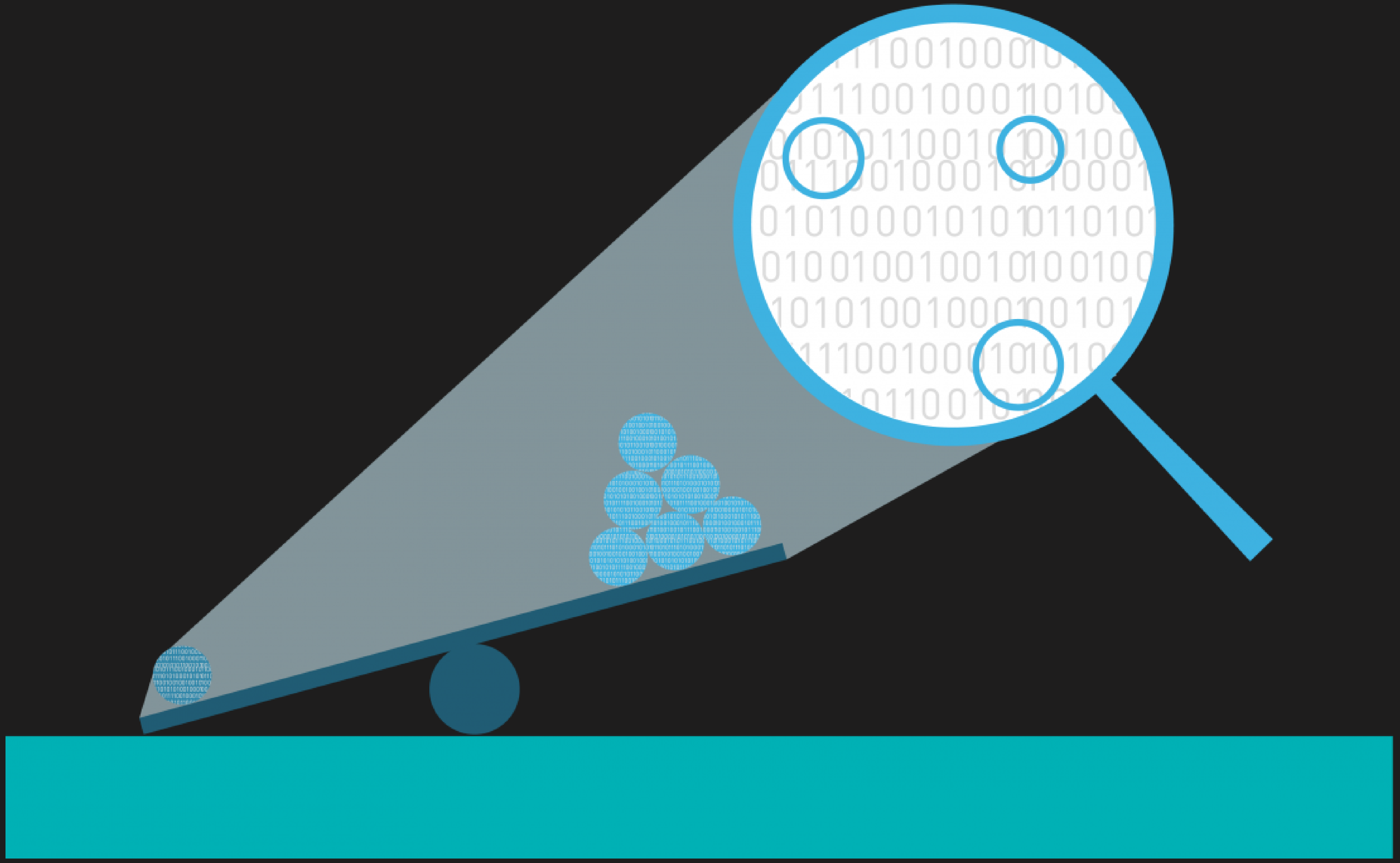
Our Mission
and Goals

Realtime Anomaly Detection with CDN



What we aim to accomplish *by the end of the term*
week 7-13 November

Table of Contents

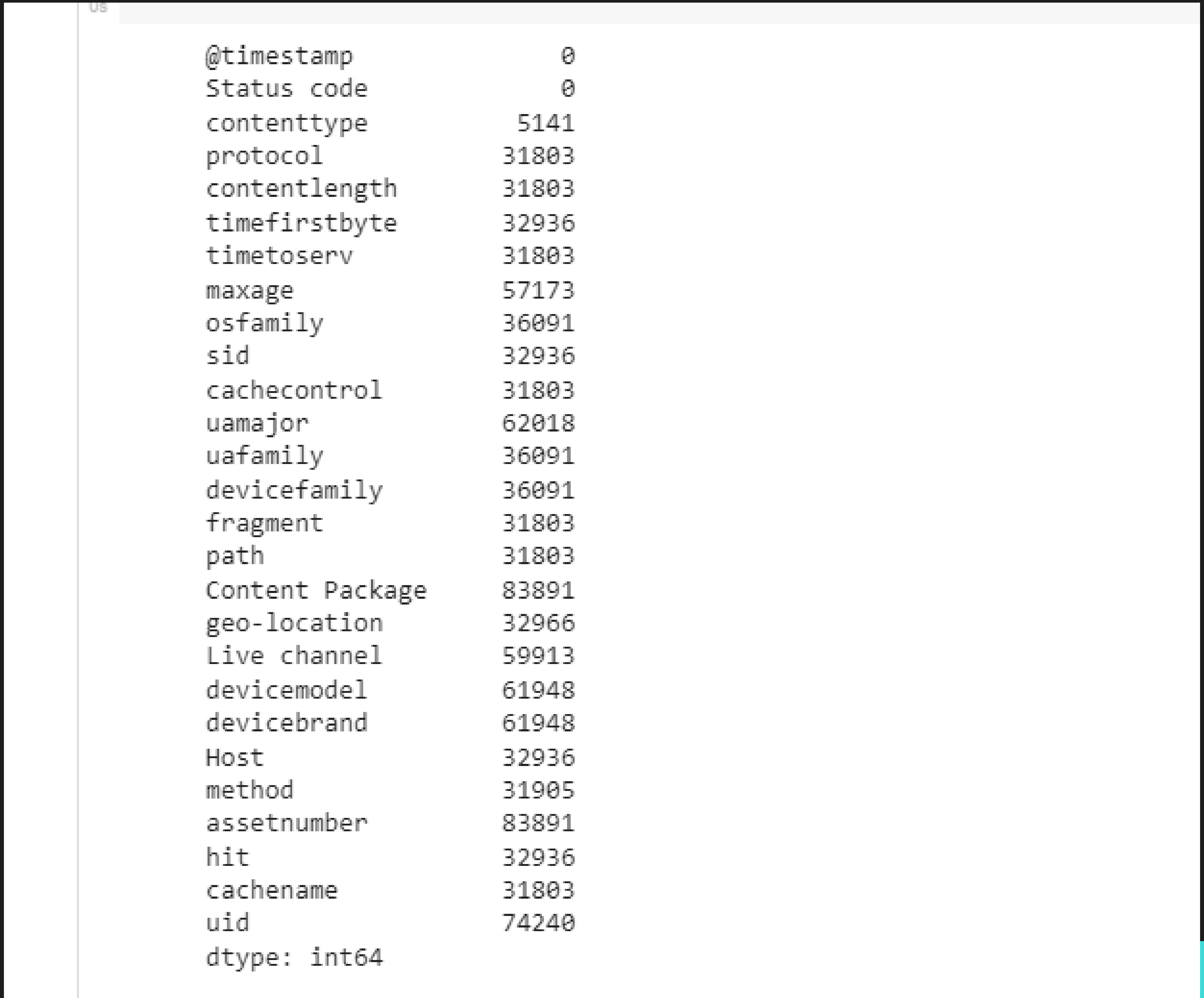
#01 Data Preprocessing

#02 Algorithms

#03 Architecture

#04 Technologies

The general first checking of the null values:



```
@timestamp             0
Status code            0
contenttype         5141
protocol           31803
contentlength      31803
timefirstbyte      32936
timetoserv         31803
maxage             57173
osfamily           36091
sid                32936
cachecontrol       31803
uamajor            62018
uafamily           36091
devicefamily       36091
fragment           31803
path               31803
Content Package    83891
geo-location       32966
Live channel       59913
devicemodel        61948
devicebrand        61948
Host               32936
method             31905
assetnumber        83891
hit                32936
cachename          31803
uid                74240
dtype: int64
```

How did we handle null values and features selection?

We dropped the columns which consist extremely more null values than others.

