CHAPTER 3

WHY BIG DATA IN CYBERSECURITY

3.1 Role of Big Data in Cybersecurity

The era of Internet of Things with billions of connected devices has created an ever larger surface for cyber attackers to exploit, which has resulted in the need for fast and accurate detection of those attacks.

The ability to process these massive amounts of data in real time using **big data analytics** tools brings along many benefits that could be utilized in cyber threat analysis systems. By making use of big data collected from networks, computers, sensors, and cloud systems, cyber threat analysts and intrusion detection/prevention systems can discover useful information in real time.

The synergy between Big Data and AI enables proactive risk assessment by continuously analyzing network traffic, user behavior, and historical attack data. This approach facilitates early **detection** of potential security breaches, significantly reducing the window of vulnerability.

Nowadays, big data is becoming an important topic for research in almost every field, especially cyber security. The main sources of generation of this data are social media sites and smart devices. Generation of data at this speed leads to the various concerns regarding the security of the data that has been created as it is very important to keep this data safe because this data also contains some important and sensitive data such as bank account number passwords, credit card details etc so it is important to keep this datasecure. Also, advanced Big Data analytics provide tools to extract and utilize this data, making violations of privacy easier.

3.2 Big Data Trends

Big data has become a major topic or the theme of the technology media, it has also made its way into many compliances and in internal audits. In EY's Global Forensic Data Analysis Survey 2014, 72% of respondents believe that emerging big data technologies can play a key role in fraud prevention and detection yet only few about 7% of respondents were aware about any specific big data technologies, and only very few about 2%of them were actually using them. FDA (Forensic data analysis) technologies are available to help the companies to maintain the pace with increasing data at very high speed (volumes), as well as business complexities.

3.3 Big Data Analytics in Cybersecurity

A. Big Data Analytics used in Fraud Detection

1. Data Pre-processing Techniques

- Before analyzing data for fraud, it must be cleaned and organized. This means fixing incorrect information, filling in missing details, and making sure all data follows a consistent format.
- For example, if a phone number is missing a digit or a transaction has an incorrect date, these issues need to be corrected before further analysis.

2. Statistical Parameters Calculation

- Fraud detection systems analyze numbers to find unusual activities. They calculate averages (normal spending habits), quintiles (data divided into equal parts), and probability distributions (chances of certain behaviors happening).
- For example, if most customers spend between $10-$100 daily but one suddenly spends $10,000, it may be flagged as suspicious.

3. Models and Probability Distributions

- Businesses create models to predict how transactions should normally look. These models are based on historical data and patterns. If a transaction significantly deviates from the expected model, it might indicate fraud.
- For example, an e-commerce site might have a model where users typically order 1-3 items per purchase, so an order of 50 items might raise suspicion.

4. Computing User Profiles

- Each user has a spending or behavioral pattern that helps define a profile. This includes factors like how much they usually spend, where they shop, and their typical transaction times.
- If a user suddenly starts making large purchases at odd hours in different countries, it might indicate fraudulent activity.

5. Time-Series Analysis

- This technique looks at data over time to detect trends or anomalies. If a customer normally shops online once a week but suddenly makes ten purchases in a single day, that unusual activity might be flagged.
- Similarly, it can detect fraud in stock markets by analyzing sudden price changes.

6. Clustering and Classification

- Clustering means grouping similar data together, while classification assigns new data to a specific group. These techniques help detect fraud by identifying patterns.

- For instance, if a group of credit card users from one region always spends within a certain range, but a new transaction falls outside this range, it could indicate fraud.

7. Matching Algorithms for Anomaly Detection

- Matching algorithms compare current transactions with past data to detect anything unusual.
- If a user usually buys groceries but suddenly purchases expensive electronics overseas, the system flags it as suspicious.
- These algorithms reduce false alarms by refining their models over time and improve accuracy in detecting real fraud cases.

