Use of big data in hacking and social engineering

Abstract: Nowadays, in the fast-paced world of Google and Facebook, every detail of human being could be considered as a set of data or array of data that can be stored, verified, and processed in several ways for the benefits of users. Big data would be perfectly described with humongous large and complex data entities, where classic approach application software is incompetent for them. Big data epitomizes the evidence chattels classified by a high volume, velocity, and variability to require precise technology and analytical approaches for its transformation into value. Big data include netting data, search, data stowing, transmission, updating, data scrutiny, visualization, sharing, querying, data source, and information confidentiality. Big data can castoff in innumerable sectors like defense, health care, and Internet of things. The most famous example probably being Palantir, which was primarily sponsored by the CIA (Central Intelligence Agency). Its primary function was to deliver analytics sway in the war against terrorism of any kind but with accumulative dependency on big data, the menace of exploitation of this data also arises. The prominence of big data does not gyrate around data magnitude or dimensions rather it revolves around how you process it. You can consider stats from whichever cradle and analyze it to discover answers that facilitate cost diminutions, interval time declines, fresh product development and elevated offerings, and smart management. When you conglomerate big data with efficient and dynamic analytics, you can achieve business-correlated tasks such as detecting fraudulent behavior, recalculating entire risk portfolios in shorter span of time, determining root causes of failures, disputes, and blemishes in near real time. Few instances such as Cambridge Analytica lighten the insight of the exploitation of the big data. There are several instances where large amount of data has been stolen like in 2014, Yahoo Inc., where 3 billion accounts were effectively compromised according to official sources or in 2016, Adult Friend Finder where 412.2 million accounts were effected with credit card details compromised as well.

Deprived of the encompassing span of big data, it is taut to perceive a consequence where dappled endeavors and marginal verboten deeds would be a newsflash. It is merely with the inclusion of big data, does the sheer extent of this statistics turn heads. If one individual cheats during a test, it is just earnest of a quip from the instructor. If the entire class collaborates and cultivates a structure of cheating, it becomes newsworthy. The Panama Papers are an exceptional specimen...
of an event that is not a requisite illegal, but sketchy to say the least. The statistic that several sets of international figures were acknowledged in this bulk data set is what marks the news. With the evolution of big data, it makes treasured visions for hackers invariably tempting, but it also provides a big structure of data that converts it to payload utmost necessary to protect.

In such a scenario, the security of big data is very important. This chapter shares sheer insight of how big data can be used in hacking and social engineering. This chapter will try to list down the ways big data is mined from various sources such as Google Services of Android and Facebook. It will list the various ways the big data is used in day-to-day life by the given companies and other advertising companies. This chapter will try to enlist all the major ill ways this data can be used against us and the ways the important and private data can be protected from the data-collecting companies.

**Keywords:** big data, ethical hacking, social engineering, Cambridge Analytica, big data security, data privacy, risk and threat

### 4.1 Introduction to big data

Big data is humongous information collections, multifarious that customary processing application software for data set that is derisory to pact. Big data encounters comprise several functionalities such as netting, exploration, and examination of information along with features like sharing and visualization. It also involves querying and updating information along with confidentiality. There are numerous theories concomitant with big data that are veracity, volume, velocity, variety, and value.

The nomenclature big data implies to the practice of prognostic analytics, user behavior analytics, or few other unconventional data analytic techniques that excerpt value from statistics, and seldom to a certain magnitude of dataset. Analysis of datasets can reveal new correlations to highlight business trends, practitioners of medicine, prevent diseases, combat crime, and so on. Often issues related to several portfolios like government, researchers, marketing, and business executives alike are met with the help of enormous datasets in ranges, including Internet search, urban informatics, financial technology, and business informatics. Even scientists encounter several limitations in various topics, including genomics, meteorology, and environmental research comprising complex physics simulations.

The motivation of big data can be best demarcated as quoted by Carly Fiorina, “The Goalmouth is to convert Facts into information which in-turn can be converting to insight.” There is a massive growth in datasets because Internet of things devices such as mobile devices, aerial (remote sensing), radiofrequency identification readers, and wireless sensor networks, they are gradually gathered by cheap and abundant information sensing. Big data exemplifies the asset of information with features
such as volume, variety, and velocity to oblige explicit technology and analytical techniques for its transformation into value. Additionally, a new $V$, veracity, is added by some officidom to describe it, revisionism challenged by some industry authorities. The three $V$’s have been primarily expanded to other harmonizing characteristics of big data [1].

The systematic study of big data can lead to:

- **Tuning according to target audience** – Big data is used by business today for scrutinizing gushes of the target audience and entertain them with optimized services to upsurge the business.
- **Cost cutting in various sectors** – Scrutiny of such mammoth bulk of data has also aided business in cutting down their overhead expenses in various sectors. Several bucks are being saved by enhancements in operational efficiency and more.
- **Intensification in operating boundaries in different sectors** – Big data also aids businesses in increasing operational brims in different sectors. With the help of big data, lot of blue-collar labor can be converted into machine task and this helps in growing operating precincts.

Big data can be described by the following characteristics:

- **Volume**: Volume can be defined as the magnitude of generated data and stowed data. The volume of the data regulates the value and potential insight and whether it can be deliberated as big data or not.
- **Variety**: Variety can be defined as the category and attitude of the data. This is usually for the people who scrutinize the data to increase the efficiency of the resultant insight. Big data concludes the missing or the omitted pieces from data fusion; it derives its information from sources like text to video anything.
- **Velocity**: This can be termed as the promptness at which the facts and figures are bred and treated to fulfill the requirements and dares that lies in the route of progress and expansion. Big data is frequently usable, reachable, and affordable in real time.
- **Veracity**: This can be defined as data eminence of netted data that can vary prominently, manipulating the precise analysis.

Cyber-physical and workshop systems may have a 6C system:

- Connection (sensor and networks)
- Cloud (computing and data on demand)
- Customization (personalization and value)
- Content/context (meaning and correlation)
- Cyber (model and memory)
- Community (sharing and collaboration)
Data as a requisite should be treated with cutting-edge analytics and algorithms to reveal evocative statistics. For example, to achieve success in a factory one must contemplate both visible and invisible concerns with various components. Information generation algorithms must distinguish and address obscure issues such as machine degradation and component wear.

