**Part A:**

**Question 1:** What is an Intrusion Detection System? Is it possible to implement an Intrusion Detection System on this dataset? Explain the workflow as described in the paper for implementing Intrusion Detection System.

An intrusion detection system (IDS) is a system that monitors and analyzes data to detect any intrusion in the system or network. There are three methods of detecting attack, signature-based techniques (known attacks), anomaly-based detection (compares current user activities against predefined profiles), and hybrid-based detection which is a mix of the first two to reduce false positive rates and provide accurate IDS.

It is possible to implement an IDS on the KDDCUP99 dataset. However, due to the high volume and variety of the dataset, it is very difficult to detect attacks by traditional techniques. Big data techniques are required for efficient and accurate IDS implementation on this dataset. Spark-Chi-SVM model is used for intrusion detection on the KDDCUP99 dataset.

The workflow of the Spark-Chi-SVM model is as below:

1. Load KDDCUP99 dataset and export it into Resilient Distributed Datasets (RDD) and DataFrame in Apache Spark.
2. Data preprocessing. Large scale datasets contain noisy, redundant, and different types of data, which cannot be directly fed into machine learning algorithms (SVM can only deal with numerical data). Data preprocessing helps get the data ready for data analysis. Data preprocessing includes string indexer, encoding, and standardization. The string indexer and encoding convert categorical data into numerical data in which the data undergoes standardization, which is a key technique to obtain reliable results. Standardization uses the unit variance method to ensure the features with big numbers do not dominate thus the results are more reliable. This will improve the classification efficiency.
3. Feature selection. Redundant and irrelevant features slow down the processing speed and make no improvement on the model performance. ChiSqSelector is used for feature selection and used to reduce the dimensionality of the dataset and determine the optimal feature subset. Chi-Squared test is used as a test of feature independence and helps decides which features to select. The numTopFeatures method is used to determine the optimal number of features in terms of accuracy and efficiency.
4. Train the Spark-Chi-SVM model with the training dataset. Support Vector Machine (SVM) classifier (i.e. SVMWithSGD) is a supervised training method that analyzes data in use of classification and regression, and used to train the intrusion detection model to classify data into normal or attack. SVM is computationally expensive due to its high generalization and learning ability, however, Apache Spark can be used to reduce the execution time.
5. Test and evaluate the model with the KDD dataset. The training time, predict time, Area under curve (AUROC), and Area under Precision-Recall Curve (AUPR) are listed for three different classifiers: SVM only (without Chi-selector technique for features selection), Chi-SVM, and logistic regression (with Chi-selector technique for features selection). The model built with Chi-SVM classifier achieves the highest AUROC and AUPR with the shortest training and predict time.

Question 2: Import Data

Graphical user interface, text, application

Description automatically generated

Question 3: Set data into RDDs

Graphical user interface, application, Word

Description automatically generated

Question 4: Split the Data

Table

Description automatically generated

Question 5: Extract the 6 columns and the label column

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Question 6: Total number of connections in ascending order for protocol\_type and service

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, chart, application

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Question 7:

1. Average Duration and label activity type – Scatter Plot

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Portsweep label has a long average duration but does not occur often. Smurf and Neptune occurs very often but does not last for a long duration.

1. Average port\_rate and label activity type:

Graphical user interface, text, email

Description automatically generated

Chart

Description automatically generated

Multiple label activities have an average dst\_host\_same\_src\_port\_rate of just under 1.0. Smurf, nmap, and ipsweep have high port rates.

Average duration and service type:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

We can observe that ecr\_i, private, and http are the most common services with the most occurrences.

Question 8:

Table

Description automatically generated

Need to preprocess the data, use stringindexer to convert categorical data to numerical data. Will focus on df2 (extracted columns from question 5). We can also perform one hot encoding but will choose not to keep it simple.

Graphical user interface, text, application, email

Description automatically generated

Dropped the categorical columns since we already converted to numerical.

Table

Description automatically generated

Extract data target column and standardized the data.

Graphical user interface, text, application, email

Description automatically generated

Data was split into train/test, performed chi-squared method for feature selection. Found out at too late that we need to use the ML lib instead of Scikit learn. Will reduce the number of features used to improve efficiency. We can see that feature 2, src\_bytes have a low chi-square value; therefore, feature is not significant in determining whether the label will be normal.

Graphical user interface, text

Description automatically generated

Train using linear SVM model to predict/determine if the label will be normal or not normal, can use SVM with gradient descent for better results but will keep the model simple. We train the model using the training data and test against our test data.

Graphical user interface, text, application

Description automatically generated

Can see that our model has an accuracy of 0.978, the precision is 0.956, and the recall is 0.931. We can probably improve the score if we decided to use gradient descent along with the SVM model. We can also improve the score if we decided to use other features than the ones extracted and then performed feature selection, but that’s computationally expensive as well.

**Part B:**

**Question 1:**

1. A platform as a service (PaaS) solution that hosts web apps in Azure provides professional development services to continuously add features to custom applications.
   1. Yes, PaaS provides a framework with professional development services/tools that allows developers to customize or add features to cloud-based applications. The professional development service/tools include infrastructure (servers, storage, networking) as well as middleware, development tools, business intelligence services, database management systems, etc. It is designed to support the complete web application lifecycle (including updating). It also offers services such as workflow, directory security, and scheduling for apps.
2. A platform as a service (PaaS) database offering in Azure provides built in high availability.
   1. Yes, high availability is a general cloud feature which is included in an Azure PaaS database, Azure has a zone-redundant architecture, in which the platforms automatically replicate the resource and data across zones. So, if a failure ever occurs, the application can continue using the resource and data from the backup zone.

**Questions 2-3:**

1. A relational database must be used when:

D, Strong consistency guarantees are required. The schema of a relational database must be defined before any data can be added; therefore, the schema is not dynamic. Key-value pairs are stored in a non-relational database. Large images and videos are also stored in a non-relational database.

1. When you are implementing a Software as a Service solution, you are responsible for:

D, the user is only responsible for configuring the applications for his/her needs. The rest three are the responsibilities of the service provider.

**Question 4:**

1. To achieve a hybrid cloud model, a company must always migrate from a private cloud model
   1. No, a company can also start off as a public cloud and then combined with an on-premises infrastructure to create a hybrid cloud model
2. A company can extend the capacity of its internal network by using public cloud
   1. Yes, a company can extend the capacity of its internal network by configuring its cloud environment and connect the on-premises network to the cloud environment. This is much cheaper than expanding its existing on-premises infrastructure.
3. In a public cloud model, only guest users at your company can access the resources in the cloud
   1. No, anyone with an account in Azure Active Directory can be given access to the cloud resources

**Question 5:**

a. A cloud service that remains available after a failure occurs: Fault tolerance, the ability to prevent failure and remain available after a failure occurs

b. A cloud service that can be recovered after a failure occurs: Disaster recovery, the ability to recover from a failure and to prevent the loss of data

c. A cloud service that performs quickly when demand increases: Dynamically Scalability, the ability to increase the capacity dynamically when the demand increases

d. A cloud service that can be accessed quickly from the internet: Low latency, the ability to provide service with minimum delay from the internet