: What is machine data?

by Dan Power

Editor, DSSResources.com

Machine data can be used to provide decision support for relevant decision tasks in organizations. Machine refers to a computing machine and its associated operating system and application software. So machine data is all of the data generated by a computing machine while it operates. Companies like Splice Machine and Splunk are innovating and developing tools and methods to capture, organize and analyze machine data. Examples of machine data include: application logs, clickstream data, sensor data and Web access logs. Machine data is helping to create the "big data" phenomenon for decision support and analytics.

Splice Machine (http://www.splicemachine.com/) helps companies build Big Data applications. Their technology provides the benefits of NoSQL databases and standard SQL. They claim a number of benefits including: "massive scalability, flexible schema, fault tolerance and high availability along with easy integration to BI tools."

Splunk (http://www.splunk.com) defines machine data as "the data generated by all the systems running in data centers, the 'internet of things', and the new world of connected devices. It's all of the data generated by the applications, servers, network devices, security devices and remote infrastructure that power your organization (Kwang, 2012)."

According to the Splunk website "Machine data contains a definitive record of all activity and behavior of your customers, users, transactions, applications, servers, networks, factory machinery, and so on. And it's more than just logs. It's configuration data, data from APIs and message queues, change events, the output of diagnostic commands and call detail records, sensor data from remote equipment, and more."

There are many distinct machine data formats so analysis of the data is difficult. Machine data can potentially help with diagnosing computing and web site service problems, detecting security breaches and threats, tracking customers and transactions, understanding the status of remote sensors and demonstrating compliance for transactions.

The following is a list and explanation of some major machine data sources from the Splunk website (http://www.splunk.com/view/machine-data/SP-CAAACDC):

Application Logs

Many applications write local log files to document events. They are often according to Splunk "the best way to report on business and user activity and detect fraud scenarios, since they have all the details of transactions. When developers put timing information into their log events, they can also be used to monitor and report on application performance."

Business Process Logs

Complex events processing and business process management system logs have both business and IT relevant data. These logs generally include records of customer activity and events across multiple transaction channels.

Call Detail Records

"Call detail records (CDRs), charging data records, event data records are some of the names given to events logged by telecoms and network switches. CDRs contain useful details of the call or service that passed through the switch, such as the number making the call, the number receiving the call, call time, call duration, type of call, etc. ... Splunk software can quickly index the data and combine it with other business data to enable users to derive new insights from this rich usage information." For example, Cisco Unified Communications Manager keeps extensive data in call detail records (CDRs). More than 50 examples are reviewed at http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/service/7_1_2/cdrdef/cdrex.html#wp1047910.

Clickstream Data

User activity on the Internet is captured in clickstream data, a record of clicks or input actions. A clickstream is a detailed time sequenced record of activity of all user actions captured in that file. This data can provide information about a user's website and web page activity, including user visits, how long a user was on a page or site, and in what order pages were visited. This data is potentially valuable for usability analysis and marketing research. According to Splunk, "existing web analytics and data warehouse products often sample the data, missing the complete view of behavior and provide no real-time analysis." Both ISPs and individual Web sites can track a user's clickstream.

Database and File System Audit Logs and Tables

Databases contain sensitive corporate data—customer records, financial data, patient records and more. Splunk website argues "Audit records of all database queries are vital to have in order to understand who accessed or changed what data when. Database audit logs are also useful to understand how applications are using databases to optimize queries. Some databases log audit records to files, while others maintain audit tables accessible via SQL."

Operating System Metrics, Status and Diagnostic Commands

Operating systems record metrics like CPU and memory utilization and status information. This data is "potentially invaluable for troubleshooting, analyzing trends to discover latent issues and investigating security incidents."

SCADA Data

Splunk notes "Supervisory Control and Data Acquisition (SCADA) refers to a type of industrial control system (ICS) that gathers and analyzes real-time data from equipment in industries such as energy, transport, oil and gas, water and waste control. These systems produce significant quantities of data about the status, operation, utilization, and communication of components. This data can be used to identify trends, patterns, anomalies in the SCADA infrastructure and used to drive customer value." SCADA data can help monitor and control industrial processes.

Sensor Data

A sensor is a mechanical device that is sensitive to light, temperature, radiation level, or other physical variable. A sensor is a measurement device. According to Splunk, the "growing network of sensor devices generate data based on monitoring environmental conditions, such as temperature, sound, pressure, power, water levels, etc. This data can have a wide range of practical applications if collected, aggregated, analyzed and acted upon. Examples include, water level monitoring, machine health monitoring and smart home monitoring."

Web Access Logs

Finally, Splunk explains "Web access logs report every request processed by a web server--what client IP it came from, what URL was requested, what the referring URL was, and data regarding the success or failure of the request. They're most commonly processed to produce web analytics reports for marketing—daily counts of visitors, most requested pages, and the like. ... The only

challenge is sheer volume with busy websites experiencing billions of hits a day as the norm." For example, the Apache HTTP Server provides comprehensive and flexible logging capabilities, including the server access log record of all requests processed by the server.

According to Kevin Kwang(2012), Splunk "CEO Godfrey Sullivan reckons data generated from machine-to-machine communication 'most relevant' and 'biggest component' of big data, and identifies telecom, financial and government sectors as biggest customers ... " Ravi Kalakota "Machine data provides a definitive, time-stamped record of current and historical activity and events within and outside an organization, including application and system performance, user activity, system configuration changes, electronic transaction records, security alerts, error messages and device locations." Kalkota argues the analytics gap is that existing IT and BI solutions are unable to handle machine data.

The machine data software vendors are offering tools to store, query and search data previously difficult to store and access. The claim is that analyzing machine data can provide “operational intelligence”. As Carnelley notes, Splunk "needs a 2-page white paper to explain what it means." Companies need more decision support use cases providing examples of decision support goals and how using machine data can reach the goal. Tapping new data sources is possible and may be helpful, but beware of the hyperbole.

References

Carnelley, P. "Using Machine Data To Create Business Value," blog Pac-Online, January 19, 2012 at URL http://blog.pac-online.com/2012/01/using-machine-data-to-create-business-value/.

Darrow, B. "Machine data is for people too," Gigaom, March 21, 2012 at URL http://gigaom.com/2012/03/21/mining-machine-generated-data-structure-data-2012/

Devlin, B. "With Big Data Out of the Box, Maintaining Order is a Must," BeyeNetwork, January 29, 2013 at URL http://www.b-eye-network.com/view/16817 ,

Kalakota, R. "Machine Data Analytics: Splunk," DZone, July 28, 2012 at URL http://java.dzone.com/articles/machine-data-analytics-splunk

Kwang, K. "Splunk: Machine data defines big data hype," ZDNet, March 2, 2012 at URL http://www.zdnet.com/splunk-machine-data-defines-big-data-hype-2062304078/

Log Files documentation, http://www.apache.org/, at URL http://httpd.apache.org/docs/2.2/logs.html

Splunk video "Splunk + Machine Data = Operational Intelligence," YouTube at URL http://youtu.be/3CxlM_tFdys, 2:15 minutes.

Stackpole, B. "'Big data' strains, stresses real-time business intelligence systems," SearchBusinessAnalytics January 2013 at URL http://searchbusinessanalytics.techtarget.com/feature/Big-data-strains-stresses-real-time-business-intelligence-systems .

Author: Daniel Power
Last update: 2013-02-27 08:22