4.1.1 Application of big data

Big data helps in transmuting cream commercial progressions by appropriate and precise analysis of accessible statistics. These processes generally embrace:

i. **Procurement with big data:** Ultimatum of requirements or necessities can be appropriately conjectured as per various conditions and features offered with big data.

ii. **Big data in product improvement:** It can approximately predict the type of invention compulsory to intensify sales.

iii. **Data warehousing in manufacturing industry:** Data warehousing is a major analytical methodology for categorizing apparatus or measures the practice deviance from the quality benchmark.

iv. **Data warehousing system for product dissemination:** Grounded depending on statistics presented; records scrutiny is considered useful to confirm symmetric circulation in arcade.

v. **Data warehousing system in product advertisement:** Data warehousing system aids in significant advertisement stratagem that could upsurge sale by several folds.

vi. **Price administration using data warehousing system:** Data warehousing system helps business in studying market chart. This is an important part to sustain position in arcade and price management.

vii. **Merchandising:** Retail arcade relies majorly on data warehousing system and analytics to identify the recent trends of the goods.

viii. **Data warehousing system in sales:** Data analytics assists in optimizing product mix. It helps in aggregating sale for the commerce. It is also consignment of sales resources and accounts, and other operations.

ix. **Store maneuvers using data warehousing system:** Stored procedures can be observed by various analytical tools that lead to shrink in manual work. It regulates several factors like training of demographics or inventory echelons based on predicted procurement patterns.

x. **Data warehousing system in HRs:** Data warehousing system has an altered way of recruitment and other human resource maneuvers. You can also discover the physiognomies and behaviors of efficacious employees, as well as other employee insights to accomplish talent better.
xi. **Data warehousing system in banking:** Data warehousing system has provided major prospect to corporations to visualize the larger scenario due to harmonizing the delicate trend of the records for prioritizing the privacy and shielding of information along with conveying value adds for customers. It has been fully embraced by several companies to drive business and advance the services they offer to customers.

xii. **Data warehousing system in finance:** Financial amenities have extensively espoused data warehousing system analytics to advise enhanced investment assessments with constant returns.

xiii. **Data warehousing system in telecom:** According to reports in “Global Data Warehousing System Analytics Market in Telecom Industry 2014–2018,” it was found that the usage of data analytic tools in telecom segment is predicted to propagate at a compound annual growth rate of nearly 28% over the next four years.

xiv. **Data warehousing system in retail:** Retailers hitch data warehousing system to suggest that consumer has personalized shopping experiences. Evaluating customer is one-way data warehousing system technology in making a spot in retail. Two-thirds of retailers have made financial gains in customer management and CRM through data warehousing system.

xv. **Data warehousing system in healthcare:** Data warehousing system is used for scrutinizing data in the electronic medical record system with the objective of sinking costs and refining patient care. This data includes the amorphous data from physician notes, pathology reports, and so on. Data warehousing system and healthcare analytics have the technological advancement to predict, prevent, and cure diseases.

xvi. **Data warehousing system in media and entertainment:** Data warehousing system is altering the broadcasting and entertainment industry, providing users and viewers a much more tailored and enriched experience. Data warehousing system is utilized for growing revenues, analyzing real-time patron sentiment, increasing promotion effectiveness, ratings, and viewership.

xvii. **Data warehousing system in tourism:** Data warehousing system is renovating the global tourism. Information about the world is easily available than ever before. People have detailed itineraries with the help of data warehousing system.

xviii. **Data warehousing system in airlines:** Data warehousing system analytics provides with necessary tactics to the aviation industry. An airline now knows where each and every plane is heading, where any passenger is sitting in any of the flight, and what a passenger is watching on the IFE (In-flight Entertainment) or connectivity system.

xix. **Data warehousing system in social media:** Data warehousing system is a motivating influence behind every marketing resolution made by social media houses and it is driving personalization to the highest extent possible (Figure 4.1).
4.1.1.1 Why big data is a lucrative target

As we race into the future, a swelling amount of modules concomitant to the infrastructure of our realm and enterprises are reliant on an Internet assembly. The probability of devastating cyberattacks from aggressive states, cyberterrorists, and hacktivists becomes much more real: This can be visualized pretty well in movie named Die Hard 4.0, where several unmanned cars crashing or rerouting of energy and electricity on a large scale thereby leading to blackout or tampering in traffic signal leading to accidents.

Few technological loopholes would never lead to an efficacious kinematic assault in a large scale. As an alternative to get access to the system, the invaders use several diverse but fundamental methodologies over the time. Data sabotage, that is, altering of data records can be considered one such cyberattack that seems to be minute but could be used by invaders for major advantages. Small manipulation in data could affect a lot in major sectors like stock market or defense agencies. A small manipulation of rating of a particular fake product in retail market could lead to its perception as a original product and major sale boost in retail sector or a simple tickle in financial figure of a company’s remuneration could provide a major boost in stock market.

US agencies such as CIA and FBI are perceived as major fronts in 2016 for cybercrimes.
Several open confab concerning cyberterrorizations have been dedicated to the concealment and accessibility of information. In near future, we might also visualize several online maneuvers of manipulating major governmental decision, investors of stock market, or corporate decisions due to alterations and manipulation in veracity of the electronic figures provided to them.

4.1.1.1 New concerns for cybersecurity connoisseurs
Numerous sectors in recent years have seen ascending trends of data integrity outbreaks. A false news of President Obama’s injury by Syrian hackers through Twitter account of Associated Press, leading to a sharp 150-point dip in stock market, can be seen as a simple but direct example of the same. The similar example can also be seen as minute altercation in a cooling system by Stuxnet worm, which lead to rescind Iranian nuclear program [2].

“Data veracity outbreaks have a number of dimensions to them,” said Eddie Schwartz, universal vice president at ISACA, an international cybersecurity association. “If you get hold of a meticulous system like the power grid or water system that encompasses machinery operated by workstations and make minute alteration in the operational directives for that equipment, it can lead to some cataclysmic consequences – power outages or deviations in chemical balance.”

