THE ESSENTIAL GUIDE TO MACHINE DATA

Infrastructure Machine Data

splunk>

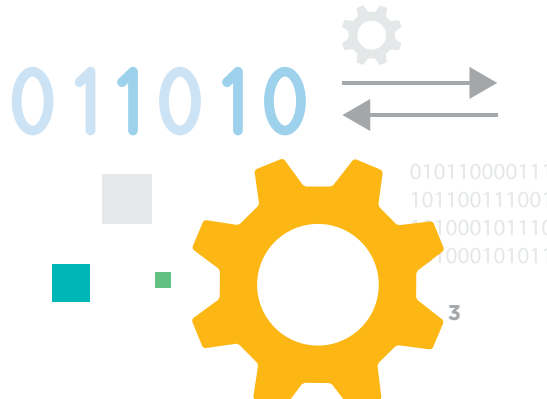
DIGITAL EXHAUST. TIME-SERIES DATA. BIG DATA.

Whatever you call it, machine data is one of the most underused and undervalued assets of any organization. And, unfortunately, it's usually kept for some minimum amount of time before being tossed out and never looked at again.

But some of the most important insights you can gain—across IT and the business—are hidden in this data: where things went wrong, how to optimize the customer experience, the fingerprints of fraud. All of these insights can be found in the machine data generated by the normal operations of your organization.

Machine data is valuable because it contains a definitive record of all the activity and behavior of your customers, users, transactions, applications, servers, networks and mobile devices. It includes configurations, data from APIs, message queues, change events, the output of diagnostic commands, call detail records and sensor data from industrial systems and more.

The challenge with leveraging machine data is that it comes in a dizzying array of unpredictable formats, and traditional monitoring and analysis tools weren't designed for the variety, velocity, volume or variability of this data. But there's a tremendous upside for organizations that take advantage of this data—including quickly diagnosing service problems, detecting sophisticated security threats, understanding the health and performance of remote equipment and demonstrating compliance.



USING MACHINE DATA IN PRACTICE

Using machine data requires three (seemingly) simple steps:

The organizations that get the most value from machine data are able to take disparate data types, link them together, and gain value from the result. But one of the biggest challenges is understanding what data you should ingest.

By defining the use cases you're attempting to resolve – be it security, IT operations, business analytics or the Internet of Things – you can start to identify the data sources you should ingest and begin correlating.

So how does machine data provide value? See the example to the right.

In this example, analyzing the machine data makes the story clear:

1. A customer's order didn't go through
2. The customer called Support to try to resolve the issue
3. After some time on hold, the customer sent a negative tweet about the company

By linking together the machine data, the company can see the original issue and get a full view of the customer experience.

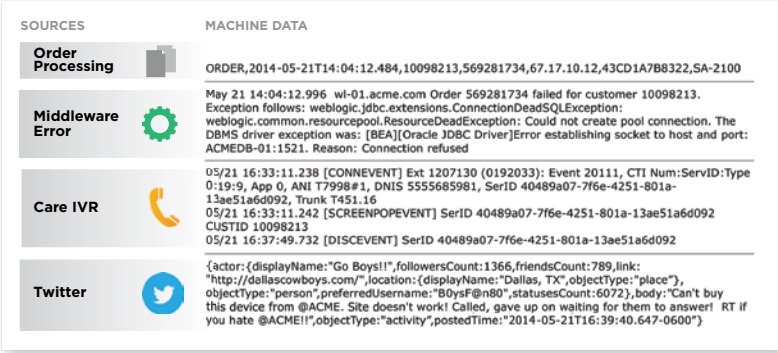


Figure 1: Machine data can come from any number of sources, and at first glance, can look like random text.

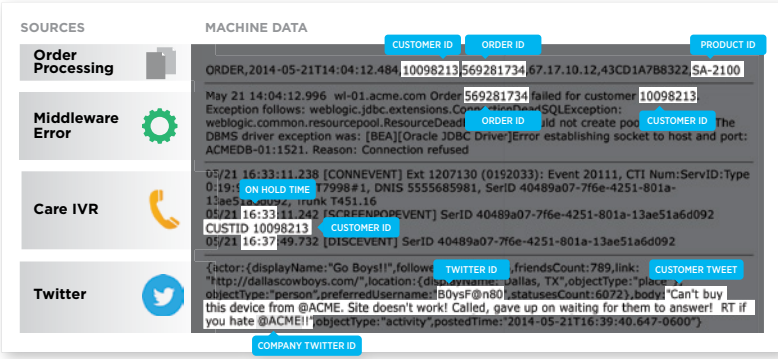


Figure 2: The value of machine data is hidden in this text.

