https://github.com/rzli6/ML-Storage.git (private)

Target Paper: Predicting Disk Replacement towards Reliable Data Centers

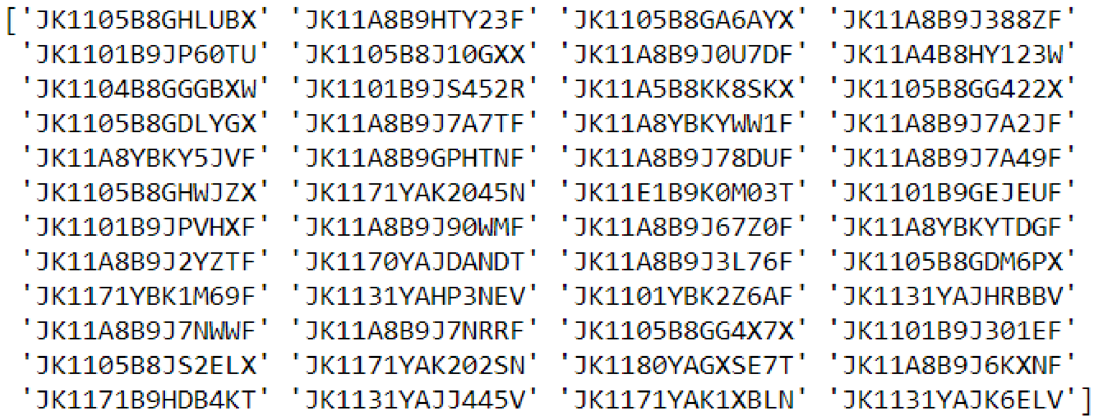
Goal: Realize the method descripted in the paper.

Dataset: 2016 Q1 – 2018 Q1

Sub-goal: Data processing & Attributes Selection

Procedure Details and Outcome:

The first step is to download the data from blackblaze website, and then use *pandas* package to combine and group the data from different dates. To simplify the problem and also for the sake of memory usage, I started with the model HDS722020ALA330 (HitA) with the 12 features described in the paper. I extracted the failed disks during this period.



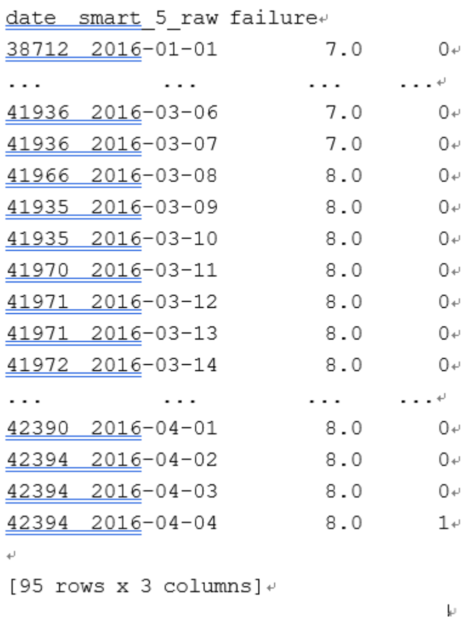
```
['JK1105B8GHLUBX' 'JK11A8B9HTY23F' 'JK1105B8GA6AYX' 'JK11A8B9J388ZF'
 'JK1101B9JP60TU' 'JK1105B8J10GXX' 'JK11A8B9J0U7DF' 'JK11A4B8HY123W'
 'JK1104B8GGGBXW' 'JK1101B9JS452R' 'JK11A5B8KK8SKX' 'JK1105B8GG422X'
 'JK1105B8GDLYGX' 'JK11A8B9J7A7TF' 'JK11A8YBKYWW1F' 'JK11A8B9J7A2JF'
 'JK11A8YBKY5JVF' 'JK11A8B9GPHTNF' 'JK11A8B9J78DUF' 'JK11A8B9J7A49F'
 'JK1105B8GHWJZX' 'JK1171YAK2045N' 'JK11E1B9K0M03T' 'JK1101B9GEJEUF'
 'JK1101B9JPVHXF' 'JK11A8B9J90WMF' 'JK11A8B9J67Z0F' 'JK11A8YBKYTDGF'
 'JK11A8B9J2YZTF' 'JK1170YAJDANDT' 'JK11A8B9J3L76F' 'JK1105B8GDM6PX'
 'JK1171YBK1M69F' 'JK1131YAHP3NEV' 'JK1101YBK2Z6AF' 'JK1131YAJHRBBV'
 'JK11A8B9J7NWWF' 'JK11A8B9J7NRRF' 'JK1105B8GG4X7X' 'JK1101B9J301EF'
 'JK1105B8JS2ELX' 'JK1171YAK202SN' 'JK1180YAGXSE7T' 'JK11A8B9J6KXNF'
 'JK1171B9HDB4KT' 'JK1131YAJJ445V' 'JK1171YAK1XBLN' 'JK1131YAJK6ELV']
```

The failed disks during the 27 months' period. 48 in total.

