BIG DATA

PRESENTED BY:

JOANNA MARIE DIAZ

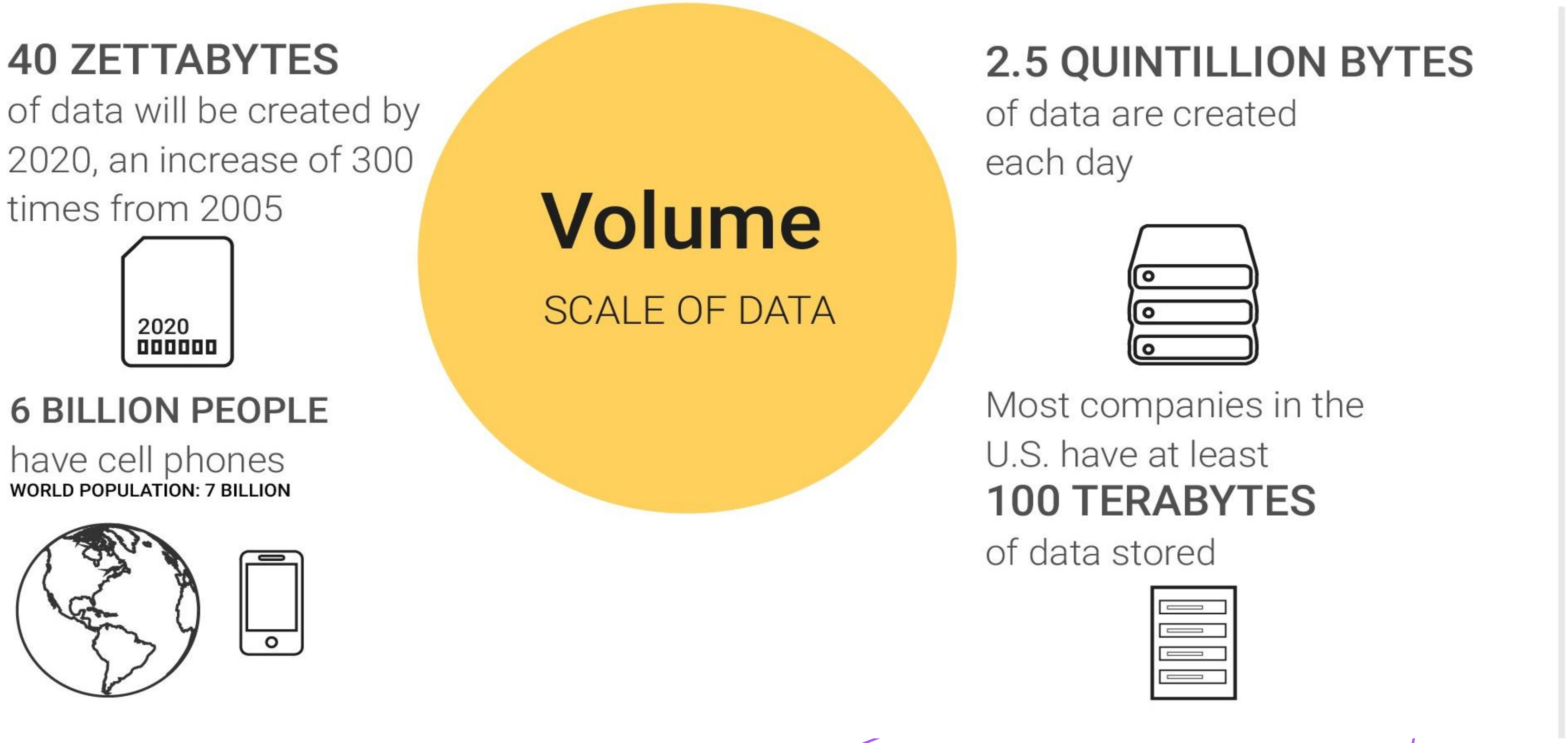
- Big Data can be defined as a large amount of accumulated data that is unstructured, semi-structured, and unstructured data that came from different resources.

- According to IBM, Big Data has been primarily characterized by its 4 V's. Which are Volume, Velocity, Variety, and Veracity.

