



(19) **United States**

(12) **Patent Application Publication**
Pawar et al.

(10) **Pub. No.: US 2014/0181038 A1**

(43) **Pub. Date: Jun. 26, 2014**

(54) **SYSTEMS AND METHODS TO CATEGORIZE UNPROTECTED VIRTUAL MACHINES**

Publication Classification

(71) Applicant: **CommVault Systems, Inc.**, Oceanport, NJ (US)

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(72) Inventors: **Rahul S. Pawar**, Marlboro, NJ (US); **Rajiv Kottomtharayil**, Marlboro, NJ (US); **Parag Gokhale**, Marlboro, NJ (US); **Sumer Dilip Deshpande**, Ocean, NJ (US); **Henry Wallace Dornemann**, Eatontown, NJ (US)

(52) **U.S. Cl.**
CPC **G06F 17/30091** (2013.01)
USPC **707/650**

(57) **ABSTRACT**

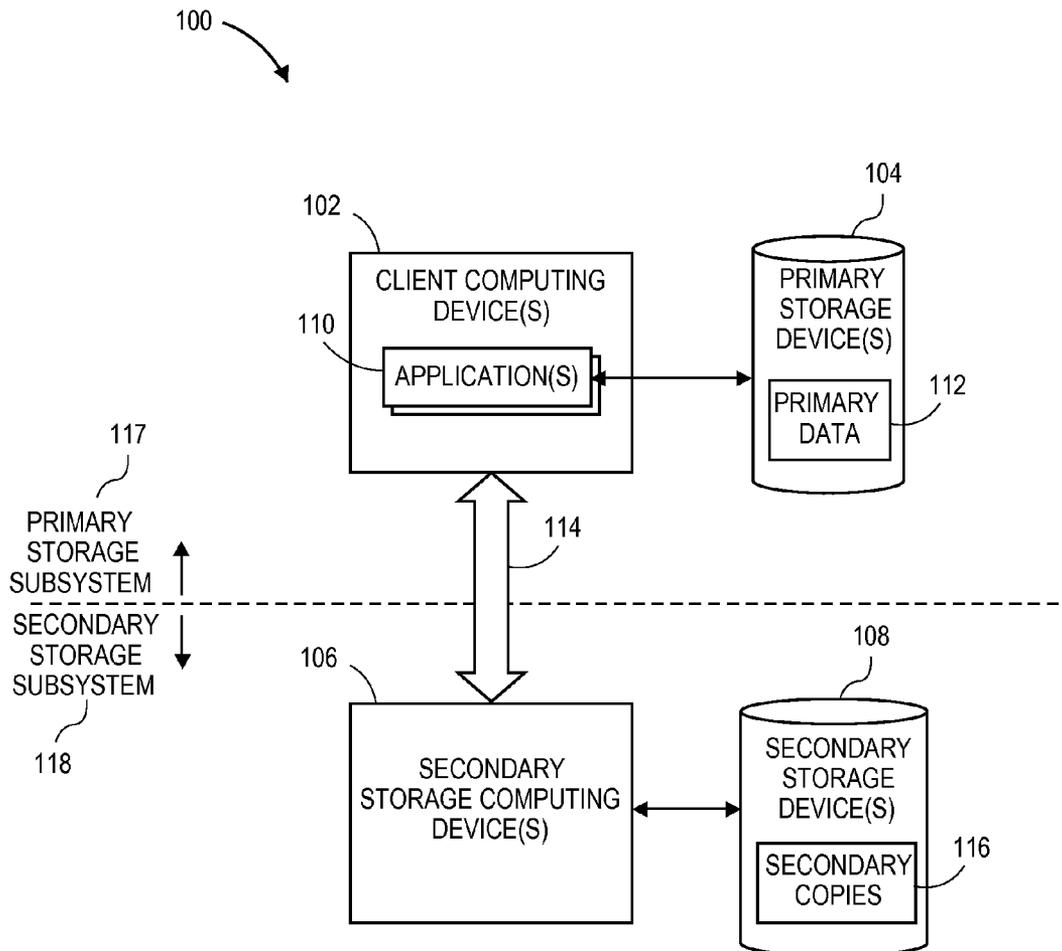
Virtualization sprawl can lead to virtual machines with no designated periodic backup. If the data associated with these unprotected virtual machines is not backed up, it cannot be restored if needed, leading to system failures. A data storage system identifies and protects the unprotected virtual machines. For instance, the system automatically determines whether characteristics of the unprotected virtual machines meet characterization rules for backup policies and assigns backup policies to the unprotected virtual machines whose characteristics meet the characterization rules for a certain backup policy. The system further identifies those unprotected virtual machines that remain uncategorized and applies a default backup policy to the uncategorized and unprotected virtual machines to provide protected virtual machines.

(21) Appl. No.: **13/930,918**

(22) Filed: **Jun. 28, 2013**

Related U.S. Application Data

(60) Provisional application No. 61/740,929, filed on Dec. 21, 2012.



