Lab: Suspected DPRK IT Worker (Insider Threat)

Background

ACME Corp. recently hired a remote GenAI developer to help build internal AI-driven tools. The employee was onboarded through standard hiring channels and is officially recorded as being hired in San Jose, California.

The role requires frequent research, experimentation, and collaboration using a mix of developer tools, documentation sites, and productivity software. The company supports remote work and uses Zoom for teleconferencing.

Approved productivity and development tools include:

- Evernote (note-taking)
- Adobe Creative Suite
- Gemini (GenAI research and development)

Concern

Over the last year, multiple governments and private-sector security teams have reported North Korean (DPRK) IT workers obtaining remote technical roles under false identities. These individuals are often highly capable developers but may:

- Work on behalf of the DPRK regime
- Funnel income to sanctioned entities
- Conduct secondary malicious activity such as data theft, IP exfiltration, or enabling future intrusions
- Working for several companies simultaneously
- Using stolen identities and forging documentation with tools like Canva and Adobe Creative Suite
- Leverage remote access tools and Raspberry Pi devices to access corporate laptops in a Laptop Farm
- Using VPN tools to disguise their source location

Unlike traditional insider threats, DPRK IT workers rarely trigger obvious malware alerts. Instead, they tend to exhibit subtle behavioral signals spread across:

- Research habits
- Unusual tool usage
- Data handling patterns
- Financial and cryptocurrency-related activity
- Cultural or language artifacts
- Remote access and monitoring
- Exfiltrate data via RMM and/or unsanctioned note taking tools
- Use of tools unrelated to their role as they complete work for another employer
- Security tool bypass and evasion

No single artifact is definitive. The risk emerges only when multiple weak signals are correlated.

Lab Objective

In this lab, you will analyze Zscaler Internet Access (ZIA) web proxy telemetry to determine whether there is sufficient behavioral evidence to justify escalating the case to formal incident response for a suspected DPRK IT Worker.

You are not hunting for known bad indicators or confirmed malware.
Your task is to:

- Identify patterns and inconsistencies
- Assess policy violations vs. benign curiosity
- Decide whether the aggregate risk crosses a response threshold

How the Lab Data Is Structured

Facts about the dataset:

- Time window of logs: ~2:30pm–3:00pm, January 27, 2026
- Activity represents one user suspected user who triggered DLP alerts
- User is a remote GenAI developer
- Hired location: San Jose, CA
- Dataset includes normal work-related activity mixed with subtle anomalies

Within the provided CSV file (lab_dprk.csv), you will observe a mix of:

- Legitimate development and research activity
- Searches, downloads, and site visits that may be unexpected or misaligned with the role and approved tooling

Analyst Guidance

This lab is intentionally ambiguous.

You should:

- Avoid binary “good vs. bad” thinking. Think like a detective.
- Focus on behavioral consistency
- Ask:
  - *Does this activity align with the stated role and approved tooling?*
  - *Why might someone need multiple parallel tools for the same function?*
  - *What alternative explanations exist?*
  - *How could this activity I’m observing be related to the hypothesis that this is a DPRK IT Worker?*

Hunt Themes to Explore (No Single Indicator Is Enough)

You are encouraged to explore clusters of behavior, not individual events. Questions to help you investigate:

Question	**Answer**
Are there any search results that raise suspicion? (PRO TIP: Don’t assume traditional search engines are the only means of finding information. People consume information in different formats (e.g. audio, video, etc.)	
What software / tools are searched for or downloaded?	
Is the GitHub activity what you’d expect for a user in this role?	
Are there any tools researched or downloaded that could be considered offensive in nature?	
Are there any tools researched or downloaded that could be used for remotely interacting with the system if it were in a	

Laptop Farm?	
How could the suspected DPRK IT Worker join Zoom meetings on camera if they were not physically at the laptop?	
Were any executable files interacted with? If so, are any suspicious?	
Does any of the traffic indicate consumption of Korean-language media?	
Does any of the traffic indicate potential DPRK funding mechanisms or non-conventional transfer of money?	
Are there signs of intentional policy evasion versus curiosity?	
Do you believe there is sufficient evidence to initiate incident response? Why or why not?	