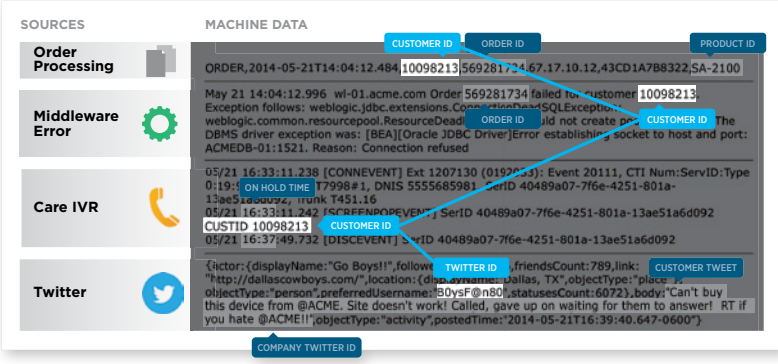


Figure 3: By correlating different types of machine data together, you can start to gain real insight into what's going on in your infrastructure, see security threats or even use the insights to drive better business decisions.

INFRASTRUCTURE MACHINE DATA

This book provides a high-level overview of the value you can get from the machine data created by your virtual and physical infrastructure as part of normal operations. This data can support a variety of use cases, ranging from monitoring your cloud deployments to identifying breach attempts and plugging vulnerabilities.

While each organization's needs and data sources will vary by vendor, product and infrastructure, this book details where you should look for type of machine data and the value it can provide to IT, security, IoT and business analytics use cases.

Many of the data sources listed in this book can support multiple use cases – this is a major part of what drives machine data's tremendous value. The use cases supported by each data source can be easily identified with the icons below.

SECURITY & COMPLIANCE

IT OPS, APP DELIVERY & DEVOPS

INTERNET OF THINGS

BUSINESS ANALYTICS

Table of Contents

Virtual Infrastructure Data **8**

Amazon Web Services (AWS) 8

Microsoft Azure 10

VMware Server Logs, Configuration Data and Performance Metrics 12

Physical Infrastructure Data **14**

Physical Card Readers 14

Sensor Data 16

Server Logs 18

Backup 20

Storage 22

Mainframe 24

Patch Logs 26

Telephony 28

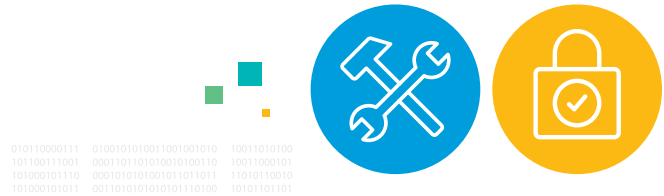
Metric Line Protocols 30

AMAZON WEB SERVICES

Use Cases: IT Operations, Security & Compliance
Examples: CloudTrail, CloudWatch, Config, S3

AWS is the largest and most widely used public cloud infrastructure, providing on-demand compute, storage, database, big data and application services with consumption-based pricing. AWS can be used to replace traditional enterprise virtual server infrastructure in which software runs on individual virtual machines (VM) or to host cloud-native applications built from a collection of AWS services. AWS includes a host of service management, automation, security, network and monitoring services used to deploy, scale, decommission, audit and administer one's AWS environment, subscriptions and hosted applications.

VIRTUAL INFRASTRUCTURE DATA



Use Cases:

IT Ops: AWS services provide similar types of system and service data as traditional IT infrastructure, much of which is consolidated by the CloudWatch service. These include service monitoring, alarms and dashboards for metrics, logs, and events generated by other AWS resources and applications. Typical events and measures include when instances are instantiated and decommissioned, CPU usage, network traffic and storage consumption.

Security & Compliance: Security data from AWS services includes login and logout events and attempts, API calls and logs from network and web application firewalls.