THE 4 V'S OF BIG DATA

VOLUME

THE SIZE OF DATA THAT IS BEING GATHERED AND PROCESSED



40 ZETTABYTES of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Volume

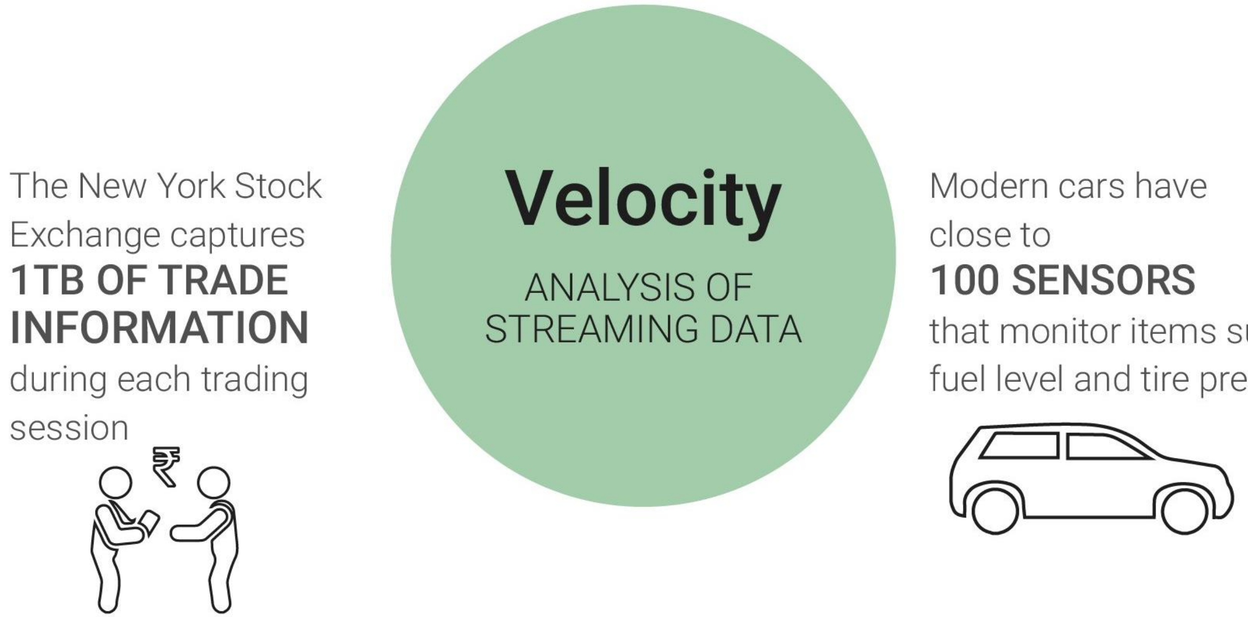
SCALE OF DATA

2.5 QUINTILLION BYTES of data are created each day

Most companies in the U.S. have at least **100 TERABYTES** of data stored

VELOCITY

SPEED AND RATE OF GROWTH OF DATA



The New York Stock Exchange captures **1TB OF TRADE INFORMATION** during each trading session

Velocity

ANALYSIS OF STREAMING DATA

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

VARIETY

REFERS TO HETEROGENEITY, REPRESENTATION, AND INTERPRETATION OF THE UNVEILED DATA

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES

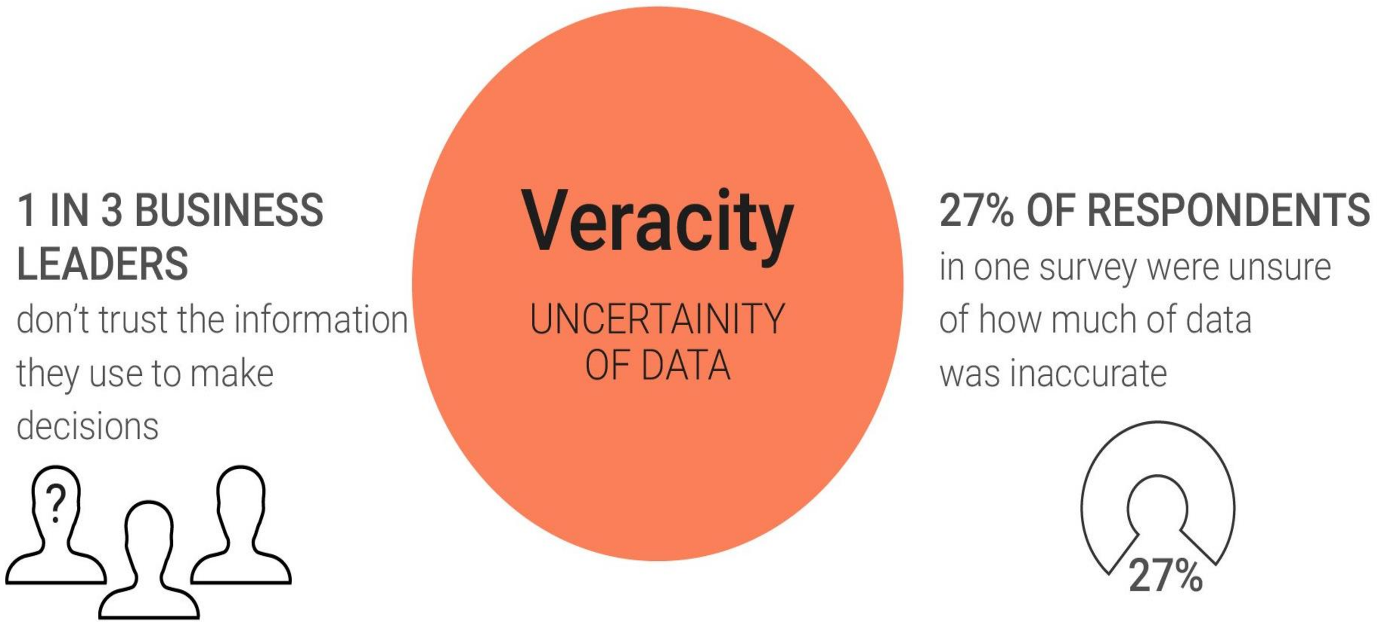
**30 BILLION
PIECES OF CONTENT**
are shared on facebook every month

**4 BILLION +
HOURS OF VIDEO**
are watched on You Tube each month

4 MILLION TWEETS
are sent per day by about 200 million monthly active users

VERACITY

CONFIDENCE OR TRUST IN THE DATA, I.E. THE PROVENANCE OR RELIABILITY OF THE DATA SOURCE, ITS CONTEXT AND HOW MEANINGFUL IT IS FOR THE ANALYSIS



1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Veracity
UNCERTAINITY OF DATA

27% OF RESPONDENTS
in one survey were unsure of how much of data was inaccurate

IMAGINE HOW WILL YOU HANDLE THIS HUGE AMOUNT OF DATA

BIG DATA ANALYTICS

The unstoppable accumulation of data will be chaos and never be of value unless it will be controlled and analyzed. Analytics application will certify the refined exploitation of the escalating volumes of data for diverse versions of business purposes, not only for yielding simple data-driven insight but also for a precise visual assumption of future trends and events (Sedkaoui, 2018).

This analytical tool will respond to questions such as:

"What has happened?" (Descriptive analytics),

"What could happen?" (Predictive analytics), and

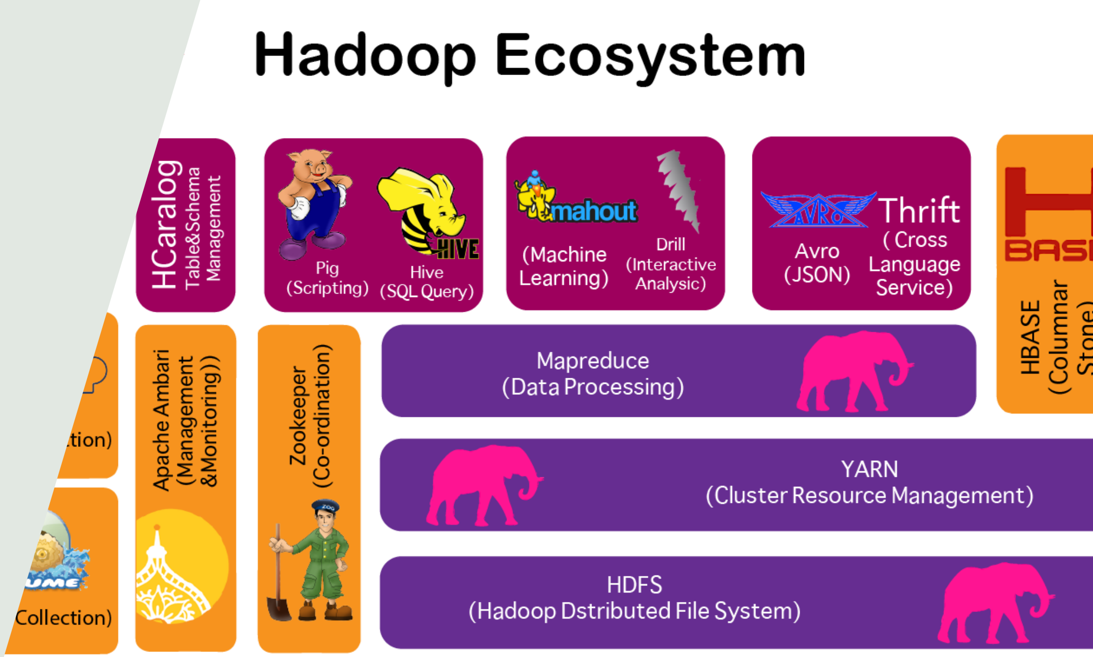
"What should we do?" (Prescriptive analytics).

APACHE
hadoop
TM

APACHE HADOOP

- IS AN OPEN-SOURCE SOFTWARE THAT PROVIDES STORAGE OF HUGE AMOUNTS OF DATA AND RUNS APPLICATIONS ON A CLUSTER OF COMMODITY SOFTWARE.
- ORIGINALLY BUILT AS A BATCH PROCESSING FRAMEWORK.
- ITS MAIN TASK IS TO SUPPORT GROWING BIG DATA TECHNOLOGIES THAT PROVIDE DATA ANALYSIS, MACHINE LEARNING, AND DATA MINING.

HADOOP ECOSYSTEM



Hadoop Ecosystem is a platform of frameworks that collaborate in handling Big Data problems. It composes of different data-processing software whose tasks comprise ingesting, storing, analyzing, and maintaining data.

The main components of Hadoop's ecosystem are:

- HDFS
- MapReduce
- Yarn
- Hive
- Apache Pig
- HBase
- Mahout
- Zookeeper
- Oozie Component
- Sqoop
- Flume
- Ambari
- Apache Drill
- Spark
- Solr And Lucene Component

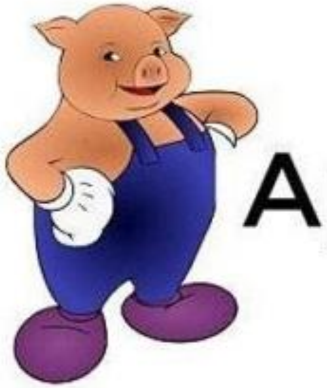
- A distributed file system that depict as a primary storage for Hadoop application
- It replicates and distributes data on different clusters by means of smaller blocks which makes it Fault-tolerant. even if one DataNode crushes the data is not lost.
- Runs in a Java programming environment.
- Has two components which are NameNode and DataNode which execute to store large data across multiple nodes in the Hadoop cluster.

- A programming model inside Hadoop that employs a parallel distributed programming paradigm.
- It enables writing applications that process large data sets using distributed and parallel algorithms in a Hadoop environment.
- It has two features which are:
  1. Map function which converts one set of data into another, where individual elements are broken down into tuples. (key /value pairs).
  2. Reduce function: It takes data from the Map function as input. Reduce function aggregates & summarizes the results produced by the Map function.

hadoop YARN

- Which stands for Yet Another Resource Negotiator
- Its responsibility is to monitor and allocate computing resources needed by the application executions.
- Has two main components:
  1. Resource Manager which is responsible for tracking resources in the cluster and scheduling executed task.
  2. Node Manager that runs on a slave machine and monitors the resource usage like CPU, memory, etc. of the local node. It communicates to the Resource Manager providing updates on the resources.

- A data warehouse project which is built on top of Hadoop to provide ease of query to the user. It has a SQL-like language called HiveQL that yields data queries and analysis.
- It provides a language abstraction and produces a MapReduce job under the hood.



Apache Pig

- Is a scripting language for analyzing and querying large data sets that are stored in HDFS.
- Users can write their own functions using their preferred scripting language.
- It evaluates both structured and unstructured data
- Interprets a load command to load the data in the pig. Then perform various functions such as grouping data, filtering, joining, sorting, etc. After the executions, data can be dumped on the screen or stored the result back into HDFS according to the user's requirement.

APACHE HBASE

- A NoSQL database that is built on top of HDFS and is considered as Hadoop database.
- It mimics Google's Bigtable and supports various types of data.
- It is built in Java and its application can be written in Avro, REST, and Thrift APIs. Enterprises use this for real-time data analysis.
- It equips sparse data sets which are common in **Big Data use case**
- It has two components:
  1. HBase Master – It maintains and monitors the Hadoop cluster, manages the database, handles DDL operations, and controls failovers
  2. Regional Server – it is a process that controls reads, writes, updates, and deletes requests from clients. It runs on every node in a Hadoop cluster which is HDFS DataNode.

Provides a platform for creating machine learning application that is scalable. These machine learning models can be collaborative filtering, clustering, and classification.

Apache Zookeeper

- It collaborates multiple services in the Hadoop ecosystem and keep records on all transactions.
- Zookeper manages to organize and maintain services in a distributed environment with its simple APIs and Architecture.
- It saves the time required for synchronization (which solves deadlock that occurs when two or more tasks fight for the same resource), configuration maintenance, grouping, and naming.
- Enable developers to focus on core applications rather than concentrating on a distributed environment of the application.

OOZiE

- It works as a scheduler system to run and manage Hadoop jobs and merge multiple complex jobs to run into a single sequential order of work.
- It consist of two jobs:
  1. **Workflow** – It stores and runs a workflow composed of Hadoop jobs and stores the job as Directed Acyclic Graph to determine the sequence of actions that will get executed.
  2. **Coordinator** – It runs workflow jobs based on predefined schedules and the availability of data.

