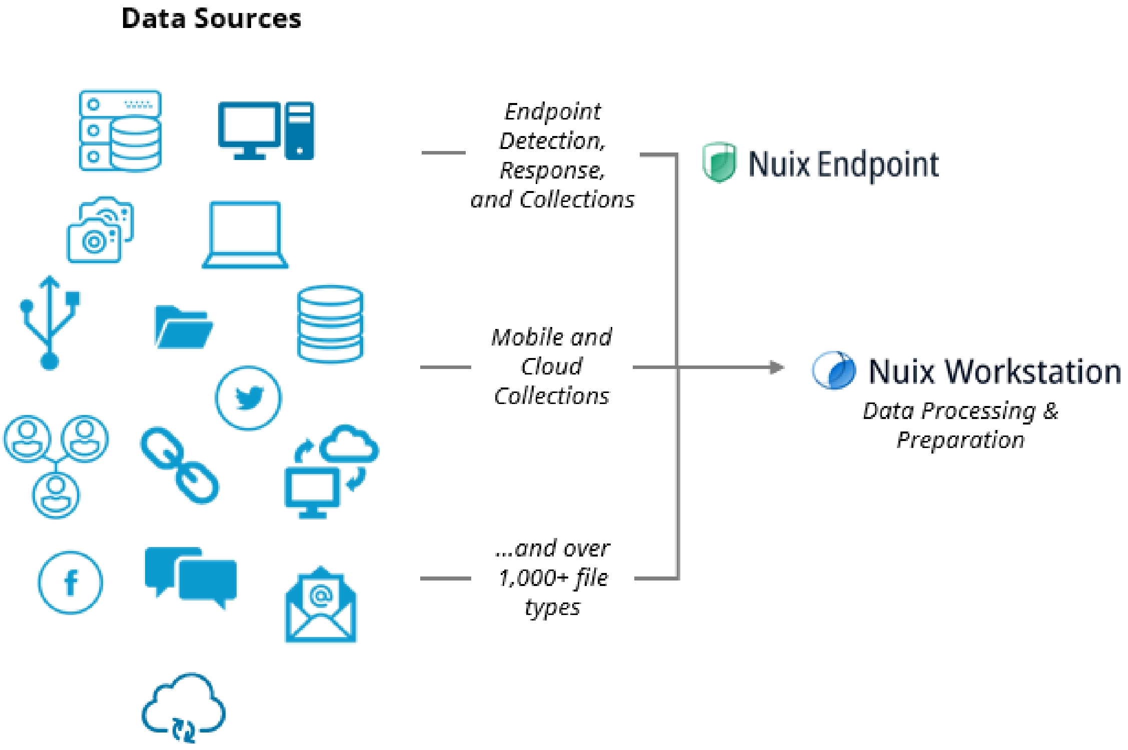
Building Unstructured Data Solutions for Today and Tomorrow: Part 3 –

Processing & Preparation

Written By:
Alex Chatzistamatis

Now that we've covered collections—what I believe is a foundational component to an unstructured data solution—I think it's time to focus on easily the most important requirement: processing.

Continuing with the concept of building a structure (haha, see what I did there?) one could argue that processing is the frame which everything else is ultimately built around. As I'm sure most of you reading this are aware, processing is a make or break stage. The key thing to remember is your processing tool is only as good as the data you plan to consume, regardless of how narrow or broad your scope may be.


Data Sources
Endpoint Detection, Response, and Collections
Mobile and Cloud Collections
...and over 1,000+ file types
Nuix Endpoint
Nuix Workstation
Data Processing & Preparation

FRANKENSTEIN PROCESSING

As I discussed in Part 2 about collections, processing is one of the other steps that can suffer from the *Frankenstein effect*—various technology components being used to achieve the same result. Even though I've seen this time and time again, it still shocks me that many technologists believe there's no immediate solution for 'Frankensteined' technology.

During a recent consultation, I ran out of fingers counting the number of separate tools the organization was using, each being a point solution for a specific data type. Within a few minutes of discussing some of the capabilities that Nuix processing has to offer, the client was blown away with the variety of data that Nuix can process, regardless of total volume, with unmatched velocity, thus proving there IS a better way.

OUR PROCESSING 'WAY'

Some of the most common requirements my clients have when trying to address their processing challenges include:

- Consolidating processing tools in order to do 'more with less'
- Replacing multiple internal or external workflows with a single platform to reduce TCO
- More functionality to handle disparate data types (including forensic images, emails, logs, cloud data, and more)
- Scalability to meet the demand of growing data volumes.

