

24 DE ENERO DE 2025



**Universidad
Europea**

DISEÑO DE UNA INGESTA EN HDFS
SISTEMAS INTELIGENTES

JOSE LUIS MEZQUITA JIMENEZ y JESUS PEREZ DE MIGUEL

En esta práctica del módulo 3 de la asignatura Sistemas Inteligentes, se ha desarrollado un diseño para la ingesta de datos relacionados con los nuevos reportes de **vulnerabilidades detectadas en el ámbito de la ciberseguridad**.

A continuación, vamos a resolver las cuestiones indicadas en el enunciado de la práctica:

1. Indicar la temática elegida para el proyecto ficticio de analítica de datos.

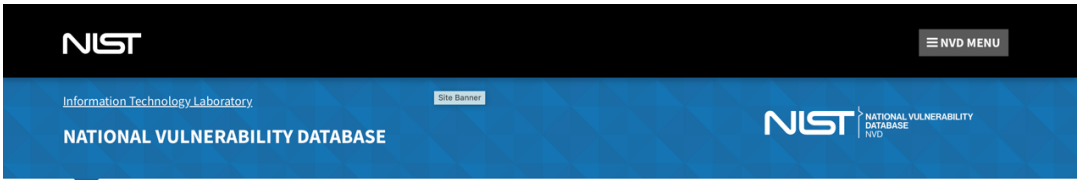
La temática seleccionada para la analítica de datos es sobre las nuevas vulnerabilidades en sistemas informáticos que se van descubriendo.

2. Indicar un mínimo de tres fuentes diferentes (al menos una debe ser real, el resto pueden ser ficticias) y formato de datos de cada una de las cuales se extraería la información necesaria para dicho proyecto.

A continuación, se muestran las fuentes (todas de sitios webs reales):

National Vulnerability Database (más conocida como NIST):

- Fuente: <https://nvd.nist.gov/vuln/data-feeds>
- Datos disponibles: Listado de vulnerabilidades con identificadores CVE, descripción, clasificación CVSS y métricas asociadas.
- Formato: Ficheros en formato JSON.



VULNERABILITIES

NVD Data Feeds

APIs and Data Feed Types

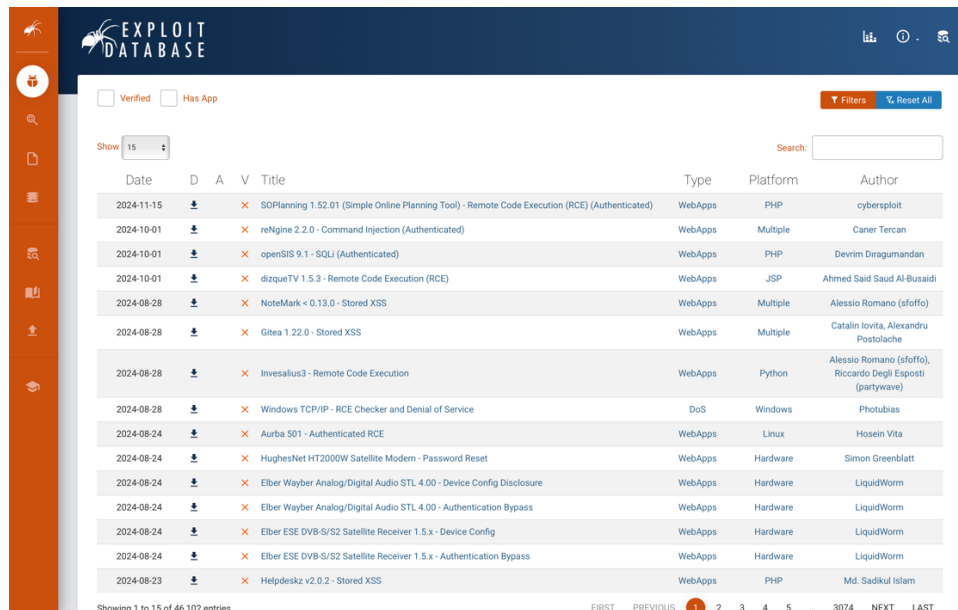
The following table contains links and short descriptions for each API or data feed the NVD offers. Please read how to keep up-to-date with NVD data when using the traditional data feeds.