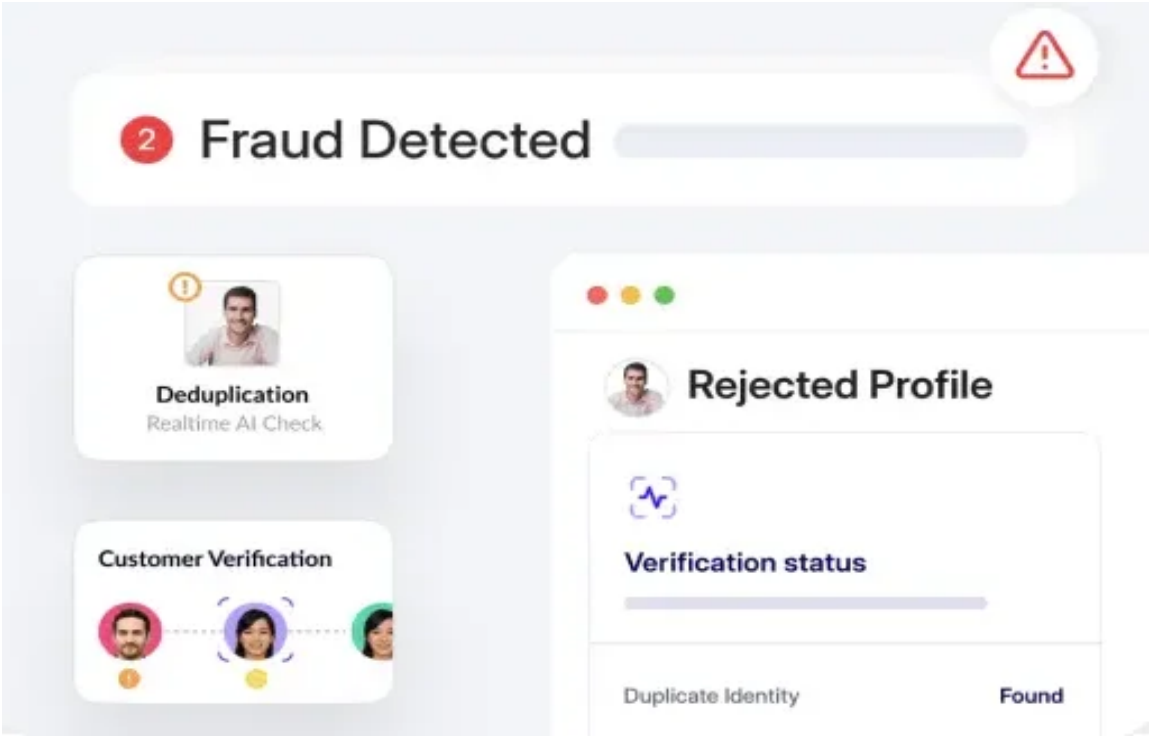

Fig. Fraud Detection Example(Source- Google,Hyperverge)

B. Big Data Analytics used to detect Anomaly Based Intrusion

- Security systems track normal behavior in a network (e.g., how users log in, how much data is transferred).
- Certain key performance indicators (KPIs) are chosen, like the number of login attempts or data download size.
- Thresholds (limits) are set. For example, if a user normally downloads 100MB per day but suddenly downloads 5GB, it crosses the threshold.
- If a threshold is exceeded, the system flags the event for investigation.

- Example: Imagine an employee usually logs in from their office computer between 9 AM - 5 PM. Suddenly, there's a login attempt from another country at 2 AM.
- The system detects this as an anomaly and alerts security teams to check if it's a cyberattack.

C. Provide Security Intelligence

- It collects and processes security logs from different sources (firewalls, servers, emails, network traffic).
- It automatically connects different pieces of information to identify a possible cyberattack.
- It provides security teams with reports and alerts for immediate action.
- Example:If multiple employees receive phishing emails and one person clicks on a suspicious link, the system can track:

  →Where the email came from.

  →Who else received the same email?

  →Whether the link in the email led to a known malicious website.

  →If any data was stolen.