4.1.1.2 Previous data warehousing system breaches in recent times

i. Yahoo

**Date:** 2013–14  
**Impact:** 3 billion user accounts  
**Details:** “In September 2016, the past prevailing Internet colossal, while in parleys to peddle itself to Verizon, indicated that it had been the prey of the humongous data breach in recent antiquity, probable by ‘a state-sponsored artiste.’ The outbreak compromised the original appellations, dates of birth, email addresses, and handset no. of Five hundred million patrons. The corporation published that the preponderance of the passwords had been hashed via the robust crypt algorithm.

Few months later, it buried that previous record with the revelation that a breach in 2013, by a different set of black hat hackers had compromised 1 billion records with names, dates of birth, security questions and answers, email addresses and passwords that were not well secured as those involved in 2014. In October 2017, Yahoo reviled that, all 3 billion-user accounts were being compromised.

The breaches bashed a probable $350 million off from Yahoo’s sale amount. Verizon eventually remunerated $4.48 billion for Yahoo’s core Internet industry. The pact stated that the two corporations to share regulatory and legal obligations from the breaches.”
ii. Adult Friend Finder

**Date:** October 2016  
**Impact:** More than 412.2 million accounts  
**Details:** “The Friend Finder website, which comprised spontaneous hookup and adult content network like Adult Friend Finder, Cams.com, iCams.com etc., were penetrated in mid-October 2016. Hackers unruffled two decades of data from six databases that include complete details like names, email addresses and passwords.

The feeble SHA-1 hashing algorithm fortified most of the passwords, which was predestined that almost cent percent of them were been decoded by the time LeakedSource.com circulated its scrutiny of the entire data next month.

CSO Online stated at the stage that, ‘a canvasser who has online Twitter identity as 1x0123 and as Revolver in other networks displayed images of Adult Friend Finder displaying an issue called LFI (Local File Inclusion vulnerability) which was being targeted. The ID stated said that the vulnerability was exposed in a service method on the production servers deployed by Adult Friend Finder. AFF Vice President Diana Ballou acknowledged the same and confirmed, that the issue was fixed which was prevailing due to injection vulnerability.’”

iii. eBay

**Date:** May 2014  
**Impact:** Nearly 150 million user data  
**Details:** “The Internet Giant known for its online auction testified a cyber-attack in May 2014 that is said to have exposed all the details and hashed passwords of all of its users. The company said, it was a scenario of Social Engineering where hackers accessed the company intranet using the credentials of three internal employees, and had a completed backdoor access for almost a year. In this tenure they paved their way to the consumer database.

As mitigation, it requested its users to update their passwords, but alleged that the financial statistics, such as credit card info, was stowed disjointedly and was not compromised. The venture was condemned for a dearth of communication notifying its consumers and pitiable implementation of the password-renewal procedure.”

iv. Equifax

**Date:** July 29 2017  
**Impact:** 143 million consumers along with Credit Card info for 209,000 consumers  
**Details:** “Equifax that is one of the dominant credit agencies in America revealed on Sept. 7, 2017 that an application susceptibility on one of their networks led to a records breach that exposed roughly 147.9 million users. The breach was exposed on July 29, but the enterprise stated that it possibly happened in mid-May.”

v. Uber

**Date:** Late 2016  
**Impact:** 57 million along with 600,000 drivers exposed.
Details: “The Corporation came to know about the breach in late 2016 wherein couple of hackers was able to retrieve personal details of 57 customers of the Uber. They were also able to retrieve the driver license details of 600,000 Uber drivers. Credit card or Social Security numbers were secured as per the company. The hackers got access Uber’s GitHub code repository account, where they retrieved user credentials to Uber’s AWS account. Those authorizations should certainly not be on GitHub.

The Breach was made public a year later by Uber. They compensated the hackers with $100,000 to rescind the records with no clause or way to authenticate that same. The paid them stating it was a ‘bug bounty.’ Uber also sacked its CSO and placed the responsibility on him.

The breach is said to have affected Uber in both reputation and money. At the time that the break was announced, the business was in discussions to trade a stake to Softbank. Uber’s valuation declined from $68 billion to $48 billion by the time the deal was signed.”

4.1.1.3 Vulnerabilities in data warehousing system for big data

Data warehousing system for big data or big data analytics was defined by the connoisseurs with the help of terminology like value of the data, volume, and variety of the same, along with velocity and veracity of the data. This is also defined by 5V’s. Recently, an additional V is gaining the focus of the market, let alone the experts of big data analysis. Vulnerability, as it gains focus in the market, distresses entire enterprise sector and is urging for major attention since, if this is not handled, rest all will be at stake. Due to numerous proficiency, it has now received the consideration of entire domain.

Due to its capabilities of further optimizing the business by better understanding of the habitué and enhanced productivity suggestion, it has made the life of decision makers a lot easier. Then a clause of confidentiality also comes into the picture, which mandates the enterprise to secure patron’s data from any unauthorized scrutiny and due to this the vulnerability dispute needs to be addressed as an important contemplation.

4.1.1.3.1 Reason for all the 6V’s

The data confidentiality is the major dimension several syndicates are dealing with, still there are numerous unquestionable cradle for procurement of user’s personal data. As stated by Marr, “Vulnerability addresses the information that a mounting number of people are becoming comfortable on to the element that their delicate data, the sensitive data of many commercial initiatives, is being gulped up by the gigabyte, used to pry into their comportment and, eventually, peddle things.”
Like several research organizations, an organization used for credit referencing named as Experian mentions similar views in their white papers and other research documents. To mien at the data susceptibility trait, one could mien at a sociological facet at the issue. Several experts like head strategist of Experian named as John Roughley states, “We think about things emotionally, and the emotion that’s associated with data is sometimes one of nervousness, anticipation or liability. That’s partly because it's new but it’s also because everyone’s seen stories in the various sources of Media about data breaches, and record number of individuals have experienced their records being tainted in some shape or form – the phone ringing off the hook with people asking about payment fortification indemnification.”

