

**STORING LOG DATA EFFICIENTLY WHILE SUPPORTING QUERYING
TO ASSIST IN COMPUTER NETWORK SECURITY**

Inventors:

Wei Huang

Wenting Tang

Christian F. Beedgen

BACKGROUND

1. CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of U.S. Provisional Application No. 60/882,289, filed December 28, 2006, which is hereby incorporated by reference herein in its entirety.

2. FIELD OF THE INVENTION

[0002] This invention pertains in general to security information/event management (SIM or SIEM) and in particular to storing security information/events efficiently while supporting querying.

3. DESCRIPTION OF THE RELATED ART

[0003] The field of security information/event management (SIM or SIEM) is generally concerned with 1) collecting data from networks and networked devices that reflects network activity and/or operation of the devices and 2) analyzing the data to enhance security. For example, the data can be analyzed to identify an attack on the network or a networked device and determine which user or machine is responsible. If the attack is ongoing, a countermeasure can be performed to thwart the attack or mitigate the damage caused by the attack. The data that is collected usually originates in a message (such as an event, alert, or alarm) or an entry in a log file, which is generated by a networked device. Exemplary networked devices include firewalls, intrusion detection systems, and servers.

[0004] Each message or log file entry ("event") is stored for future use. Stored events can be organized in a variety of ways. Each organizational method has its own advantages and disadvantages when it comes to writing event data, searching event data, and deleting event data.

[0005] Consider the following scenario: Each event includes an attribute called event receipt time. Since the value of the event receipt time attribute is frequently used for searching, store events based on their event receipt times. For example, create one file for each minute of the day. In order to store an event, determine that event's event receipt time. Append the event to the file that corresponds to that minute of event receipt time.

[0006] When subsequent events arrive, their event receipt times will always increase monotonically. This means that writing the subsequent event data will require only append operations. No seeking of the storage medium is necessary. This makes for good efficiency in writing the event data. In order to search the event data based on event receipt times, once the first event has been

identified, the subsequent events are available by reading the storage medium in order. Again, no seeking is necessary. This makes for good efficiency in searching the event data based on event receipt time. In order to delete the oldest event data, the oldest files are deleted. If the oldest file is always deleted first, then the storage medium will not become fragmented. This makes for good efficiency in deleting the event data.

[0007] The problem with this approach is that searching the event data based on any attribute other than the event receipt time is very time consuming. For example, assume that each event also includes an attribute that indicates the device or application that generated the event (“event source”). In order to search the event data for events that indicate a particular event source (i.e., events that include a particular value for the event source attribute), the entire storage medium will have to be reviewed. This is very inefficient.

[0008] What is needed is a way to store security information/events efficiently while supporting querying for different event attributes (e.g., by supporting multi-dimensional indexing).

BRIEF SUMMARY OF THE INVENTION

[0009] A logging system stores security information/events efficiently while supporting querying for different event attributes. The logging system can be used in conjunction with a security information/event management (SIEM) system. Log data, which can be generated by various sources (including devices and applications), can be in any format. Log data is comprised of one or more data instances called “events.” An event can be, for example, an entry in a log file, an entry in a syslog server, an alert, an alarm, a network packet, an email, or a notification page. In general, an event is generated once and does not change afterwards.

[0010] In one embodiment, the logging system includes an event receiver, a storage manager, and a communication mechanism. The event receiver receives log data, processes the log data, and outputs a data “chunk.” The event receiver includes a control system, a set of buffers, and a metadata structure. The control system controls operation of the event receiver. The set of buffers stores one or more events. The metadata structure stores metadata about the contents of the set of buffers. In one embodiment, the metadata includes a unique identifier associated with the event receiver, the number of events in the set of buffers, and, for each of one or more “fields of interest,” a minimum value and a maximum value that reflect the range of values of that field over all of the events in the set of buffers. The metadata structure acts as a search index when querying event data.

[0011] The storage manager receives data chunks and stores them so that they can be queried. The storage manager includes a control system, a datafiles table, a chunks table, and one or more datafiles. The control system controls operation of the storage manager. The datafiles table stores information about the one or more datafiles. In one embodiment, this information includes, for each datafile, a unique identifier associated with the datafile and the location of the datafile. The chunks table stores information about the one or more chunks that are stored in the storage manager

(specifically, stored in the one or more datafiles). In one embodiment, this information includes, for each chunk, the metadata stored in the chunk and the location of the chunk. A datafile stores multiple chunks. The communication mechanism communicatively couples the event receiver and the storage manager.

[0012] The event receiver and the storage manager jointly perform a method for storing log data. Before the method begins, the set of buffers and the metadata structure are initialized. The event receiver receives log data. The event receiver control system separates the log data into one or more events and determines when each event was received by the event receiver. The control system stores in the set of buffers the events and, for each event, a time/date stamp that reflects when the event was received. The control system also updates the metadata structure. At some point in time, the control system generates a data chunk based on the metadata structure and the contents of the set of buffers. In one embodiment, a chunk includes the metadata structure and a compressed version of the contents of the set of buffers. The set of buffers and the metadata structure are re-initialized, thereby flushing the set of buffers. The control system sends the chunk to the storage manager. The storage manager receives the chunk, stores the chunk in a datafile, and updates the chunks table.

[0013] The storage manager performs a method for reclaiming storage. The oldest datafile associated with a particular retention policy is identified. Information regarding all of the chunks contained in the identified datafile is removed from the chunks table. The entry in the datafiles tables that represents the identified datafile is deleted. A new entry is created in the datafiles table. The newly reclaimed datafile is added to the list of available pre-allocated datafiles and is ready to receive new chunks.

[0014] After a chunk has been stored in a datafile, the events within the chunk can be queried. A query is represented as an expression that can be evaluated against an event. The expression includes one or more search terms. In order to perform a query, data chunks are identified that could contain an event that satisfies the query. Specifically, search terms within the query are identified that contain information that was contained in the metadata structure. The “metadata search terms” are used to search the chunks table. In this way, a search can be constrained based on particular values for information that was stored in the metadata. The identified chunks are disassembled into their constituent events. Events that satisfy the query are identified.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] FIG. 1 is a block diagram illustrating an environment having a security information/event management system, according to one embodiment.