And in the following steps, I use the JK1105B8GHWJZX with feature smart_5_raw as a tentative example. The attempt has been made to find the changepoint of the smart_5_raw of the abovementioned disk. Normal distribution was used. Though it is extremely likely that the smart_5_raw is not normal distributed, as the true distribution may be very complexed and hard to explore, normal distribution is fitted to simplify the question. The following pictures show that there are 95 observations in total, and on the 67th day, the value changed from 7.0 to 8.0.

In fact, this example is not so representative, because the start date of the disk is 2016-01-01, which means part of its life-time is hidden. And 7 to 8 is not a big change, as latter we can see that this value of some failed disk can reach 150+.

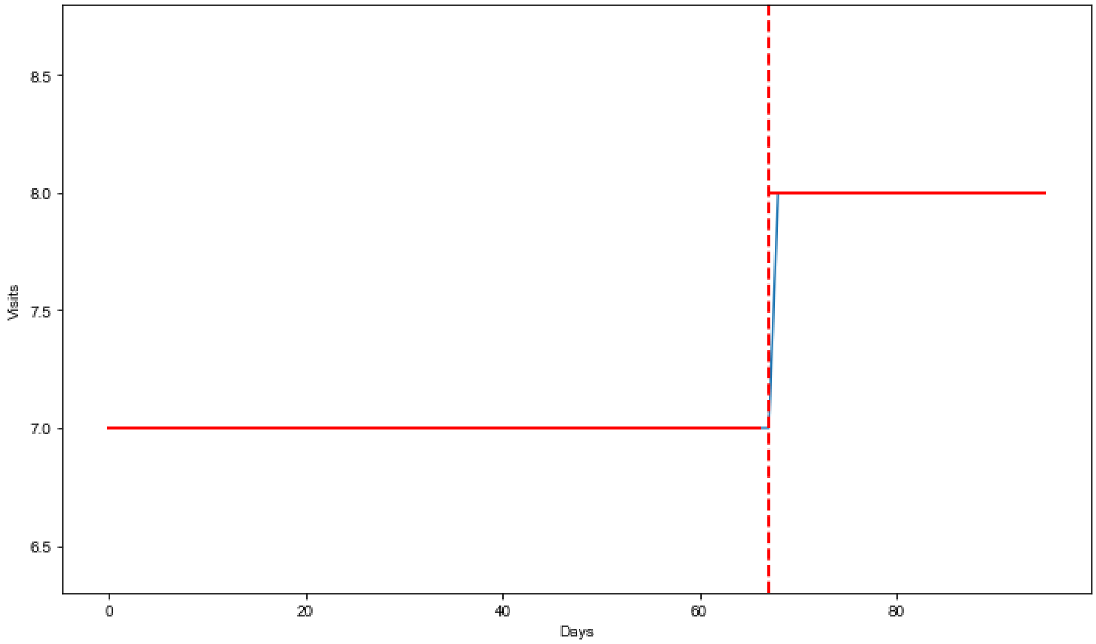
Smart_5_raw: reallocated sectors count



```
date  smart_5_raw failure
38712  2016-01-01        7.0      0
...        ...          ...     ...
41936  2016-03-06        7.0      0
41936  2016-03-07        7.0      0
41966  2016-03-08        8.0      0
41935  2016-03-09        8.0      0
41935  2016-03-10        8.0      0
41970  2016-03-11        8.0      0
41971  2016-03-12        8.0      0
41971  2016-03-13        8.0      0
41972  2016-03-14        8.0      0
...        ...          ...     ...
42390  2016-04-01        8.0      0
42394  2016-04-02        8.0      0
42394  2016-04-03        8.0      0
42394  2016-04-04        8.0      1

[95 rows x 3 columns]
```

Digital representation of changepoint on smart_5_raw



Smart_5_raw of JK1105B8GHWJZX has a change point on the 67th day.

I applied the same code to all the failed disks'. Among 48 failed disks, 21 (43.75%) are related to the smart_5_raw. Compared to the 31% from the paper, the increment in the value is reasonable in the sense that I didn't test whether all the changepoints are permanent. Some of changepoints may be temporary, thus fake.

The median of the days of changepoints before failure of smart_5_raw among all the failed disks is 5, which is very far from 12 described in the paper. However, the mean (11.714) is rather close. I think this could be caused by the limited sample size - the sample size in paper is 2 times more than what I did. I will expand my sample size latter to check the difference.

Smart_5_raw: reallocated sectors count