DARKOWL

UNDERSTANDING DARKNET DATA AT SCALE

DarkOwl employs a modified model of "Big Data" often depicted by the "V's" of Big Data.

Intro to 'Big Data'

The NIST Data Interoperability Framework defines "Big Data" as a large amount of data in the networked, digitized, sensor-laden, information-driven world.[1] The authors of that framework describe "Big Data" and "data science" as buzzwords that are essentially composites of many other concepts across computational mathematics and network science.

Data can appear in "structured" and "unstructured" formats. According to IBM, not all data is created equal. Structured data is often quantitative, highly organized, and easily decipherable, while unstructured data is more often qualitative, and not easily processed, and analyzed with conventional tools.[2]

2.5 QUINQUINTILLION BYTES OF DATA IS PRODUCED EVERY DAY, **80-90%** OF WHICH IS UNSTRUCTURED DATA.[3][4]

In the last decade the amount of unstructured data available to an individual has skyrocketed. Think about the amount of raw data a person consumes or generates on any given day, through mediums like SMS text messaging, watching, and/or creating YouTube videos, editing, and sharing digital photographs, interacting with dynamic web pages, and keeping up with the demands of social media.

Darknet 101

The darknet is a layer of the internet that was designed specifically for anonymity. It is more difficult to access than the surface web, and is accessible with only via special tools and software – specifically browsers and other protocols. You cannot access the darknet by simply typing a dark web address into your web browser. There are also darknet-adjacent networks, such as instant messaging platforms like Telegram, the deep web, some high-risk surface websites.

Quick Definitions

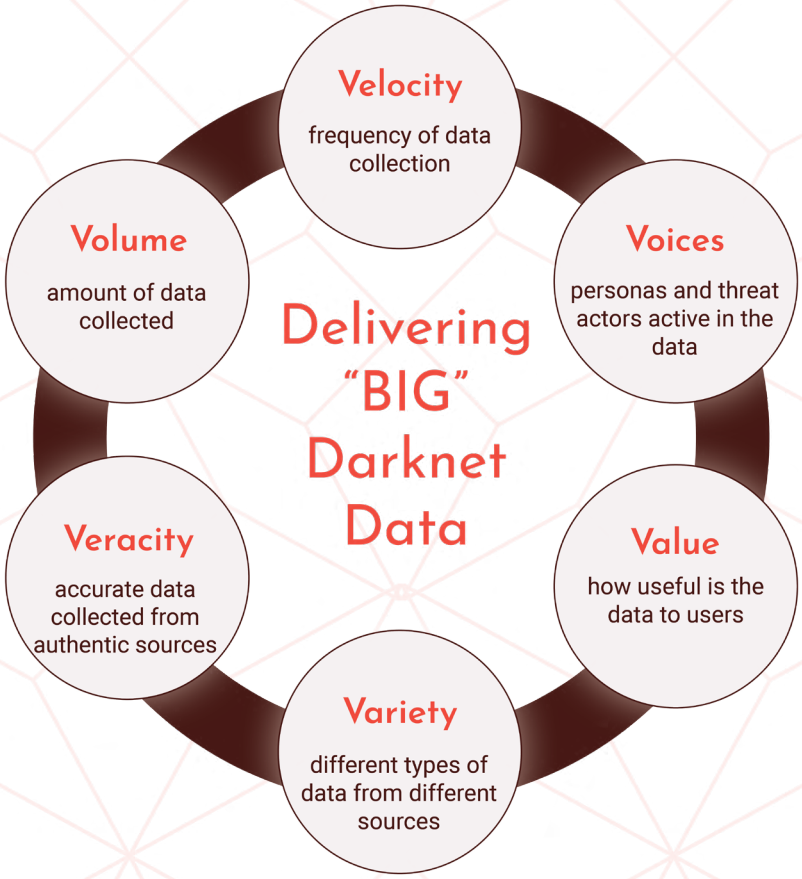
darknet: Also referred to as the "dark web." A layer of the internet that cannot be accessed by traditional browsers, but requires anonymous proxy networks or infrastructure for access. Tor is the most common.

deep web: Online content that is not indexed by search engines, such as authentication required protected and paste sites and can be best described as any content with a surface web site that requires authentication.

high-risk surface web: consists of areas of the surface web (or "regular" internet) that have a high degree of overlap with the darknet community. This includes some chan-type imageboards, paste sites, and other select forums.