Of course, this is just scratching the surface, but the Nuix Engine is up for the challenge. The end-to-end Nuix platform is an extremely powerful solution for organizations, but the unique and extensible Nuix Engine is the driving force that makes it all possible.

The Nuix Engine enables endless possibilities including the ability to:

- Gain one window into all the data in a single location where you can use advanced search, culling, and filtering capabilities to understand the content and its context to make informed decisions
- Scale to the largest data volumes, no matter the size of the data set; we even can you let begin working immediately while data is processing, saving you time and effort
- Include all possible content and metadata across thousands of different file types, better known as the Ten Dimensions of Data
- Provide flexible workflows with multiple deployment options, extensible APIs, and unique workflows that help maximize productivity.

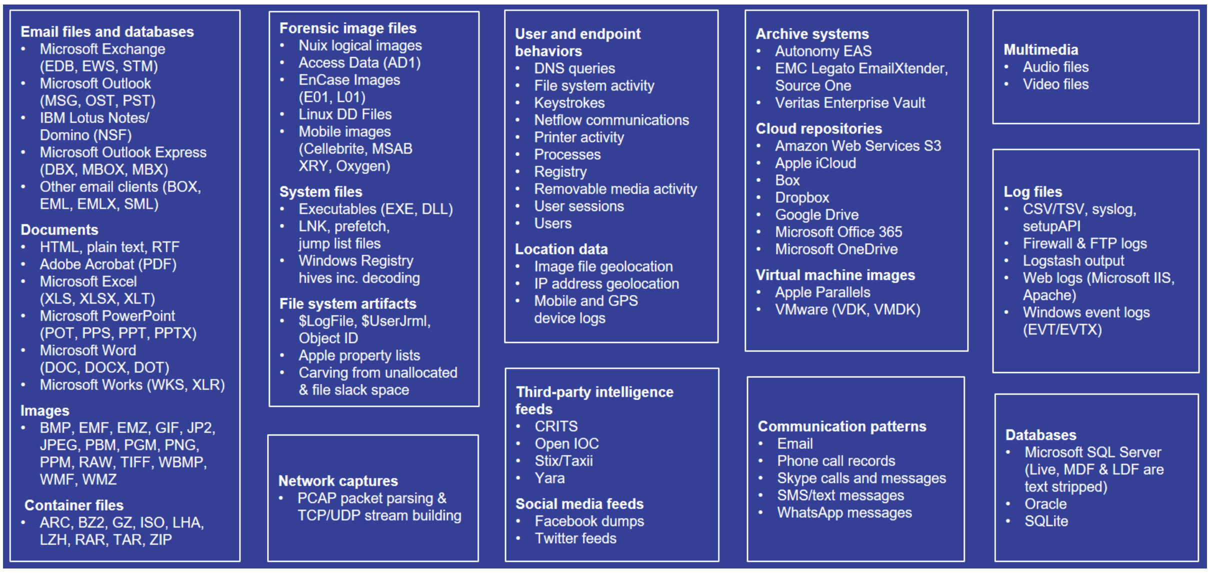
Now, allow me to geek out a bit and highlight some of the additional benefits Nuix processing can provide an organization:

- Native support for over 1,000+ different file types with a growing list of supported file types each release
- Built-in, API-based connectors for Microsoft 365 and Amazon S3, among many others
- Extract text and metadata from items, making both easily searchable
- Expanding contents of containers (PSTs, ZIPs, E01s, and more…)
- Providing granular control over the depth of extractions.

TAKING ON THE BREADTH OF YOUR DATA

Take a minute to review the list of file types below. That's right … Nuix can consume all these different data sources at unparalleled speeds, all while making the data fully searchable from a single pane of glass!

Imagine being able to combine data from endpoints, DLPs, web servers, disk images, user email, mobile device, and chat data from a single location. Pretty cool, huh?



Email files and databases
- Microsoft Exchange (EDB, EWS, STM)
- Microsoft Outlook (MSG, OST, PST)
- IBM Lotus Notes/ Domino (NSF)
- Microsoft Outlook Express (DBX, MBOX, MBX)
- Other email clients (BOX, EML, EMLX, SML)

Documents
- HTML, plain text, RTF
- Adobe Acrobat (PDF)
- Microsoft Excel (XLS, XLSX, XLT)
- Microsoft PowerPoint (POT, PPS, PPT, PPTX)
- Microsoft Word (DOC, DOCX, DOT)
- Microsoft Works (WKS, XLR)

Images
- BMP, EMF, EMZ, GIF, JP2, JPEG, PBM, PGM, PNG, PPM, RAW, TIFF, WBMP, WMF, WMZ

Container files
- ARC, BZ2, GZ, ISO, LHA, LZH, RAR, TAR, ZIP

Forensic image files
- Nuix logical images
- Access Data (AD1)
- EnCase Images (E01, L01)
- Linux DD Files
- Mobile images (Cellebrite, MSAB XRY, Oxygen)