The principal fears around data warehousing system susceptibility could be addressed through rudimentary questions like, how did my data reach these advertising companies contacting me? What extent of access they have to my data? What around the financial info? How much easily hackers can access my data? Will all of that be whipped by hackers to siphon off money from my account? Nevertheless, to address these disquiets, there needs to be some key steps such as reassuring customers, whose whatsoever information they offer to the company will be securely stored, will not be misplaced, or used for malevolent purposes.

As John Roughley explained, “It's about doing what you promise you will do, and as officialdoms we have a prerequisite to mark certain that we always perform with integrity and with regards to the custodianship of someone’s data. It’s about keeping it secure, keeping it safe, and not breaking any promises with regards to what we will do with it.”

Data garage such as NoSQL have several security susceptibilities, which cause confidentiality issues. A conspicuous security blemish is that it is incompetent to encrypt records during the cataloging or logging of data or while dispensing it into diverse groups, when it is streamed or unruffled [2].

Out of all the majority of data warehousing system vulnerabilities being faced by business, common six can be mapped as follows:

**a) Pitable authentication for records**

With the cradle of informative data records being flowing in and out of a company’s data warehousing unit specially if discussed about Retails Department, and the ease of getting into the database with poor authentication system, it can act as an entry point for any malicious person. Through Rouge scanners, this can open the doors for fabricated transaction, improper rating systems, and so on into the functional system.

This can be mitigated through a granular level of control on the data with 5W’s questions to maintain a trail for all the inbound and outbound data flow in the system endpoints along with a context-driven dogma podium for proper role settings.

As articulated in several security seminars, if you let the flow of improperly articulated and unstructured data with garbage-type data security, the same will be
haunting you for the rest of your life span in the same company. Hence, it needs to be ensured that the incoming records are from reliable sources, and that it is not tampered.

b) Apprehensive web consoles
Front-ends act as another security exposure for data warehousing systems. Considerable amount of interaction with data warehousing uses Internet-based web interface, which act as doors for cybercriminals due to their mostly unaddressed security loopholes. Using techniques like eves dropping, data can be captured easily, which flows in and out using Internet and can be altered to complete their causes. These types of scenarios do not depend on the size of data.

Authentication-related issues could also be instigated via techniques like SQL injection, where web front ends with less or no authentication are at high stakes. Stored procedures can be considered as a mandate to safeguard the data along with parameterized queries for data statements as engagement in modus operandi.

c) Rudimentary security controls unavailability
The lack or unavailability of robust security incorporation in major firms acts as another liability in data warehousing systems. Most of the mid-level or small-level firms do not consider security as a part of fundamental design of any solution, which leads to data leakage since several times these firms act as third-party source for any functionality.

Security deployment should be both preemptive and responsive to safeguard their data from getting into malicious hands. Firms should properly scrutinize for the same. Security threats or various other anomalies can be identified using threat scanning, which can be deployed along with perimeter defenses that can scrutinize in real time.

d) Pitiful encryption system clubbed with derisory masking protocols
Masking can also be represented as a small manipulation of data. This is mostly needed to cover the loopholes in the poorly defined encryption algorithm that can occur in the data flow during the integration points of systems.

Another issue with customer data is anonymity. The use of automated technologies like machine learning can lead to uncovering the user’s identity via simple derivation or direct accidental visibility. These data need to be anonymized for the security of customer. Conceded concealment is a noteworthy area of data security while still industries contemplate of safekeeping with the value of missing data.

Ensuring that a given syndicate has comprehensive set of algorithms and guidelines in place for masking, and encryption is a major scenario in the current market. In ingestion points, to avert any further security concerns, apt dogmas need to be placed as well.
e) Record’s improper lineage and respective audits
Records that are misplaced or scattered can be considered as another epitome of data warehousing system vulnerability. Improper or lost trail data creates an exposure to further details of the customer along with the configurations of security control shielding the data. Mostly, these scenarios can also arise due to records storage in multiple locations like On-premise and Cloud without any proper documentation.

As stated by Security Expert Morrell, security of data warehouse is an integral part for being innocuous. However, one should also take precautionary measure of developmental cybersecurity disputes. This can be done by data tracing with proper documentation. This is also critical for rules of compliance of confidentiality regulatory.

As a continuation of Morrell’s statement, every single strand of data and its lineage requires a continuous audit trail to latent complications. It is very important to maintain detailed log of records. Starting from ingestion to usage, and from usage to authentication details and processing details, all should be maintained in a log to avoid missing any unknown threat. Pinpointed authentication information about the people acts as a precautionary measure to backtrack in case of security breach as well as real-time security to avoid any major security breach.

f) Huge dataset
Data warehousing system is already a set with huge data processing in a data warehouse. In such a scenario, data redundancy or data replication are not stringently required for future system, and failure can act as an overhead security concern. Rather securing networking with remote data and securing the same could provide better security.

As stated by Morrell, keeping several replicas of data floating for the system without proper trail cannot be considered as a way to assure data security. Hence, secured system and definitive measures are applied on cloistered data used for analysis to minimize Online surfacing and allow minimum public functions access to it. This eradicating security concerns due to insecure replicas.

4.1.1.4 Precautions needed to be taken
Will the above-mentioned issues and procedures thwart all data warehousing security susceptibilities? Perhaps not. They can only act as a base for a good start but it is rather important that all the firms should consider and focus seriously on data warehousing system security. Ease of data management along with several cost-saving processing methodology is the major reason for expanding popularity from large organizations to smaller and medium sized as well. Data analytics can be further broken down into service of data digging and data collection, which is nowadays being assisted by cloud-based stowing services. Nevertheless, the issue of
confidentiality and other security-related threats is still surrounding the integrated system of data warehousing and cloud-based storage system.