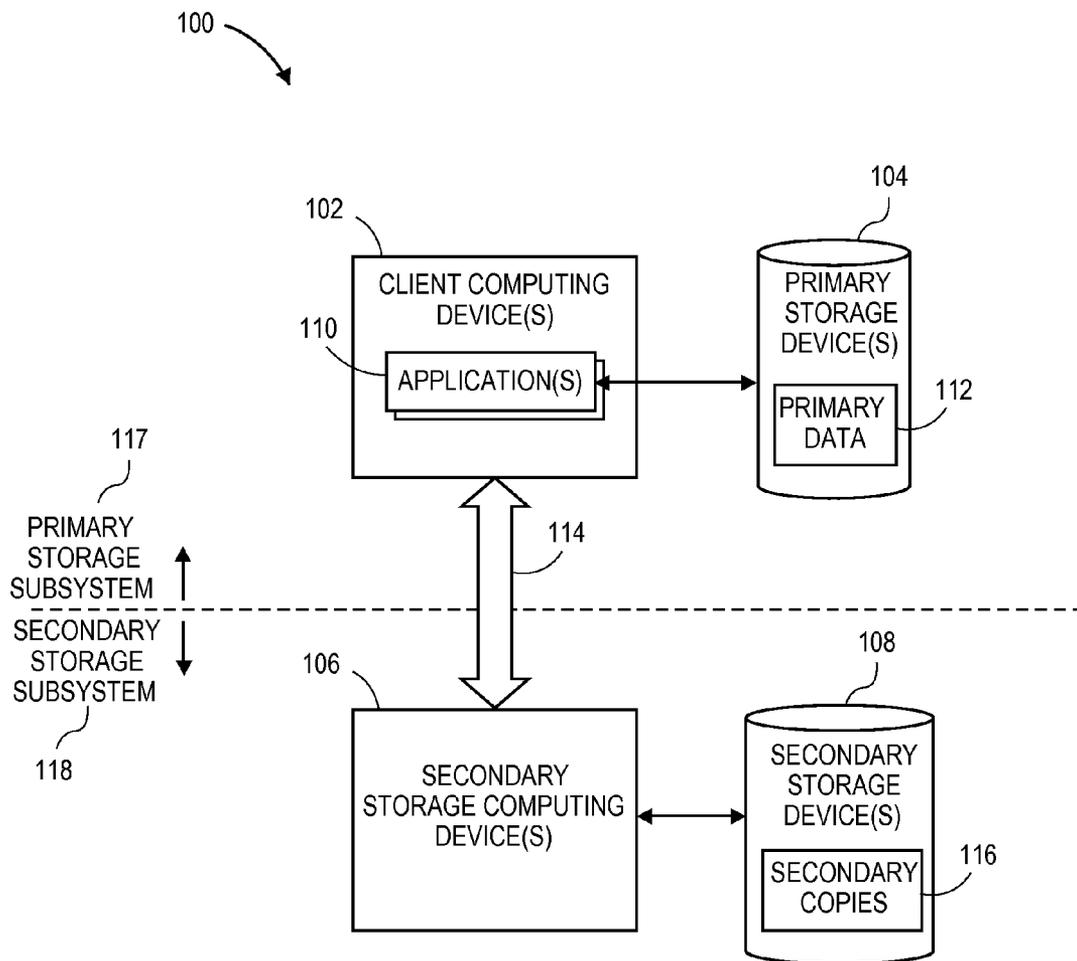


FIG. 1A

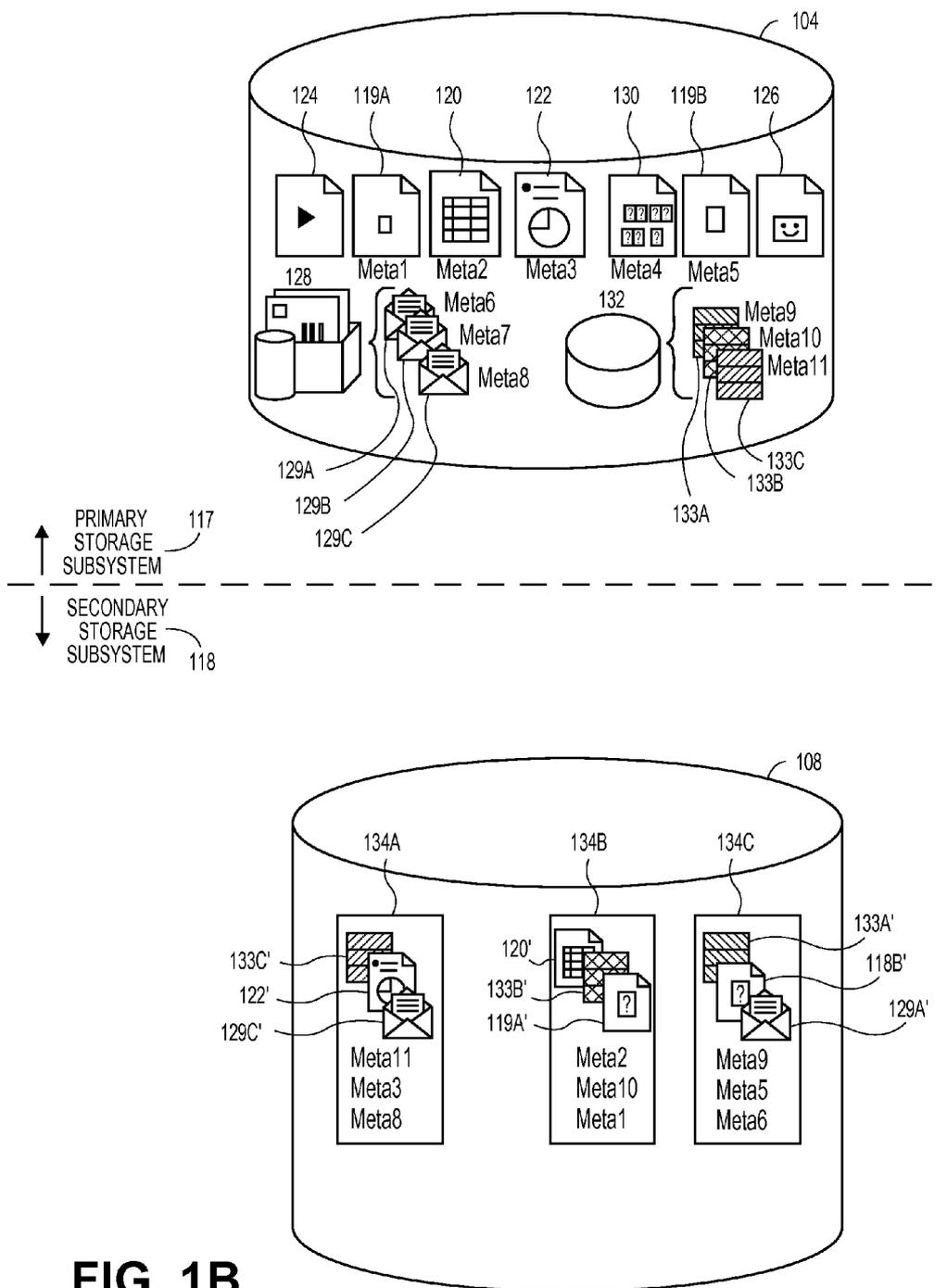


FIG. 1B

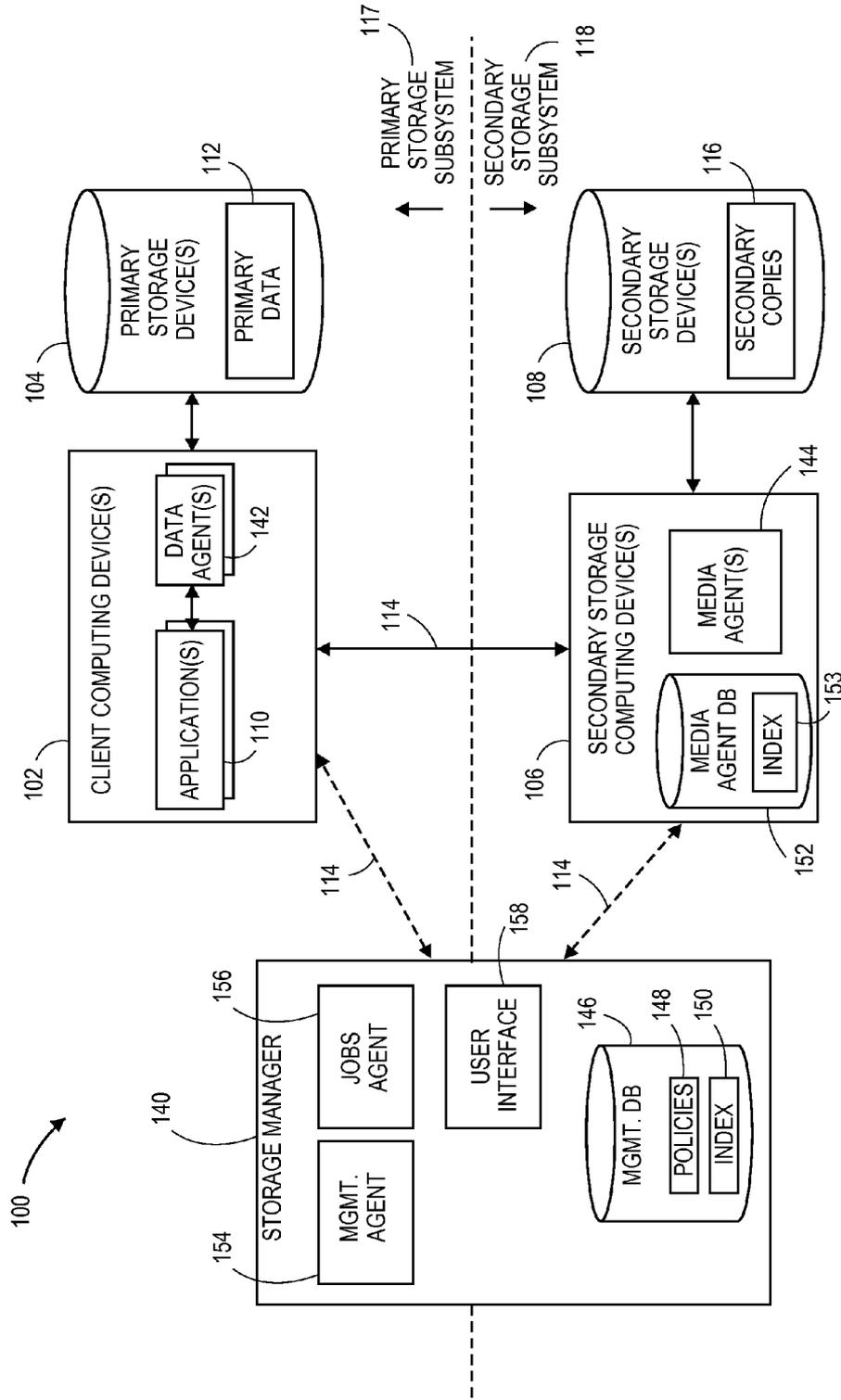


FIG. 1C

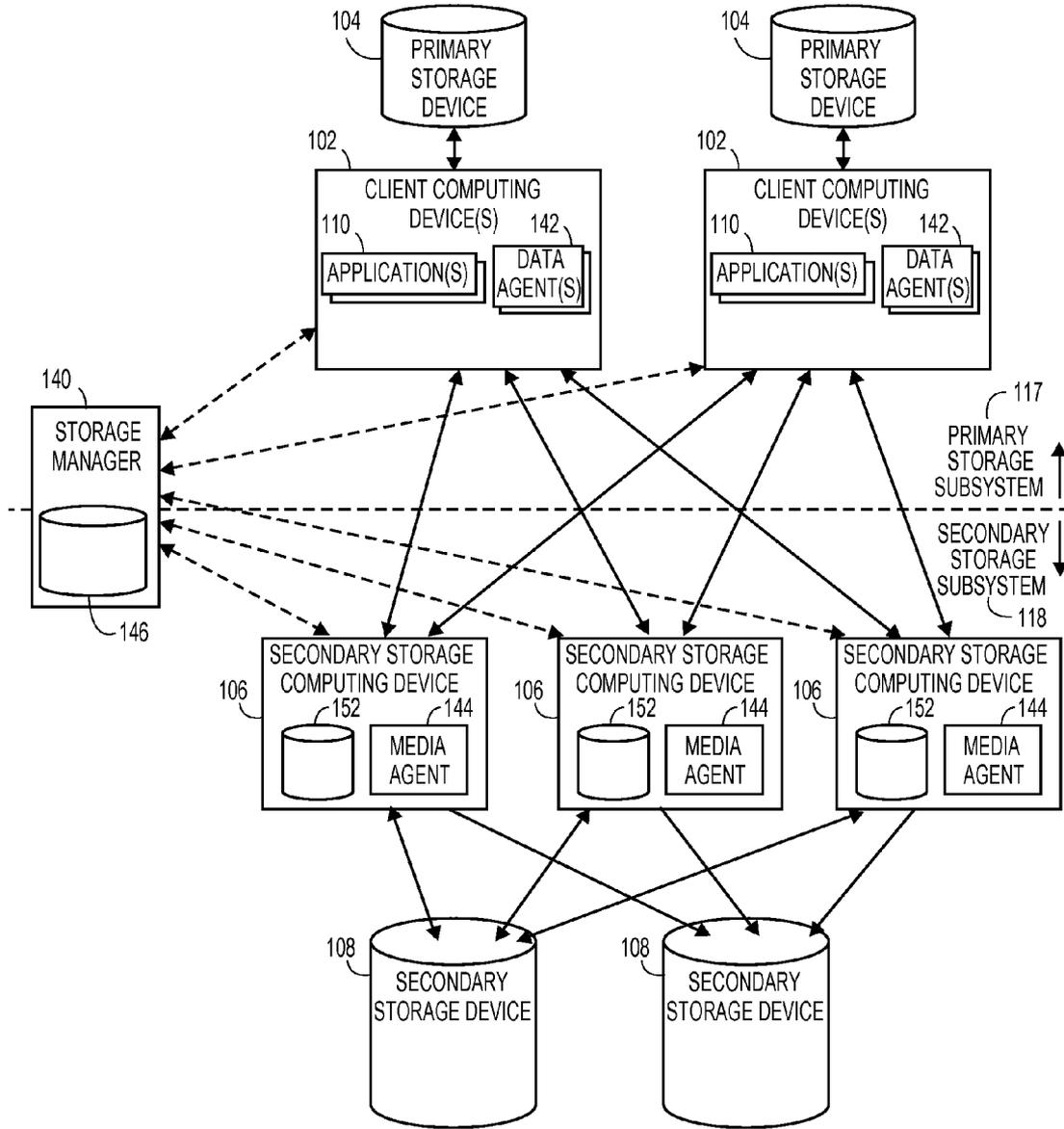


FIG. 1D

