Project Deliverable 1
Group 10
Sept 24, 2022

1. Introduction

1.1 Overview

With the introduction of the Internet in the 80s, the amount of information available has increased tremendously. This led to an increasing need to store and organize data. This saw the rise of Databases and Database Management Systems(DBSM), which allowed the user to extract data from that stored in a database. In a study conducted, it was discovered that 5 exabytes of data had been created at Google by 2003. By the end of 2010, the same amount of data was being created every 2 days and by 2022, it is being believed to be created every 30 minutes. The total amount of data being captured by industries is doubled every 1.2 years, but only an approx of 0.5% of data created is ever used or analyzed. We can easily see the need for a more reliable way of storing and analyzing the data available to us. Database system overcomes all the shortcomings and helps us better utilize the enormous data we have access to.

1.2 Project Description

With the introduction of mobile based taxis like Uber and Lyft, Yellow Cabs in NYC are facing a downward trend in terms of number of trips and passengers. Covid'19 didn't help their case, as passengers looking for safer options preferred mobile based taxi services. Our project is meant to help the drivers as well as passengers decide if operating or taking a taxi is the best possible option from location A to location B at specific or random time intervals or particular periods across a calendar year. The user has the flexibility to choose the parameters like time, location, payment method, trip fare etc. in order to make an informed decision whether it be a taxi driver, a tourist or tech giants like Google Pay looking for their user demographic.

We are planning to build a user interface which would allow various users to see and analyze trends from data consisting of records from early 2005 to 2022.

1.3 Motivation for the project

The data about Yellow Cabs in NYC is easily available on the internet. But there is no computerized application which acts as a one-stop for everything related to trip data based on Yellow Cabs. Building an effective system that could manage such data requires significant time and investment. For an average user, looking at the Excel files is counterproductive as they can see what the data consists of, but deriving useful conclusions for over a million of records requires a useful tool that can comprehend, visualize data and give users a reliable architecture that could be accessed from anywhere at anytime. Our goal is to fill this gap and develop an application based on a database management system which is not just storing and retrieving data, but also represent the data in order to generate more effective trends.

The dataset consists of over 20 attributes and over 2+ millions records every month, which gives us ample data to derive trends from and show a graphical representation of that particular trend, enabling users to better visualize the vast amount of data. Hence, a Database System that can handle this data requires performing millions of computations to generate results.