Old or traditional algorithm-based classic security application, which are designed to perform only on a particular volume of data, cannot handle this large volumes and leads to data leakage and other security and confidentiality threat. Dynamic data like that in stock markets that can be considered as of major importance in a nation’s economy, also cannot be handled by these traditional security applications. Therefore, just a consistent security check will not be able to detect security blotches for constant streaming data. For this scenario, you need full-time seclusion while data streaming and data warehousing system analysis [3].

i. Security of transactional logs and shielding
Due to unavailability of data trails of their storage location, which is generally a fault of auto-tiering technique, more and more such encounters happen. Even though sensitive records like logs and other transactional records have erratic level of security but without shielding it is all inadequate. Instances like necessitating of auto-tiering system to manage the amplified accessible and scalable data due to huge transmission of data being stimulated by IT executive create an issue that is tough to handle at later stage.

ii. Percolation of input from end-points and validation
For data warehousing systems, front-end or end-point devices are the primary factors for preservation. Input data received from end-point are stowed and treated, and other obligatory tasks are accomplished. This is the major reason for acquiring legitimate end-point that is reliable to the maximum extend.

iii. Framework for fortifying dispersed calculation and procedures
Security arrangements and other fortifications in framework for dispersed calculation like MapReduce (which is a utility in Hadoop) is mostly lacking for computational safekeeping and other digital resources. Major preclusions are shielding the data and safeguarding the mappers in the manifestation of an unauthorized attacker.

iv. On premise of real-time data safeguarding and shielding
Large volumes of data generation were the major reason for most groups that were incompetent to preserve consistent validations. Nevertheless, favorable condition would be to execute security checks and scrutiny in real time or virtually in real time.

v. Encryption and defending access control method communication
To defend data from theft, a fortified data storage device can be considered as a smart step. Yet, encryption is still one important parameter in data stowing devices as vulnerability can arise anytime in any device.
vi. Data provenance
To categorize data facts, it is obligatory to be cognizant of its basis in mandate to regulate the data source precisely; authentication, endorsement, and access regulator could be achieved.

vii. Microlevel assessing and access controls
Low-level authentication control of data warehousing system by databases like NoSQL D.B or the Hadoop Distributed F.S entails a proper and updated validation methodology and obligatory authentication. Scrutinizing diverse records could be expedient and these statistics could be obliging in distinguishing any kind of security breach or other malevolent activity.

4.1.2 The result
Roughley stated that, “We can initiate to aid individuals with the methodology to extract the utmost value from their data, the way to treasure an economical supply of electricity or other energy source, benefits while acquire a bank mortgage, even advance and enhanced medic services using the data shared by their fitness tracker. However, we need to acknowledge the actual point that we have all become entities in the Data Warehousing system with gradually being habituated to data analytics and data sharing.”

To put it simply, there are numerous ways that data exposure can be secured, let it be from securing and updating the servers up to date with security servers. Essentially following a virtuous line of disclosure for customer records, and using it for real worth ought to find a solution to the modern present-day problem.

4.2 Hacking
Hacking is an endeavor to abuse a computer system or a remote grid inside a larger network. In simple words, it is the illicit access to or control over computer grid-safety systems for some forbidden drive. The party engaged in hacking deeds are branded as a hacker. These hackers may alter structure or security topographies to achieve an objective that diverges from the original drive of the system. Hacking can also denote to nonmalicious actions, generally concerning scarce or improvised variations to equipment or processes.

Ethical hacking signifies the act of tracing flaws and vulnerabilities of workstation and information engines by duplicating the resolution and activities of malevolent hackers. Ethical hacking is also acknowledged as penetration testing or intrusion testing. An ethical hacker is a security expert who smears his/her hacking
abilities for defensive tenacities on behalf of the possessors of information systems. By piloting penetration tests, an ethical hacker gazes for answer to the following four basic questions like information/locations/systems: can an attacker gain access, can an attacker see on the target, the value of available information to attacker, and is the attempted hack recorded in the targeted system?

Hackers deploy the range of modus operandi for hacking, including

- Spoofing attack: It comprises of sites that fabricate information by imitating legitimate websites, and they are consequently treated as trustworthy sites by users or further programs.
- Vulnerability scanner: These types of programs scan remote computers on grids for known vulnerabilities.
- Password cracking: It can be termed as the method of retrieving passwords from information stockpiled or communicated by computer systems.
- Viruses: These are self-replicating set of codes that spread by injecting replicas of themselves into other executable programs or documents.
- Packet sniffer: These can be defined as those applications that seize data packets with intentions of viewing information and passwords in transit over the networks.
- Root kit: They epitomize a set of code packages that grind to sabotage functionality of an operating system from legitimate operators.
- Trojan horse: It functions as a back-door in a computer system to permit an intruder to achieve access to the system in future.
- Key loggers: These are the tools deliberated to record every keystroke on the infested machine for later retrieval.

Certain organizations hire hackers as part of their upkeep staff. These authentic hackers also recognized as ethical hackers or white hat hackers that use their capabilities to discover faults in the syndicate’s security system, thus averting distinctiveness in individuality larceny and further computer-linked delinquencies. White hat hackers are typically perceived as hackers who use their expertise to assist people. They may be rehabilitated black-hat hackers or may merely be well proficient in the procedures and practices used by hackers. An organization can employ these professionals to test and implement best methodologies that make them less susceptible to malicious hacking efforts in the future.

4.2.1 Big data versus ethical hacking

While the syndicates currently are converging on exploring data warehousing system and analytics because of economical stowage, reachability, usability, and conception of distributed computing, they unknowingly also create a prospect for hackers in social engineering as well. A technique wherein the hacker can know the
inclinations and interests of employee in the enterprise that can assist in constructing an efficacious social engineering attack. For example: With the predisposed data warehousing system of the employee, the records can be excavated easily, whose sites are frequently logged by the employees and the frequency of stopover to the given site (Facebook, YouTube, etc.). With this information, a naive hyperlink in a spam e-mail can be twisted to disclose not only his individual minutiae but can also be enticed into providing corporate authorizations and thus providing numerous accesses to the hacker.

Currently, data warehousing system and networks deliver “Just-in-time” backing for governments, syndicates, and officialdoms during crises. It will also protect forthcoming scenario of national and international network security, new procedures of sovereignty. It also enriches the thoughtfulness of use, abuse, and networking of broad topical. These statistics if in mischievous hand can be a base point for taking down an entire region or government off-guard.