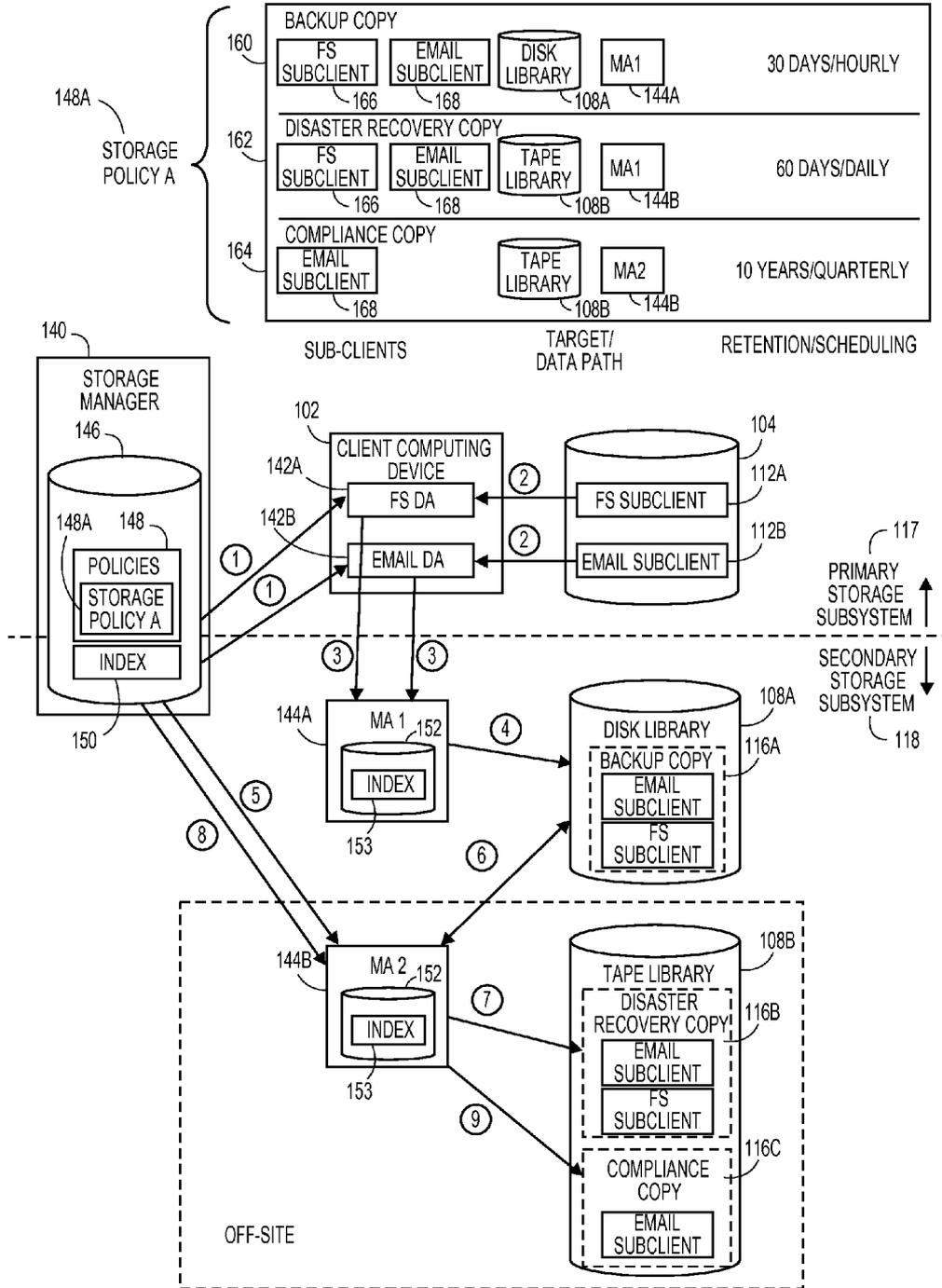


FIG. 1E

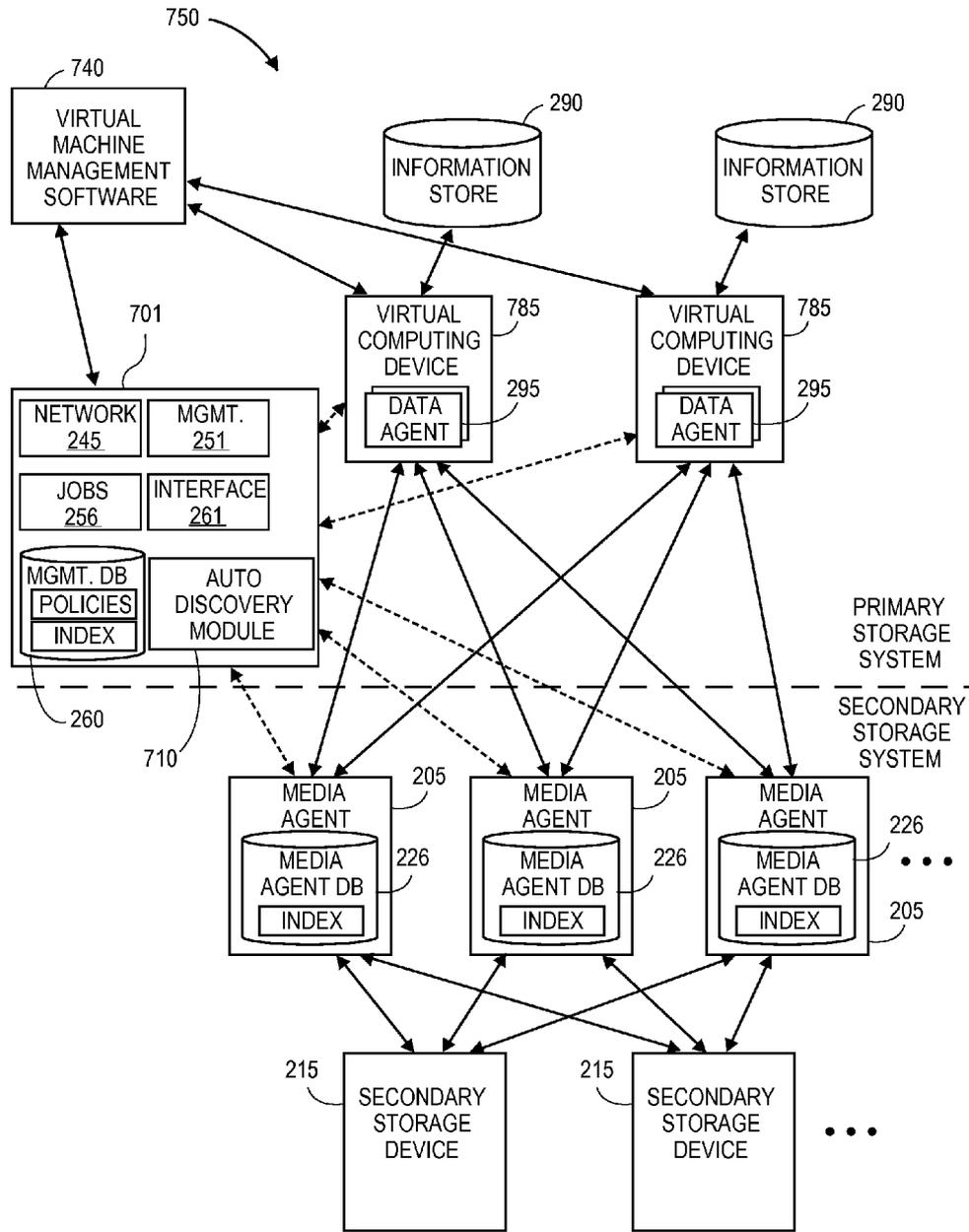


FIG. 2

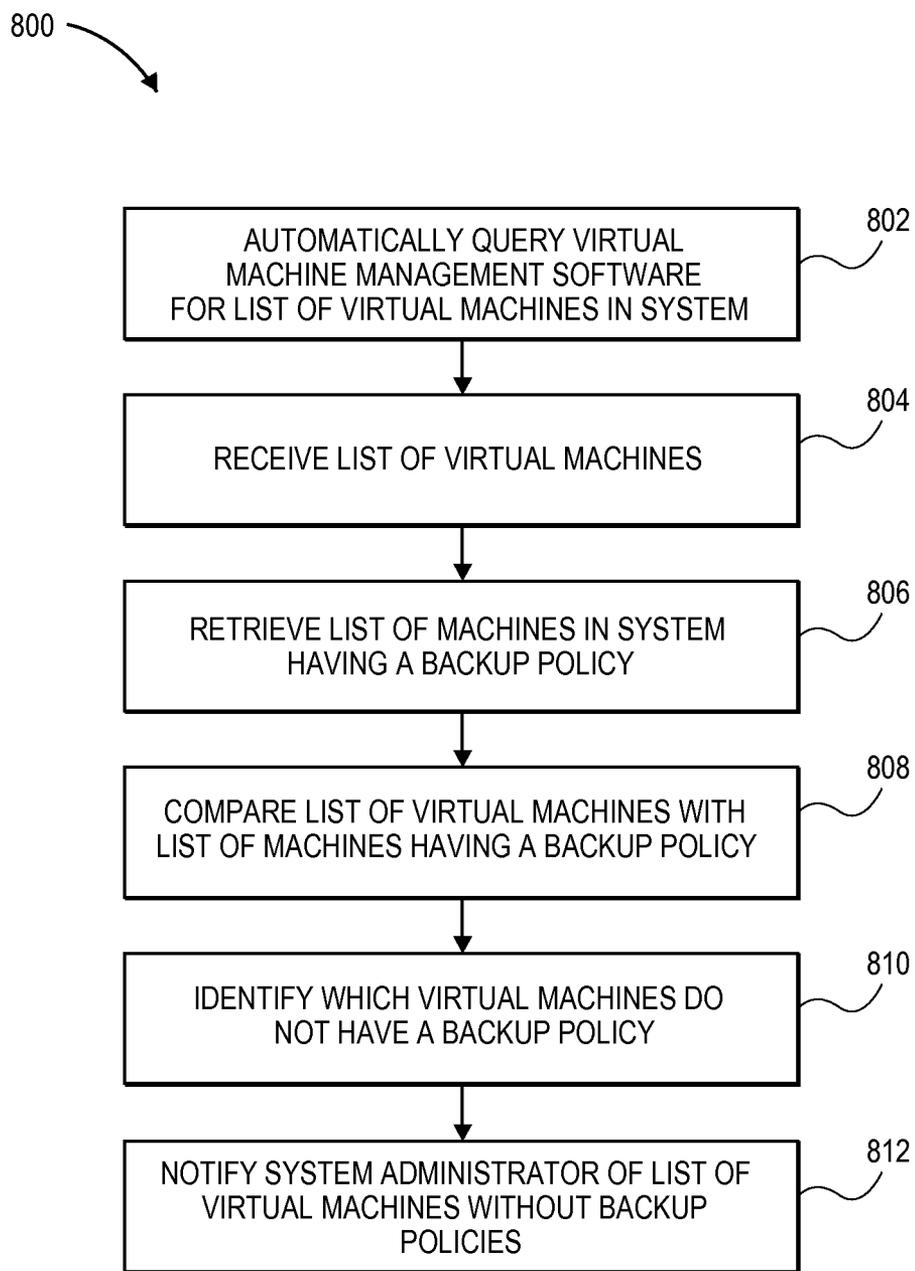


FIG. 3

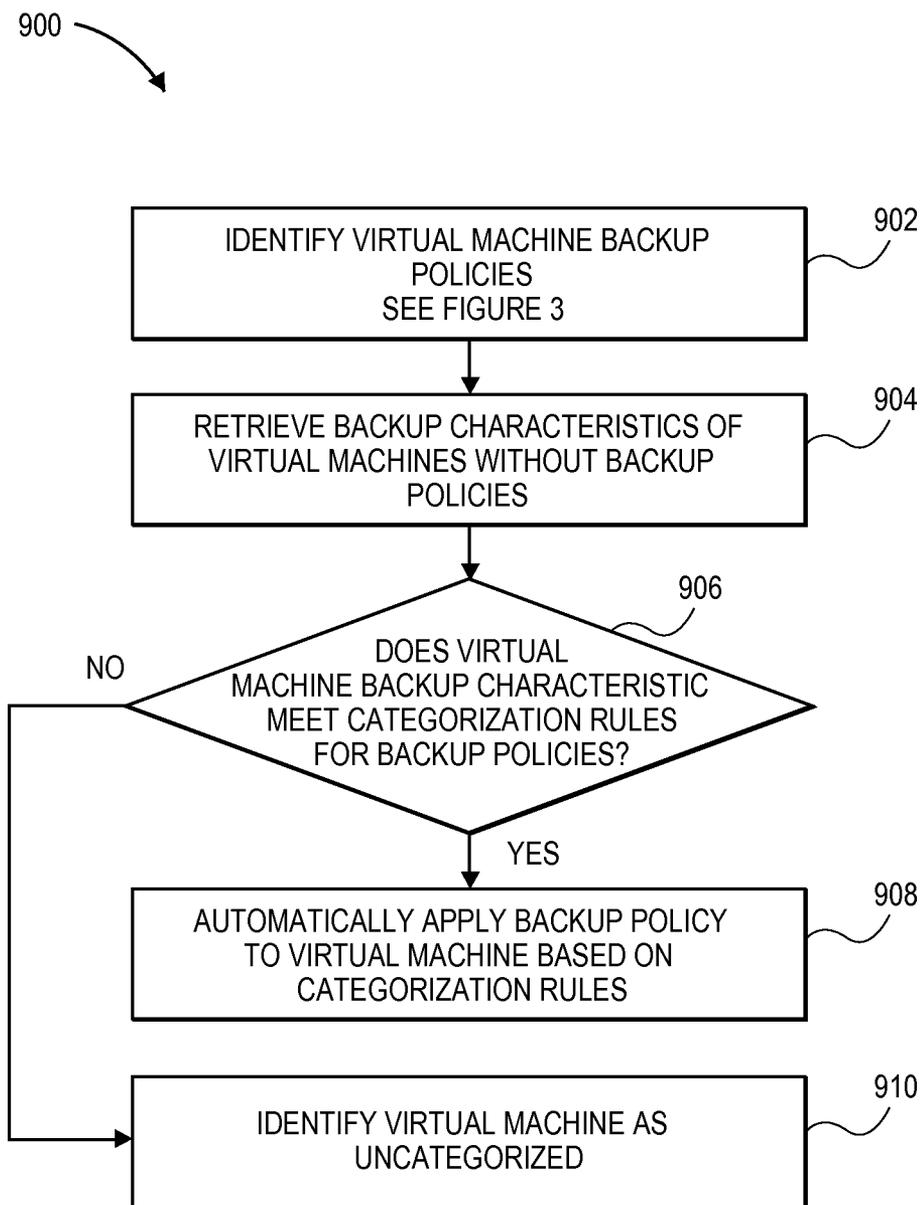


FIG. 4

1000

Rule based discovery

Discover By

DataStore Affinity
Cluster Affinity
Data Center Affinity
DataStore Affinity
ESX Server Affinity
Match guest host names by regular expression
OS Group
ResourcePool Affinity
VApp Affinity

