# Preparation

“Data” is everything. Without data, there is nothing to investigate, nothing to hunt for. Reliability of the data determines how productive and efficient will be your hunt team. It won’t be wrong to say Data Science is a key element of Threat Hunting. Therefore, it is important to discuss data, data reliability and data dictionary.

## The data

Threat hunting involves analyzing data from variety of sources to recognize unusual patterns. There is no standard on how much data will be needed or what data sources will be needed for specific threat hunt.

These logs would come from network activity and the data from operating systems of endpoints and applications. First, it is important to understand, based on the operational environment, what fundamental data sources are available at your disposal. The data sources can be categorized as network, end points or security relevant. The security event data would come from:

1. Physically controlled areas: Data centers, Server Racks / Cabinets, or Control Centers have gates for access control, CCTV cameras, or heat sensors.
2. Networking devices: this would include network access control devices such as web proxies, firewalls and intrusion prevention systems, detective controls such as intrusion prevention systems.
3. Network traffic: packet capture
4. Operating system on End Points: Different end points such as servers, workstations, or Human Machine Interaces (HMIs) provide system events (operations performed by OS) and audit events.
5. Applications and Services: Customized or Commercial Off The Shelf (COTS) applications, services, Application Programming Interfaces (APIs) will provide data request and responses, usage information and other significant events.
6. Security software: Data from Antimalware software, and vulnerabilities management software

The information should be documented in detail with standardized metadata in a central repository. This metadata repository is also called as event data dictionary. It is important because different fields / attributes available in the event data could provide contextual information for data analysis and correlation.

## Developing and Maintaining Data Dictionary

Threat hunting requires a thorough understanding of normal operations. That means, it is essential to know every single data source logged and collected for development of analytics

e.g. Security, Sysmon or PowerShell logs. Beyond log sources, team should also know the attributes behind every log data collected. Why?, data scientists or analysts spend 80% of their time just finding, cleaning and organizing data. Threat hunting effort is no different. The threat hunter must have an easy way to discover, access and share the data.

“Data Dictionary” solves this problem. The Data Dictionary is a collection of names, definitions, and attributes about data elements that are being used or captured. Data dictionary helps avoides data inconsistencies, provides consistency in the collection and use of data, and enforces the use of data standardization. A data dictionary should become the go-to tool to understand everything about a data set and check data quality at a glance.

If you have recently initiated or kicked-off the Threat Hunting program or have no data governance team, it is likely that Data Dictionary doesn’t exist within your organization. Creating Data Dictionary is not a small endeavor. Effort required to develop the data dictionary largely depends on the technology landscape of the organization, organizational processes and available human resources. Organizations may choose to develop it all at once (focused project) or gradually develop and improve this library during different security initiatives.

Specific contents in a data dictionary can vary. In general, these components are various types of metadata, providing information about data. When planning to create a data dictionary, it is important to consider all available data management resources, including databases and spreadsheets. Online templates are useful for creating this type of data dictionary . Dragos introduced “Collection Management Framework”[[1]](#footnote-1) that provides guidance on keeping track of tools used and data being collected.

Initiate the development by assessing and documenting the tools implemented and configured in your organization. Start by listing the data sources at high-level and then collect more detailed information from each data source Following resources could be useful for development.

* Existing commercial security tools within your ecosystem (i.e EDR solutions) that collect data for you, your vendor should be providing data dictionaries for every security event.
* Open-Source Security Events Metadata (OSSEM)[[2]](#footnote-2). OSSEM defines and share a common data model to improve the data standardization and transformation of security event logs. It also allows you to define and share data structures and relationships identified in security events logs.
* MITRE CAR[[3]](#footnote-3) provides the dictionary of data objects that may be monitored based on MITRE ATT&CK framework

## Data Reliability

Let us focus on the use of data now. Threat hunter will use scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Therefore, data reliability crucial. It is an aspect of data quality that defines how much data is complete and accurate. This improves the data trust. It eliminates the guesswork, gives accurate analysis and insights.

Reliable data is:

* Complete – datasets must contain all required information. It should not be limited to high value assets.
* Accurate – data must conform with reality.
* Timely – data must be accurate in a specific period.
* Validated – data must have right values for the attributes.
* Consistent – data may get stored and transported to different applications, quality of data must be maintained.
* Unique – the data set shouldn’t be recorded more than once.