Search History

Most search engines use the `q=` parameter in the URL to pass along the search phrase. Make a list of search engines, for example:

- Google
- Bing
- Yahoo
- DuckDuckGo
- Yandex
- Brave
- Startpage
- Ecosia
- Etc.

Test searching phrases and observing how the URL changes. For example:

Google
Test Phrase: TEST123
URL: `https://www.google.com/search?q=TEST123&...`

Bing
Test Phrase: HELLO WORLD!
URL: `https://www.bing.com/search?q=HELLO+WORLD%21&...`

In Bash, you could use awk to match one of the search engines hostnames, then extract the search phrase, however the command gets complex and the output isn't easy to read.

Shell

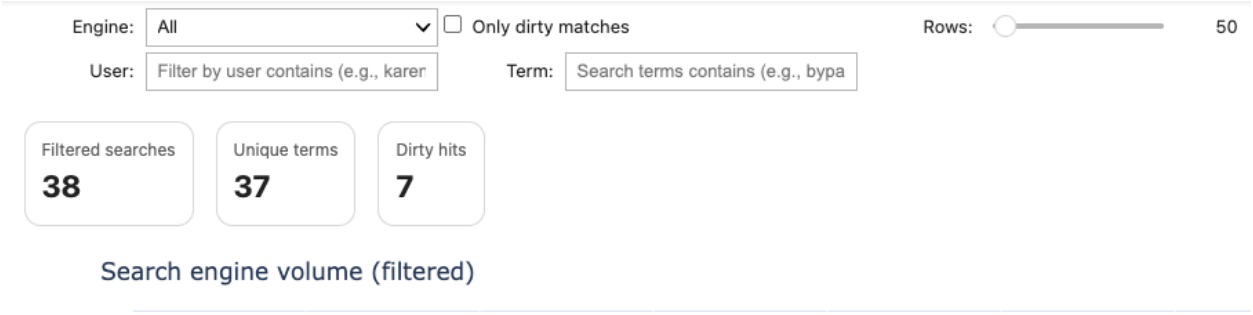
```
awk -F',' '
NR>1 && tolower($8) ~
/(google\.com\/search|bing\.com\/search|yahoo\.com\/search|yandex\.com\/search)
/ {
    url=$8

    if (match(url, /[?&]q=([^&]+)/, m)) {
        query=m[1]

        gsub(/\+/, " ", query)

        while (match(query, /%[0-9A-Fa-f]{2}/)) {
            hex=substr(query, RSTART+1, 2)
            char=sprintf("%c", strtonum("0x" hex))
            query=substr(query,1,RSTART-1) char substr(query,RSTART+3)
        }

        print url "," query
    }
}
' data/lab_dprk.csv
```

This is where Jupyter Notebooks come in handy. In the `/notebooks/` directory, there's a sample Jupyter Notebook (`search_engine_hunter.ipynb`) already set up to identify search engines and extract keywords.

As an added feature to show the flexibility and capabilities of a well crafted hunting notebook, there's also a wordlist (`dirty_wordlist.txt`) in the `/data/` directory with a list of search phrases that might help you catch a DPRK IT Worker. Review the text file and modify / add keywords to the `dirty_wordlist.txt` file. What phrases would you want to highlight in users' search engine activity that could indicate an insider threat?

When you run the `search_engine_hunter.ipynb`, you'll not only see the search terms parsed out for you, you'll also see if any of the transactions match to suspicious search phrase in the `dirty_wordlist.txt` file.

At the bottom of the notebook, you can use the interactive user interface to filter the search engine, users, rows, search terms, or only show searches that match to a phrase on the `dirty_wordlist.txt` file.