[0016] FIG. 2 is block diagram illustrating a computer for acting as a logging system of a security information/event management system, according to one embodiment.

[0017] FIG. 3 is a block diagram illustrating a logging system of a security information/event management system, according to one embodiment.

[0018] FIG. 4 is a flowchart illustrating a method for storing log data, according to one embodiment.

[0019] FIG. 5 is a flowchart illustrating a method for reclaiming storage, according to one embodiment.

[0020] FIG. 6 is a flowchart illustrating a method for querying, according to one embodiment.

[0021] The figures depict an embodiment for purposes of illustration only. One skilled in the art will readily recognize from the following description that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles described herein.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0022] Described herein is a computer-based system for collecting data from disparate devices across a computer network, normalizing the data to a common schema, and consolidating the normalized data. The data ("events") can then be monitored, analyzed, and used for investigation and remediation in a centralized view. Events can be cross-correlated with rules to create meta-events. Correlation includes, for example, discovering the relationships between events, inferring the significance of those relationships (e.g., by generating meta-events), prioritizing the events and meta-events, and providing a framework for taking action. The system (one embodiment of which is manifest as computer software) enables aggregation, correlation, detection, and investigative tracking of suspicious network activities. The system also supports response management, ad-hoc query resolution, reporting and replay for forensic analysis, and graphical visualization of network threats and activity.

[0023] Although the present system will be discussed with reference to various illustrated examples, these examples should not be read to limit the broader spirit and scope of the present invention. For example, the examples presented herein describe distributed agents, managers and consoles, which are but one embodiment of the present invention. The general concepts and reach of the present invention are much broader and may extend to any computer-based or network-based security system. Also, examples of the messages that may be passed to and from the components of the system and the data schemas that may be used by components of the system are given in an attempt to further describe the present invention, but are not meant to be all-inclusive examples and should not be regarded as such.

[0024] Some portions of the detailed description that follows are presented in terms of algorithms and symbolic representations of operations on data within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the computer science arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not

necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared and otherwise manipulated. It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers or the like. It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise, it will be appreciated that throughout the description of the present invention, use of terms such as "processing", "computing", "calculating", "determining", "displaying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission or display devices.

[0025] As indicated above, one embodiment of the present invention is instantiated in computer software, that is, computer readable instructions, which, when executed by one or more computer processors/systems, instruct the processors/systems to perform the designated actions. Such computer software may be resident in one or more computer readable media, such as hard drives, CD-ROMs, DVD-ROMs, read-only memory, read-write memory and so on. Such software may be distributed on one or more of these media, or may be made available for download across one or more computer networks (e.g., the Internet). Regardless of the format, the computer programming, rendering and processing techniques discussed herein are simply examples of the types of programming, rendering and processing techniques that may be used to implement aspects of the present invention. These examples should in no way limit the present invention, which is best understood with reference to the claims that follow this description.

1. Security Information/Event Management (SIEM) System Architecture

[0026] FIG. 1 is a block diagram illustrating an environment having a security information/event management system, according to one embodiment. FIG. 1 includes a security information/event management (SIEM) system 100 and one or more data sources 110. A data source 110 is a network node, which can be a device or a software application. Exemplary data sources 110 include intrusion detection systems (IDSs), intrusion prevention systems (IPSs), vulnerability assessment tools, firewalls, anti-virus tools, anti-spam tools, encryption tools, application audit logs, and physical security logs.

[0027] Types of data sources 110 include security detection and proxy systems, access and policy controls, core service logs and log consolidators, network hardware, encryption devices, and physical security. Exemplary security detection and proxy systems include IDSs, IPSs, multipurpose security appliances, vulnerability assessment and management, anti-virus, honeypots, threat response technology, and network monitoring. Exemplary access and policy control systems include access and

identity management, virtual private networks (VPNs), caching engines, firewalls, and security policy management. Exemplary core service logs and log consolidators include operating system logs, database audit logs, application logs, log consolidators, web server logs, and management consoles. Exemplary network hardware includes routers and switches. Exemplary encryption devices include data security and integrity. Exemplary physical security systems include card-key readers, biometrics, burglar alarms, and fire alarms.

[0028] In the illustrated embodiment, the SIEM system 100 includes one or more agents 120, one or more managers 130, one or more databases 140, one or more online archives 150, one or more user interfaces 160, and one or more logging systems 170. In some embodiments, these modules are combined in a single platform or distributed in two, three, or more platforms (such as in FIG. 1). The use of this multi-tier architecture supports scalability as a computer network or system grows. The SIEM system 100 is further described in U.S. Application No. 10/308,415, filed December 2, 2002, which is hereby incorporated by reference herein in its entirety.

[0029] An agent 120 provides an interface to a data source 110. Specifically, the agent 120 collects data (“raw events”) from a data source 110, processes the data, and sends the processed data (“events”) to a manager 130. The agent 120 can operate anywhere, such as at a separate device communicating via a protocol such as simple network management protocol (SNMP) traps, at a consolidation point within the network, or at the data source 110. For example, if the data source 110 is a software application, the agent 120 can be co-hosted on the device that hosts the data source. In one embodiment, the agent 120 is the Connector product from ArcSight, Inc. of Cupertino, CA.

[0030] Processing can include normalization, aggregation, and filtering. For example, individual raw events are parsed and normalized for use by the manager 130. Normalization can involve normalizing values (such as severity, priority, and time zone) into a common format and/or normalizing a data structure into a common schema. Events can be categorized using a common, human-readable format. This format makes it easier for users to understand the events and makes it easier to analyze the events using filters, rules, reports, and data monitors. In one embodiment, the common format is the Common Event Format (CEF) log management standard from ArcSight, Inc. Normalization is further described in U.S. Application No. 10/308,941, filed December 2, 2002, which is hereby incorporated by reference herein in its entirety.