Automatic

vcbmounter

FIG. 5

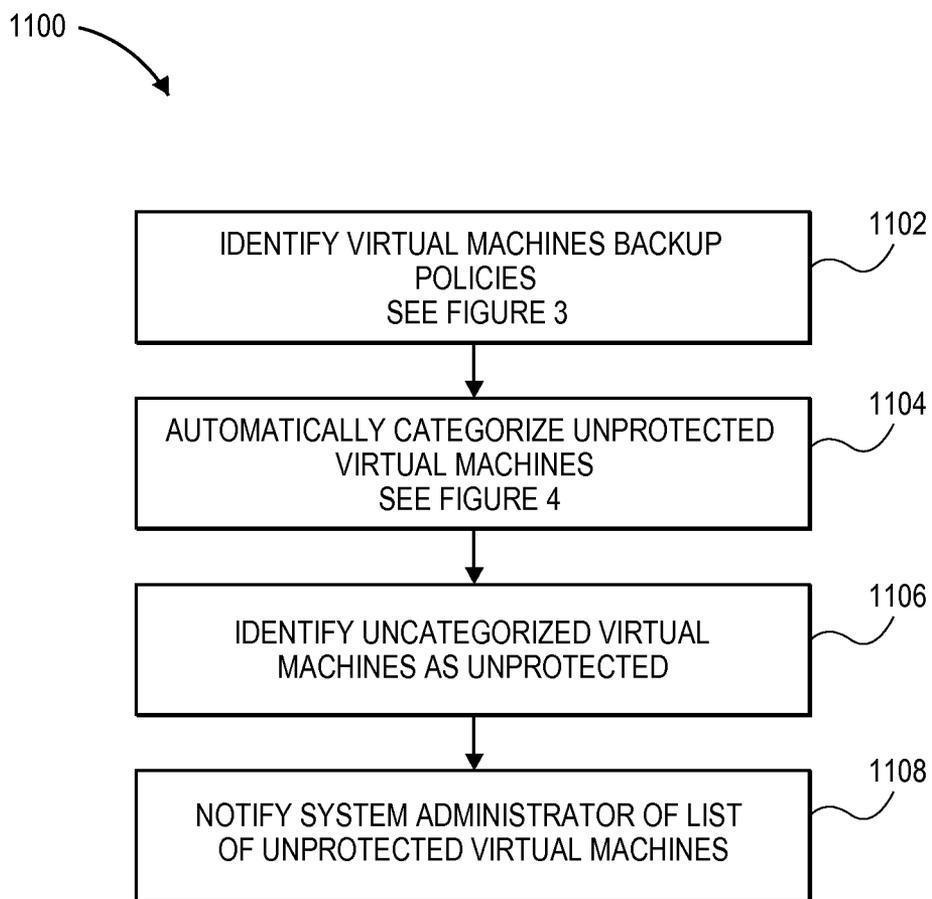


FIG. 6

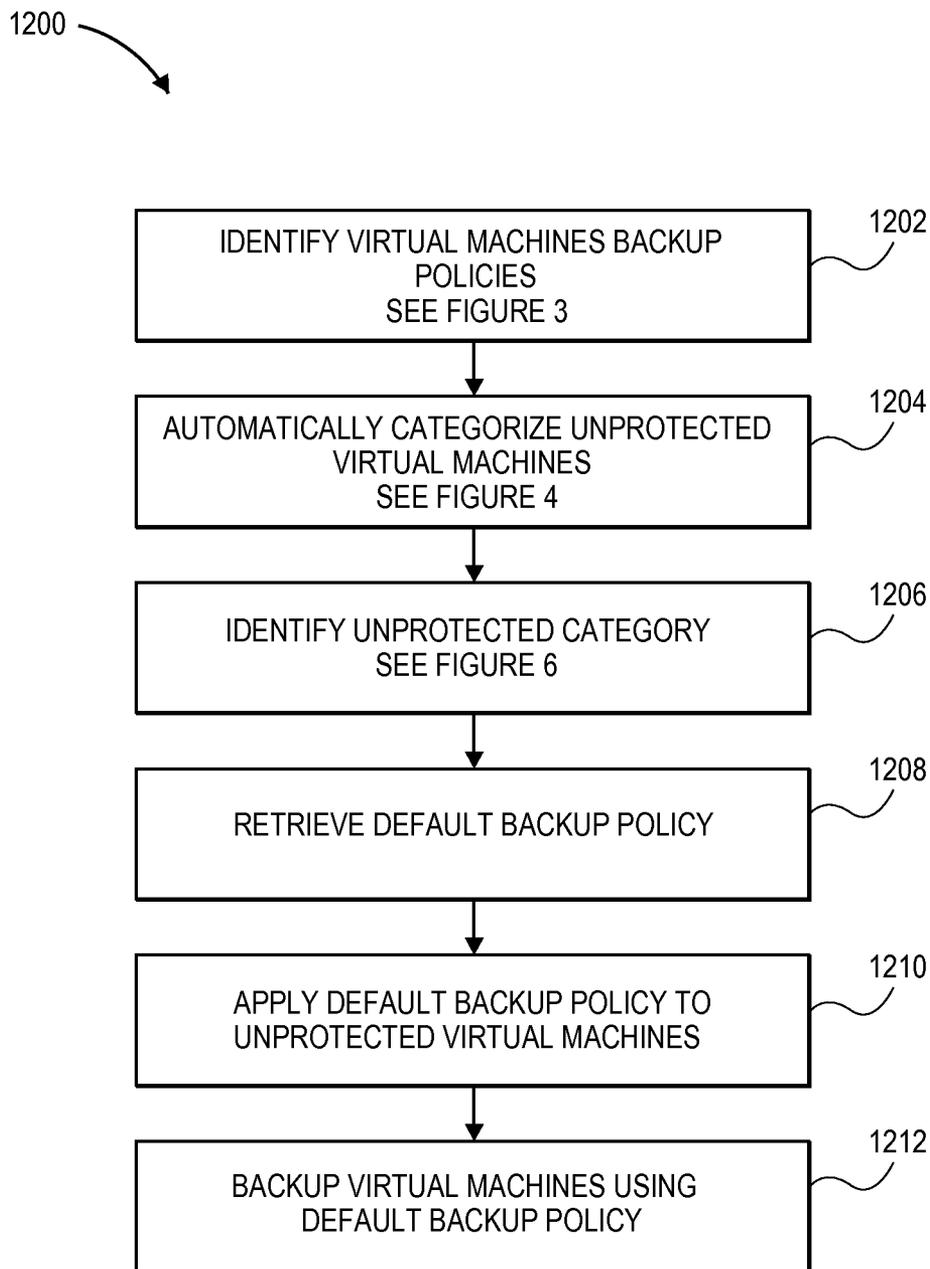


FIG. 7

SYSTEMS AND METHODS TO CATEGORIZE UNPROTECTED VIRTUAL MACHINES

INCORPORATION BY REFERENCE TO ANY PRIORITY APPLICATIONS

[0001] Any and all applications for which a foreign or domestic priority claim is identified in the Application Data Sheet as filed with the present application are hereby incorporated by reference under 37 CFR 1.57.

BACKGROUND

[0002] Businesses worldwide recognize the commercial value of their data and seek reliable, cost-effective ways to protect the information stored on their computer networks while minimizing impact on productivity. Protecting information is often part of a routine process that is performed within an organization.

[0003] A company might back up critical computing systems such as databases, file servers, web servers, and so on as part of a daily, weekly, or monthly maintenance schedule. The company may similarly protect computing systems used by each of its employees, such as those used by an accounting department, marketing department, engineering department, and so forth.

[0004] Given the rapidly expanding volume of data under management, companies also continue to seek innovative techniques for managing data growth, in addition to protecting data. For instance, companies often implement migration techniques for moving data to lower cost storage over time and data reduction techniques for reducing redundant data, pruning lower priority data, etc.

[0005] Enterprises also increasingly view their stored data as a valuable asset. Along these lines, customers are looking for solutions that not only protect and manage, but also leverage their data. For instance, solutions providing data analysis capabilities, improved data presentation and access features, and the like, are in increasing demand.