MICROSOFT AZURE

Use Cases: IT Operations, Security & Compliance
Examples: WADLogs, WADEventLogs, WADPerformanceCounter, WADDiagnostInfrastructure

Azure is a popular and widely used public cloud infrastructure, providing on-demand compute, storage, database, big data and application services with consumption-based pricing. Azure can be used to replace traditional enterprise virtual server infrastructure in which software runs on individual VMs, or to host cloud-native applications built from a collection of Azure services. Azure includes a host of service management, automation, security, network and monitoring services used to deploy, scale, decommission, audit and administer one's Azure environment, subscriptions and hosted applications.

VIRTUAL INFRASTRUCTURE DATA

Use Cases:

IT Ops: Azure services provide detailed logs for monitoring one's infrastructure across entire technology stack, VMs, containers, storage and application services. The data is useful in maintaining application delivery quality and service levels, measuring user behavior, resource utilization and for capacity planning and cost management.

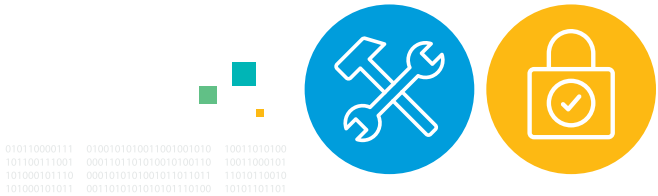
Security & Compliance: Security teams can use Azure service logs to audit and attest to compliance with established policies. Log data also is invaluable for incident forensic analysis, such as identifying unauthorized access attempts from access logs, tracking resources and configuration change events and identifying vulnerabilities in hosts or firewalls.

VMware SERVER LOGS, CONFIGURATION DATA AND PERFORMANCE METRICS

Use Cases: IT Operations, Security & Compliance
Examples: vCenter, ESXi

VMware vSphere ESXi is the most commonly used enterprise server virtualization platform. The VMware management platform, whether one of the vSphere products, vCloud or standalone hypervisor, produce a variety of data, fall into four main categories:

- **vCenter logs** - vCenter is the "control center" of a vSphere environment. The vCenter logs show information including: who is logging in to make changes, which individuals made changes and authentication failures.
- **ESXi logs** - every vSphere environment includes one or more ESXi hypervisors; these are the systems that host the virtual machines. ESXi logs contain information that is useful when troubleshooting hardware and configuration issues.
- **Inventory information** - the vCenter environment tracks configuration about a number of items including: hypervisors, virtual machines, datastores, clusters and more. This includes the configuration of each item, and how a given item relates to any other. This information is not represented in the log files from either the vCenter or ESXi servers. This information can be viewed using the vSphere client or by using vSphere APIs to pull this information. In both cases this information is pulled from the vCenter servers.
- **Performance information** - for each configuration item, the vCenter server tracks a number of performance metrics about that item. Datastore latency, virtual or physical CPU utilization,



and over 100 other metrics fall into this category. As with the inventory information, this information is not present in the log files and must be viewed through the vSphere client or polled through the vSphere API.

Use Cases:

IT Ops: Operations teams can use VMware data to measure the health of the overall hypervisor environment and underlying guest operating systems. Admins can use this data for capacity planning and for troubleshooting of ongoing performance issues, such as datastore latency issues.

This data also records hardware resource usage that can be used to optimize VM deployments across a server pool to maximize resource consumption without having workloads overwhelm any given server.

Security & Compliance: The uncoupled nature of virtual resources and underlying physical hardware can cause complex challenges during incident investigations, capacity analyses, change tracking and security reporting. One common security use case for VMware data comes from the vCenter logs, which audit the activity of individuals using the vSphere interface to re-assign user permissions within the VMware environment.

PHYSICAL CARD READERS

Use Cases: Security & Compliance

Most organizations use automated systems to secure physical access to facilities. Historically, these have been simple magnetic strips affixed to employee badges; however, locations with stringent security requirements may use some form of biometric reader or digital key. Regardless of the technology, the systems compare an individual's identity with a database and activate doors when the user is authorized to enter a particular location. As digital systems, badge readers record information such as user ID, date and time of entry and perhaps a photo for each access attempt.