[0031] Aggregation and filtering reduce the volume of events sent to the manager 130, which saves network bandwidth and storage space, increases the manager’s efficiency and accuracy, and reduces event processing time. Aggregation is further described in U.S. Application No. 10/308,584, filed December 2, 2002, which is hereby incorporated by reference herein in its entirety. The agent 120 sends events to the manager 130 in batches based on the expiration of a time period or based on a threshold number of events being reached. Batching events for transmission to the manager 130 is further described in U.S. Patent No. 7,219,239, issued May 15, 2007, which is hereby incorporated by reference herein in its entirety.

[0032] The agent 120 can also send commands to the data source 110 and/or execute commands on the local host, such as instructing a scanner to run a scan. These actions can be executed manually or through automated actions from rules and data monitors. Command support is further described in U.S. Application No. 10/308,417, filed December 2, 2002, which is hereby incorporated by reference herein in its entirety. The agent 120 can also add information to the data that it has collected, such as by looking up an Internet Protocol (IP) address and/or hostname in order to resolve IP/hostname lookup at the manager 130.

[0033] The agent 120 is configured via an associated configuration file (not shown). The agent 120 can include one or more software modules including a normalizing component, a time correction component, an aggregation component, a batching component, a resolver component, a transport component, and/or additional components. These components can be activated and/or deactivated through appropriate commands in the configuration file. During configuration, the agent 120 is registered to a manager 130 and configured with characteristics based on its data source 110 and desired behavior. The agent 120 is further configurable through both manual and automated processes. For example, the manager 130 can send to the agent 120 a command or configuration update. Agent components are further described in U.S. Application No. 10/308,548, filed December 2, 2002, which is hereby incorporated by reference herein in its entirety.

[0034] A manager 130 provides analysis capabilities, case management workflow capabilities, and services capabilities. Communications between the manager 130 and an agent 120 can be bi-directional (e.g., to enable the manager 130 to transmit a command to the platform hosting the agent 120) and encrypted. In some installations, the manager 130 can act as a concentrator for multiple agents 120 and can forward information to other managers 130 (e.g., managers deployed at a corporate headquarters). To perform its tasks, the manager 130 uses a variety of filters, rules, reports, data monitors, dashboards, and network models. In one embodiment, the manager 130 is a Java-based server such as the Enterprise Security Manager (ESM) product from ArcSight, Inc.

[0035] Analysis can include detection, correlation, and escalation. For example, the manager 130 cross-correlates the events received from the agents 120 using a rules engine (not shown), which evaluates each event with network model and vulnerability information to develop real-time threat summaries. Correlation is further described in U.S. Application No. 10/308,767, filed December 2, 2002, which is hereby incorporated by reference herein in its entirety. Regarding case management, the manager 130 can maintain reports regarding the status of security incidents and their resolution. Incident reports are further described in U.S. Application No. 10/713,471, filed November 14, 2003, which is hereby incorporated by reference herein in its entirety. Services can include administration, notification, and reporting. The manager 130 can also provide access to a knowledge base.

[0036] As events are received by the manager 130, they are stored in a database 140. Storing the events enables them to be used later for analysis and reference. In one embodiment, the database 140

is a relational database management system such as a database from Oracle Corporation of Redwood Shores, CA.

[0037] In one embodiment, the database 140 stores data in partitions, which are chronological slices of the database. For example, one new partition is created each day to store that day's events. A partition can be compressed and stored in an online archive 150 for later retrieval. Partition management is further described in U.S. Application No. 10/839,563, filed May 4, 2004, which is hereby incorporated by reference herein in its entirety. In one embodiment, partition management is provided by the SmartStorage archiving and retrieval component of the Security Lifecycle Information Management (SLIM) product from ArcSight, Inc.

[0038] A user interacts with the manager 130 via a user interface 160. The user interface 160 enables the user to navigate the features and functions of the manager 130. A single manager 130 can support multiple user interface instances. The features and functions that are available to the user available can depend on the user's role and permissions and/or the manager's configuration. In one embodiment, access control lists enable multiple security professionals to use the same manager 130 and database 140 but each professional has his own views, correlation rules, alerts, reports, and knowledge bases appropriate to his responsibilities. Communication between the manager 130 and the user interface 160 is bi-directional and can be encrypted.

[0039] In one embodiment, there are two types of user interfaces 160: a workstation-based interface and a web browser-based interface. The workstation interface is a standalone software application that is intended for use by full-time security staff in a Security Operations Center (SOC) or similar security monitoring environment. The workstation interface includes an authoring tool for creating and modifying filters, rules, reports, pattern discovery, dashboards, and data monitors. The workstation interface also enables a user to administer users, database partitions, and workflow (e.g., incident investigation and reporting). For example, the workstation interface enables a user to perform routine monitoring, build complex correlation and long sequence rules, and perform routine administrative functions. In one embodiment, the workstation interface is the ESM Console product from ArcSight, Inc.

[0040] The web interface is an independent and remotely installable web server that provides a secure interface with the manager 130 for web browser clients. The web interface is intended for use as a streamlined interface for customers of Managed Service Security Providers (MSSPs), SOC operators, and users who need to access the manager 130 from outside the protected network. Because the web server can be installed at a location remote from the manager 130, the web server can operate outside the firewall that protects the manager 130. The web interface provides event monitoring and drill-down capabilities. In one embodiment, as a security feature, the web interface does not enable authoring or administrative functions. In one embodiment, the web interface is the ArcSight Web product from ArcSight, Inc.

[0041] In one embodiment, a logging system 170 is an event data storage appliance that is optimized for extremely high event throughput. The logging system 170 stores security events (sometimes referred to as “log data”). In one embodiment, the security events are stored in compressed form. However, the logging system 170 can retrieve these events on demand (unmodified) for forensics-quality data. Multiple logging systems 170 can work together to scale up to support high sustained input rates when storing events. Event queries can be distributed across a peer network of logging systems 170. A user can configure the logging system 170 via a user interface (not shown). In one embodiment, the logging system 170 is the Logger product from ArcSight, Inc.

