**ADA Project Summary-Report**

Create samples using MIMIC-II ICU data and Icentia 11k ECG data with the help of random sampling and reservoir sampling and find the optimal sample size.

**Taniya Adil(IIT2019054), Alka Trivedi(IIT2019055)**

***VI Semester, B.Tech, Department of Information Technology,***

***Indian Institute of Information Technology Allahabad, Prayagraj, India***

1. **Importance:**

With the advancement in technology we are having a lot of not only static data but dynamic data also. There is a growing need to analyze both the data at rest and data in motion to drive the applications in today’s technical world. Our project can help in streaming the real world dynamic data and in random sampling the data for further analysis.

1. **Objective:**

The objective of our project is to create samples using MIMIC-II ICU data and Icentia 11k ECG data with the help of random sampling and reservoir sampling and find the optimal size.

1. **Design of the study:**

The design of our project is divided into three parts:

* Data Processing: Data processing is done to send them for streaming.
* Data Streaming: After processing the data is converted into datastream and then using kafka and zookeeper, we are streaming the data via producer and consumer.
* Data Sampling: Streamed data is sampled using Random Sampling and Reservoir Sampling.

1. **Introduction:**

Spark Streaming is an API that can be connected with a wide variety of sources including kafka to deliver high scalability, throughput, fault-tolerance, and other benefits for a high-functioning stream processing mechanism. These are some features that benefits processing live data streams and channelizing them accurately.

Spark Sampling is also very important when we have a very large dataset and we only wanted to test or analyse a subset of the data.

This project uses kafka and zookeeper for data preprecessing and data streaming. And for data sampling we have used random sampling and reservoir sampling.

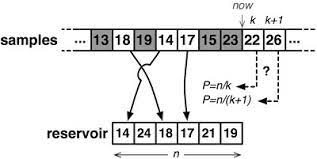
Sampling is a general technique for tackling massive amounts of data. A sample is a subset of data that has approximately the same properties of the original data.

**(i) Random sampling**

Simple random Sampling - Every case of the population has an equal probability of inclusion in sample.

Stratified random Sampling - Stratified sampling is where the population is divided into strata (or subgroups) and a random sample is taken from each subgroup.

**(ii) Reservoir sampling** - It is a family of randomized algorithms for randomly choosing k samples from a list of n items, where n is either a very large or unknown number.



1. **Dataset:**

The dataset used for this project is MIMIC-II ICU admissions.csv dataset for streaming purpose. For the data sampling purpose we are making an input.txt file from the data consumer consumed data.

**The admissions.csv file from MIMIC-II ICU dataset contains following parameters:**

**i. hadm\_id:** This field contains the admitting id of the patient

**ii. subject\_id:** This field contains patient id

**iii. admit\_dt:** This field contains the admit date and time of the patient

**iv. disch\_dt:** This field contains the discharge date and time of the patient

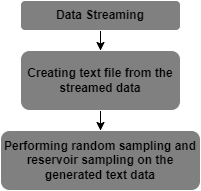
1. **Methodology:**

Before beginning with the methodology of the project there are some pre-requisite associated with it. These prerequisites are listed below:

**Prerequisites:**

* Kafka
* Zookeeper
* Spark
* Python3
* Linux Operating System

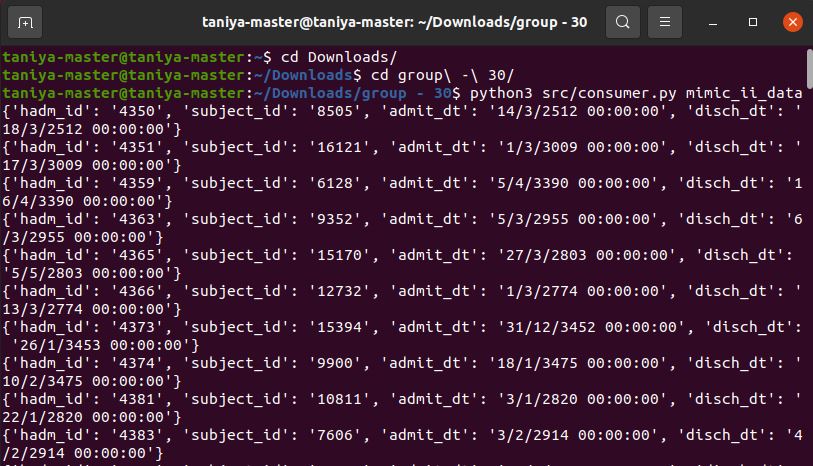
The flow of execution of the project is:

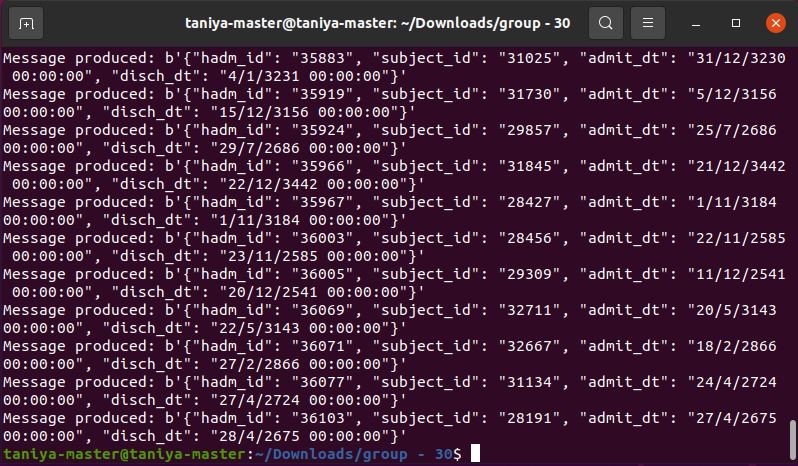


The detailed step of all the processes are listed below:

**Data Streaming:**

* Producer will read the csv file extract its data
* Extracted data will be send to the consumer

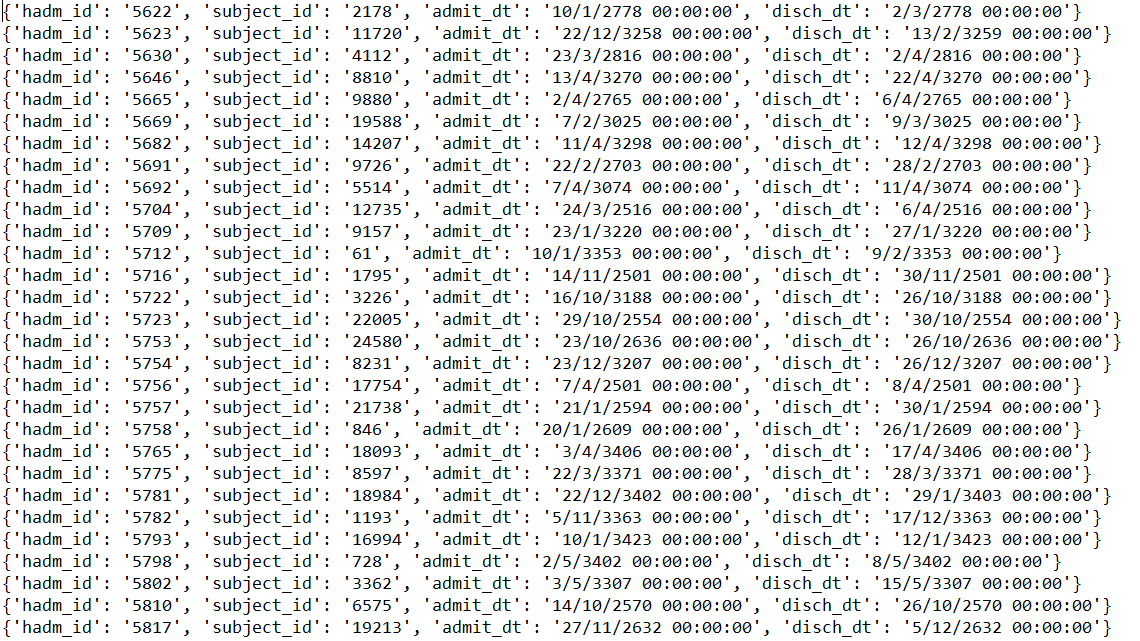




**Creating text file from the streamed data:**

* The data which the consumer will consume will be written on to a text file name ‘input.txt’ which will serve as an input for the sampling

