GROUP 10 PEER REVIEW REPORT

Group1- Impact of GDPR on Database System

The presented paper has a good selection of topic which is based on General Data Protection Regulation (GDPR), also the presentation was lucid and, easy to understand. The authors provide us with elaborated explanation as well as real world examples of companies who violated the rules and regulations of GDPR is being provided which was informative. Importance of personal data has been highlighted in the paper. GDPR is a fundamentally a regulation which entrusts the companies who own user data to safeguard it, and not distribute it to various other organization which can use this data for malicious activities, etc. If done so, GDPR penalizes them which monetary penalty which is the maximum of 4% turnover of the company or 20 million pounds.

Flaws of GDPR workloads on relational versus non-relational databases has been compared with the mentioning of the time efficiency. Through this paper, we got to learn about the Importance of Individual data, measures taken to protect the data and actions executed in the case of violation of GDPR regulations. The process to built GDPR Compliant database system for storing data is explained thoroughly using Redis and PostgreSQL databases.
Finally, the authors conclude by listing the GDPR compliance benefits:

1. The compliance results in very high-performance overheads
2. Compliance is easier in RDMS than in NoSQL
3. While dealing with Compliance you need to examine the tradeoff between full compliance and efficiency.

Group13: A review for Joins over Encrypted Data with Fine Granular Security

Our Understanding:
The Paper and presentation talked about encrypting data on join and security in general.
The goal of the paper is clearly described as a method to provide better encryption which can be beneficial in terms of security, data leaks and time complexity.
The paper exposes use to two encryption algorithms: Key-Policy Attribute Based Encryption (KP-ABE) and Searchable Symmetric Encryption (SSE). Fundamentally, both encryption algorithms work on encrypting the data and generating the key to match the data on decryption.
The authors further put a claim that no algorithm can have all

1. Non-interactive, single singer
2. Efficient O(n^2)
3. No leakage of data beyond the output size,

So there exists a careful trade off among the above points that can achieve an optimum secure system as compared to the existing system.
CRITIQUE:

The clarity of subject is clearly explained in the paper and presentation with appropriate examples and tables. It would have been more beneficial if the authors gave a working example which portrayed a real-life application or a potential application of this subject.

Overall, if this paper were graded on a scale of 10 I would give 10/10 as after reading the paper, it gave a clear understanding of the existing problem and the solution to it through the above mentioned encrypting algorithm, illustrative figures and diagrams in the paper and presentation were helpful for me to visualize the concept clearly.

Group2: On Predicting Suicidal Risk from Social Media Posts

Our Understanding:

The Paper and presentation talk about a machine learning approach to predict suicide risk.
The abstract and introduction have a clear picture about the problem i.e., the increasing risk of people commit suicide each year, the solution i.e., an algorithm called Suicide Ideation detection on Social Media using an Ordinal formulating (SISMO) that uses a machine leaning algorithm that employs certain improvements over existing algorithms on the same topic.
These improvements are

1. Considering posts with respect to the user and sequentially over-time.
2. Penalizing misclassification differently
3. Transforming the problem into a multi-class problem.

Due to these improvements, SISMO is able to problem more details about the magnitude of risk the user may be at in the suicide ideation process.
The machine learning algorithm uses a dataset containing approximately 270,000 user data from Reddit posts from 9 mental health and suicide related subreddits.
The algorithm has 5 classification stages as mentioned in the paper:

1. Supportive
2. Suicide Indicator
3. Suicide Ideation
4. Suicidal behavior
5. Actual Attempt.

CRITIQUE:

The paper clearly presents to issues, problems, talks about existing solution and finally presents a solution which beats existing solution. The presented algorithm SISMO is clearly explained with appropriate diagrams and mathematical expressions. The training stage and classification stage of the algorithm is clear and easy to understand. The working example provides a potential algorithm logic which uses this algorithm to detect suicide tendencies.