Without the encircling span of data warehousing system, it’s tough to conceive a scenario where dappled ventures and borderline illegal acts would make news. It’s only with the application of data warehousing system does the utter scale of this evidence turn heads. If one individual gazes at another individual’s sheet during a test, it’s commendable of a red mark from the professor. If the whole class cooperates in an organized way and develops a coordination of cheating, it becomes newsworthy. The Panama Papers for illustration are an admirable example of something that is not obligatory illegally, but sketchy to say the least. The element that hundreds of high-profile global figures were acknowledged in this mass dataset is what makes the news. With the evolution of data warehousing system, it makes opportunities for hackers even more appealing, but it also creates a pool of data that becomes even more necessary to protect [4] (Table 4.1).

Table 4.1: Instances of harms versus benefits of data warehousing system.

<table>
<thead>
<tr>
<th>Scenarios of issue</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incursion of cloistered communications</td>
<td>Shared and political engrossment on very enormous scale</td>
<td>Very truncated hurdles to intervention</td>
</tr>
<tr>
<td>Unrestricted revelation of anecdotal cloistered specifics</td>
<td>Analytics illuminates to enhanced and well-timed treatments in the healthiness domain/commercials that might be concerned to you, and so on</td>
<td>Analytics can conjecture peculiar actualities from innocuous feedback data</td>
</tr>
<tr>
<td>Tracing, stalking</td>
<td>Location sharing can be used for triangulation, judging shorter routes, proximate allies, even evading natural calamities, and others</td>
<td>A criminal can utilize the info to raid house when empty</td>
</tr>
</tbody>
</table>
With data warehousing system set, hackers may possibly destruct or yank data warehousing system sets with reasonably trivial alterations in instruction to achieve benefit. Certain techniques might be anodyne to the community but hackers might even exploit annual economic corporate reports for individual advantage. Such vicissitudes in monetary reporting models might also alter the policymaking of management, investors, dealers, and further people who build their verdicts on these monetary reports.

Industries like Equifax, which is one of the distinct consumer credit agencies, functions on multibillion-dollar statistics advisor industry, which acts as a perfect example. They decorate an exhaustive depiction of an individual’s life and that sketch is utilized to style resolutions with direct impressions. As a corporation swells to its stockpile of data, the worth matures exponentially; so, the imperative of dataset traders is to uninterruptedly hoard as much data as conceivable.

In nearby impending, hackers might have the capability to intrude into workstations that pedal vital technological paraphernalia that regulates water distribution, rail networks, gas distribution, and so on. By gaining admin access of this workstation, hackers can alter the operational configurations or manually construct anarchy. This would have a disparaging consequence. Grounded on the research steered by specialists, a point was established that this was undeniably conceivable. As per reports, events of such potential have not achieved the public news, yet it’s a possibility that it could have occurred already.

Thankfully, the good people are keeping up and developing strategies to thwart modern cyberattacks. Let us compare how cyberattacks have traditionally been detected and how data-centric menace revealing system is updating the cybersecurity sphere, leading safekeeping enterprises to design a contextualized and analytical slant to threat recognition system.

4.2.1.1 Scalability and data amalgamation: to detect infringements, you have to validate each piece of data

Customary security incident and evidence management software was not capable enough to accumulate ample and adequate information to perceive up-to-date, erudite infiltrations. Furthermore, although they utilize chronological data, most of them do not have the storage or handling competences to scrutinize data later than 30 days, which leads to overlook significant idiosyncrasies. Additionally, these tools scrutinize diverse cradles of data discretely rather than in conjunction with one another [5–7].

Updated tools that have occurred take into account the speed, size, variety, and complexity of data in a mandate to distinguish the new era of cyberattacks. The fresh paradigm appeals for layering predictive analytics and machine learning systems on cream layer of all cradles of data in an organization’s cyberinfrastructure (Figure 4.2).
4.2.1.2 Well-designed conception is crucial

Pictorial illustrations of infrastructure statistics can assist in making security exposures visible. Conversely, present-day safekeeping mavens are not well proficient in data conception. Stereotypically, their prescribed training includes just statistics, computer science, and security. In such circumstances, if records are detained...
across much longer time horizons and from several disparate sources, well-de-
dsigned visualization becomes indispensable to threat scrutiny.

Companies that use data conceptualization tools have customarily utilized
them for post-destruction design and not for real-time threats monitoring. If plat-
forms are integrated and paired with streamlined visualization, users can swiftly
and accurately pinpoint system susceptibilities [5, 6].

4.2.1.3 Smaller companies more exposed due to unaffordability of cybersecurity

Traditionally hackers used to hit substantial establishments with comprehensive
cyberattacks envisioned to disorder huge number of systems and make headline
news. The modern cyberattack, however, has more low-profile outbreak on confi-
dential records with the intent to go undetected. Small-scale corporations are most
exposed, as they can’t afford to implement and manage tech that traces the foot-
print of the data warehousing system over the endpoints of their organizations.

The artificial intelligence and human expertise for monitoring are not necessar-
ily booming costly, but the hardware for treating of such gigantic volumes of data
might be exceedingly exorbitant. Thus, the security tactic should majorly rely on
the worth of the chattels that need to be protected [10, 11, 12].

4.2.1.4 Fundamental challenges to combating cybersecurity coercions

Day by day, malware outbreaks intensify in volume and intricacy; they are grim for
traditional diagnostic tools and arrangement to tackle them because of majorly two
factors: scalability and data density.

For example, each day at Sophos Labs, more than 300,000 new potentially mis-
chievous files require scrutiny, and SQL-dependent infrastructure will not scale
well and has high maintenance cost [7, 8].

4.2.1.5 Data warehousing system analytics as a path forward to cybersecurity

Detection of hacking attempts and countermeasures to instantly respond is a major
focus area. Prevent, detect, and respond are collectively called as PDR paradigm. This
can be considered as the doors where data warehousing system analytics comes in.

Corporations and analytical firms are now confirming that these encounters might
probably be overwhelmed with data warehousing system analytics. Investigative cor-
porations have been scripting reports and counseling their patrons about the impres-
sions of data warehousing system analytics on cybersecurity across diligences: For
example:
– IDC pinpoints that cloud and data warehousing system will avert cyberthreats to the health organizations.
– According to Gartner, one-fourth of universal corporations has already adopted methodology of data warehousing system processing [13].