System files
- Executables (EXE, DLL)
- LNK, prefetch, jump list files
- Windows Registry hives inc. decoding

File system artifacts
- $LogFile, $UserJrml, Object ID
- Apple property lists
- Carving from unallocated & file slack space

Network captures
- PCAP packet parsing & TCP/UDP stream building

User and endpoint behaviors
- DNS queries
- File system activity
- Keystrokes
- Netflow communications
- Printer activity
- Processes
- Registry
- Removable media activity
- User sessions
- Users

Location data
- Image file geolocation
- IP address geolocation
- Mobile and GPS device logs

Third-party intelligence feeds
- CRITS
- Open IOC
- Stix/Taxii
- Yara

Social media feeds
- Facebook dumps
- Twitter feeds

Archive systems
- Autonomy EAS
- EMC Legato EmailXtender, Source One
- Veritas Enterprise Vault

Cloud repositories
- Amazon Web Services S3
- Apple iCloud
- Box
- Dropbox
- Google Drive
- Microsoft Office 365
- Microsoft OneDrive

Virtual machine images
- Apple Parallels
- VMware (VDK, VMDK)

Communication patterns
- Email
- Phone call records
- Skype calls and messages
- SMS/text messages
- WhatsApp messages

Multimedia
- Audio files
- Video files

Log files
- CSV/TSV, syslog, setupAPI
- Firewall & FTP logs
- Logstash output
- Web logs (Microsoft IIS, Apache)
- Windows event logs (EVT/EVTX)

Databases
- Microsoft SQL Server (Live, MDF & LDF are text stripped)
- Oracle
- SQLite

How exactly is this possible? It all comes down to the secret sauce in the Nuix Engine and how our users decide to present the data for their organizational use cases. For example, in preparation for litigation, it may be a good idea to perform a proportionality analysis to determine the depth of the extraction required.

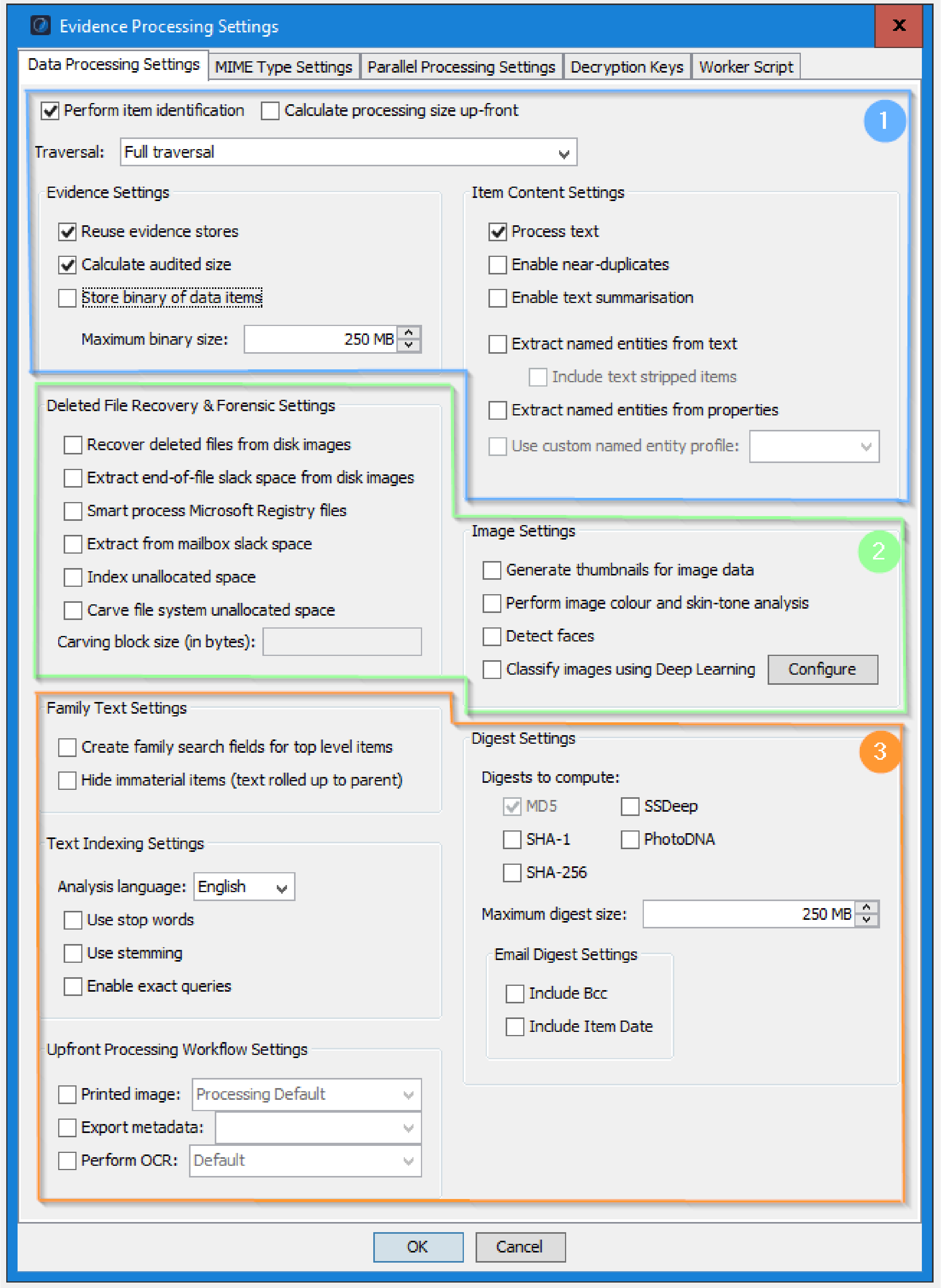
Doug Austin from eDiscovery Today recently referenced a Craig Ball blog in which the topic of advanced settings are considered . Whether you're preparing data for litigation, beginning your forensic investigation, or working on data governance initiatives, Nuix has you covered with a wide range of options.