[0006] Increasingly, companies are turning to virtualized computing devices to store, manipulate, and display the information that is constantly subject to change. The term virtualization in the computing arts can refer to the creation of a virtual instance of an entity (e.g., a hardware platform, operating system, storage device or network resource, etc.) that behaves like a physical instance. For instance, a virtual machine can be a software representation of a physical machine. Companies currently use virtualization for a variety of purposes, such as to reduce the number of physical servers or other computers by instantiating multiple virtual machines on a single physical host computer. In this manner, virtualization can be used to centralize administrative tasks while improving scalability and work loads, and can be an important tool for maximizing hardware utilization.

[0007] As such, enterprises are generating and backing up or otherwise protecting increasing volumes of data, presenting significant data management challenges.

SUMMARY

[0008] The benefits of server virtualization are compelling and are driving the transition to large scale virtual server deployments. From the cost savings recognized through server consolidation or business flexibility and agility inherent in cloud architectures, virtualization technologies are being rapidly deployed.

[0009] However, virtual server deployment is having unintended consequences for data protection. Virtual machines are easy to create. They can pop up over night with no warning and little communication with network administrators. Network administrators should know how many virtual machines are running on their network, which physical machines are hosting them, where their storage is located, what software is associated with them, why they were originally created, and who created them. But often they do not. This leads to virtual machines with no designated periodic backup or, in other words, unprotected virtual machines. If the data associated with these unprotected virtual machines is not backed up, it cannot be restored if needed, leading to system failures.

[0010] Virtual machines have the same support, security, and compliance issues as physical machines. Traditional backup tools require manual intervention to configure a backup schedule and policy for each newly created virtual machine. Not only is this process time consuming in terms of manual configuration, but also requires time to determine where each virtual machine came from, and then determine to most appropriate backup schedule to apply to each virtual machine. Virtualization sprawl occurs when the number of virtual machines on a network reaches a point where the administrator can no longer manage them effectively.

[0011] Due to the above challenges, it can be important to provide efficient, user-friendly tools for facilitating effective backup policies and for providing access to backup data. Systems and methods are provided herein to automatically discover and protect unprotected virtual machines.

[0012] Certain embodiments relate to a method to categorize unprotected virtual machines in a data management system. The method comprises automatically identifying with one or more computer processors one or more virtual machines not associated with a backup policy in a data management system, automatically retrieving with one or more computer processors backup characteristics of each virtual machine not associated with a backup policy, automatically retrieving with one or more computer processors one or more categories, each category associated with criteria and a backup policy, automatically comparing with one or more computer processors the backup characteristics of each virtual machine not associated with a backup policy with the criteria associated with each category, and when the backup characteristics of the virtual machine match the criteria associated with the category, automatically applying with one or more computer processors the associated backup policy to the virtual machine.

[0013] According to certain embodiments, a system to categorize unprotected virtual machines in an information management cell is disclosed. The system comprises computer hardware including one or more computer processors, and computer-readable storage comprising computer-readable instructions that, when executed by the one or more processors, cause the computer hardware to perform operations defined by the computer-readable instructions. The computer-readable instructions are configured to automatically identify one or more virtual machines not associated with a backup policy in an information management cell, automatically retrieve backup characteristics of each virtual machine not associated with a backup policy, automatically retrieve one or more categories, each category associated with criteria and a backup policy, automatically compare the backup characteristics of each virtual machine not associated with a backup policy with the criteria associated with each category,

and when the backup characteristics of the virtual machine match the criteria associated with the category, automatically apply the associated backup policy to the virtual machine.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0014] FIG. 1A is a block diagram illustrating an exemplary information management system.
- [0015] FIG. 1B is a detailed view of a primary storage device, a secondary storage device, and some examples of primary data and secondary copy data.
- [0016] FIG. 1C is a block diagram of an exemplary information management system including a storage manager, one or more data agents, and one or more media agents.
- [0017] FIG. 1D is a block diagram illustrating a scalable information management system.
- [0018] FIG. 1E illustrates certain secondary copy operations according to an exemplary storage policy.
- [0019] FIG. 2 is a block diagram illustrating another exemplary information management cell.
- [0020] FIG. 3 illustrates a flow chart of an exemplary embodiment of a process to identify virtual machine backup policies usable by the system of FIG. 2.
- [0021] FIG. 4 illustrates a flow chart of an exemplary embodiment of a process to automatically categorize unprotected virtual machines usable by the system of FIG. 2.
- [0022] FIG. 5 is an exemplary screen shot of available auto-discovery rules usable by the system of FIG. 2.
- [0023] FIG. 6 illustrates a flow chart of an exemplary embodiment of a process to identify an unprotected category of virtual machines usable by the system of FIG. 2.
- [0024] FIG. 7 illustrates a flow chart of an exemplary embodiment of a process to protect unprotected virtual machines usable by the system of FIG. 2.

DETAILED DESCRIPTION

Information Management System Overview

- [0025] Systems and methods are described herein to automatically discover and protect unprotected virtual machines. Further examples of systems and methods for 1) identifying virtual machine backup policies; 2) automatically categorizing unprotected virtual machines; 3) identifying uncategorized virtual machines as unprotected; and 4) protecting unprotected virtual machines are described below with respect to FIGS. 2-7.
- [0026] Moreover, it will be appreciated that data generated by information management systems such as those that will now be described with respect to FIGS. 1A-1E can be protected as well. And, as will be described, the componentry for implementing secondary data operations can be incorporated into such systems.
- [0027] With the increasing importance of protecting and leveraging data, organizations simply cannot afford to take the risk of losing critical data. Moreover, runaway data growth and other modern realities make protecting and managing data an increasingly difficult task. There is therefore a need for efficient, powerful, and user-friendly solutions for protecting and managing data.
- [0028] Depending on the size of the organization, there are typically many data production sources which are under the purview of tens, hundreds, or even thousands of employees or other individuals. In the past, individual employees were sometimes responsible for managing and protecting their

data. A patchwork of hardware and software point solutions have been applied in other cases. These solutions were often provided by different vendors and had limited or no interoperability.

[0029] Certain embodiments described herein provide systems and methods capable of addressing these and other shortcomings of prior approaches by implementing unified, organization-wide information management. FIG. 1A shows one such information management system 100, which generally includes combinations of hardware and software configured to protect and manage data and metadata generated and used by the various computing devices in the information management system 100.

[0030] The organization which employs the information management system 100 may be a corporation or other business entity, non-profit organization, educational institution, household, governmental agency, or the like.

[0031] Generally, the systems and associated components described herein may be compatible with and/or provide some or all of the functionality of the systems and corresponding components described in one or more of the following U.S. patents and patent application publications assigned to CommVault Systems, Inc., each of which is hereby incorporated in its entirety by reference herein:

- [0032] U.S. Pat. Pub. No. 2010/0332456, entitled "DATA OBJECT STORE AND SERVER FOR A CLOUD STORAGE ENVIRONMENT, INCLUDING DATA DEDUPLICATION AND DATA MANAGEMENT ACROSS MULTIPLE CLOUD STORAGE SITES";
- [0033] U.S. Pat. No. 7,035,880, entitled "MODULAR BACKUP AND RETRIEVAL SYSTEM USED IN CONJUNCTION WITH A STORAGE AREA NETWORK";
- [0034] U.S. Pat. No. 7,343,453, entitled "HIERARCHICAL SYSTEMS AND METHODS FOR PROVIDING A UNIFIED VIEW OF STORAGE INFORMATION";
- [0035] U.S. Pat. No. 7,395,282, entitled "HIERARCHICAL BACKUP AND RETRIEVAL SYSTEM";
- [0036] U.S. Pat. No. 7,246,207, entitled "SYSTEM AND METHOD FOR DYNAMICALLY PERFORMING STORAGE OPERATIONS IN A COMPUTER NETWORK";
- [0037] U.S. Pat. No. 7,747,579, entitled "METABASE FOR FACILITATING DATA CLASSIFICATION";
- [0038] U.S. Pat. No. 8,229,954, entitled "MANAGING COPIES OF DATA";
- [0039] U.S. Pat. No. 7,617,262, entitled "SYSTEM AND METHODS FOR MONITORING APPLICATION DATA IN A DATA REPLICATION SYSTEM";
- [0040] U.S. Pat. No. 7,529,782, entitled "SYSTEM AND METHODS FOR PERFORMING A SNAPSHOT AND FOR RESTORING DATA";
- [0041] U.S. Pat. No. 8,230,195, entitled "SYSTEM AND METHOD FOR PERFORMING AUXILIARY STORAGE OPERATIONS";
- [0042] U.S. Pat. Pub. No. 2012/0084268, entitled "CONTENT-ALIGNED, BLOCK-BASED DEDUPLICATION";
- [0043] U.S. Pat. Pub. No. 2006/0224846, entitled "SYSTEM AND METHOD TO SUPPORT SINGLE INSTANCE STORAGE OPERATIONS";

[0044] U.S. Pat. Pub. No. 2009/0329534, entitled “APPLICATION-AWARE AND REMOTE SINGLE INSTANCE DATA MANAGEMENT”;

[0045] U.S. Pat. Pub. No. 2012/0150826, entitled “DISTRIBUTED DEDUPLICATED STORAGE SYSTEM”;

[0046] U.S. Pat. Pub. No. 2012/0150818, entitled “CLIENT-SIDE REPOSITORY IN A NETWORKED DEDUPLICATED STORAGE SYSTEM”;

[0047] U.S. Pat. No. 8,170,995, entitled “METHOD AND SYSTEM FOR OFFLINE INDEXING OF CONTENT AND CLASSIFYING STORED DATA”; and

[0048] U.S. Pat. No. 8,156,086, entitled “SYSTEMS AND METHODS FOR STORED DATA VERIFICATION”.

