CSCI 622 Project Report: Phase-1

Aravind Vicinthangal Prathivaathi
Griffin Dunn
Steven Simmons

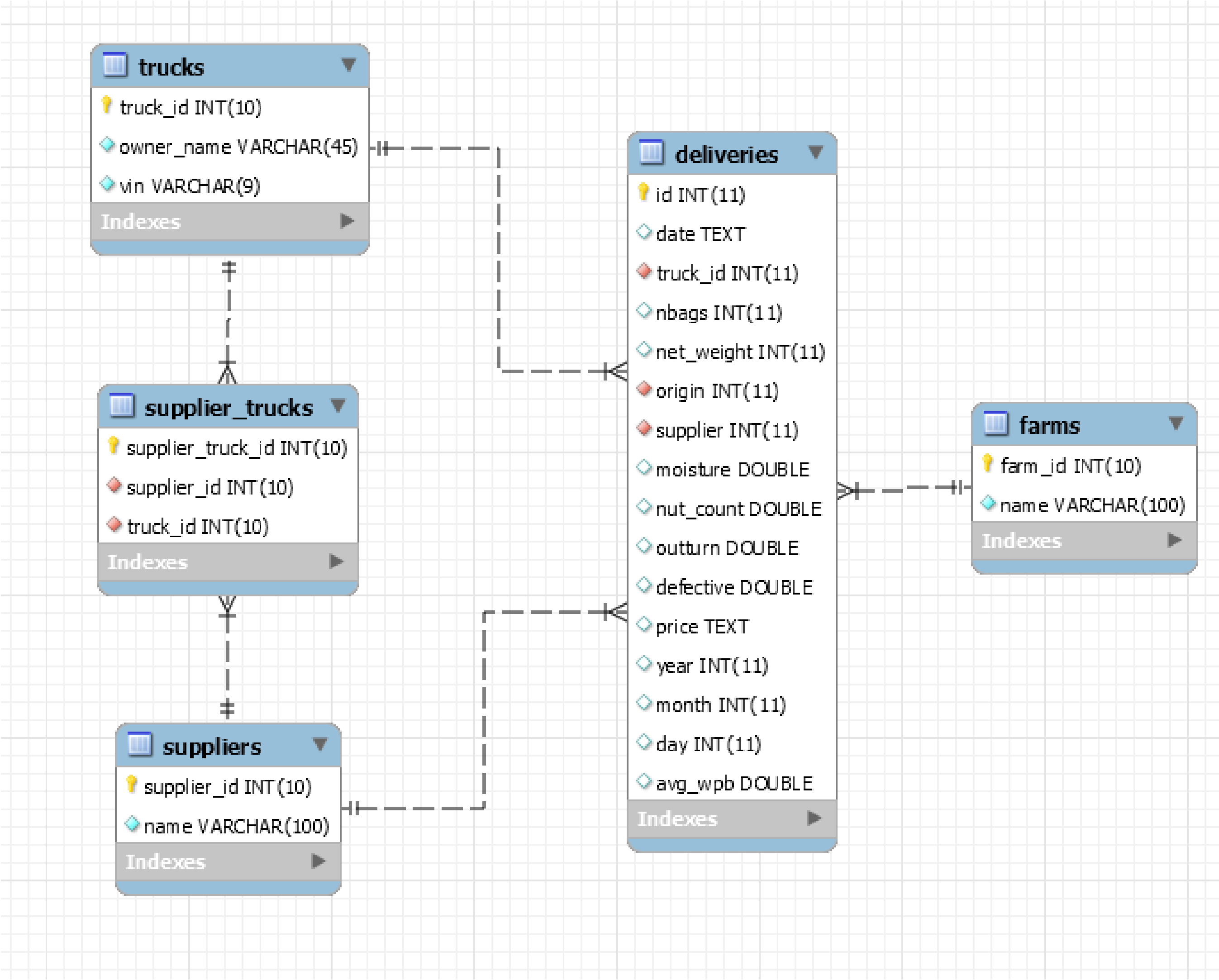

Figure 1: Relational schema for Cashew Truck Arrivals

ABSTRACT

The two most important cornerstone in any big data application is data security and privacy. Data security is a way of technically safeguarding the data from unauthorized use, while data privacy involved how this safeguarded data is being handled and the sensitivity of the individual attributes that make up the data. Our project's main focus is to explore new and interesting concepts involved in the security and privacy of our data and ultimately design an application that implements these concepts. We also intend to cover and analyze some prevalent issues surrounding the security and privacy of data. This phase-1 review of our project will be a quick introduction to our project, the data, it's schema and the software we use to store and access the data and the privacy and security concepts we intend to explore.

KEYWORDS

datasets, anonymization, AWS, cloud, access control

1 INTRODUCTION

All organizations in this world have sensitive data that should be securely stored and have restricted access. With the world well into the digital age, the majority of the companies hold store data in clouds, privately owned servers and other distributed systems. Data protection, as a result, has become the top priority for any company. Breach of sensitive data could wreak havoc among a company's user-base, it's partners and it's employees.