Users of the data feeds provided on this page must have an understanding of the XML and/or JSON standards and XML or JSON related technologies as defined by www.w3.org.

Type	Description
CVE and CPE APIs	An alternative to the traditional vulnerability data feed files. The APIs are far more flexible and offer a richer dataset in a single interface compared to the JSON Vulnerability Feeds and CPE Match Feed.
JSON Vulnerability Feeds	Each vulnerability in the file includes a description and associated reference links from the CVE* dictionary feed, as well as CVSS base metrics, vulnerable product configuration, and weakness categorization.
CPE Match Feed	A feed that provides the product/platform applicability statement to CPE URI matching based on the CPEs in the official CPE dictionary.
Vulnerability Translation Feeds	Translations of vulnerability feeds.
Vulnerability Vendor Comments	Comments provided by vendors regarding a particular flaw affecting within a product.
CPE Dictionary	dictionary containing a list of products.
Common Configuration Enumeration (CCE) Reference Data	Reference data for common configuration items.

Exploit Database (Exploit-DB):

- Fuente: <https://www.exploit-db.com/>
- Datos disponibles: Información sobre exploits conocidos, incluyendo código fuente, detalles técnicos y plataformas afectadas.
- Formato: Ficheros en formato CSV.

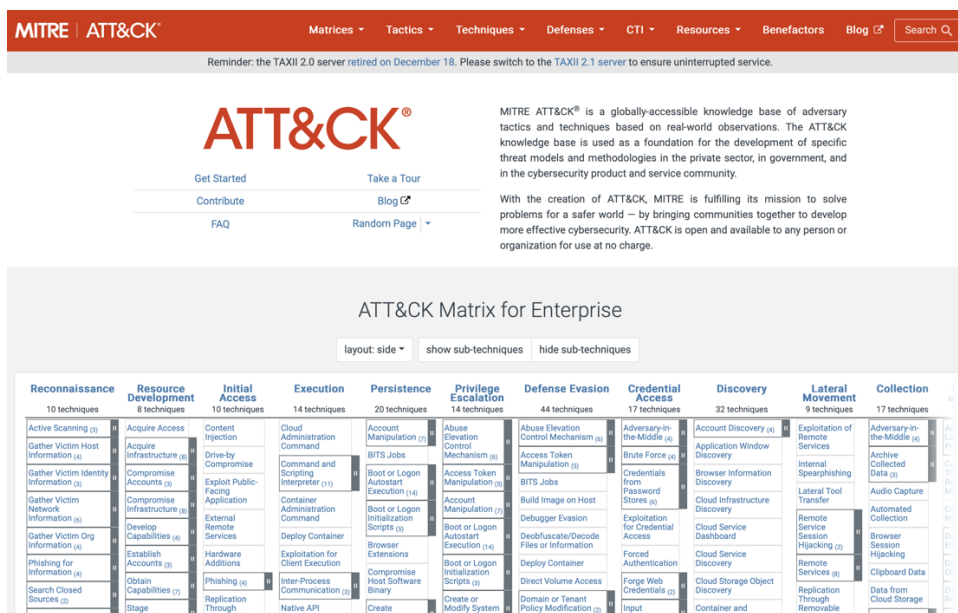


The screenshot shows the Exploit Database interface. At the top, there's a navigation bar with the logo and search filters. Below, a table lists various exploits with columns for Date, D (Download), A (Add), V (Verify), Title, Type, Platform, and Author. The table is paginated, showing 15 entries per page. The first few entries include exploits like 'SOPlanning 1.52.01', 'reNgine 2.2.0', and 'openSIS 9.1 - SQLi'.