Use Cases:

Security & Compliance: For IT security teams, the data from card readers provide the same sort of access information for physical locations as a network firewall log. The data can be used to detect attempted breaches and be correlated to system and network logs to identify potential insider threats and provide overall situational awareness. It can also be used to detect access at unusual times and locations or for unusual durations.

SENSOR DATA

Use Cases: IT Operations, Security, Business Analytics, Internet of Things
Examples: Binary and numeric values including switch state, temperature, pressure, frequency, flow, from MQTT, AMQP and CoAP brokers, HTTP event collector

---

Industrial equipment, sensors and other devices often have embedded processors and networking that allows them to record and transmit a vast array of information about operating conditions. Regardless of device, their data provides unprecedented detail about performance parameters and anomalies that can indicate larger problems—for example, a device ready to fail or issues with another system. Aggregating and correlating data from multiple devices and subsystems provides a complete picture of equipment, system, factory or building performance.

PHYSICAL INFRASTRUCTURE DATA

Use Cases:

IT Ops: Some of the most important parameters for operations teams to monitor are environmental conditions such as temperature, humidity, airflow and voltage regulation in a data center. Similar readings are available from individual servers and network equipment that, when correlated, can highlight problems in the facility or equipment ready to fail.

Security & Compliance: Sensor data can help protect mission-critical assets and industrial systems against cybersecurity threats by providing visibility into system performance or set points that could put machines or people at risk. Data can also be used to satisfy compliance reporting requirements.

Internet of Things:

Preventative Maintenance and Asset Lifecycle Management: Sensor data can provide insights into asset deployment, utilization and resource consumption. Operational data can also be used to proactively approach long-term asset management, maintenance and performance.

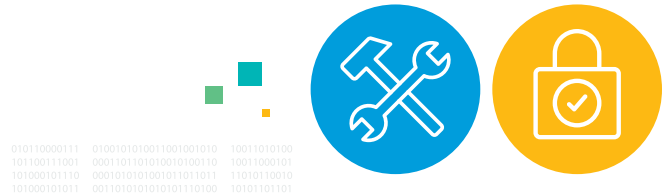
Monitoring and Diagnostics: Monitoring sensors can help ensure that equipment in the field operates as intended, for example, monitoring and tracking unplanned device or system downtime. The data can also be used to understand the cause of failure on a device to improve efficiency and availability, and to identify outliers and issues in device production or deployment.

SERVER LOGS

Use Cases: IT Operations, Security & Compliance, Application Delivery

Server operating systems routinely record a variety of operational, security, error and debugging data such as system libraries loaded during boot, application processes open, network connections, file systems mounted and system memory usage. The level of detail is configurable by the system administrator; however, there are sufficient options to provide a complete picture of system activity throughout its lifetime. Depending on the subsystem, server logs are useful to system, network, storage and security teams.

PHYSICAL INFRASTRUCTURE DATA



Use Cases:

IT Ops & Application Delivery: Server logs provide a detailed record of overall system health and forensic information about the exact time of errors and anomalous conditions that are invaluable in finding the root cause of system problems.

Security & Compliance: Server logs include data from security subsystems such as the local firewall, login attempts and file access errors that security teams can use to identify breach attempts, track successful system penetrations and plug vulnerabilities. Monitoring server logs such as file access, authentication and application usage can help secure infrastructure components.

BACKUP

Use Cases: IT Operations

Despite the use of data replication to mirror systems, databases and file stores, data backup remains an essential IT function by providing for long-term, archival storage of valuable information, much of which has legal and regulatory requirements regarding its preservation. Backups also can be used to store multiple versions of system images and data, allowing organizations to reverse changes, accidental deletions or corrupted data quickly, restoring the system or database to a known good state. Backup software can use different types of storage media depending on the likelihood of needing the data: external disks or virtual tape libraries for active data and tape, optical disks or a cloud service for long-term storage.

Use Cases:

IT Ops: Backup systems routinely log activity and system conditions, recording information such as job history, error conditions, backup target and a detailed manifest of copied files or volumes. This data allows operations teams to monitor the health of backup systems, software and jobs; triggers alerts in the case of errors; and assists in debugging backup failures. It also allows teams to locate where specific data may be stored, when a recovery is required.