Group 11: Review on Repeat Buyer Prediction for E-Commerce

Our Understanding:

The chosen paper focusses on a very interesting topic about repeat buyers after the sales promotion. We didn't know that there is downside to the sales promotion to it. This review paper highlights the downside to it and what can be done to minimize the cost of sales promotion and how can it be reduced to targeted audience (repeat buyers) which the members of the group could clearly elude with clear understanding of the paper and gave example using the algorithm. It gives us a proper understanding about how E-commerce companies deal with such problems. This review paper studies the feature engineering and model training which helps the merchant how to reduce the promotion cost and how to find the target audience (the buyers who will be loyal in the future) and hence the reduced expenditure can be used in various other necessary investments by the merchant.

One of the main tasks is to find the apt feature set for model training and testing phase. However, that particular task is outside of the scope of the course and hence, group members have accurately stated the feature set by the said companies. have Following the feature set task, it is just a classification problem where group tests the obtained feature on various Machine Learning models so as to get the best accuracy. For the said classification problem, the methodology used is infamous comprehensive feature engineering.

The initial data set contains a user-merchant-label relation. Class label for a particular user-merchant group is 1, if user has purchased something within the 6 months of the sale, else it was 0. However, using only the said information would not give the accurate results. Hence user-demographic data (age and gender in this case) and user-activity log data was used.

Using the said data produced exceptional results because the activity log data even kept the track of things like number of clicks, add to cart, purchase and add to favorites, which gave more vivid view of the buying trends of the user.

The said data is first used to create interactive profiles like: User-Merchant profile, User-Brand profile, Merchant-brand profile, User-Category profile, Brand-Category profile. These profiles give more in-depth understanding of the user-buying patterns. The most common one is User-Merchant profile. After getting all the features, all of them are ranked known as Feature ranking. Following this is Model training and at the end the performance analysis of various models is done.

The Working example gave even more clear understanding of the process. It seems that the members have really gave their time to create the working example. The group is working on the subset of the bigger problem. Here they find the relational model to solve this, by joining the tables to get the relevant training and testing data.

Critical Review:

The Whole review paper was well structured. Abstract gave just the right amount of information to understand the problem being tackled with. The members gave a clear picture of the work done by the authors as it was important to understand their work before getting to their own working example. They also showed how their work was relevant to our course work. The steps they followed were easy to understand to get the gist of the topic.

The group members have specified the steps to do that in a very easy-to-understand the workflow diagram. Also, it seems with their review paper they have really well analyzed the authors' work and incorporated their own understanding with them. There is always scope of improvement.

Improvements that can be made:

In working example, the members could have given understanding of the process via work-flow diagrams as they did for authors' work. They could have highlighted the course relevance in a detailed way in their presentation.

All in all, we believe that the team has really worked hard in understanding the paper and have done quite good research on the particular paper and other related sources. They have made us realize that how buying pattern are important when we even didn't have enough information about the problem being a problem.

Group 12: Looking Ahead makes Query Plans Robust

Understanding:

The team has made a commendable effort to introduce an efficient solution for handling the query optimizer failures. The paper initially talks about the limitation of the traditional query optimizers that is a crucial part of the relational database, then leads on concerning the performance in execution time before and after applying the proposed solution. It explained how the currently employed optimizer lacks in achieving the robustness and the suitable query performance since the cost calculation of a query is associated with considering a single join at a time and choosing them based on the lowest current cardinality and selectivity regardless of the size of the data. This causes a huge difference in performance of the best and the worst-case scenario of data loading.

The Lookahead information passing (LIP), the algorithm introduced to overcome the issues of the general query optimizers has been very well explained with a working example, which is very closely related to our coursework. The algorithm considers selectivity of all the joins ahead of the time and reordering them based on the count of occurrence to achieve robustness and lesser probes in hash table. The algorithm comprises of 4 phases namely, build phase, Bloom Filter Probe Phase, Adaptive Reordering Phase and Hash Table Probe Phase. The first phase mentions how hash tables and bloom filters are created and used to determine the optimal ordering of join statements. The second phase consists of analyzing the earlier formed filters utilizing the fact table and managing hit/miss statistics for each. Third phase uses the statistics for adaptive reordering phase to optimize the join orders.

Review:

The working example makes the understanding of the paper much more lucid since the algorithm is implemented on the IMDB database from our coursework. The paper is extremely informative and well-designed covering the drawbacks and limitations of the research paper.

Since, the paper and presentation has been very well written and structured, that there is a very small room for improvement.

We are thoroughly impressed by the hard work put in and the efforts made by the team.