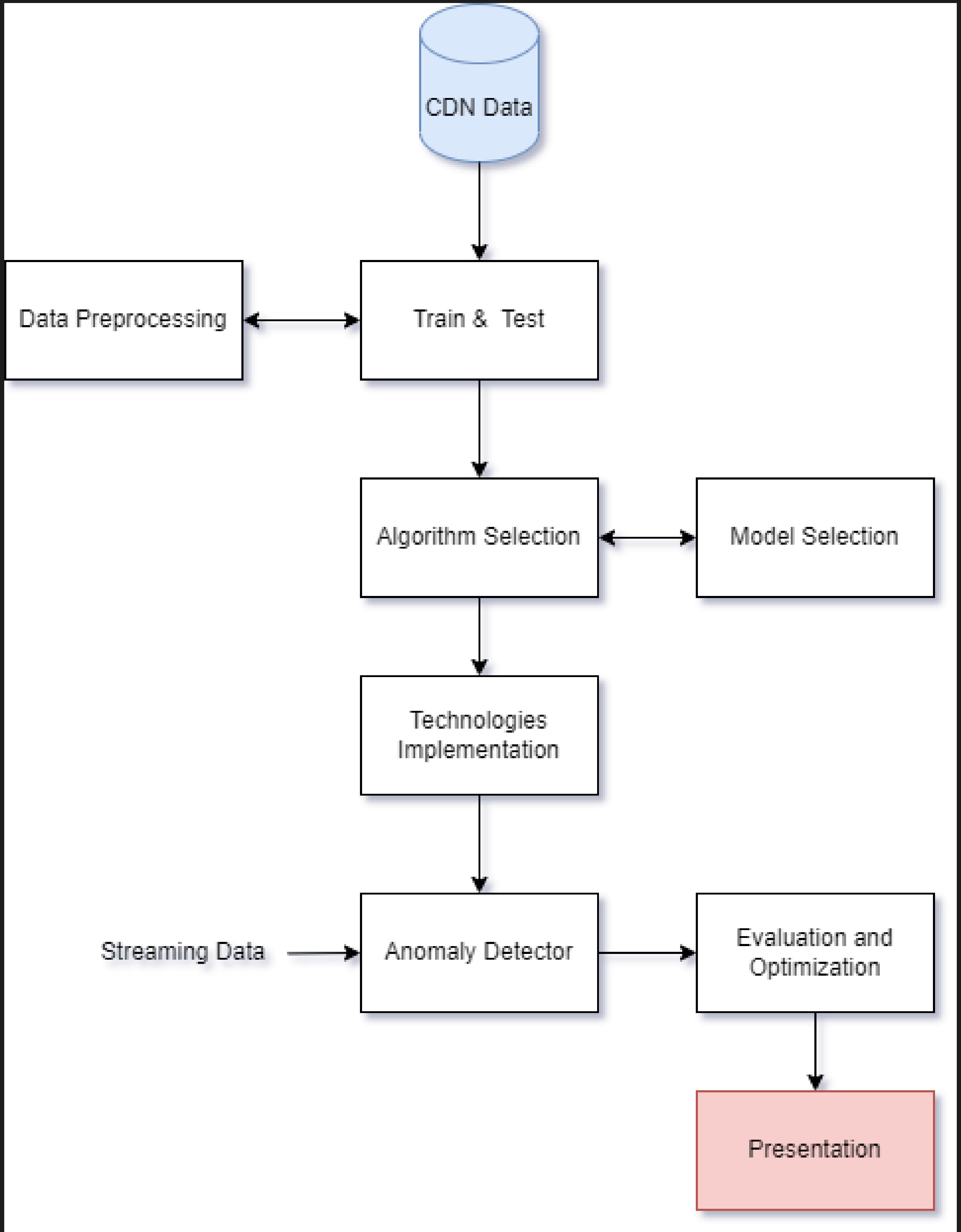
Most features are correlated with each other to some degree but some have very low correlations such as timetoserv , osfamily,and timefirstbyte. We removed them.

For numerical values which had null values, we used and filled with their median.

For categorical values which had null values we used mode to fill.

We looked for low variance features, to be able to remove them. There were some potential features to remove, but as min and max value difference was not that much we did not remove them.

Architecture

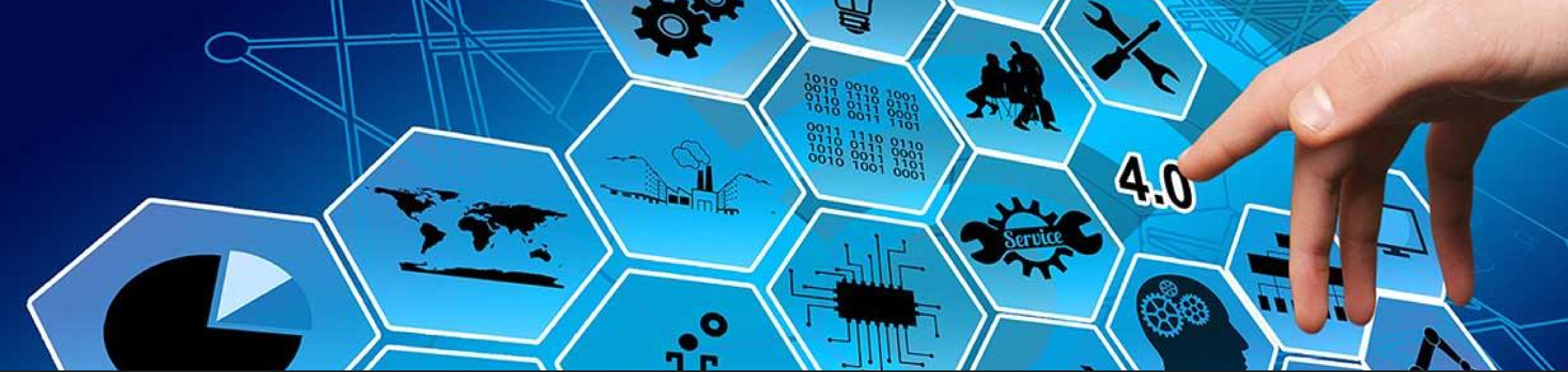


Algorithms

Local Outlier Factor #01

LSTM #02

#03



Technologies

Docker

Kafka

Spark

influx DB

Useful links:

https://www.overleaf.com/1871192315vsxmzygrtxjv

https://colab.research.google.com/drive/13pO40ueV5nMZnT2yzp69UmK39jbgM6d4?usp=sharing