Below is a screenshot of Nuix Workstation data processing settings, which enable the aforementioned benefits. Specifically, I’d like to call out:

1. **Basic Index Settings** – Control the depth of details
   a. Full content (text), metadata, and the family/hierarchy of files
   b. Flexibility to control storing the item’s binary and other settings to enable near duplicates, topic modelling, and identifying sensitive data.

2. **Forensic Settings** – Control the level of forensic vigour
   a. Recovering deleted files, working with slack space, and file carving
   b. Creating photo thumbnails, skin-tone analysis, face detection, and deep learning.

3. **Advanced Index Settings** – Enrich your index with additional options specific to use-cases
   a. Grouping families together for search, enabling exact queries, leverage stemming, or stop words
   b. Controlling the digests and hash calculations
   c. Streamlining imaging, OCR, and metadata workflow.



Evidence Processing Settings

Data Processing Settings | MIME Type Settings | Parallel Processing Settings | Decryption Keys | Worker Script

☑ Perform item identification □ Calculate processing size up-front

Traversal: Full traversal

Evidence Settings

☑ Reuse evidence stores

☑ Calculate audited size

□ Store binary of data items

Maximum binary size: 250 MB

Item Content Settings

☑ Process text

□ Enable near-duplicates

□ Enable text summarisation

□ Extract named entities from text

□ Include text stripped items

□ Extract named entities from properties

□ Use custom named entity profile:

Deleted File Recovery & Forensic Settings

□ Recover deleted files from disk images

□ Extract end-of-file slack space from disk images

□ Smart process Microsoft Registry files

□ Extract from mailbox slack space

□ Index unallocated space

□ Carve file system unallocated space

Carving block size (in bytes):

Image Settings

□ Generate thumbnails for image data

□ Perform image colour and skin-tone analysis

□ Detect faces

□ Classify images using Deep Learning

Family Text Settings

□ Create family search fields for top level items

□ Hide immaterial items (text rolled up to parent)

Text Indexing Settings

Analysis language: English

□ Use stop words

□ Use stemming

□ Enable exact queries

Upfront Processing Workflow Settings

□ Printed image: Processing Default

□ Export metadata:

□ Perform OCR: Default

Digest Settings

Digests to compute:

☑ MD5 □ SSDeep

□ SHA-1 □ PhotoDNA

□ SHA-256

Maximum digest size: 250 MB

Email Digest Settings

□ Include Bcc

□ Include Item Date

Once you’re locked in with the appropriate data processing settings (which can be turned into an easily reusable template) buckle your seatbelt! Nuix can consume data of all sizes and take advantage of the horsepower it's running on to scale from GBs per hour to TBs per day.

In fact, one of my new clients in the pharmaceutical space was able to process a rather large and complex data set with Nuix in 36 hours as opposed to its older technology, which couldn't even move the needle after close to 1.5 months.

To make it clear, the client recognized an immediate ROI within days of implementing Nuix!