**Generated text file:**

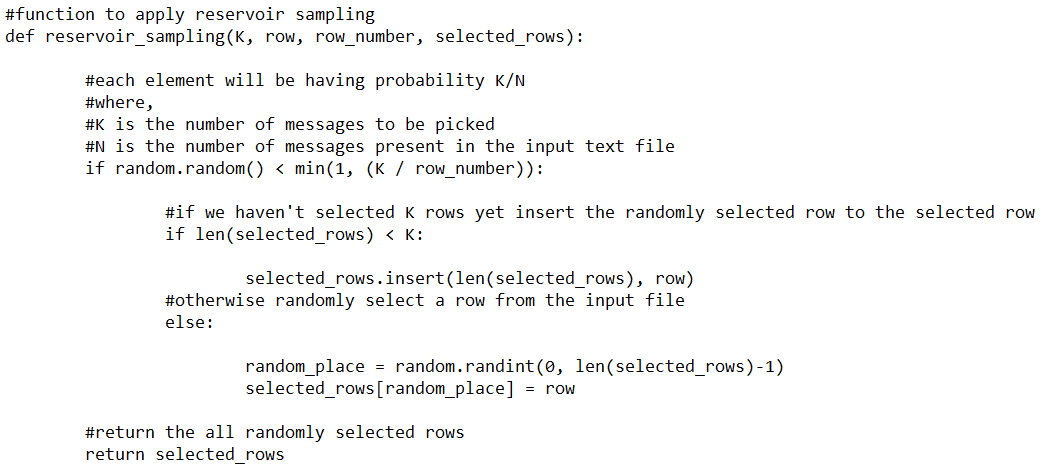


**Performing random sampling:**

* From the imputed text file random number of messages will be sampled. Since, in random sampling all the messages have same probability for sampled so we are taking any random message without replacement.

**Performing reservoir sampling:**

* Here we will be given as input a number ‘K’ which will be the number of messages we wanted to extract from the text file. In reservoir sampling the probability of each message will be K/N where N is the number of messages in the text file. Also, one thing to note is that N will be changing every time as the data is streaming continuously into the text file. That is why in this sampling all messages will not be having same probability to be sampled. In this sampling also we are sampling messages without replacement.