[0042] The logging system 170 can receive both processed events (e.g., events adhering to the Common Event Format) and raw events. In one embodiment, raw events are received directly from data sources 110 (such as syslog messages and log files), and processed events are received from agents 120 or managers 130. The logging system 170 can also send both raw events and processed events. In one embodiment, raw events are sent as syslog messages (to any device; not shown), and processed events are sent to the manager 130. The logging system 170 will be further described below.

[0043] Through the above-described architecture, the SIEM system 100 can support a centralized or decentralized environment. This is useful because an organization may want to implement a single instance of the SIEM system 100 and use an access control list to partition users. Alternatively, the organization may choose to deploy separate SIEM systems 100 for each of a number of groups and consolidate the results at a “master” level. Such a deployment can also achieve a “follow-the-sun” arrangement where geographically dispersed peer groups collaborate with each other by passing primary oversight responsibility to the group currently working standard business hours. SIEM systems 100 can also be deployed in a corporate hierarchy where business divisions work separately and support a rollup to a centralized management function.

2. Log Data

[0044] Described herein are systems and methods for storing log data efficiently while supporting querying. “Log data,” as used herein, can be generated by various sources, including both devices and applications. These sources include, for example, the data sources 110 described above as well as network systems, computers, operating systems, anti-virus systems, databases, physical infrastructure, identity management systems, directory services, system health information systems, web traffic, legacy systems, proprietary systems, mainframes, mainframe applications, security systems, physical devices, and SIEM sources (such as agents 120 and managers 130).

[0045] A system can obtain log data in many ways. For example, log data can be received (e.g., according to the syslog protocol). Alternatively, log data can be accessed (e.g., by reading a file that is stored locally or remotely). Other methods include, for example, Open Database Connectivity (ODBC), Simple Network Management Protocol (SNMP) traps, NetFlow, and proprietary

Application Programming Interfaces (APIs). Log data can also be input by a user (e.g., using a command line interface (CLI)).

[0046] Log data can be in any format. One such format is, for example, Common Event Format (described above). Other formats are, for example, specific to the data sources 110 that generated the log data.

[0047] Log data is comprised of one or more data instances called “events.” An event can be, for example, an entry in a log file, an entry in a syslog server, an alert, an alarm, a network packet, an email, or a notification page. In general, an event is generated once and does not change afterwards.

[0048] In one embodiment, an event includes implicit meta-data and a message. Implicit meta-data can include information about, for example, the device or application that generated the event (“event source”) and when the event was received from the event source (“receipt time”). In one embodiment, the receipt time is a date/time stamp, and the event source is a network endpoint identifier (e.g., an IP address or Media Access Control (MAC) address) and/or a description of the source, possibly including information about the product’s vendor and version.

[0049] The message represents what was received from the event source and can be in any form (binary data, alphanumeric data, etc.). In one embodiment, the message is free-form text that describes a noteworthy scenario or change. In another embodiment, the message also includes explicit meta-data. Explicit meta-data is obtained, for example, by parsing the message. When an event source generates an event, the event usually includes information that indicates when the event occurred (“event occurrence time”). The event occurrence time, which is usually a date/time stamp, is an example of explicit meta-data and is frequently used for analysis. Different event sources often produce non-uniform explicit meta-data (e.g., priority or criticality of event, devices/applications/users affected by event, and which user triggered event).

[0050] In one embodiment, if an event does not include an occurrence time, an implicit timestamp generated by an event receiver when it received the event (described below) is treated as the original occurrence timestamp. As an event is processed and potentially forwarded through various systems, each system usually has an implicit notation of event receipt time.

[0051] In one embodiment, an event represents a data structure that includes one or more fields, where each field can contain a value. The size of this data structure usually falls within the range of 100 bytes to 10 kilobytes.

3. Logging System Architecture

[0052] FIG. 2 is a high-level block diagram of a computer 200 for acting as a logging system 170 of a security information/event management (SIEM) system 100 according to one embodiment. Illustrated are at least one processor 202 coupled to a bus 204. Also coupled to the bus 204 are a memory 206, a storage device 208, a keyboard 210, a graphics adapter 212, a pointing device 214,

and a network adapter 216. In one embodiment, the functionality of the bus 204 is provided by an interconnecting chipset. A display 218 is coupled to the graphics adapter 212.

[0053] The storage device 208 is any device capable of holding data, like a hard drive, compact disk read-only memory (CD-ROM), DVD, or a solid-state memory device. The memory 206 holds instructions and data used by the processor 202. The pointing device 214 may be a mouse, track ball, or other type of pointing device, and is used in combination with the keyboard 210 to input data into the computer 200. The graphics adapter 212 displays images and other information on the display 218. The network adapter 216 couples the computer 200 to a local or wide area network.

[0054] As is known in the art, a computer 200 can have different and/or other components than those shown in FIG. 2. In addition, the computer 200 can lack certain illustrated components. For example, a computer 200 acting as a logging system 170 can lack a keyboard 210, pointing device 214, graphics adapter 212, and/or display 218. Moreover, the storage device 208 can be local and/or remote from the computer 200 (such as embodied within a storage area network (SAN)).

[0055] FIG. 3 is a block diagram illustrating a logging system 170 of a security information/event management (SIEM) system 100, according to one embodiment. In the illustrated embodiment, the logging system 170 includes an event receiver 310, a storage manager 320, and a communication mechanism 330. Although only one event receiver 310 is shown for clarity, the system 170 can support a large number of concurrent sessions with many event receivers 310. In one embodiment, each event receiver 310 is associated with a unique identifier.

[0056] The event receiver 310 receives log data 340, processes the log data 340, and outputs a data “chunk” 350. The event receiver 310 includes a control system 355, a set of one or more buffers 360, and a metadata structure 365. The control system 355 is communicatively coupled to the set of one or more buffers 360 and the metadata structure 365.