Apache Sqoop

- A tool that Interacts between RDBMS and Hadoop ecosystem application. It mainly assists in moving data from an enterprise database to a Hadoop cluster and performing the ETL process.
- It also acts as a load balancer by reducing extra storage and processing loads to other devices.

Is a distributed ingestion tool for collecting structured and unstructured streaming data from various sources on the web, aggregating them, and moving them into HDFS. The streaming data can be a log file, email messages, sensors, etc. It is a part of the Hadoop ecosystem that ingests streaming data into HDFS.

Apache Ambari

- It comprises software that makes Hadoop amendable. It is capable of provisioning, managing, and monitoring Apache Hadoop clusters.
- It provides:
  1. **Hadoop cluster provisioning:** It gives a step-by-step procedure on how to install Hadoop services across many hosts. It also handles the configuration of Hadoop services across all clusters.
  2. **Hadoop Cluster management:** It acts as a central management system for starting, stopping and reconfiguring of Hadoop services across all clusters.
  3. **Hadoop cluster monitoring:** Has a dashboard for monitoring health and status. The Ambari framework notifies the user when anything goes wrong. For instance, if a node goes down or has low disk space, etc.

APACHE DRILL

- A schema-free SQL query of which the main purpose is to process a large-scale data set of either structured or unstructured data.
- Apache Drill is a low latency-distributed query engine that is designed to measure several thousands of nodes and query petabytes of data. The Drill has a specialized skill to eliminate cache data and release space.
- Drill gives faster insights without ETL overheads like loading, schema creation, maintenance, transformation, etc.

APACHE Spark™

- Is a unified processing platform that provides streaming, SQL, machine learning, and graph processing.
- It has an in-memory computation that makes data processing faster for both batch and stream processing.
- It supports most popular programming languages (Python, Java, Scala, and R)

Apache Solr™

- A fast query searching platform.
- It search and index the Hadoop ecosystem
- It offers distributed indexing, automated failover and recovery, load-balanced query, centralized configuration, and much more.

DATA ANALYTICS IN ACTION

THE UNSW_NB15 DATASET

UNSW-NB15 was developed in 2015 (Moustafa and Slay, 2015) to provide a more comprehensive dataset for Network Intrusion Detection System (NIDS) which can mirror modern network traffic scenarios, extensive variation of low footprint intrusions, and the depth structure of information about the network traffic.

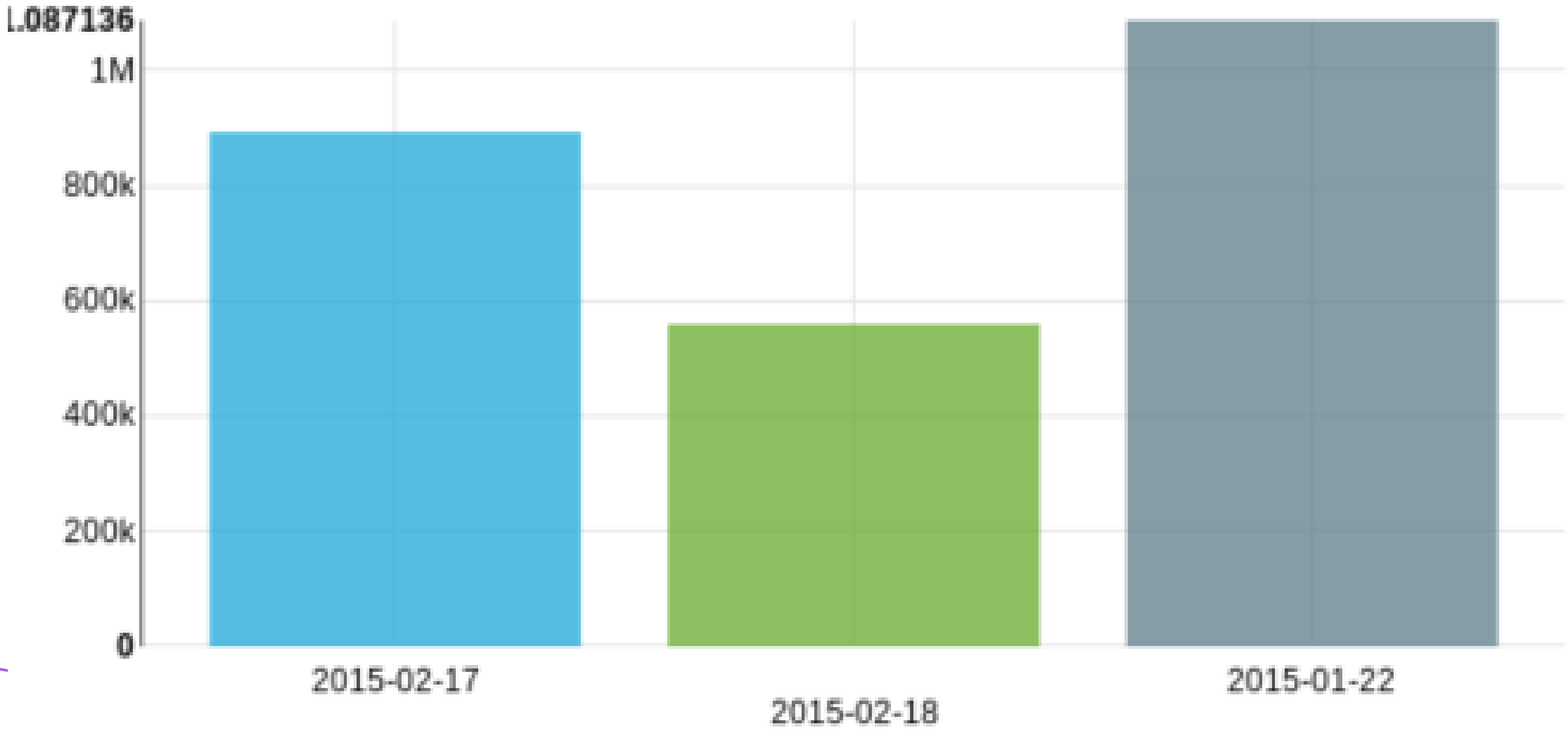
For this activity, UNSW-NB15 datasets were utilized as an experiment for big data analysis and visualization using Hive and Spark queries.
The provided dataset composed of 49 columns and 2,539,738 rows.