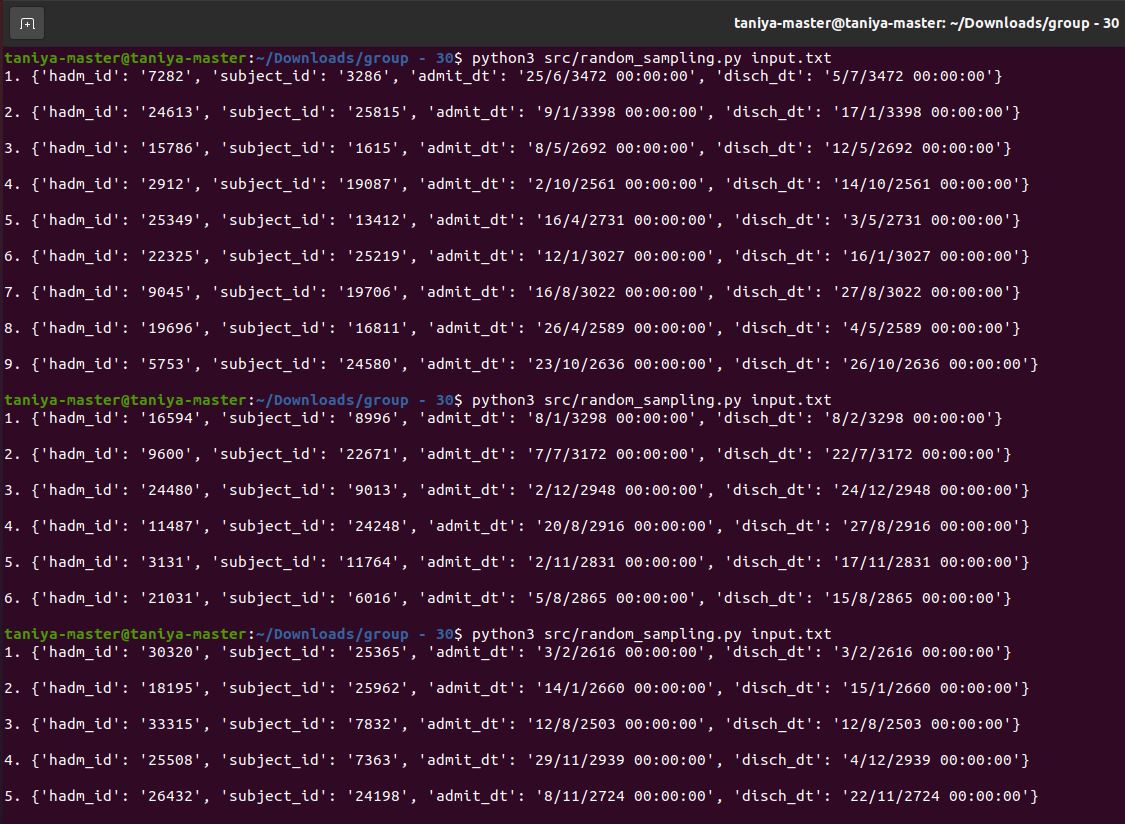


1. **Result:**

The objective of the project was divided into 2 parts.

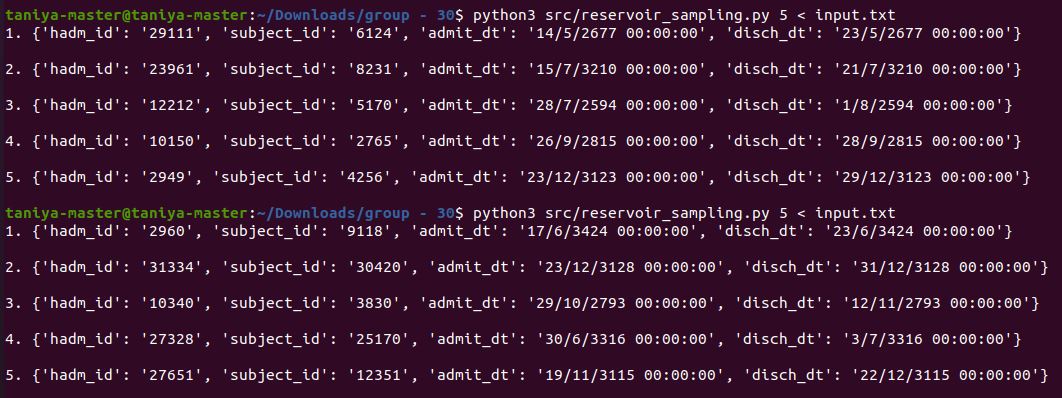
* To implement Random Sampling Technique using Given Dataset.

In random Sampling, Every case of the population has an equal probability of inclusion in sample. Hence, In below snippet, we can see the data is generated in a randomised fashion and is always printing unique number of rows at a given time.