STORAGE

Use Cases: IT Operations
Examples: EMC, Netapp, IBM

Data center storage is provisioned in two general ways: built into servers and shared using various network storage protocols, or via a dedicated storage array that consolidates capacity for use by multiple applications that access it using either a dedicated storage area network (SAN) or ethernet LAN file-sharing protocol. The activity of internal, server-based storage is typically recorded in system logs, however storage arrays have internal controllers/storage processors that run a storage-optimized OS and log a plethora of operating, error and usage data. Since many organizations have several such arrays, the logs often are consolidated by a storage management system that can report on the aggregate activity and capacity.

Use Cases:

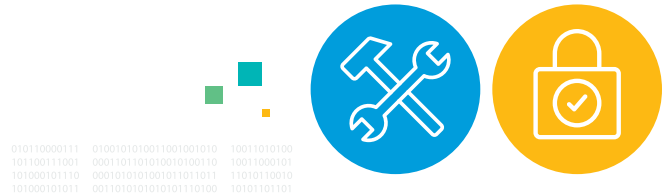
IT Ops: Shared storage logs record overall system health (both hardware and software), error conditions (such as a failed controller, network interface or disks) and usage (both capacity used per volume and file or volume accesses). Collectively, the information can alert operations teams to problems, the need for more capacity and performance bottlenecks.

MAINFRAME

Use Cases: IT Operations, Security & Compliance

Mainframes are the original business computer: large, centralized systems housing multiple processors, system memory (RAM) and I/O controllers. Despite their 60-year legacy, mainframes still are widely used for mission-critical applications, particularly transaction processing. Although they usually run a proprietary OS, mainframes also can be virtualized to run Unix and Linux or, with add-on processor cards, Windows Server. Mainframes are valued for their bulletproof reliability and security, using highly redundant hardware and resilient, stringently tested software. As such, they appeal to organizations wanting to consolidate workloads onto a small number of systems and that need the added reliability and versatility.

PHYSICAL INFRASTRUCTURE DATA



Use Cases:

IT Ops: Like other servers, mainframes measure and log numerous system parameters that show their current status, configuration and overall health. Since most mainframe subsystems are redundant, system logs also show non-disruptive hardware failures or anomalous behavior that is predictive of an impending failure. Due to their use for critical applications, mainframes often record application performance data such as memory usage, I/O and transaction throughput, processor utilization and network activity.

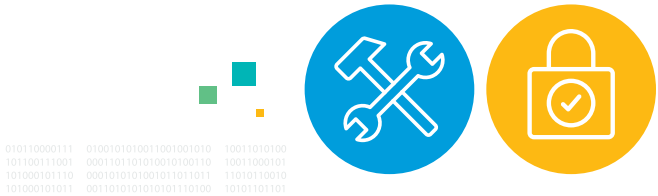
Security & Compliance: Mainframes contain critical operational data. In a security context, mainframe data is treated like any other enterprise data that requires visibility and monitoring for data confidentiality and integrity, compliance and audits with regulatory requirements, and for access monitoring.

PATCH LOGS

Use Cases: IT Operations, Security & Compliance

Keeping operating systems and applications updated with the latest bug fixes and security patches is an essential task that can prevent unplanned downtime, random application crashes and security breaches. Although commercial apps and operating systems often have embedded patching software, some organizations use independent patch management software to consolidate patch management and ensure the consistent application of patches across their software fleet and to build patch jobs for custom, internal applications.

Patch management software keeps a patch inventory using a database of available updates and can match these against an organization's installed software. Other features include patch scheduling, post-install testing and validation and documentation of required system configurations and patching procedures.



Use Cases:

IT Ops: Operations teams use patch logs to verify the timely and correct application of scheduled patches, identify unpatched systems and applications, and alert to errors in the patching process. Correlating errors to patch logs can indicate when an error is due to a patch.

Security & Compliance: Security teams can use patch logs to monitor system updates and determine which assets could be at risk, due to failed or out-of-date patches.