HIVEQL DATA ANALYSIS

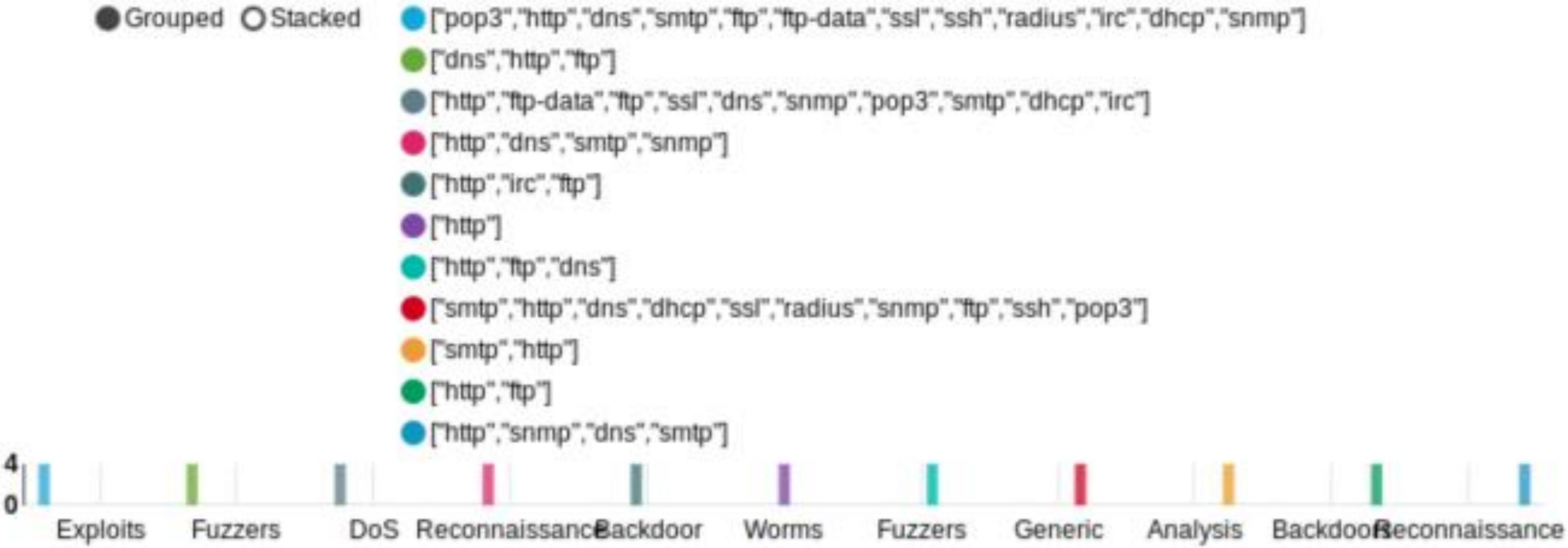
HiveQL was used to perform exploratory analysis on known columns and rows of the UNSW_NB15 dataset. It includes the exploration of Hive’s built-in functions: Date, Math, Conditional, and String methods.

USING HIVEQL THE FOLLOWING INFORMATION IS EXTRACTED:

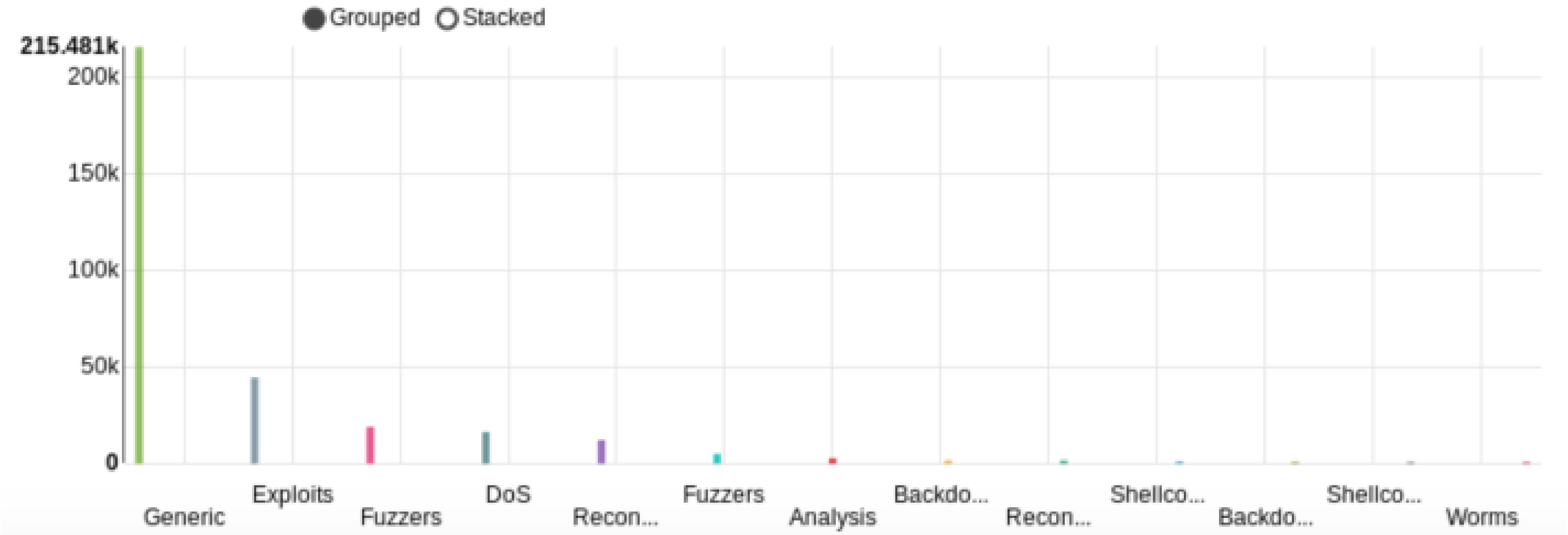
The date and hours of the dataset show the span of creation of the dataset, and prove the specific date of accumulation.