Data on the Darknet

The darknet and deep web are vast sources of structured, semi-structured and unstructured data that requires advanced architecture to collect, process, analyze, and distribute meaningful and targeted datasets to clients and users across diverse industry verticals. This includes FinTech, InsureTech, Identity Protection and Threat Intelligence providers. DarkOwl employs a modified model of "Big Data" often depicted by the "V's" of Big Data.



Volume

DarkOwl delivers petabytes of data processed in real time, with crawlers operating across different anonymous networks, deep websites, and platforms. As of this week, our Vision UI has collected and indexed over 215 million documents of darknet data across Tor, I2P, and Zeronet in the last year. Our Entity API has uncovered and archived over 8.8 billion emails, 15 billion credit card numbers, 1.8 billion IP addresses, and over 387 million cryptocurrency addresses.

Velocity

DarkOwl's resources are designed to provide fast and frequent data updates by collecting from real-time instant messaging sources and capturing live discussions between users on darknet forums. In the last 24 hours, our system crawled and indexed over 1 million new documents of data.

Veracity

DarkOwl collects data in its original, raw-text format from legitimate and authentic sources discovered in the darknet, deep web, and high-risk surface web. DarkOwl scrapes darknet data without translation in its native language to avoid contextual loss from automated in-platform translation services.

Variety

The data DarkOwl discovers is disparate from diverse and distributed data sources such as Tor, I2P, ZeroNet, open FTP sites, and chat platforms with instant or new real-time messaging. We collect everything from darknet marketplace listings for drugs and malware to user contributions to forums and Telegram channel messages.

Value

DarkOwl delivers its data in a variety of delivery mechanisms along with our expert insights to help drive high-value business decisions for our clients and stakeholders. Darknet data in this raw format helps provides valuable evidence for qualitative investigations to quantitative risk calculations.

Voices

Darknet data centralizes around the voices of the various personas and threat actors conducting criminal operations in the underground. DarkOwl's Lexicon helps users easily decipher and filter by marketplace, vendors, forums, threat actor pseudonyms, and ransomware-as-a-service (RaaS) operators.

Delivery Mechanisms of Scalable Data

Data Warehouse

A data warehouse consists of mostly structured data that is typically accessed via SQL. Data warehouses are traditionally based on RDBMS technologies such as Oracle, DB2, Postgres etc., and they take a ton of resources to build and maintain, hence the drop in popularity over time.

Data Lake

Data lake consists of a combination of structured AND unstructured data. Mostly unstructured data – as in medical transcriptions, court documents, audio, video, screen shots and so on. The structured data is mostly to tag and link the unstructured data. Data lakes are more popular now due to the ease of creating lakes. Data lakes are supported by cloud native vendors such as Amazon AWS, Google Cloud, Microsoft Azure, etc. DarkOwl can set up custom data lakes that contains a subset of our data, that we give customers access to.

Data Feeds

Data feeding describes the process of pushing parts of our Big Data over to the customer side. For example, we feed only credentials to some customers, or only credit cards to another, and in some cases, provide a daily snapshot of everything that a data provider has visibility into directly to the customer for their own business use case.