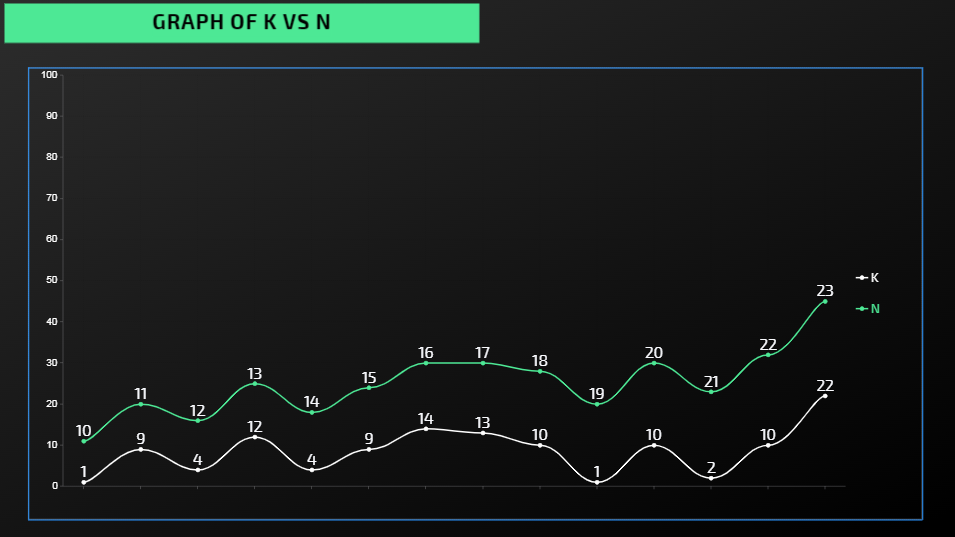
* To implement Reservoir Sampling technique using Given Dataset.

In reservoir Sampling, we randomly choose k samples from a list of n items, where n is either a very large or unknown number. Hence, In below snippet, we can see that everytime ‘k’ number of data is generated from a total of ‘n’ data-rows contained in ‘input.txt’ and is always printing random ‘k’ rows at a given time.



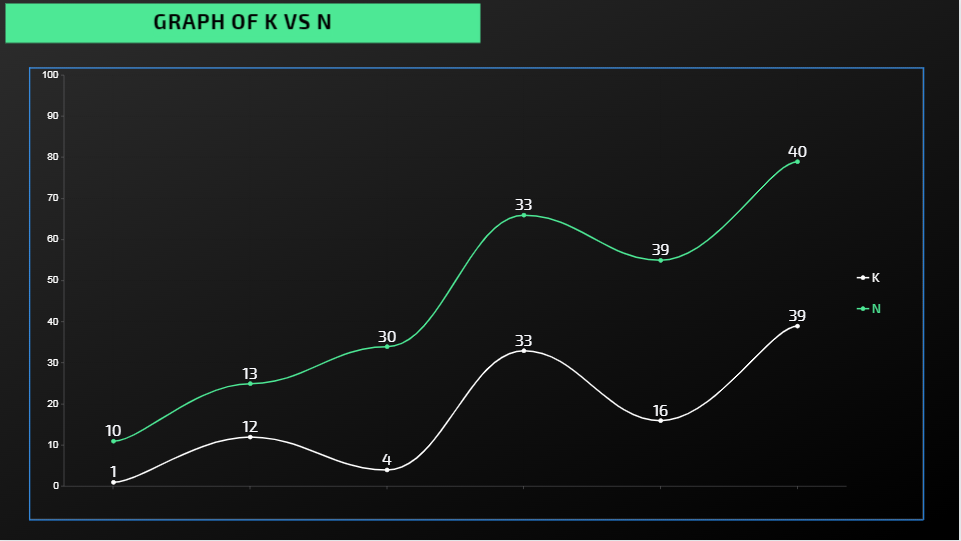
**Graph Representation for both the Algorithms:**

(i) ***Random Sampling***



(ii) ***Reservoir Sampling***

The value of ‘K’ is fixed in Reservoir Sampling whereas in Random Sampling it is a random value between (1,n), Hence the difference in both the algorithms will only be the sampled number of messages at a given time.



1. **Future Work:**

The future work in this project includes the following elements:

* Integrating the other attributes from the dataset to create a more robust pattern to help health Experts in more Accurate Analysis.
* Integrating the whole project into a website or app to monitor the changes in real-time data that will make it more easier to draw accurate conclusions in given circumstances.
* Integrating a deep learning model into the project so that the system keeps learning from the current data and changing data as well.

1. **References:**

[1] <https://github.com/gstamatelat/random-sampling>

[2] <https://journals.sagepub.com/doi/pdf/10.1177/1550147718773999>

[3] <https://www.ece.iastate.edu/snt/files/2019/01/sss-edbt-2019.pdf>