The classification and number of different attack categories versus the services and protocol that have been used during the attack simulation.



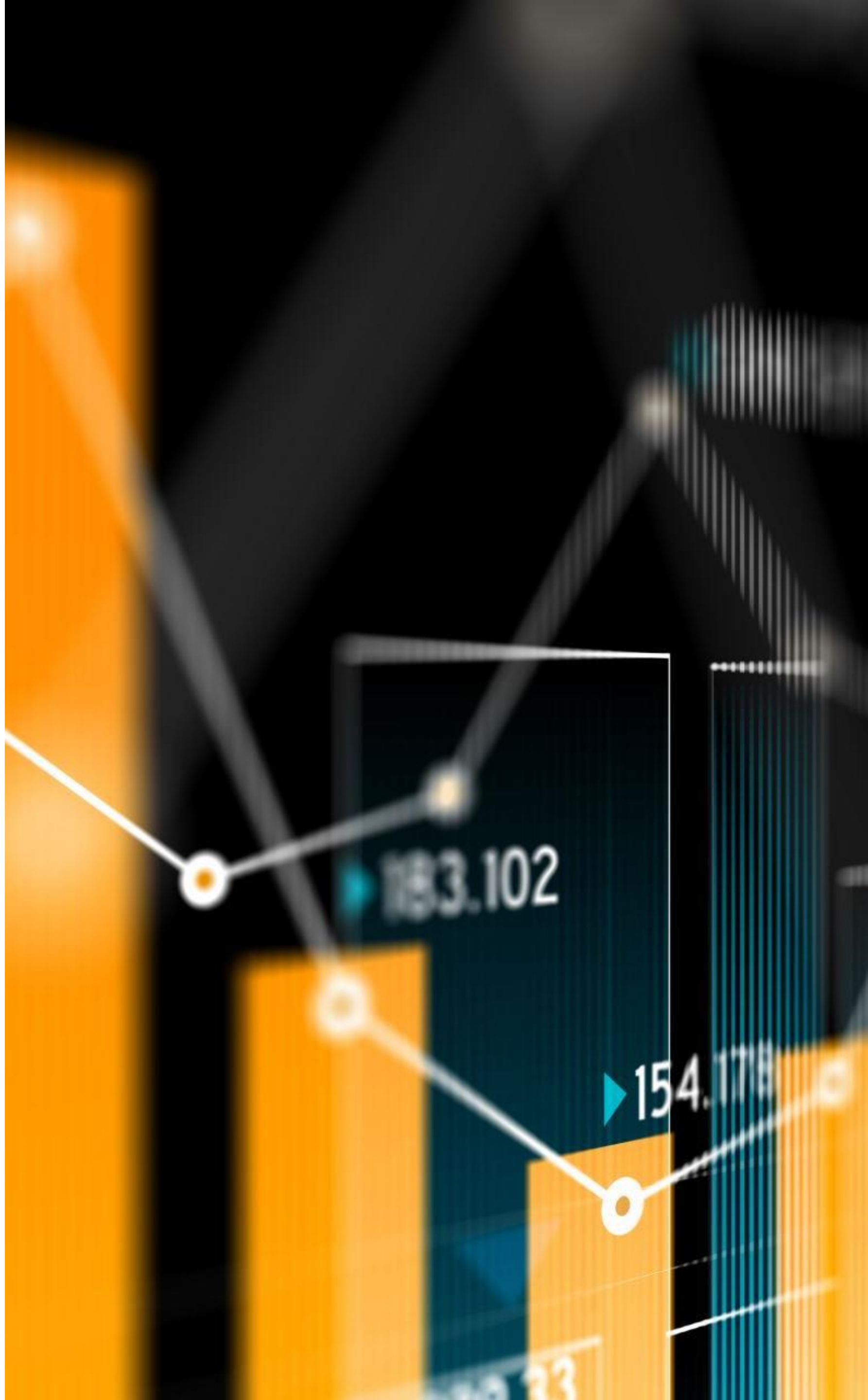
The great difference in amount between normal and abnormal connections.



PYSPARK DATA ANALYSIS

For this part of the assessment, Google's Colaboratory was used to set up the Spark environment where data analysis and ML training was performed. The following step were performed:

- Preprocessing
- Correlation Analysis
- Training the Model



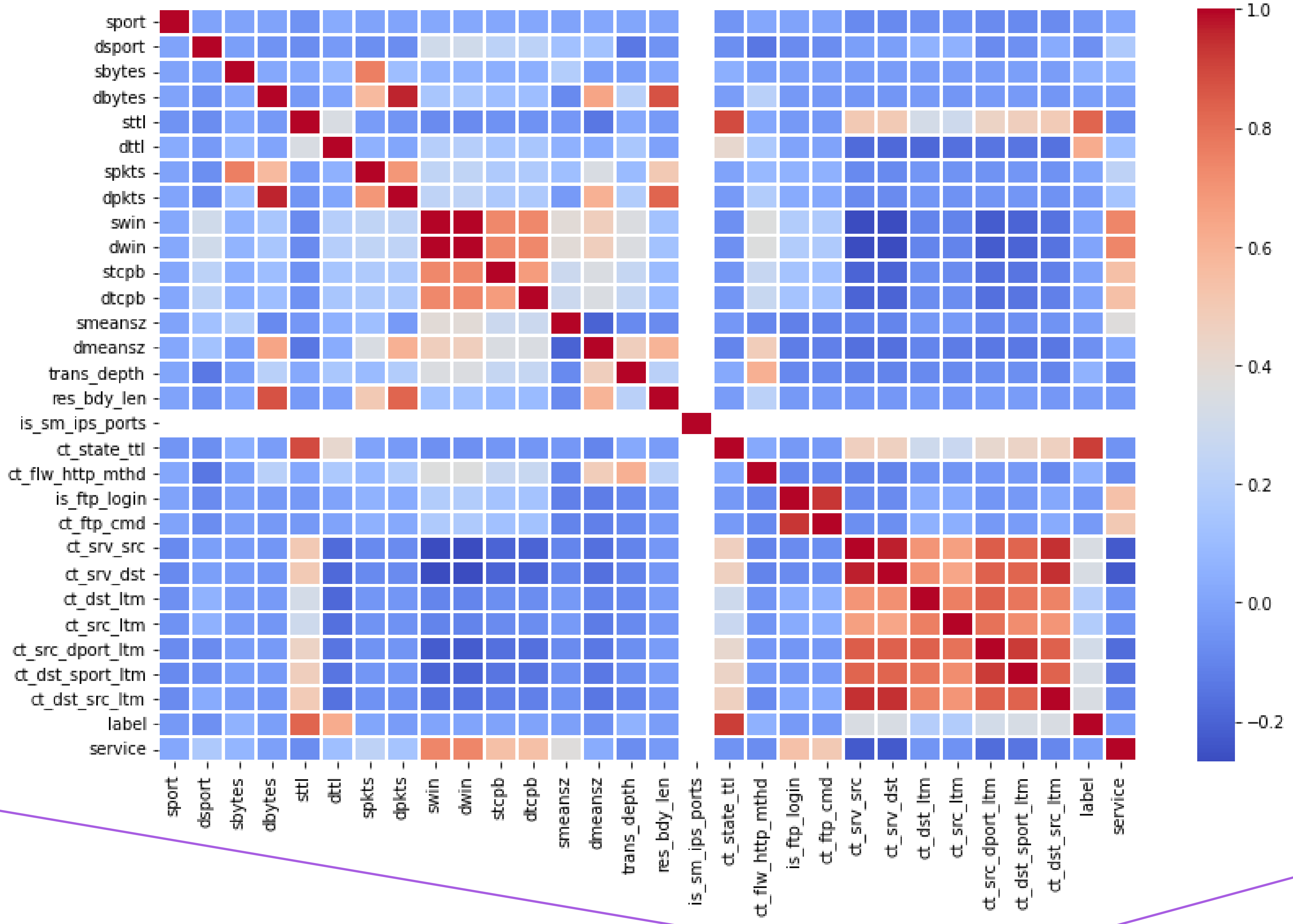
PREPROCESSING

- Dropping of duplicate rows
- Filling missing values
- Selecting relevant features
- One hot encoding / Normalization
- Feature extraction

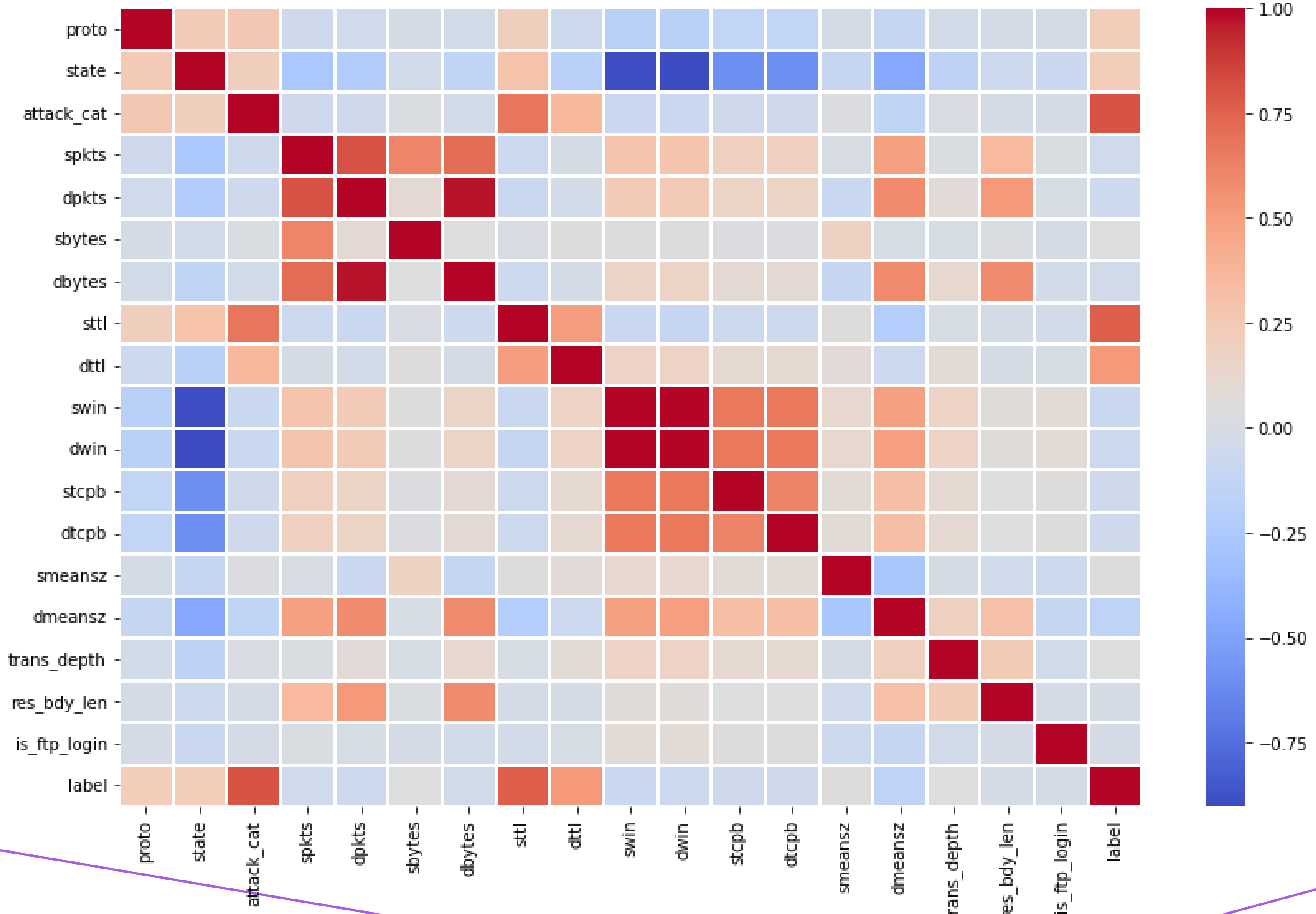
CORRELATION ANALYSIS

On this step, the selected and extracted features undergone the correlation method to find the linear relationship between the feature and label column. I set the value of selecting the correlated value between two variable from more than zero and less than one. I hereby after use the column with positive relationship with the label as the selected features for training the model.