4.3 Social engineering

Human user interface being used in various manipulative ways to accomplish mischievous activities with vivid range can be explained as a definition for social engineering. Social engineering, in the milieu of information security, discusses disclosure of confidential info with psychosomatic influence of people. Category comprising assurance tricks for the system access, information congregation tenacity, or deceit diverges from a customary classic “con” in that it is frequently one of several steps in a more intricate fraud structure.

Social engineering can also be explained as concomitant with the social sciences, which are deed of psychosomatic influence of a human, but this recent security vulnerability has surrounded the information security experts since few decades. The cognitive biases are the basis of entire social engineering procedures that are pinpointed on explicit characteristics of human judgment, sporadically referred to as human hardware bugs [14].

Numerous combinations of mix and match are used to generate attack techniques. The assaults cast off for this attack are utilized by the hackers to snip secluded data of the users. There are several examples of this type of attack like the one in which the user is called in their cell phones by people posing as bank employees to fetch their card details for malicious transaction or those in which a mail is sent to the user with a link to click which will in turn take the user to a malicious infected page to load virus into the system, and so on.

Generally, these types of attacks are transcripted with one or several bookmarks. Out of those steps, for a well-defined carefully planned attack, the first step is generally the homework on the victim to collect circumstantial records that range from user’s likes/dislikes to chalking out entry points via carefully validating the security guidelines and protocols.

In the next step mostly, the preposterous person tries to gain victim’s trust to provide him the stimuli which will trigger the actual attacking actions like breach of security protocols and capturing subtle information.

4.3.1 Lifecycle of an well-organized attack

This can be considered as one of such attacks with no real-time protections except maintaining logs, which will lead to finding out later. It basically acts in manipulation
or other humanoid faults, which let them as insider’s access without directly acting upon the intrusion system for the hackers. Since inside attacks cannot be fully predicted based on human manipulation of authentic users, it is more tauter to recognize such attacks, let alone the real-time protection. These types of attacks can only be tackled when users are trained thoroughly with all the modes of attack and how to be precautious for them (Figure 4.3).

4.3.2 Types of social engineering

There are several and diverse methodologies, which along with human collaboration give shape to the attacks. Generally, these can be classified into five broad types of assaults [15].

i. Baiting

As a fish is caught with the help of a bait or a rat catcher uses a bait to trap it, same is this type of attack where the greed or curiosity of the victim is used as a bait to provide a false assurance. This greed or curiosity either lands them right into the deceptive trap compromising their personal information or wide opens their workstation for viruses. The baits are generally having an authentic look, which provides the victim with false assurance. Physical media is the mostly used form to disperse these types of malwares.

Such a scenario like a malware infected flash drive can be considered as an example, which contains the bait suitable for the target user. Due to inquisitiveness,
victim uses the flash drive in any workstation, thereby providing straight route for the malware to infest the system.

These types of attacks are not confined to physical world only; advertisements and other lucrative links to download any software act as a form of online bait. The baits are mostly generalized form and not targeted to any particular user.

ii. Scareware

All the online users have generally seen or faced scenarios where multiple alarms suddenly pop up in the browser or system. Series of fictitious threats are bombarded in the system. This type of attack is termed as scareware, where the victim is threatened in a cyberway to make them believe that their system is compromised and/or is infested with malware. This leads the user to actually download a software suggested by the attacker, which is the real payload for the attacker to compromise the system. So in short, a rogue scanner software or deceptive software that threatens the user to act according to the attacker can be termed as scareware.

Figure 4.4 can be considered as one of the most common scenarios being encountered by almost every Internet user, where popup banners with utmost legitimate looking banners are bombarded in the browser. These popups generally have threatening messages or texts like the one in Figure 4.4. The users are forced to install malicious software or click a link that redirects them to a payload containing site to compromise the system [9].

![Figure 4.4: Example of a scareware.](image-url)
Spam emails with threat and warnings are a mode of operandi for this type of attack, which lures the user to spend on worthless products.

### iii. Pretexting

In this type of attack, series of well-planned manipulations are crafted by an invader to acquire information of the victim. The perpetrator often instigates the attack by pretending as someone else to the victim to requisite classified data to accomplish the assignment. All varieties of apposite information and records are congregated utilizing this swindle like as SSNs (Social Security Number) can be considered as input or output for this type of attacks.

In a classic mode, invader kicks off the attack according to the following steps:

- Imitates as colleague, law enforcement agency, bank and tax officials, or other entities that under specific circumstances having authority-level access.
- Enquires about classified, important but partial information of the victim to avoid major doubts to the victim.
- Uses the data received to data mine the rest of the classified and more important data that can harm the victim in a major way.

### iv. Phishing

One of the online’s most prominent and prevalent type of manipulation attack dependent on directly reaching the user via mailbox or messaging services can be defined as attack style of phishing. It depends majorly on the human tendency of receiving free services or earnestness or sense of distress. It focuses on a better form of a lie in which subtle info is spit out to the victim to generate the sense of curiosity or urgency, thereby leading them to clicking the malevolent link in the mails or chats, which redirect them to payload pages or attachments.

As shown in Figure 4.5, using an electronic mail false sense of affection or caring is injected in the user along with curiosity of knowing the identification of the source showing the affection. The link that shows a greetings being shared by an unknown user actually leads to a payload-containing website that is to be triggered as soon as the user navigates into the webpage. Once the payload is installed, the user falls on the mercy of predator only.

These types of attacks are generally send in a mass to huge set of receivers, with almost similarity to the original links, and regularly updating the mail servers with information from security platforms can actually help the admins to obstruct these types of attacks.

### v. Spear phishing

Since the phishing attack is more generalized and can easily be obstructed, it does not have any specific target. The modified version is also available in manipulation attacks where phishing is specifically directed according to a chosen victim that can be an individual or a member of any large syndicate. They follow the below steps:
Selection of a victim

Datamining more information about the victim like hobbies or interests, and job-related information to make attack less suspiciously.

Closely monitoring the victim to initiate attack in a proper time to attack with maximum success rate.

These types of attacks are generally long duration attacks but are ample tough to sense out and have enhanced triumph rates.