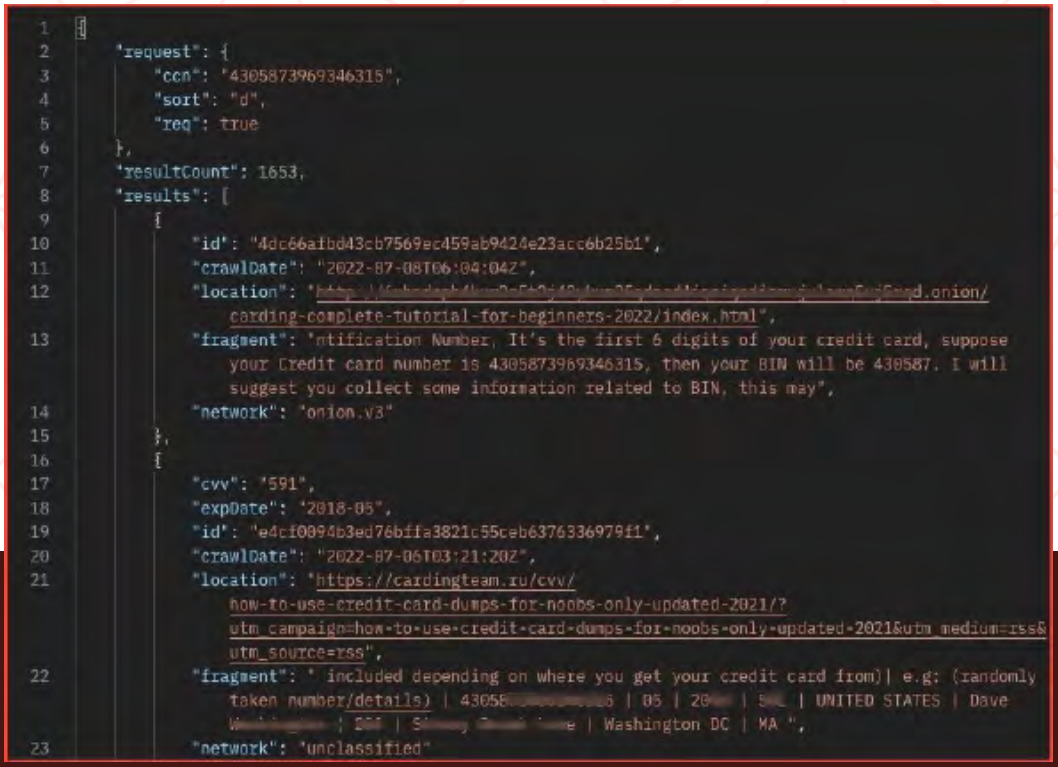


Figure 1 Screenshot of an API response from DarkOwl's Entity API Credit Card Endpoint

Data Streaming

To process data rapidly, DarkOwl uses open-source technologies such as Kafka. Such services are mostly for internal use, but we could easily set up our customer as one of the subscribers to our data stream. This especially makes sense when the velocity of data is very high, which is often the case for darknet data.

For a full list of darknet terms, check out our Glossary of Darknet Terms >>

The darknet is home to a diverse group of users with complex lexicons that often overlap with the hacking, gaming, software development, law enforcement communities, and more. DarkOwl's Glossary of Darknet Terms is a continually evolving resource that defines the common vernacular, slang terms, and acronyms that our analysts find in places like underground forums, instant messaging platforms (such as Telegram), as well as in information security research pertaining to the darknet.

Product Highlights

Vision UI

Search and Monitor the most comprehensive Darknet Dataset

The Vision app is the industry leading platform for Analysts to simply, safely, and comprehensively search the largest commercially available source of Darknet data. Vision provides a user friendly interface with powerful querying capabilities to search, monitor, and create alerts for critical information.

Entity API

Safely Access Discrete Darknet Data Points

With Entity API, users can quickly and efficiently identify, monitor, and target particular threats in the darknet that are relevant to their particular needs and use-cases, including clients who need to see curated content that comes directly from recent darknet posts.

SOURCES

1 https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf
2 https://www.ibm.com/cloud/blog/structured-vs-unstructured-data
3 https://seedscientific.com/how-much-data-is-created-every-day
4 https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data

DARKOWL DATA SOURCES

Tor, I2P, ZeroNet, authenticated forums, darknet marketplaces, IRC, high-risk paste sites, encrypted chat services, and open FTP servers.

DARKOWL

ABOUT DARKOWL

DarkOwl uses machine learning to automatically, continuously, and anonymously collect, index and rank darknet, deep web, and high-risk surface net data that allows for simplicity in searching.

The platform collects and stores data in near realtime, allowing darknet sites that frequently change location and availability, be queried in a safe and secure manner without having to access the darknet itself.

For more information, visit www.darkowl.com