[0057] The control system 355 controls operation of the event receiver 310 and is further described below with reference to FIG. 4.

[0058] The set of one or more buffers 360 stores one or more events. The set of buffers 360 also stores, for each event, a time/date stamp that reflects when the event was received by the event receiver 310. For example, the set of buffers 360 appends to each event this time/date stamp value (thereby adding a “ReceiptTime” field).

[0059] The metadata structure 365 stores metadata about the contents of the set of buffers 360. In one embodiment, this metadata includes the unique identifier associated with the event receiver 310 that received the events, the number of events in the set of buffers, and, for each of one or more “fields of interest,” a minimum value and a maximum value that reflect the range of values of that field over all of the events in the set of buffers. The metadata structure 365 acts as a search index when querying event data (described below).

[0060] For example, assume that an event includes a field called OccurrenceTime whose value reflects the time that the event occurred. If OccurrenceTime were a field of interest, the metadata

structure 365 would include a minimum value for OccurrenceTime and a maximum value for OccurrenceTime. The minimum value of OccurrenceTime would be the OccurrenceTime for the event in the set of buffers 360 that occurred first. The maximum value of OccurrenceTime would be the OccurrenceTime for the event in the set of buffers 360 that occurred last.

[0061] In one embodiment, ReceiptTime is also a field of interest. In this embodiment, therefore, the metadata structure 365 also stores a minimum value and a maximum value that reflect the range of values of receipt times over all of the events in the set of buffers. The minimum value of ReceiptTime would be the ReceiptTime for the event in the set of buffers 360 that was received first. The maximum value of ReceiptTime would be the ReceiptTime for the event in the set of buffers 360 that was received last. In one embodiment, only the minimum value of ReceiptTime is stored. In this embodiment, the maximum value of ReceiptTime is not stored; this decreases storage requirements. If a buffer 360 is flushed often (which happens when a chunk is generated, described below), the maximum value of ReceiptTime will be close to the minimum value of ReceiptTime (e.g., one second later).

[0062] In one embodiment, a field of interest is not an event field per se. Instead, it is a “derived” value that is determined based on the values stored in one or more fields of an event.

[0063] The storage manager 320 receives data chunks 350 and stores them so that they can be queried. The storage manager 320 includes a control system 370, a datafiles table 375, a chunks table 380, and one or more datafiles 385. The control system 370 is communicatively coupled to the datafiles table 375, the chunks table 380, and the one or more datafiles 385.

[0064] The control system 370 controls operation of the storage manager 320 and is further described below with reference to FIG. 4.

[0065] The datafiles table 375 stores information about the one or more datafiles 385. In one embodiment, each entry in the datafiles table 375 represents one datafile 385 for which space has been allocated, and the entry includes a unique identifier associated with the datafile and the location of the datafile (e.g., a file system, a path therein, and a file name). A datafile 385 listed in the datafiles table 375 may or may not contain data (e.g., chunks 350). The datafiles table 375 is stored, for example, in a database (not shown). In one embodiment, datafiles 385 are allocated before they are needed. In this embodiment, a list of these pre-allocated datafiles 385 (called a “free list”) is maintained.

[0066] The chunks table 380 stores information about the one or more chunks 350 that are stored in the storage manager 320 (specifically, stored in the one or more datafiles 385). In one embodiment, this information includes, for each chunk 350, the metadata stored in the chunk (described below) and the location of the chunk (e.g., the unique identifier associated with the datafile that stores the chunk and the location within the datafile where the chunk is stored (e.g., as an offset)). The chunks table 380 is stored, for example, in a database (not shown).

[0067] A datafile 385 stores multiple chunks 350. In one embodiment, all datafiles are the same size (e.g., 1 gigabyte) and are organized in time order. The datafile 385 is stored, for example, on a

raw disk or in a data storage system such as a file system (not shown). If the datafile 385 is stored on a raw disk, data can be accessed faster, since additional layers of indirection are not required. Also, security can be increased.

[0068] The communication mechanism 330 communicatively couples the event receiver 310 and the storage manager 320. In one embodiment, the communication mechanism 330 includes a partially-public or wholly-public network such as the Internet. In other embodiments, the communication mechanism 330 includes a private network or one or more distinct or logical private networks (e.g., virtual private networks or local area networks). Communication links to and from the communication mechanism 330 can be wired or wireless (e.g., terrestrial- or satellite-based transceivers). In one embodiment, the communication mechanism 330 is a packet-switched network such as an IP-based wide or metropolitan area network that uses the Ethernet protocol.

[0069] In another embodiment, the communication mechanism 330 is local to a single computer system (e.g., if a portion of the event receiver 310 and a portion of the storage manager 320 are executing on the same device). In this embodiment, the communication mechanism 330 is implemented, for example, through a local, software-only loopback device. For example, the data is copied to various locations in memory, and communication occurs via an API.

[0070] In yet another embodiment, the communication mechanism 330 is local to a single process (e.g., if a portion of the event receiver 310 and a portion of the storage manager 320 are executing on the same device and in the same process). In this embodiment, the communication mechanism 330 is implemented, for example, through shared memory and/or pointers thereto.

4. Initial Storage

[0071] FIG. 4 is a flowchart illustrating a method for storing log data, according to one embodiment of the invention. In one embodiment, the method 400 of FIG. 4 is performed jointly by the event receiver 310 (e.g., its control system 355) and the storage manager 320 (e.g., its control system 370).

[0072] In one embodiment, before the method 400 begins, the set of buffers 360 and the metadata structure 365 are initialized. For example, the control system 355 stores in the metadata structure 365 the unique identifier associated with the event receiver 310.

[0073] The method 400 begins when the event receiver 310 receives 410 log data 340. In one embodiment, the log data 340 is received in the form of a stream.

[0074] The control system 355 separates 420 the log data into one or more events and determines 420 when each event was received by the event receiver 310.