Date	D	A	V	Title	Type	Platform	Author
2024-11-15				SOPlanning 1.52.01 (Simple Online Planning Tool) - Remote Code Execution (RCE) (Authenticated)	WebApps	PHP	cybersploit
2024-10-01				reNgine 2.2.0 - Command Injection (Authenticated)	WebApps	Multiple	Caner Tercan
2024-10-01				openSIS 9.1 - SQLi (Authenticated)	WebApps	PHP	Devrim Diragumandan
2024-10-01				dizqueTV 1.5.3 - Remote Code Execution (RCE)	WebApps	JSP	Ahmed Said Saud Al-Busaidi
2024-08-28				NoteMark < 0.13.0 - Stored XSS	WebApps	Multiple	Alessio Romano (sfofo)
2024-08-28				Gitea 1.22.0 - Stored XSS	WebApps	Multiple	Catalin Iovita, Alexandru Postolache
2024-08-28				Invesalus3 - Remote Code Execution	WebApps	Python	Alessio Romano (sfofo), Riccardo Degli Esposti (partywave)
2024-08-28				Windows TCP/IP - RCE Checker and Denial of Service	DoS	Windows	Photobias
2024-08-24				Aurba 501 - Authenticated RCE	WebApps	Linux	Hossein Vita
2024-08-24				HughesNet HT2000W Satellite Modem - Password Reset	WebApps	Hardware	Simon Greenblatt
2024-08-24				Elber Wayber Analog/Digital Audio STL 4.00 - Device Config Disclosure	WebApps	Hardware	LiquidWorm
2024-08-24				Elber Wayber Analog/Digital Audio STL 4.00 - Authentication Bypass	WebApps	Hardware	LiquidWorm
2024-08-24				Elber ESE DVB-S/S2 Satellite Receiver 1.5.x - Device Config	WebApps	Hardware	LiquidWorm
2024-08-24				Elber ESE DVB-S/S2 Satellite Receiver 1.5.x - Authentication Bypass	WebApps	Hardware	LiquidWorm
2024-08-23				Helpdesk v2.0.2 - Stored XSS	WebApps	PHP	Md. Sadikul Islam

Mitre ATT&CK:

- Fuente: <https://attack.mitre.org/>
- Datos disponibles: Información sobre tácticas, técnicas y procedimientos (TTPs) utilizados por actores maliciosos.
- Formato: Ficheros JSON y CSV (se necesita en este caso una API pública).

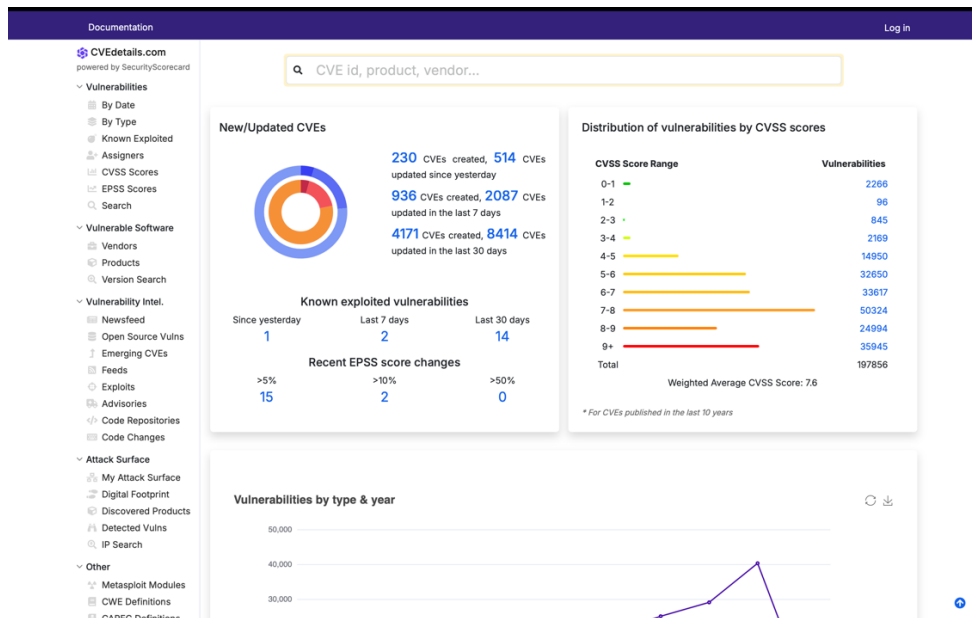


The screenshot shows the MITRE ATT&CK website. At the top, there's a navigation bar with the logo and search filters. Below, there's a section for 'ATT&CK Matrix for Enterprise' with a table of techniques and tactics. The table is organized into columns for different categories of attacks, such as Reconnaissance, Resource Development, Initial Access, Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Lateral Movement, and Collection. Each category contains a list of specific techniques and tactics, along with their descriptions and references.