Engine: All ▾ □ Only dirty matches Rows: 50

User: Filter by user contains (e.g., karen Term: Search terms contains (e.g., bypa

Filtered searches 38

Unique terms 37

Dirty hits 7

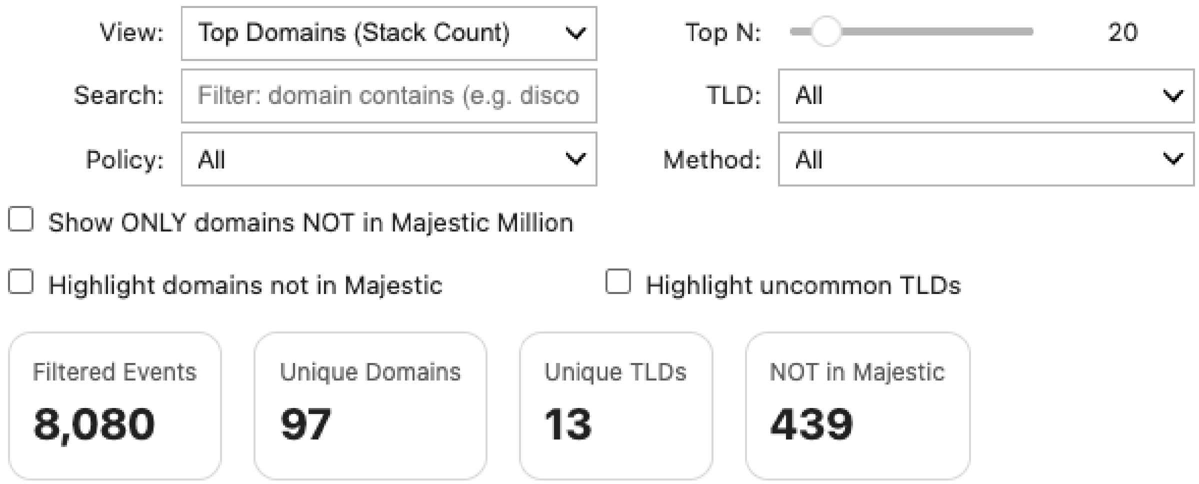
Search engine volume (filtered)

Uncommon Domains

Leaning into the Least Frequency of Occurrence (a data analysis technique, often called "stacking," that identifies items appearing the fewest times within a dataset – used to find outliers, such as rare domains) hunting methodology, we can enrich the telemetry with additional data sources such as the Majestic Million (a list of the top 1 million websites in the world).

In the `/notebooks/` directory, you'll find the pre-configured Notebook (`majestic_million_enrichment.ipynb`), which joins the `lab_dprk.csv` dataset with the `majestic_million.csv` enrichment file.

The Notebook parses out the hostname and top-level domain (TLD) for further filtering and sorting capabilities. Pre-configured views will help visualize the most and least occurring hostnames. Checkboxes can highlight domains NOT in the Majestic Million or uncommon TLDs. This can make spotting outliers quick and easy. Take a few minutes to play around with the Notebook and look for outliers that could indicate an insider threat. Pivot on interesting findings.



View: Top Domains (Stack Count)

Top N: 20

Search: Filter: domain contains (e.g. disco

TLD: All

Policy: All

Method: All

☐ Show ONLY domains NOT in Majestic Million

☐ Highlight domains not in Majestic

☐ Highlight uncommon TLDs

Filtered Events	Unique Domains	Unique TLDs	NOT in Majestic
8,080	97	13	439

Final Decision Exercise

You are briefing leadership.

You must choose one:

- ☐ Continue monitoring, no escalation
- ☐ Targeted internal inquiry (HR / IT validation)
- ☐ Formal incident response for suspected DPRK IT Worker

Be prepared to defend your decision using evidence from the dataset, not assumptions.