[0049] The illustrated information management system 100 includes one or more client computing device 102 having at least one application 110 executing thereon, and one or more primary storage devices 104 storing primary data 112. The client computing device(s) 102 and the primary storage devices 104 may generally be referred to in some cases as a primary storage subsystem 117.

[0050] Depending on the context, the term “information management system” can refer to generally all of the illustrated hardware and software components. Or, in other instances, the term may refer to only a subset of the illustrated components.

[0051] For instance, in some cases information management system 100 generally refers to a combination of specialized components used to protect, move, manage, manipulate and/or process data and metadata generated by the client computing devices 102. However, the term may generally not refer to the underlying components that generate and/or store the primary data 112, such as the client computing devices 102 themselves, the applications 110 and operating system residing on the client computing devices 102, and the primary storage devices 104.

[0052] As an example, “information management system” may sometimes refer only to one or more of the following components and corresponding data structures: storage managers, data agents, and media agents. These components will be described in further detail below.

Client Computing Devices

[0053] There are typically a variety of sources in an organization that produce data to be protected and managed. As just one illustrative example, in a corporate environment such data sources can be employee workstations and company servers such as a mail server, a web server, or the like. In the information management system 100, the data generation sources include the one or more client computing devices 102.

[0054] The client computing devices 102 may include, without limitation, one or more: workstations, personal computers, desktop computers, or other types of generally fixed computing systems such as mainframe computers and mini-computers.

[0055] The client computing devices 102 can also include mobile or portable computing devices, such as one or more laptops, tablet computers, personal data assistants, mobile phones (such as smartphones), and other mobile or portable computing devices such as embedded computers, set top boxes, vehicle-mounted devices, wearable computers, etc.

[0056] In some cases, each client computing device 102 is associated with one or more users and/or corresponding user accounts, of employees or other individuals.

[0057] The term “client computing device” is used herein because the information management system 100 generally “serves” the data management and protection needs for the data generated by the client computing devices 102. However, the use of this term does not imply that the client computing devices 102 cannot be “servers” in other respects. For instance, a particular client computing device 102 may act as a server with respect to other devices, such as other client computing devices 102. As just a few examples, the client computing devices 102 can include mail servers, file servers, database servers, and web servers.

[0058] The client computing devices 102 may additionally include virtualized and/or cloud computing resources. For instance, one or more virtual machines may be provided to the organization by a third-party cloud service vendor. Or, in some embodiments, the client computing devices 102 include one or more virtual machine(s) running on a virtual machine host computing device operated by the organization. As one example, the organization may use one virtual machine as a database server and another virtual machine as a mail server. A virtual machine manager (VMM) (e.g., a Hypervisor) may manage the virtual machines, and reside and execute on the virtual machine host computing device.

[0059] Each client computing device 102 may have one or more applications 110 (e.g., software applications) executing thereon which generate and manipulate the data that is to be protected from loss.

[0060] The applications 110 generally facilitate the operations of an organization (or multiple affiliated organizations), and can include, without limitation, mail server applications (e.g., Microsoft Exchange Server), file server applications, mail client applications (e.g., Microsoft Exchange Client), database applications (e.g., SQL, Oracle, SAP, Lotus Notes Database), word processing applications (e.g., Microsoft Word), spreadsheet applications, financial applications, presentation applications, browser applications, mobile applications, entertainment applications, and so on.

[0061] The applications 110 can include at least one operating system (e.g., Microsoft Windows, Mac OS X, iOS, IBM z/OS, Linux, other Unix-based operating systems, etc.), which may support one or more file systems and host the other applications 110.

[0062] As shown, the client computing devices 102 and other components in the information management system 100 can be connected to one another via one or more communication pathways 114. The communication pathways 114 can include one or more networks or other connection types including as any of following, without limitation: the Internet, a wide area network (WAN), a local area network (LAN), a Storage Area Network (SAN), a Fibre Channel connection, a Small Computer System Interface (SCSI) connection, a virtual private network (VPN), a token ring or TCP/IP based network, an intranet network, a point-to-point link, a cellular network, a wireless data transmission system, a two-way cable system, an interactive kiosk network, a satellite network, a broadband network, a baseband network, other appropriate wired, wireless, or partially wired/wireless computer or telecommunications networks, combinations of the same or the like. The communication pathways 114 in some cases may also include application programming interfaces

(APIs) including, e.g., cloud service provider APIs, virtual machine management APIs, and hosted service provider APIs.

Primary Data and Exemplary Primary Storage Devices

[0063] Primary data **112** according to some embodiments is production data or other “live” data generated by the operating system and other applications **110** residing on a client computing device **102**. The primary data **112** is stored on the primary storage device(s) **104** and is organized via a file system supported by the client computing device **102**. For instance, the client computing device(s) **102** and corresponding applications **110** may create, access, modify, write, delete, and otherwise use primary data **112**.

[0064] Primary data **112** is generally in the native format of the source application **110**. According to certain aspects, primary data **112** is an initial or first (e.g., created before any other copies or before at least one other copy) stored copy of data generated by the source application **110**. Primary data **112** in some cases is created substantially directly from data generated by the corresponding source applications **110**.

[0065] The primary data **112** may sometimes be referred to as a “primary copy” in the sense that it is a discrete set of data. However, the use of this term does not necessarily imply that the “primary copy” is a copy in the sense that it was copied or otherwise derived from another stored version.

[0066] The primary storage devices **104** storing the primary data **112** may be relatively fast and/or expensive (e.g., a disk drive, a hard-disk array, solid state memory, etc.). In addition, primary data **112** may be intended for relatively short term retention (e.g., several hours, days, or weeks).

[0067] According to some embodiments, the client computing device **102** can access primary data **112** from the primary storage device **104** by making conventional file system calls via the operating system. Primary data **112** representing files may include structured data (e.g., database files), unstructured data (e.g., documents), and/or semi-structured data. Some specific examples are described below with respect to FIG. 1B.

[0068] It can be useful in performing certain tasks to break the primary data **112** up into units of different granularities. In general, primary data **112** can include files, directories, file system volumes, data blocks, extents, or any other types or granularities of data objects. As used herein, a “data object” can refer to both (1) any file that is currently addressable by a file system or that was previously addressable by the file system (e.g., an archive file) and (2) a subset of such a file.

[0069] As will be described in further detail, it can also be useful in performing certain functions of the information management system **100** to access and modify metadata within the primary data **112**. Metadata generally includes information about data objects or characteristics associated with the data objects.

[0070] Metadata can include, without limitation, one or more of the following: the data owner (e.g., the client or user that generates the data), the last modified time (e.g., the time of the most recent modification of the data object), a data object name (e.g., a file name), a data object size (e.g., a number of bytes of data), information about the content (e.g., an indication as to the existence of a particular search term), to/from information for email (e.g., an email sender, recipient, etc.), creation date, file type (e.g., format or application type), last accessed time, application type (e.g., type of application that generated the data object), location/network (e.g.,

a current, past or future location of the data object and network pathways to/from the data object), frequency of change (e.g., a period in which the data object is modified), business unit (e.g., a group or department that generates, manages or is otherwise associated with the data object), and aging information (e.g., a schedule, such as a time period, in which the data object is migrated to secondary or long term storage), boot sectors, partition layouts, file location within a file folder directory structure, user permissions, owners, groups, access control lists [ACLs]), system metadata (e.g., registry information), combinations of the same or the like.

[0071] In addition to metadata generated by or related to file systems and operating systems, some of the applications **110** maintain indices of metadata for data objects, e.g., metadata associated with individual email messages. Thus, each data object may be associated with corresponding metadata. The use of metadata to perform classification and other functions is described in greater detail below.