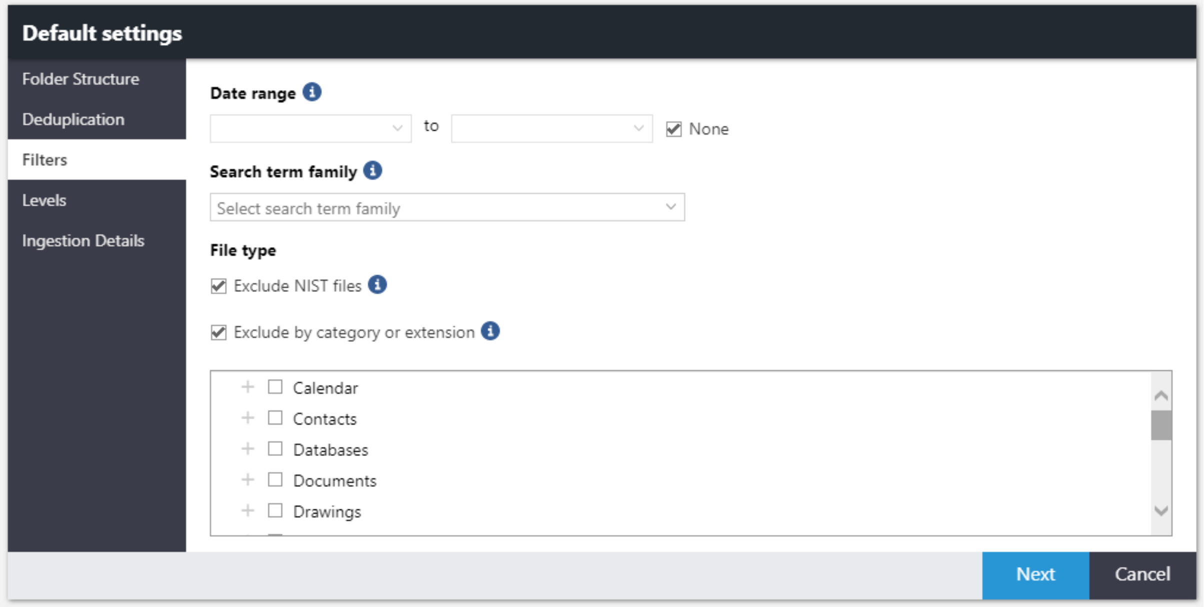
MAKING $@#$ SEARCHABLE

"While processing is a key capability of the Nuix Engine, it is so much more powerful," said Nuix CTO Stephen Stewart. "The Nuix Engine's ability consume all 10 dimensions of data, normalize all the text and metadata into a consistent schema is amazing. Said more straight—Nuix makes $@#$ searchable."

Stephen is absolutely right! The fun really starts after the data has been processed and it is normalized and fully searchable. Nuix provides simple yet powerful ways to prepare the data for downstream analysis and review.

Again, regardless of use case, Nuix has you covered:

- Comprehensive data reduction techniques that include keyword (and advanced/complex) searches, threading, file types, date ranges, de-NISTing, and more
- Granular deduplication by item, family, custodian, and even more customizable logic
- Early data assessment including number crunching, pivot reports, and metadata.



Date range
to
☑ None

Search term family
Select search term family

File type
☑ Exclude NIST files
☑ Exclude by category or extension

□ Calendar
□ Contacts
□ Databases
□ Documents
□ Drawings

WRAPPING UP AND LOOKING AHEAD

Between the volume, velocity, and variety that the Nuix Engine brings to the table to quickly consume various types of data, making it fully enriched and searchable, coupled with the ability to filter and cull hundreds of different ways, it should be pretty apparent why building an unstructured data solution around the Nuix platform makes sense.

Looking back at the some of the processing requirements, here are a few ways Nuix can help solve these challenges:

- Get to the big picture quickly and comprehensively, answer the fundamental questions of any matter, and make early data/case assessments with all the facts at your fingertips.

- Mitigate risks with consistent, repeatable, and forensically defensible process across each item and data source.
- Empower your experts with insights like no other technology, revealing and contextualizing the stories hidden in the data.
- Capture the content, metadata, and context of each item for more than 1,000 file formats and turn it into meaningful information.
- Use a single platform for all your use cases ranging from litigation, forensic investigations, governance, risk, compliance, incident response, and more.

Taking the critical endpoint data capture using Nuix Endpoint (covered in Part 2), we can layer it next to additional unstructured data sources to be fully processed and searchable using the Nuix Engine. Regardless of data type or size, Nuix is flexible enough to scale based on your organization's requirements.

Stayed tuned for Part Four, where I'll continue our unstructured data solution journey and understand how the data that's been collected and processed can be explored and analysed using our software.