[0075] The control system 355 stores 430 in the buffer 360 the events and, for each event, a time/date stamp that reflects when the event was received. The control system 355 also updates 430 the metadata structure 365. For example, the number of events in the buffer will have increased. The minimum and maximum values for the field(s) of interest may also need to be updated. In one

embodiment, data write operations and metadata write operations are synchronized in order to avoid possible inconsistency if a system crash occurs. For example, a transactional database system is used so that if an event is stored in the buffer 360, the metadata structure 365 is guaranteed to be updated accordingly, even if the underlying system crashes in between the two steps.

[0076] At some point in time (see below), the control system 355 generates 440 a data chunk 350 based on the metadata structure 365 and the contents of the buffer 360. In one embodiment, a chunk includes the metadata structure 365 and a compressed version of the contents of the buffer 360. The compressed version can be generated using any data compression algorithm (e.g., a lossless compression algorithm such as that used by GNU zip (gzip)). Compressing the buffer contents makes this approach a cost-effective choice for long-term storage of data. In one embodiment, different chunks can have different sizes, and a maximum size can be specified.

[0077] In one embodiment, the chunk 350 also includes a “magic number” and a version identifier. The magic number, sometimes called a file signature, is a short sequence of bytes that identifies the data type of the chunk. For example, the magic number is reasonably unique (i.e., unique with a high probability) across other data and file formats, including other chunks. Thus, when a chunk is read, it is easy to determine whether the chunk is in the expected format. If the chunk’s actual magic number differs from the expected magic number, then the chunk is “wrong” (e.g., corrupted). If the actual magic number matches the expected magic number, then data that occurs later in the chunk might still be wrong. However, the matching magic number excludes this possibility for the majority of common situations. The version identifier enables the accommodation of data and file formats that have changed. For example, when a chunk is read, the version identifier can be used in conjunction with the magic number to indicate additional information about the data or file format.

[0078] In another embodiment (also not shown), the control system 355 also generates a message digest of the contents of the buffer 360. For example, the control system 355 applies a cryptographic hash function to a string that represents the contents of the buffer 360. Any cryptographic hash function can be used, such as Message-Digest algorithm 5 (MD5) or an algorithm in the Secure Hash Algorithm family (e.g., SHA-256). In one embodiment, the digest value is stored in the metadata structure 365 before the chunk is created. This value can later be used to determine whether the buffer data that is stored in the chunk (in compressed form) has been changed or tampered with. This helps guarantee the integrity of stored events by making it noticeable when events have been changed.

[0079] The buffer 360 and the metadata structure 365 are then re-initialized 440, thereby flushing the buffer 360. In one embodiment, the set of buffers 360 includes multiple buffers. This embodiment enables one buffer to be used to store incoming events while another buffer is full or is being flushed.

[0080] In one embodiment, step 440 is performed when the buffer 360 is full. In another embodiment, step 440 is performed when a particular period of time (a “timeout window”) has elapsed, during which no events were received by the buffer 360.

[0081] The control system 355 sends 450 the data chunk 350 to the storage manager 320.

[0082] The storage manager 320 receives 460 the chunk 350. The control system 370 stores 470 the chunk in a datafile 385 (see below). In one embodiment, the chunk is encrypted before it is stored for security purposes. The control system 370 also updates 470 the chunks table 380. For example, the control system 370 adds to the table information regarding the chunk 350 that it just stored in the datafile 385.

[0083] The control system 370 writes chunks 350 in “appending” order inside each datafile 385. This is sometimes referred to as “write-once journaled.” In one embodiment, the control system maintains a “write pointer” that indicates a location within a datafile where a chunk can be written. After a chunk has been written to a datafile, the write pointer is modified to indicate a location within the same datafile (specifically, at the end of the chunk that was just written). If writing a chunk fills a datafile, the write pointer is modified to indicate a location within a different datafile (specifically, at the beginning) that can be used to store chunks. In one embodiment (not shown), chunk writes are deferred by first caching chunks in memory. Multiple continuous chunks are then combined into one write operation in order to optimize full-stripe writes on RAID 5 disk storage systems. By using large sequential input operations such as writes, the hardware is driven at a high speed, throughput, and concurrency.

[0084] If a pre-allocated datafile exists (e.g., as listed in the free list described above), the control system 370 uses the datafile and removes that datafile’s unique identifier from the free list (since that datafile is no longer available). If no pre-allocated datafile exists, the control system 370 creates a new one by locating available space and updating the datafiles table 375. For example, the control system 370 adds to the table information regarding the new datafile 385 that it just created. In one embodiment, the unique identifier assigned to the new datafile 385 is equal to the sum of 1 and the unique identifier associated with the datafile 385 that was most recently allocated.

[0085] The method 400 has many desirable characteristics. For example, it is highly scalable, since it can support receiving a very high number of events-per-second (EPS). Multiple event receivers 310 can be used, and the writing of event data is fast because it involves only append operations, not seek operations. The method 400 also features high availability, since it provides continuous access to data. Deleting old events does not fragment the storage medium, which means that no defragmentation process is required and therefore no maintenance window is required, either. Implicit downtime for cleanup tasks is not required. Also, since disk write operations are efficient, they avoid overhead in order to leave room for handling queries.

5. Storage Reclamation

[0086] At some point in time (discussed below), storage being used by one or more datafiles 385 is reclaimed for future use. FIG. 5 is a flowchart illustrating a method for reclaiming storage, according to one embodiment. In one embodiment, the method 500 of FIG. 5 is performed by the storage manager 320 (e.g., its control system 370).

[0087] The oldest datafile 385 associated with a particular retention policy (described below) is identified 510. Since datafiles have unique identifiers based on monotonically increasing numbers, it is easy to query the datafiles table 375 to find the oldest datafile (i.e., the datafile that has the lowest unique identifier) associated with the retention policy.

[0088] Information regarding all of the chunks 350 contained in the identified datafile 385 is removed 520 from the chunks table 380.

[0089] The entry in the datafiles table 375 that represents the identified datafile 385 is deleted 530.