Low reliability of the data can result in aggregations of incorrect data that can lead to wrong decisions that is, incomplete or incorrect hunt results.

Therefore, investing in data reliability consistently will yield faster and reliable threat hunts. It will allow the automation of detection, essential for effective security monitoring.

Mature security organizations likely have security data governance team to improve the reliability of the data required for security monitoring. However, that does not guarantee the data reliability. It is feasible that team runs into the issues such as coverage, missing data, missing standard naming conventions, parsing issues, and timestamp.

Therefore, it is essential that Data Governance and Threat Hunt teams work side by side, share the hunt objectives and findings. This feedback process will ensure the continuous improvement needed for improving data reliability.

In the absence of data governance team, Threat Hunt team should assess and improve data reliability.

## Security Technology Stack

The event data generated by different data sources within the organization is continuous. The Size of an organization greatly influences the volume of data it generates. On any given day, a large organization can generate hundreds of gigabytes of log data. When dealing with that much data, there are some common issues. The automated data collection, storage and analytic tools is ideal while analyzing the large volume of data and correlating multiple data sources.

Log management systems allow log data collection, data retention, log indexing, reporting, and searching capabilities. Whereas, Security Incident and Event Management (SIEM) system is characterized by Security Event Management (SEM), Eecurity Information Management (SIM), and Security Event Correlation (SEC). SIEM automatically correlates, including all your log data, better than what humans can do alone. SIEM approaches log analysis with a security focus.

Organizations have finite resources. Therefore, organizations prioritize the log sources and data management. Threat hunter must be aware of what data has been automated and is made available at a centralized location (log management system) and what data must be collected and extracted manually.

Threat hunters use a variety of tools to support their methodologies. Tools can include the following:

* Log management system: this allows threat hunter to query, analyze and correlate large volume of data. Allows data analysis and correlation at scale.
* Advanced analytics and statistical analytics tools: If organization has not implemented log management or SIEM, open-source tools such as MongoDB or Redis tools could support ingestion and analysis of large volume of data. Although, feasible, this manual process would require the team to collect and load the data manually. This could be very time consuming as, the team must:
  + develop custom parsers for different log sources for data ingestion
  + develop custom queries for search and correlation of data
* Spreadsheets: This is well known and universally adopted tool for data analysis and statistical analysis. However, it is not optimal while handling large volume of data and correlation.

# Hunt Tempo

Threat hunts could be triggered by different security initiatives such as:

* New threat intelligence data – e.g. new threat actor, updates to the TTPs from known threat actor(s).
* Cybersecurity Incidents: findings from cybersecurity incidents or lessons learned
* Vulnerabilities reported after red team exercises / penetration testing
* Findings from tabletop exercises: tabletop exercises are excellent way to explore some process and data deficiencies for cybersecurity incident investigations.
* Other threat hunts: Threat hunt may result in findings that may trigger additional hunting projects.
* Identification and analysis of Crown jewels: The crown jewels could be critical processes, assets, or data, if compromised; may result in severe consequences.
* Threat modeling: Threat modeling is process that allows organizations to identify, enumerate and prioritize the threats based on the absence of security controls for the system.
* Regulatory requirements: new or changes in regulatory requirements may influence the need for new hunts. E.g. Changes in detection and reporting regulatory requirements could provide feed in hunt backlog.

It is important to keep the backlog of threat hunt ideas so that team can prioritize and schedule.

## Prioritization

With finite resources within the team, it is essential to prioritize the hunts. The prioritization should depend combination on factors as :