MULTI-CLASS FEATURES AND LABEL CORRELATION



BINARY CLASS FEATURES AND LABEL CORRELATION



TRAINING THE MODEL

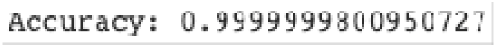
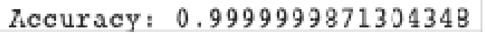
"The process of training an ML model involves providing an ML algorithm (that is, the *learning algorithm*) with training data to learn from. The term *ML model* refers to the model artifact that is created by the training process." - AWS Developer's Guide

For this step, two ML models were created. The discrepancy in the service column was considered for Multi-class Classification and determining normal and abnormal connection was chosen for Binary Classification.

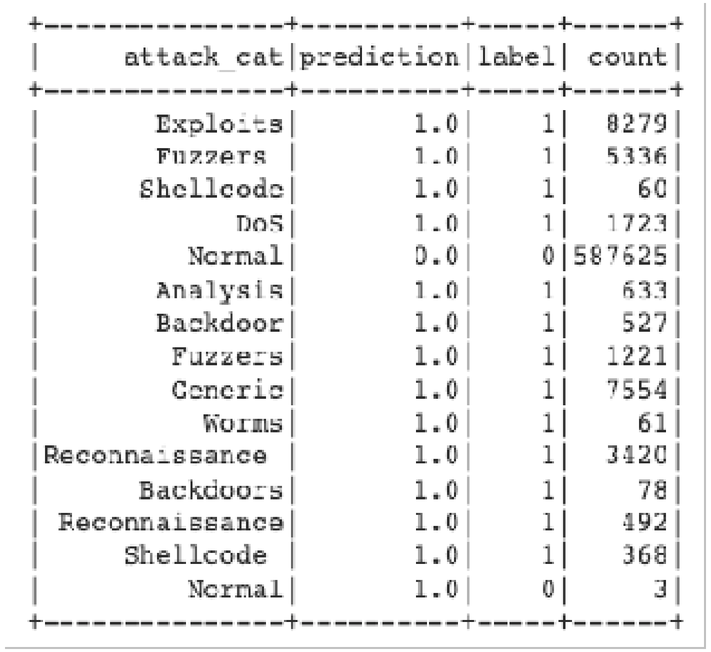
MULTI-CLASS CLASSIFICATION MODEL

Classifier	Evaluation Result	With Cross Validation
Decision Tree	97.24%	99.98%
Random Forest	99.45%	99.71%

BINARY CLASSIFICATION MODEL

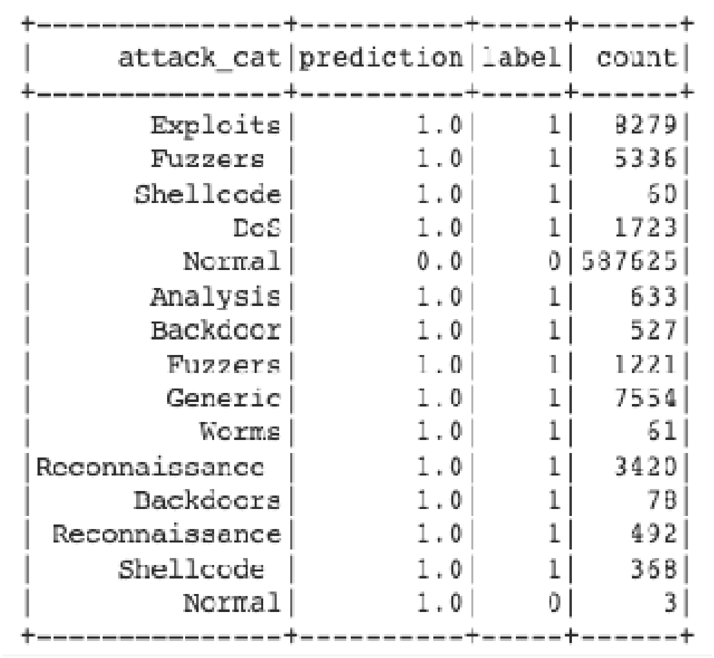
Classifier	**Accuracy Result**
Logistic Regression	 Accuracy: 0.9999999800950727
One vs Rest	 Accuracy: 0.9999999871304348

Logistic Regression



```
+---------------+----------+-----+------+
|     attack_cat|prediction|label| count|
+---------------+----------+-----+------+
|       Exploits|       1.0|    1|  8279|
|       Fuzzers |       1.0|    1|  5336|
|      Shellcode|       1.0|    1|    60|
|            DoS|       1.0|    1|  1723|
|         Normal|       0.0|    0|587625|
|       Analysis|       1.0|    1|   633|
|       Backdoor|       1.0|    1|   527|
|        Fuzzers|       1.0|    1|  1221|
|        Generic|       1.0|    1|  7554|
|          Worms|       1.0|    1|    61|
|Reconnaissance |       1.0|    1|  3420|
|      Backdoors|       1.0|    1|    78|
| Reconnaissance|       1.0|    1|   492|
|     Shellcode |       1.0|    1|   368|
|         Normal|       1.0|    0|     3|
+---------------+----------+-----+------+
```

One vs Rest



```
+---------------+----------+-----+------+
|     attack_cat|prediction|label| count|
+---------------+----------+-----+------+
|       Exploits|       1.0|    1|  8279|
|       Fuzzers |       1.0|    1|  5336|
|      Shellcode|       1.0|    1|    60|
|            DoS|       1.0|    1|  1723|
|         Normal|       0.0|    0|587625|
|       Analysis|       1.0|    1|   633|
|       Backdoor|       1.0|    1|   527|
|        Fuzzers|       1.0|    1|  1221|
|        Generic|       1.0|    1|  7554|
|          Worms|       1.0|    1|    61|
|Reconnaissance |       1.0|    1|  3420|
|      Backdoors|       1.0|    1|    78|
| Reconnaissance|       1.0|    1|   492|
|     Shellcode |       1.0|    1|   368|
|         Normal|       1.0|    0|     3|
+---------------+----------+-----+------+
```

AVAILABLE AT: HTTPS://GITHUB.COM/J-ANNE/BIG_DATA_ANALYTICS.GIT

FULL PRESENTATION

END