[0090] A new entry is created 540 in the datafiles table 375, with a) a new unique identifier that is one higher than the highest used datafile identifier and b) a path attribute referring to the physical location of the previously oldest datafile (i.e., the datafile that was identified in step 510).

[0091] The newly reclaimed datafile 385 is added 550 to the list of available pre-allocated datafiles and is ready to receive new chunks.

[0092] In the illustrated embodiment, when a datafile's storage is reclaimed, that datafile is recycled (e.g., reused or written over) instead of deleted.

[0093] The details of the storage reclamation algorithm (including, for example, when to execute it and how much storage to reclaim) depend on a retention policy associated with a datafile 385. A retention policy limits the retention of a chunk 350 based on, for example, a disk-space usage threshold or a maximum time to retain the chunk. Examples of when to execute the storage reclamation algorithm are: when all of the datafiles associated with that policy are full and no more datafiles can be allocated (e.g., because there is no storage space left); when a particular threshold has been reached (e.g., in terms of the amount of free storage space left for datafiles associated with that retention policy); when a particular period of time has elapsed; when a particular number of datafiles exist that are associated with that policy; and when the oldest chunk in a datafile associated with that policy has reached a threshold age. In one embodiment, a datafile is backed up onto another system before its space is reclaimed. In this way, more storage can be made available while still maintaining existing data.

[0094] In one embodiment, all datafiles 385 are associated with the same retention policy. In another embodiment, multiple retention policies exist, and each datafile is associated with any one of the multiple retention policies. Multiple datafiles can be associated with the same retention policy. A retention policy can be created and modified by a user. In one embodiment, the storage manager 320 logically maintains one instance of the storage reclamation algorithm described above for each retention policy. For example, each datafile 385 includes metadata that indicates the retention policy that applies to that datafile, and a chunk is stored in the datafile that corresponds to that chunk's retention policy.

[0095] If multiple retention policies exist, the system 170 shown in FIG. 3 is modified slightly (not shown). Specifically, the event receiver 310 includes one set of buffers 360 and one metadata

structure 365 for each retention policy. Before an event is stored in the set of buffers and the metadata structure is updated (step 430), the control system 355 determines which retention policy should be applied to the event. This determination is based on, for example, a static mapping or an attribute of the particular event. Any attribute can be used, such as priority or event source. Based on this determination, the control system 355 stores the event in the appropriate set of buffers and updates the appropriate metadata structure. Thus, all events in a particular set of buffers will be associated with the same retention policy.

[0096] It follows that a chunk 350 generated based on that set of buffers will be associated with the same retention policy. Before the chunk is stored in a datafile 385 (step 470), the control system 370 determines the chunk's retention policy and stores the chunk in a datafile associated with that policy. Thus, all chunks in a particular datafile will be associated with the same retention policy.

[0097] In one embodiment, each retention policy has its own group of datafiles 385. Each datafile is marked with a unique number. The number decides the order of the files within one group. The data files are written in appending order. Files are not updated, and files are written once and operated in append-only mode, which prevents log data tampering. As all files within one retention group are filled up, storage is reclaimed from the first (i.e., oldest) file in the group. In one embodiment, a separate datafiles table 375 is maintained for each retention policy, which contains entries for datafiles 385 that have been allocated to that retention policy. If a free list is maintained, only one free list is used for the entire storage manager 320, regardless of how many retention policies exist.

6. Querying/Data Retrieval

[0098] After a chunk 350 has been stored in a datafile 385, the events within the chunk can be queried. A query is represented as an expression that can be evaluated against an event. The expression includes one or more search terms. In one embodiment, the query process occurs in multiple phases. The first phase identifies which data chunks 350 (if any) could contain an event that satisfies the query. The second phase disassembles the identified chunks into their constituent events. The third phase identifies which of these events (if any) satisfy the query. The first phase thereby acts as a "rough cut" for identifying which data chunks (and their events) should be investigated further and which data chunks (and their events) should be ignored. In most cases, the retention policy assigned to a chunk is not considered when events are queried or retrieved because it is not interesting which retention policy applies to a chunk that contains an event.

[0099] In the first phase, search terms within the query are identified that concern information that was contained in the metadata structure 365 (back when the event was stored as an event in the buffer 360 rather than as part of a data chunk 350 in a datafile 385). This metadata information includes the unique identifier of the associated event receiver and, for each field of interest, a minimum value and a maximum value that together reflect the range of values of that field over

multiple events (initially, events in the same buffer; later, events in the same data chunk). Recall that the metadata information was transmitted to the storage manager 320 as part of a chunk 150. Then, the metadata information was stored in the chunks table 380. Thus, in order to search the events based on this metadata, the “metadata search terms” are used to search the chunks table 380. This will yield which chunks (if any) could contain an event that satisfies the metadata search terms. In this way, a search can be constrained based on particular values (or ranges of values) for event receiver and/or fields of interest (since these values are stored in the metadata in the chunks table 380).

[0100] Because “field of interest” metadata is expressed as a range of values, the fact that a chunk satisfies a metadata search term does not necessarily mean that the chunk contains an event that satisfies the metadata search term. For example, if the metadata search term is a field value of 10 and the chunk contains events whose field values are 5 and 15, respectively, then 10 will fall within the range, and the chunk will be identified as satisfying the metadata search term. However, the chunk may not contain an event with a field value of 10. (That is why the query occurs in two phases.) What is always true, however, is that if a chunk could contain an event that satisfied the search term, then that chunk will be identified as satisfying the search term.

[0101] In the second phase, the identified chunks are disassembled into their constituent events. If the event portion of a chunk includes a compressed version of the events, then the event portion is decompressed before it is divided into its constituent events.

[0102] In the third phase, each event is compared with the complete set of search terms in order to determine whether the event satisfies the search terms. In one embodiment (not shown), the events are analyzed in a particular order. For example, the events are analyzed based on their event receipt time. Analyzing the events in a particular order and appending matching events to the search results means that the events in the search results will already be in that particular order. No sorting of the events is required.