1. The TaHiTI[[4]](#footnote-4) (**T**argeted **H**unting **i**ntegrating **T**hreat **I**ntelligence) is a threat hunting methodology that focuses on the top 3 layers of the pyramid of pain – TTPs, Tools, and Network / Host Artifacts. The hunts based on the lower three layers are based on the information received from the past attacks and campaigns – e.g. known bad domain names, IP addresses and hashes. Although important, the targeted attacks with same IOCs may not be a best use of the threat hunt team. Therefore, as team gains maturity, should focus on (or prioritize) top three layers of the pyramid as it can yield high value.
2. Other factors: Apart from general prioritization addressed above as per TaHiTI, following factors should be used for prioritization based on its importance.
   1. Regulatory Requirements: these could allow organizations to maintain their license to operate. Therefore, organizations must prioritize these if requirements change or new requirements for detection and reporting are imposed.
   2. Historical Security Incidents: Historical incidents provide important data points for hunts because those are successful violations of security controls and policies.
   3. Vulnerabilities reported after red team exercises / penetration testing: like historical security incidents, the pen tests or red team exercises provide the TTPs that have been successful.
   4. Risk evaluations of threat hunt ideas: Each threat hunt idea should be evaluated for the risks to the business. If risk to the operational continuity, safety of employees or community, financial or reputational status of the organization is sever, team must prioritize
   5. New Threat Intelligence data: Threats may go dormant or may choose to hibernate for a period. New Threats may emerge or dormant threats may suddenly become active again. The prioritization of threat hunts may get influenced due to this.
      1. A potential threat is more likely to have more impact if it involves observations supporting many different hunting hypotheses. These threats and TTPs should be prioritized above other hunts.
      2. The absence or downward trend of threat activities does not guarantee that organization won’t find it in their environment. Organizations may tend to deprioritize those hunt ideas. Reprioritization of those hunts could be considered. However, those should not be removed from backlog.

## Scheduling

Within any organization scheduling is a major pain point, simply because amount of time needed to build the schedule, manage employee availability and other organizational constraints. In the threat hunt organization, team would need two different types of scheduling focus as detailed below. In any case, team should remember, scheduling is not a one-time activity. It is continuous and on-going effort. Regular status reporting, updating of the schedule and the management of schedule changes on a regular basis ensures the schedule is “useful”.

### Resources Scheduling:

Resource scheduling is very important for the Threat Hunting group to schedule threat hunts by using organizational resources in most effective and efficient manner. Resources are the are primary sources of productivity and profitability upon which organizational strategies are frames. Organization has finite resources; therefore, resource scheduling methods must incorporate time and resource capacity into the scheduling process. There are predominantly two methods of resource scheduling[[5]](#footnote-5):

* 1. Time-constrained scheduling: assumes that time constraints are fixed and activities must be undertaken within defined time constraints. This method assumes
  2. Resource-constrained scheduling: assumes that the resources are finite. Therefore, it emphasizes that task activities must be conducted primarily within resource constraints.

Team should remember, no one technique is perfect. The managers should combine these techniques to effectively plan and schedule the hunts.

Here are few scheduling techniques that will help speed up your threat hunt scheduling process:

1. Maintain Threat Hunt Backlog: Within agile project management, product backlog term is referred often. It refers to a prioritized list of functionality which a product should contain. It is sometimes referred to as a to-do list, and is considered an 'artifact' within Agile Scrum framework. Threat hunt team should use this concept to maintain the list hunt requirements and ideas collected by the team. Establishing an appropriate backlog is very important. When teams have too little work in the backlog, there is a risk of sitting idle. That wastes time and money. On the other hand, when teams have extensive backlogs with excessive detail, the business runs the risk of having over-invested in the plans that can change. That also wastes time and money. Therefore, a middle ground is essential. The entire backbone of planning depends on the goals and resources required for the hunt. Therefore, team should maintain these requirements in the backlog.
2. Ensure staff and resources availability: Team should manage general availability by having employees mark what times and days they are not available. General availability consists of days or times when a person is normally unable to work. The availability management should be extended to other required resources as well. It is possible that organization has finite hardware and software resources needed for the.
3. Use a template: Find a template that you are comfortable using that helps you get your job done faster. A lot of companies use Excel for this, which can be kind of difficult to manage, but pick something that works well for you.
4. Create a schedule based on the employee’s skills and resources available: It is important to create a schedule that matches the right team member to the job at hand. E.g. Hunts that focus on collecting and analyzing network telemetry should be managed by networking subject matter experts. Hardware and software resources required for hunts may constrain your ability to freely schedule hunts. Therefore, scheduling process should include employees and other resources needed for hunts.
5. Evaluate the scheduling process: monitor and evaluate your schedule in real time. Use visual way to spot the gaps, overlaps, and potential errors in your scheduling process.
6. Effectively communicate: All organizations know the importance of communication, and no more important is effective communication than where employee schedules are concerned.