Reconnaissance	Resource Development	Initial Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Lateral Movement	Collection
Active Scanning (2)	Acquire Access (8)	Content Injection	Cloud Administration Command	Account Manipulation (2)	Abuse Elevation Control Mechanism (6)	Abuse Elevation Control Mechanism (6)	Adversary-in-the-Middle (4)	Account Discovery (6)	Exploitation of Remote Services	Adversary-in-the-Middle (4)
Gather Victim Host Information (4)	Acquire Infrastructure (8)	Drive-by Compromise	Command and Scripting Interpreter (15)	BITS Jobs	Access Token Manipulation (6)	Access Token Manipulation (6)	Brute Force (4)	Application Window Discovery	Internal Spearphishing	Archive Collected Data (3)
Gather Victim Identity Information (3)	Compromise Accounts (3)	Exploit Public-Facing Application	Container Administration Command	Boot or Logon Autostart Execution (14)	Account Manipulation (2)	Build Image on Host	Credentials from Password Stores (6)	Browser Information Discovery	Lateral Tool Transfer	Audio Capture
Gather Victim Network Information (6)	Compromise Infrastructure (8)	External Remote Services	Deploy Container	Boot or Logon Initialization Scripts (6)	Debugger Evasion	Debugger Evasion	Exploitation for Credential Access	Cloud Infrastructure Discovery	Remote Service Session Hijacking (2)	Automated Collection
Gather Victim Org Information (4)	Develop Capabilities (4)	Hardware Additions	Exploitation for Client Execution	Browser Extensions	Deobfuscate/Decode Files or Information	Deobfuscate/Decode Files or Information	Forced Authentication	Cloud Service Dashboard	Remote Service Session Hijacking (2)	Browser Session Hijacking
Phishing for Information (4)	Establish Accounts (3)	Phishing (4)	Inter-Process Communication (2)	Compromise Host Software Binary	Deploy Container	Deploy Container	Forge Web Credentials (2)	Cloud Storage Object Discovery	Remote Services (3)	Clipboard Data
Search Closed Sources (2)	Obtain Capabilities (2)	Replication Through Removable Media	Native API	Create Account (2)	Domain or Tenant Policy Modification (2)	Domain or Tenant Policy Modification (2)	Input Prompt (2)	Container and Resource Enumeration	Replication Through Removable Media	Data from Cloud Storage

CVE Details:

- Fuente: <https://www.cvedetails.com/>
- Datos disponibles: Base de datos de CVEs con estadísticas, gráficos y clasificación de vulnerabilidades.
- Formato: Ficheros en formato CSV.



3. Diseñar la periodicidad con la cual se extraería información de cada una de las fuentes.

La periodicidad en la recogida de información varía entre las fuentes proporcionadas debido a las diferencias en la frecuencia de actualización y el volumen de datos añadidos. Fuentes con actualizaciones diarias, como la NVD, facilitan una respuesta inmediata ante amenazas críticas, mientras que otras, como Exploit-DB, se actualizan con menor frecuencia y ofrecen datos acumulativos que no necesitan extracciones constantes. Siguiendo una práctica muy utilizada, la extracción de datos siempre se realizará durante la noche.

Es por ello por lo que se ha establecido de la siguiente manera:

- **National Vulnerability Database (NIST):** Los datos se obtienen cada día. Descarga del portal web mediante proceso automático cada día a las 02:00 y subida a HDFS.
- **Exploit Database (Exploit-DB):** Los datos se obtienen una vez a la semana. Descarga del portal web mediante proceso automático cada domingo a las 23:00 y subida a HDFS.