These types of attacks can be visualized as any assailant impersonating as an employee of the same organization as the victim but with higher authority or access to emergency services. After proper background studies and proper timing, a message is delivered by the assailant that are mostly urgent or emergency routine services which needs their authentications or other important details. The information shared by the assailant like victim’s supervisor name and all are retrieved by the assailant during the prerequisite data mining, thereby forcing the victim to believe the authenticity of the call and disclosing all classified details or dispatching them via any web link.

4.3.3 Big data versus social engineering

Social schmoozing platforms are groundbreaking platforms because of their role in transitional behavior among users and third parties with their business orientation.
Entities analyze the users’ data to operate commercial campaigns and, in lieu, foster the financial development of the platform itself, thus subsidizing to comprehend the visions of Internet pioneers, that is, to cultivate a digital grid where information is free and can be utilized for the well-being and the financial development of the entire humanity.

Not surprisingly, however, data analysis might be easily misused, for instance, by exploiting the detailed information about users toward morally questionable objectives (e.g., tailored persuasion techniques, for which we refer to another post of this blog). In addition, once disclosed to the acquiring party, data are not anymore in possession of the social network and, as such, might be illicitly forwarded to other parties. Given this scenario, we try to briefly explain what are the current capabilities and consequences of such capillary data production and analysis, that is, how much can be done starting from our digital shadow?

Nowadays, the combination of psychology and data analysis is so powerful that 70 likes on Facebook are enough to infer more about a users’ personality than what their friends know about him; 300 likes are enough to know that user more than his partner. Hence, online social networks are such privacy-invasive that there is almost a coincidence between the daily life of a person and their digital shadow. Artificial intelligence techniques are the today’s state of the art in many data analysis tasks and, while already performing excellently, their growth is not expected to stop [16].

Considering that the Internet is widespread at any level of our lives, with the online social networks acting as a giant magnifying lens on the society, and being particularly suitable to foster the political discussions, the inferences performed on our data should raise serious concerns. Data might be used to profile users, to encounter them in a much-tailored fashion, and consequently, leveraged to induce them doing something they would not do in their own to perform social engineering to the extreme, precisely. The more is known about users, the easier is also to employ persuasion techniques to propose them exactly what they like, or are scared of, thus opening the doors for a plague of our time: the widespread diffusion of fake news, which, in turn, have detrimental effects on the democracy of a country. In fact, a group of attackers with sufficient available resources can spread misconceptions and fake news on a global scale to influence the results of huge events by hacking the voters (which ironically has the same effect of vote rigging!) [9].

Very recently, the case of an alleged misuse of data carried out by a company operating in the marketing sector, named Cambridge Analytica, came under the spotlight of the media. It is a case worth discussing because it embodies much of the issues described throughout this post. First, some details about the fact: Cambridge Analytica is accused to have been involved in an illicit sharing of data with Aleksandr Kogan, a researcher who developed a Facebook-based application to gather information about users’ personalities [17, 18]. Before 2014, Facebook’s rules about data sharing were not as stricter as they are now. Specifically, a user allowing to disclose some of his/her data had also the capability to reveal pieces of
his friends’ information. In this way, from the 270K users who deliberately shared their data with the application, it had been possible to profile up to 50 million American electors. With such information in hands, Cambridge Analytica is accused to have performed microtargeting campaigns to favor the election of Donald Trump, by employing unscrupulous means, such as the spread of fake news to create a significant shift in public opinion (Figure 4.6).

![Sample of Cambridge Analytica analysis report.](image)

Figure 4.6: Sample of Cambridge Analytica analysis report.

In our view, four main lessons should be learnt from this story:

Today’s data-driven business models come at the cost of sacrificing privacy and require a high level of trust on the entities managing our data. Once data have been disclosed, in fact, there is no guarantee that the party that is entitled to use them (e.g., the legitimate application) illegally forward them to other entities or not.

Although rules are mostly imposed to limit the control that users have on their friends’ information (as Facebook did in 2014), the issue is inherently present in online social networks, since they are based on the friends/followers paradigm. Due to this model, in fact, the boundaries among users’ information spaces have become blurred. Just think of a picture where a user is inadvertently tagged. Moreover, it has been shown that a target user’s information (e.g., location) could be accurately inferred from the analysis of the profiles of his friends.

Social engineering benefits from the heterogeneity and volume of the available data, and widely employs persuasion techniques. The data-centric and all-interconnected world we live in represents the favorable scenario for the application of an extreme social engineering, that is, people can be easily profiled, contacted, and
deceived to induce effects that go far beyond the traditional industrial espionage. As a matter of fact, social engineering has the potential to spread ideologies and influence the result of huge political events by exploiting the structure of the democracy itself.

The Duolingo case, as explained in our project also, is an excellent example of how tracking of people’s behavior on a large scale and inferring their behavioral habits is one of the solutions to improve the efficiency not only of the attack patterns, but also of the training systems.

4.4 Conclusion

Data warehousing system analytics is a major boom in current cyber industry. Data warehousing system analytics if used correctly helps in identifying, understanding customers, optimizing according to their needs, science and research, military, and other defense applications. Data warehousing system analytics can help identifying illegal or hacking attempts even from minute data availability. However, on the contrary, data warehousing system can also be used in corporate espionage, spying on people and even alter their decisions (e.g., U.S Elections) and with the rise in social networking applications every details on every individual can be considered to be achieved online in some way or the other.

Due to the above factor, data warehousing system security is one of the major concerns in cyberindustry. As described by Einstein on the context of nuclear energy, tool that can provide major and sustainable development can also be the cradle of foremost devastations. Data warehousing system security can be considered as important in current cyber market.

The data warehousing system, which primarily meant 3V’s now, has been updated to 6V’s, that is, volume, value, variability, velocity, variety, and veracity. Data warehousing system analytics and the related security measures are growing every day and in this chapter an insight has been given for the same. With continuous growth in data volumes and improvement and inclusion of new tools in the market for analyzing the same, in future, data warehousing system security needs to be revamped every single moment along with other methods to identify the hacking attempts as well.

References


[21] [online] Available: https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