### Threat Hunt Project:

Above, we discussed scheduling constraints and best practices for the operational threat hunt organization. Now let us look into project management, i.e. managing threat hunt. Each threat hunt is a time bound activity with specific inputs and expected results. Therefore, threat hunts are like projects. Breaking down hunt in manageable tasks is essential. It allows the team to schedule (determine the timeline), and reality of the delivery of the hunt. Scheduling is an integral part of project management; therefore, it has been a key knowledge area in Project Management Institute[[6]](#footnote-6)’s (PMI) core publication - Project Management Body of Knowledge (PMBOK® Guide ). Here is a simple guidance for threat hunt project scheduling:

* 1. Develop a reusable Work Breakdown Structure and Work Packages: A reusable work breakdown structure should be maintained. This can be optimized for specific hunts as needed. The work breakdown allows the team to refine work packages (tasks) needed during the hunt.
  2. Schedule: Work packages can be used to assign duration, identify task interdependencies and resources needed to complete specific task.
  3. Use a template for Project Plan: Project planning templates or software tools allow you to create and maintain the detailed project timeline effectively. Microsoft Project tool is widely used. However, Microsoft Excel can be used as well. PMI has several ready to use templates [[7]](#footnote-7)that can be used for this.

# Continuous Improvements

Kaizen, a continuous and never-ending quest for improvements is essential for Threat Hunting program. Threat Hunt program cannot exist in vacuum. First and foremost, hunting is an essential component of security program continuous improvement. This section, however, aims to focus on improving the efficiency, effectiveness, and quality of each component of hunting process.

## Lessons Learned

The lessons learned activity must be carried out after each hunt. Lessons learned feedback and documentation should ensure that the team retrospectively reviews and analyzes all process areas of threat hunting. The analysis should provide the details: if objectives of threat hunt were met such as quality of threat intelligence data, workflow applicability, organizational environment, the data reliability gaps, if the team was adequately staffed with skilled resources, and if the time allocated was adequate.

## Maturity Models

The Open Group information security management maturity model (O-ISM3) describes a maturity as the measurement of an organization’s ability implement continuous improvement practices within a particular discipline. With a focus on continuous improvement, maturity models are well suited for assessing threat hunting operations.

Why maturity models are important for threat hunt teams:

* Benchmarking – Determine where the threat hunt team is in terms of current state and compare against goals and objectives for performance improvement.
* Performance Improvement – With a model a threat hunt team can directly track their operations against the best practices defined by the model. Maturity models can help organizations identify gaps which a plan that addresses specific issues.
* Unified improvement language – Implementing a maturity model ensures that the entire threat hunting organization is aligned on future goals and using the same language to discuss capabilities that would enable the organization to increase their maturity with respect to the model.

The SQRRL “A Framework for Cyber Threat Hunting”[[8]](#footnote-8) white paper is one of the most important papers released to the world of threat hunting and serves as a foundational part of the OTHF. Within SQRRL’s works, they introduce the concept of threat hunting specific maturity model and a set of criteria in which a threat hunting program can be evaluated against.

Through the introduction of a maturity model, SQRRL created a way for threat hunt teams to build short-, medium-, and long-term goals through some high-level characteristics associated with the varying levels of maturity. Depending on the organization, they may want to have more specific requirements and assessment criteria laid out within the model and so the OTHF has created a maturity model that organizations can use to assess their maturity against.