[0103] In the first phase, it is possible that none of the search terms concerns information that was contained in the metadata structure 365. If this happens, all chunks 350 will be identified as possibly containing an event that satisfies the metadata search terms (since no metadata search terms exist). The query process thereby degenerates to simply searching each stored event using all of the search terms. This is similar to the naïve, inefficient organizational method that was discussed above.

[0104] The above algorithm searches for events that are stored in chunks 350. However, the logging system 170 may contain additional events in the event receiver 310 (e.g., within the set of buffers 360) that have not yet been stored in a chunk. The algorithm above will not search these events. In one embodiment, before the algorithm is executed, the set of buffers 360 are flushed so that the events will be sent to the storage manager 320 and stored in a chunk. This way, when the algorithm is executed, the events that were formerly in the set of buffers will be searched also. In another embodiment, a separate search is executed on the event receiver 310 using the contents of the metadata structure 365 and the set of buffers 360, similar to the algorithm described above. This way,

all events will be searched, whether they are stored in the storage manager 320 or in the event receiver 310.

[0105] FIG. 6 is a flowchart illustrating a method for querying, according to one embodiment. In one embodiment, the method 600 of FIG. 6 is performed by the storage manager 320 (e.g., its control system 370). Before the method 600 begins, a search query is received. The search query includes one or more search terms.

[0106] Any metadata search terms (within the received search query) are identified 610.

[0107] The identified metadata search terms are used to search 620 the chunks table 380. Recall that each entry in the chunks table 380 corresponds to a chunk 350, and an entry includes the metadata stored in the chunk and the location of the chunk. The identified metadata search terms are used to search the metadata portion of the chunks table 380.

[0108] Each chunk 350 whose metadata satisfies the metadata search terms is retrieved 630 using the location of the chunk, which was stored in the chunks table 380.

[0109] The retrieved chunks are disassembled 640 into their constituent events.

[0110] Each event is evaluated 650 against the search query in order to determine whether the event satisfies the query. If an event satisfies the query, it is included in the search results.

7. Additional Embodiments

[0111] In one embodiment, the logging system 170 supports archiving functionality for datafiles 385. For example, a datafile 385 can be imported into and exported out of the logging system 170. As another example, a datafile 385 can be backed up onto another system and later restored into the logging system 170. Since events are stored in chunks and chunks are stored in datafiles, events are easily transferable to nearline or offline storage. Archival criteria can be similar to the criteria that are used for querying (e.g., values of information stored in metadata structures).

[0112] The above description is included to illustrate the operation of the preferred embodiments and is not meant to limit the scope of the invention. The scope of the invention is to be limited only by the following claims. From the above discussion, many variations will be apparent to one skilled in the relevant art that would yet be encompassed by the spirit and scope of the invention.

CLAIMS

1. A method for processing log data, comprising:
 - receiving log data that comprises a plurality of events, an event including one or more fields;
 - and
 - for each event in the plurality of events:
 - storing, in a buffer, the event; and
 - updating a metadata structure that comprises information about contents of the buffer, wherein information about contents of the buffer includes a first minimum value that reflects a minimum value of a first field of all of the events stored in the buffer.
2. The method of claim 1, wherein information about contents of the buffer further includes a first maximum value that reflects a maximum value of the first field of all of the events stored in the buffer.
3. The method of claim 1, wherein information about contents of the buffer further includes a second minimum value that reflects a minimum value of a second field of all of the events stored in the buffer.
4. The method of claim 1, wherein information about contents of the buffer further includes a number of events stored in the buffer.
5. The method of claim 1, further comprising:
 - in response to a first trigger condition, generating a data chunk based on contents of the metadata structure and further based on contents of the buffer
6. The method of claim 5, wherein the first trigger condition is based on a buffer usage threshold or based on a timeout window.
7. The method of claim 5, further comprising:
 - in response to a second trigger condition, reclaiming storage space used by the data chunk.
8. The method of claim 7, wherein the second trigger condition is based on a retention policy associated with the data chunk.
9. The method of claim 7, wherein the second trigger condition is based on a disk-space usage threshold or based on a maximum time to retain the chunk.
10. The method of claim 1, further comprising:
 - for each event in the plurality of events:
 - determining when the event was received; and
 - storing, in the buffer, a timestamp that reflects when the event was received.
11. The method of claim 1, wherein storing the event in the buffer comprises appending the event to content of the buffer.

12. The method of claim 1, further comprising generating a data chunk that includes contents of the metadata structure and a compressed version of contents of the buffer.
13. The method of claim 12, wherein the data chunk further includes a file signature or a version identifier.
14. The method of claim 12, wherein the data chunk further includes a message digest of contents of the buffer.
15. The method of claim 12, further comprising:
- receiving a search query that includes a set of one or more search terms;
 - identifying one or more search terms, from the set of search terms, that concern information that is contained in the metadata structure; and
 - searching one or more data chunks by comparing, for each data chunk, the identified search terms to contents of the metadata structure included within the data chunk.
16. The method of claim 15, further comprising:
- for each data chunk that satisfies the identified search terms:
 - disassembling the data chunk into a plurality of events; and
 - comparing, for each event in the plurality of events, the set of search terms to the event.
17. A computer program product for processing log data, the computer program product comprising a computer-readable medium containing computer program code for performing a method, the method comprising:
- receiving log data that comprises a plurality of events, an event including one or more fields; and
 - for each event in the plurality of events:
 - storing, in a buffer, the event; and
 - updating a metadata structure that comprises information about contents of the buffer, wherein information about contents of the buffer includes a first minimum value that reflects a minimum value of a first field of all of the events stored in the buffer.
18. An apparatus for processing log data, comprising:
- a receiving module configured to receive log data that comprises a plurality of events, an event including one or more fields;
 - a buffer module configured to store, for each event in the plurality of events, the event in a buffer; and
 - a metadata module configured to update, for each event in the plurality of events, a metadata structure that comprises information about contents of the buffer, wherein information about contents of the buffer includes a first minimum value that reflects a minimum value of a first field of all of the events stored in the buffer.