Data security usually involves maintaining the integrity and accessibility of data. It ensures that data is accessed through secure channels and every user needs some authorization to view the data. This authorization may involve restrictions in the manipulating, viewing, destroying or creating data. On the other hand, data privacy is defined as the way data is handled. Policies must be created by each organization around how data is shared and used, this policy must then open to the public view and must be followed by the company. Privacy policies usually differ with respect to the data that is being collected. Overall, we can say that the higher the sensitivity of data being collected and stored, the stronger should be the data security features and the respective policy.

In our project, we intend to build a simple application interface or API which will allow us to query a database hosted on a server or cloud. Our plan for We shall initially build a system that doesn't have any privacy policies and lacks adequate security. We will then show that private data leaks can take place without appropriate data security, and we will show the importance of implementing these concepts mentioned earlier and how they will help secure the database from attacks and unauthorized users. Finally, we will implement the concepts mentioned earlier, show the difference between the initial insecure application and the final secure one. We will be exploring data security and privacy concepts such as database auditing, data access controls, application security, database hardening, some common attacks such as SQL injections and also investigate privacy policies in our project. We shall look into the ethical issues surrounding data privacy as well.

2 PROJECT PLAN

The below subsections will provide a brief overview of our project plan, albeit prone to minor updates and alterations in the future.

2.1 Summary of our plan

We will be using a Cashew Truck Arrivals dataset [2] which contains cashew nut deliveries to port houses from their respective farm over three years. The data is around 61 KB but we can create more data if a larger dataset is needed.

The dataset contains 16 attributes and we use a relational database model to store and access the data. The schema for our relational database is shown in Figure 1. In the upcoming phase, we intend to implement integrity constraints and other security features over this schema. There is already a good potential number of sensitive data in our data set like owner_name, origin_farm and we plan to explore more.

2.2 Application and database

We plan on building a relational web interface application or an Application Programmable Interface(API) to query our data. We intend to add security and privacy policies to our application as well on how it allows users to access data. Our application will make sure only authorized users can access data.

We will be using Postgres as our data management system and a database instance hosted on AWS. Currently, our database is hosted on AWS at 54.210.93.153 where a postgres instance is running. We have created 3 separate users for each member of the team, each user is password protected and have complete permission to modify the database, as well as view, create and delete relations.

2.3 Security and privacy concepts

Our project will have an application interface, a database, and a server where the database is hosted which could potentially lead to various data security threats. We shall research security threats involving lack of data audits, poor access control, data management, data leaks, SQL injections, and other application-level threats to data confidence and integrity.

We shall naturally also focus on the three pillars of data security: 1. Confidentiality, 2. Integrity and 3. Availability. We want to make sure that authorized users can always access data at any time through a secure channel. We plan on creating user groups to implement access control and also explore different access control techniques[6].

Access control will be a major part of our system, each user will be a put in an access group with different permissions. For example, admins can create, manipulate, view or delete relations and data within the database, while regular users cannot only view data that they are authorized to but can neither delete existing data nor manipulate it. All data created by regular users must be through the interface. No regular user can be allowed to access the database directly. Different categories of users can always be added in the future with different permissions by the admins.

We also want to explore different concepts[7,8,9] which include query restrictions, anonymization, cloud computing security issues[5], new access control techniques[6], authorization, and database auditing.

1. Query Restriction: There might be users other than admins, such as software engineers or testers who may want to access the database directly. Such users should have restrictions on the queries they can run. For example, in our dataset, they cannot execute queries to view the owner_name but can always execute queries to view the number of owners.
2. Anonymization: It is a way to de-identify data records stored in the cloud. This could potentially reduce the burden on access control as it could encrypt sensitive data to users that do not have authorization.
3. New access control techniques: Role-based access control and mandatory access control [6] are some methods we plan on exploring.
4. Authorization: Every user except admins will have restricted access to data and every user will be password protected.
5. Database auditing: We plan on using software and logging systems to monitor our database, the different users who access it, what they access and potential problems related to security and privacy.

We also plan on investigating existing tools and software that will help secure our database or alert us to potential threats to security policies or the lack of it. We will also investigate the privacy issues regarding our data and the sensitive information we might need to restrict or hide from different users.

REFERENCES

[1] Elisa Bertino, *"Data mining with big data," in Data Knowledge Engineering 25 (1998) 199-216.*

[2] D. Chen and H. Zhao, *"Data Security and Privacy Protection Issues in Cloud Computing," 2012 International Conference on Computer Science and Electronics Engineering, Hangzhou, 2012, pp. 647-651.*

[3] R. Bhatia and M. Sood, *"Security of Big Data: A Review," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, 2018, pp. 182-186.*

[4] B. Nelson and T. Olovsson, *"Security and privacy for big data: A systematic literature review," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 3693-3702..*

[5] Sedayao, Jeff Bhardwaj, Rahul Gorade, Nakul. *"Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues," Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014. 10.1109/BigData.Congress.2014.92. .*

[6] E. Bertino, *"Big Data - Security and Privacy," 2015 IEEE International Congress on Big Data, New York, NY, 2015, pp. 757-761..*

[7] Jain, P., Gyanchandani, M. Khare, N. *Big data privacy: a technological perspective and review. J Big Data 3, 25 (2016). https://doi.org/10.1186/s40537-016-0059-y*