- **Mitre ATT&CK:** Los datos se obtienen una vez al mes. Descarga del portal web mediante proceso automático el primer día del mes a las 03:00 y subida a HDFS.
- **CVE Details:** Los datos se obtienen cada día. Descarga del portal web mediante proceso automático cada día a las 04:00 (después de procesar los datos de NIST) y subida a HDFS.

4. Diseñar la estructura de directorios en HDFS en la que se almacenaría dicha información.

Siguiendo el ejemplo mostrado en el enunciado de la práctica, nuestra estructura de directorios seguiría el siguiente formato:

National Vulnerability Database (NIST)

- hdfs:///usuario/ciber/vulnerabilities/nvd/day=20250101/data.json
- hdfs:///usuario/ciber/vulnerabilities/nvd/day=20250102/data.json

Exploit Database (Exploit-DB)

- hdfs:///usuario/ciber/vulnerabilities/exploit_db/week=202501/data.csv
- hdfs:///usuario/ciber/vulnerabilities/exploit_db/week=202502/data.csv

Mitre ATT&CK

- hdfs:///usuario/ciber/vulnerabilities/mitre_attack/month=202501/data.json
- hdfs:///usuario/ciber/vulnerabilities/mitre_attack/month=202502/data.json

CVE Details

- hdfs:///usuario/ciber/vulnerabilities/cve_details/day=20250101/data.csv
- hdfs:///usuario/ciber/vulnerabilities/cve_details/day=20250102/data.csv

El esquema final quedaría de la siguiente manera:

Hdfs:///usuario/ciber/vulnerabilities/

nvd/

day=20250101/ data.json

day=20250102/ data.json

day=20250103/ data.json

day=20250104/ data.json

...

exploit_db/

week=202501/data.csv

week=202502/data.csv

week=202503/data.csv

week=202504/data.csv

...

mitre_attack/

month=202501/data.json

month=202502/data.json

month=202503/data.json

month=202504/data.json

...

cve_details/

day=20250101/data.csv

day=20250102/data.csv

day=20250103/data.csv

day=20250104/data.csv

...

5. Suponiendo que se dispone de una fuente de información extra que sea una tabla de una base de datos, especificar el comando de la herramienta sqoop con el cual importaríamos diariamente datos de esta tabla a nuestros directorios HDFS (ver información del comando en:

<[https://sqoop.apache.org/docs/1.4.6/SqoopUserGuide.html#_literal_sqoop_im port_literal](https://sqoop.apache.org/docs/1.4.6/SqoopUserGuide.html#_literal_sqoop_import_literal)>).

Para importar diariamente los datos de una tabla de una base de datos a HDFS utilizando Sqoop, podemos usar el siguiente comando:

```
sqoop import \  
--connect "jdbc:mysql://internal_db/vulnerabilities_spain" \  
--username=admin --password="password123" \  
--table=vulnerabilities_spain --split-by=day \  
--target-dir="/usuario/ciber/vulnerabilities_spain" \  
--fields-terminated-by=',' \  
--as-textfile \  
--null-string '\\N' \  

```

sqoop import \: Ejecuta el comando principal de Sqoop para importar datos desde una base de datos a Hadoop HDFS.

--connect "jdbc:mysql://internal_db/vulnerabilities_spain" \: Especifica la URL de conexión a la base de datos MySQL.

--username=admin --password="password123" \: Se define el nombre de usuario y la contraseña.

--table=vulnerabilities_spain \: Indica el nombre de la tabla desde la cual se importarán los datos, en este caso vulnerabilities_spain.

--split-by=day \: Indica la columna day que se utilizará para dividir los datos en particiones.

--target-dir="/usuario/ciber/vulnerabilities_spain" \: Especifica el directorio de destino en HDFS donde se almacenarán los datos importados, en este caso /usuario/ciber/vulnerabilities_spain.

--fields-terminated-by=',' \: Define el delimitador de campos en el archivo de salida. En este caso, se usa una coma (,).

--as-textfile: Indica que los datos se deben guardar en formato de texto.

--null-string '\\N' \: Define cómo se deben manejar los valores NULL en la base de datos. En este caso, cualquier valor NULL será representado como \\N en el archivo de salida.