TELEPHONY

Use Cases: IT Operations
Examples: Cisco Unified Communications Manager, Shoretel, Twilio

Real-time business communications are no longer limited to voice calls provided by plain old telephone service (POTS); instead, voice, video, text messaging and web conferences are IP applications delivered over existing enterprise networks. Unlike traditional client-server or web applications, telephony and other communications applications have strict requirements on network quality of service, latency and packet loss, making service quality and reliability much more sensitive to network conditions and server responsiveness. Traditional POTS has conditioned people to expect immediate dial tone when picking up the phone and be intolerant of noise, echo or other problems that can plague IP telephony; as such, the systems and supporting infrastructure require careful monitoring and management to assure quality and reliability.

Use Cases:

IT Ops: Like VoIP, telephony logs provide an overview of system health along with troubleshooting and usage data similar to that of other network applications. Details include source, destination, time and duration of voice/video calls, web conferences and text messages, call-quality metrics (e.g., packet loss, latency, audio fidelity/bit rate), error conditions and user attendance at web conferences. By integrating telephony records of source/destination address with an employee database such as AD or LDAP and a DHCP database, organizations can link call records to actual user IDs and IP addresses to physical locations; information that can assist in troubleshooting and billing. Logs also can reveal any network segments experiencing congestion or other performance problems that may indicate equipment problems or the need for an upgrade.

METRIC LINE PROTOCOLS

Use Cases: IT Operations, Application Delivery, IoT
Examples: collectd, statsd

Metrics are measurements generated by a process running on a system that provide a regular data point around a given metric, such as CPU utilization. Metrics data sources generate measurements on regular intervals and generally consist of:

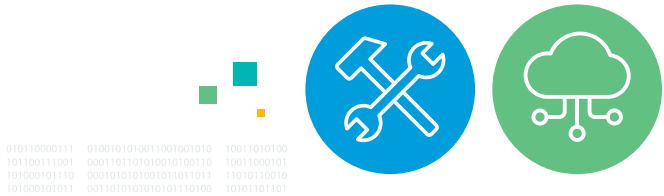
- Timestamp
- Metric Name
- Measurement (a data point)
- Dimensions (that often describe the host, kind of instance, or other attributes that you might want to filter or sort metrics on)

Metrics are typically generated by a daemon (or process) that runs on a server (OS), container, application. Each data measurement is delivered by a network protocol, such as UDP or HTTP, to a server that indexes and analyzes that information.

Metrics are particularly useful for monitoring. For example, a heart monitor that regularly checks a patient's pulse, metrics provide insight into trends or problems that affect the performance and availability of infrastructure and application. However, a heart monitor won't tell you why a patient has a sudden issue with their heart rate - you need other means to quickly identify the cause and stabilize the patient. It's the same with machine data. When combined with other data sources, usually logs, you gain insight into both what's going on, and why it's happening.

Examples of Metric Line Protocols:

collectd: Collectd is a protocol that involves an agent running on a server that is configured to measure specific attributes and transmit that information to a defined destination. Collectd is an extensible measurement engine, so you can collect a wide range of



data. Currently, collectd is most often used for core infrastructure monitoring insights, such as getting insight on the workload, memory usage, I/O, and storage of servers and other infrastructure components. Collectd is part of the open source community, and you can learn much more about collectd by visiting http://collectd.org.

statsd: is a network daemon that runs on node.js. It has gained popularity with windows administrators, application performance experts and others. Statsd provides some capabilities that allow for metrics to be delivered in batch, and while it uses the less dependable UDP network method, many administrators like how easy it is to deploy. Much like collectd, statsd is focused on collecting metrics, mostly involving the usage and performance of applications and application components, and sending them via the network to a tool that can collect and analyze that information.

Use Cases:

IT Ops & Application Delivery: Metrics Line Protocols provides usage, performance and availability data across operating systems, storage devices, applications and other components of IT infrastructure. Metrics are particularly useful for the monitoring portion of IT Operations and Application Delivery, where trends can help identify where there is a problem. Once trends and thresholds illustrate performance issues, other data sources are often correlated to determine the root cause of the problem.

IoT: As devices become more intelligent, more metrics based telemetry will be on board. Metrics line protocols represent an efficient way for these devices to report their status and performance.

ABOUT SPLUNK.

Splunk Inc. makes machine data accessible, usable and valuable to everyone. Join millions of passionate users by trying Splunk for free: **www.splunk.com/free-trials.**

splunk>

BK-Splunk-Guide-to-Machine-Data-Infrastructure-Data-102