### OTHF Maturity Model

A picture containing text

Description automatically generated

Figure 1: OTHF Maturity Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LEVEL 0  Ad Hoc | LEVEL 1  Foundational | LEVEL 2  Functional | LEVEL 3  Effective | LEVEL 4  Optimal |
| - A process is not defined and has no standardized process  - Relies primarily on automated alerting.  - Little or no routine data collection | - A process area is loosely defined  - Capabilities are based mainly on tribal knowledge and are not aligned to industry standards  - A process area lacks standardized process documentation Leverage threat intelligence to retrospectively search data sets for IOCs. Ad-hoc data searches for adversary tools (tribal knowledge) | - A process area is formally defined with supporting capabilities  -A standardized and documented process supports Capabilities  - Leveraged MITRE ATT&CK to drive change towards platforms that support real-time and metadata collections at scale. | - A process area is formally defined with supporting capabilities  -Capabilities are supported by a well-defined, standardized, and documented process.  - Formal procedures are defined for common tasks  - Incorporation of a detection framework and threat hunting development loop. Threat hunting is a driver for automated detection. Threat hunts are executed as part of a detection framework where successful hypothesis-based hunts are executed and then transitioned into automated detections. | - A process area is formally defined with supporting capabilities  - Capabilities are supported by a well-defined, standardized, and documented process.  - Formal procedures are defined for common tasks  - Centralized data and platform leveraging standardized data sources and structure supporting an intelligence-driven threat hunting framework. Strategic threat assessments provide tactical intelligence that drives the threat hunting development lifecycle.  - Processes are measured, automated, and continuously improved |

### Assessment Criteria

This section outlines the assessment criteria for each process area in scope

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **LEVEL 0** | **LEVEL 1** | **LEVEL 2** | **LEVEL 3** | **LEVEL4** |
| **PEOPLE** | - A threat hunting role does not exist or is informal - A skills catalog does not exist. No hunter training program is available | A threat hunting team exists but role description and expectations are informal documented and communicated.  - A rudimentary skills catalog exists, with no defined criteria, requirements, or a plan. No hunter training program is available. | A threat hunting team exists with dedicated roles and expectations are formally documented and communicated.  - A functional skills catalog exists, with defined criteria, requirements, or a plan. However, the skillset is assessed by the TH program leadership on an ad hoc basis.  There is no formalized plan for addressing gaps. No hunter training program a | - A effective skills catalog exists, with defined criteria, requirements, or a plan. However, skillset is assessed by the TH program leadership on a ad hoc basis. Criteria and requirements are defined, and a formalized mentoring program is in place for bridging the skills gaps  - A formalized training program is established for all levels of hunters.  - A formal recruiting plan is in place. | - A robust skills catalogue exists, with a defined criteria, requirements, or a plan. However skillset is assessed by the TH program leadership as well as through self-assessment on a regular basis. Criteria and requirements are well defined, and a formalized mentoring program is in place for bridging the skills gaps.  -A cross training program is in place for inter-departmental training  - A formal recruiting plan is in place. |
| **PROCESS** | - A hunting framework does not exist or is in its infancy  - Threat hunting does not exist | - A hunting framework is informally documented - Threat hunting is mainly a reactive service when incident response activity arises - The process area covers less than 50% of the organization | - A hunting framework is formalized and documented  - Threat hunting is proactively continued regardless of incident response activities  - No automated detection framework in place.  - The process area covers 50% to 75% of the organization | - A formalized hunting framework is regularly executed. Outcomes are consistently discussed with impacted stakeholders.  - A Threat Hunting mission statement has not been defined.  - Hunts are constantly documented and reviewed with the ability to be turned into automated detection  - The process area covers 75% to 90% of the organization | - The threat hunting frameworks is regularly reviewed and validated for efficiency. - A clear Threat Hunting mission statement has been defined and understood by the team.  - Stakeholder feedback validates that the hunt outcome meets or exceeds expectations.  - Newly developed are shared with the threat hunting community.  - Hunts are constantly documented and reviewed and turned into automated detection  - The process area covers 90% to 100% of the organization |
| **DATA SOURCE** | - Visibility on data sources is unknown  - Quality of data sources is unknown - No tools or processes to passively collect data | - Visibility on data sources is partially understood  - Data sources are informally documented - Tools are present to passively collect data  - The data sources covers less than 50% of the organization | - Visibility and quality of data sources are informally measured  - Available hunting data sources are formally documented  - Collection tools are part of the threat hunt program too actively collect data  - The data sources covers 50% to 75% of the organization | - Visibility and quality on data sources is formally measured and in place  - Data collection is executed consistently  - Hunting techniques include data science  - The data sources covers 75% to 90% of the organization including critical assets | - A standard exists for enterprise wide logging and documentation  - Standardization of hunting data sources is fully automated  - Hunt operations include data science techniques  - The data sources cover 90% to 100% of the organization across network and endpoint. |
| **THREAT INTEL** | - Threat intelligence is not a function within the organization or is still in its infancy  - Threat intelligence is never or rarely collected  - No CTI technology  - No technology integration or Data is raw and unformatted. | - Threat intelligence sharing is reported on an ad hoc basis. Expectations are informal documented and communicated.  - Threat intelligence scope is global and org specific  - Threat Intelligence platform exists with updated feeds  - Technology Integration: SIEM, Firewall/Proxy, or IDS/IPS | - Threat intelligence sharing is a separate function within the organization and expectations are formally documented and communicated.  -Threat intelligence scope is global, org specific, and industry specific.  -Teams take external and internal data input to shift from a reactive to a proactive posture.  - Threat intelligence platform exists alongside an IOC tool  - Technology Integration: TIP, SIEM, Firewall/Proxy, or IDS/IPS is being integrated within threat intel | - Threat intelligence sharing is a separate function within the organization and expectations are formally documented and communicated.  - Region-specific, global, industry-specific, org specific  - Contributors and members of organizations such as Information Sharing and Analysis Centers (ISACs) and Information Sharing and Analysis Organizations (ISAOs)   - Automation of some threat intelligence analysis tasks  - Technology Integration: TIP, SIEM, defensive tools, incident response system, and all security data is being integrated within threat intel  - Supports IR engagements based on knowledge of the adversaries involved | - Threat Intelligence is a key function that allows the business to make operationally and strategically aligned decisions. - Create tactical an strategic TI - Team has the capability to build custom applications and processes  - Majority of TI is automated  - Advanced analytics and orchestration capabilities - Region-specific, global, industry-specific, org specific  - Contributors and members of organizations such as Information Sharing and Analysis Centers (ISACs) and Information Sharing and Analysis Organizations (ISAOs)  - A sophisticated threat intelligence platform exists that allows the team to build out a SOAPA |
| **METRICS** | - Few or no metrics are identified, tracked, or reported | - Key metrics are reported on an ad hoc basis  - Key metrics are identified and measurement elements are accurate | - Performance targets such as operational metrics and key performance indicators is accurate and communicated to management | - Metrics are formally tracked and reviewed. Output is communicated and reported to management on a regular schedule.  - Improvements are discussed but not a critical priority | - Improvements are prioritized for areas where performance is not meeting target goals.  - Operational metrics are updated in real-time via automation  - Hunt outcomes included in risk assessments |

# Definitions

A data lake is a large storage repository that holds native-format raw data until it is needed. Data lakes use a flat architecture.

NoSQL databases, which are non-relational data management systems that are useful when working with large sets of distributed data. They do not require a fixed schema, which makes them ideal for raw and unstructured data.

<https://car.mitre.org/coverage/>

1. [Collection Management Frameworks – Looking Beyond Asset Inventories in Preparation for and Response to Cyber Threats](https://www.dragos.com/wp-content/uploads/CMF_For_ICS.pdf?hsCtaTracking=1b2b0c29-2196-4ebd-a68c-5099dea41ff627c19e1c-0374-490d-92f9-b9dcf071f9b5) [↑](#footnote-ref-1)
2. <https://github.com/OTRF/OSSEM> [↑](#footnote-ref-2)
3. <https://car.mitre.org/> [↑](#footnote-ref-3)
4. <https://www.betaalvereniging.nl/en/safety/tahiti/> [↑](#footnote-ref-4)
5. <https://www.pmi.org/learning/library/resource-scheduling-capacity-schedule-construction-5376> [↑](#footnote-ref-5)
6. https://www.pmi.org/ [↑](#footnote-ref-6)
7. [Project Management Templates](https://www.projectmanagement.com/templates/index.cfm?webSyncID=4c5dc84a-ec20-9d46-e29e-cde44e9598e9&sessionGUID=09c7ea06-9bd2-290e-a0ca-98baec16c30e#_=_) [↑](#footnote-ref-7)
8. [SQRRL Hunting Maturity Model](https://www.threathunting.net/files/framework-for-threat-hunting-whitepaper.pdf) [↑](#footnote-ref-8)