[0072] Each of the client computing devices **102** are associated with and/or in communication with one or more of the primary storage devices **104** storing corresponding primary data **112**. A client computing device **102** may be considered to be “associated with” or “in communication with” a primary storage device **104** if it is capable of one or more of: storing data to the primary storage device **104**, retrieving data from the primary storage device **104**, and modifying data retrieved from a primary storage device **104**.

[0073] The primary storage devices **104** can include, without limitation, disk drives, hard-disk arrays, semiconductor memory (e.g., solid state drives), and network attached storage (NAS) devices. In some cases, the primary storage devices **104** form part of a distributed file system. The primary storage devices **104** may have relatively fast I/O times and/or are relatively expensive in comparison to the secondary storage devices **108**. For example, the information management system **100** may generally regularly access data and metadata stored on primary storage devices **104**, whereas data and metadata stored on the secondary storage devices **108** is accessed relatively less frequently.

[0074] In some cases, each primary storage device **104** is dedicated to an associated client computing devices **102**. For instance, a primary storage device **104** in one embodiment is a local disk drive of a corresponding client computing device **102**. In other cases, one or more primary storage devices **104** can be shared by multiple client computing devices **102**. As one example, a primary storage device **104** can be a disk array shared by a group of client computing devices **102**, such as one of the following types of disk arrays: EMC Clariion, EMC Symmetrix, EMC Celerra, Dell EqualLogic, IBM XIV, NetApp FAS, HP EVA, and HP 3PAR.

[0075] The information management system **100** may also include hosted services (not shown), which may be hosted in some cases by an entity other than the organization that employs the other components of the information management system **100**. For instance, the hosted services may be provided by various online service providers to the organization. Such service providers can provide services including social networking services, hosted email services, or hosted productivity applications or other hosted applications).

[0076] Hosted services may include software-as-a-service (SaaS), platform-as-a-service (PaaS), application service providers (ASPs), cloud services, or other mechanisms for delivering functionality via a network. As it provides services to users, each hosted service may generate additional data and

metadata under management of the information management system **100**, e.g., as primary data **112**. In some cases, the hosted services may be accessed using one of the applications **110**. As an example, a hosted mail service may be accessed via browser running on a client computing device **102**.

Secondary Copies and Exemplary Secondary Storage Devices

[0077] The primary data **112** stored on the primary storage devices **104** may be compromised in some cases, such as when an employee deliberately or accidentally deletes or overwrites primary data **112** during their normal course of work. Or the primary storage devices **104** can be damaged or otherwise corrupted.

[0078] For recovery and/or regulatory compliance purposes, it is therefore useful to generate copies of the primary data **112**. Accordingly, the information management system **100** includes one or more secondary storage computing devices **106** and one or more secondary storage devices **108** configured to create and store one or more secondary copies **116** of the primary data **112** and associated metadata. The secondary storage computing devices **106** and the secondary storage devices **108** may be referred to in some cases as a secondary storage subsystem **118**.

[0079] Creation of secondary copies **116** can help meet information management goals, such as: restoring data and/or metadata if an original version (e.g., of primary data **112**) is lost (e.g., by deletion, corruption, or disaster); allowing point-in-time recovery; complying with regulatory data retention and electronic discovery (e-discovery) requirements; reducing utilized storage capacity; facilitating organization and search of data; improving user access to data files across multiple computing devices and/or hosted services; and implementing data retention policies.

[0080] Types of secondary copy operations can include, without limitation, backup operations, archive operations, snapshot operations, replication operations (e.g., continuous data replication [CDR]), data retention policies such as or information lifecycle management and hierarchical storage management operations, and the like. These specific types of operations are discussed in greater detail below.

[0081] Regardless of the type of secondary copy operation, the client computing devices **102** access or receive primary data **112** and communicate the data, e.g., over the communication pathways **114**, for storage in the secondary storage device(s) **108**.

[0082] A secondary copy **116** can comprise a separate stored copy of application data that is derived from one or more earlier created, stored copies (e.g., derived from primary data **112** or another secondary copy **116**). Secondary copies **116** can include point-in-time data, and may be intended for relatively long-term retention (e.g., weeks, months or years), before some or all of the data is moved to other storage or is discarded.

[0083] In some cases, a secondary copy **116** is a copy of application data created and stored subsequent to at least one other stored instance (e.g., subsequent to corresponding primary data **112** or to another secondary copy **116**), in a different storage device than at least one previous stored copy, and/or remotely from at least one previous stored copy. Secondary copies **116** may be stored in relatively slow and/or low cost storage (e.g., magnetic tape). A secondary copy **116** may

be stored in a backup or archive format, or in some other format different than the native source application format or other primary data format.

[0084] In some cases, secondary copies **116** are indexed so users can browse and restore at another point in time. After creation of a secondary copy **116** representative of certain primary data **112**, a pointer or other location indicia (e.g., a stub) may be placed in primary data **112**, or be otherwise associated with primary data **112** to indicate the current location on the secondary storage device(s) **108**.

[0085] Since an instance a data object or metadata in primary data **112** may change over time as it is modified by an application **110** (or hosted service or the operating system), the information management system **100** may create and manage multiple secondary copies **116** of a particular data object or metadata, each representing the state of the data object in primary data **112** at a particular point in time. Moreover, since an instance of a data object in primary data **112** may eventually be deleted from the primary storage device **104** and the file system, the information management system **100** may continue to manage point-in-time representations of that data object, even though the instance in primary data **112** no longer exists.

[0086] For virtualized computing devices the operating system and other applications **110** of the client computing device(s) **102** may execute within or under the management of virtualization software (e.g., a VMM), and the primary storage device(s) **104** may comprise a virtual disk created on a physical storage device. The information management system **100** may create secondary copies **116** of the files or other data objects in a virtual disk file and/or secondary copies **116** of the entire virtual disk file itself (e.g., of an entire .vmdk file).

[0087] Secondary copies **116** may be distinguished from corresponding primary data **112** in a variety of ways, some of which will now be described. First, as discussed, secondary copies **116** can be stored in a different format (e.g., backup, archive, or other non-native format) than primary data **112**. For this or other reasons, secondary copies **116** may not be directly useable by the applications **110** of the client computing device **102**, e.g., via standard system calls or otherwise without modification, processing, or other intervention by the information management system **100**.

[0088] Secondary copies **116** are also often stored on a secondary storage device **108** that is inaccessible to the applications **110** running on the client computing devices **102** (and/or hosted services). Some secondary copies **116** may be “offline copies,” in that they are not readily available (e.g. not mounted to tape or disk). Offline copies can include copies of data that the information management system **100** can access without human intervention (e.g. tapes within an automated tape library, but not yet mounted in a drive), and copies that the information management system **100** can access only with at least some human intervention (e.g. tapes located at an offsite storage site).

[0089] The secondary storage devices **108** can include any suitable type of storage device such as, without limitation, one or more tape libraries, disk drives or other magnetic, non-tape storage devices, optical media storage devices, solid state storage devices, NAS devices, combinations of the same, and the like. In some cases, the secondary storage devices **108** are provided in a cloud (e.g. a private cloud or one operated by a third-party vendor).

[0090] The secondary storage device(s) 108 in some cases comprises a disk array or a portion thereof. In some cases, a single storage device (e.g., a disk array) is used for storing both primary data 112 and at least some secondary copies 116. In one example, a disk array capable of performing hardware snapshots stores primary data 112 and creates and stores hardware snapshots of the primary data 112 as secondary copies 116.

The Use of Intermediary Devices for Creating Secondary Copies

[0091] Creating secondary copies can be a challenging task. For instance, there can be hundreds or thousands of client computing devices 102 continually generating large volumes of primary data 112 to be protected. Also, there can be significant overhead involved in the creation of secondary copies 116. Moreover, secondary storage devices 108 may be special purpose components, and interacting with them can require specialized intelligence.

[0092] In some cases, the client computing devices 102 interact directly with the secondary storage device 108 to create the secondary copies 116. However, in view of the factors described above, this approach can negatively impact the ability of the client computing devices 102 to serve the applications 110 and produce primary data 112. Further, the client computing devices 102 may not be optimized for interaction with the secondary storage devices 108.

[0093] Thus, in some embodiments, the information management system 100 includes one or more software and/or hardware components which generally act as intermediaries between the client computing devices 102 and the secondary storage devices 108. In addition to off-loading certain responsibilities from the client computing devices 102, these intermediary components can provide other benefits. For instance, as discussed further below with respect to FIG. 1D, distributing some of the work involved in creating secondary copies 116 can enhance scalability.

[0094] The intermediary components can include one or more secondary storage computing devices 106 as shown in FIG. 1A and/or one or more media agents, which can be software modules residing on corresponding secondary storage computing devices 106 (or other appropriate devices). Media agents are discussed below (e.g., with respect to FIGS. 1C-1E).

[0095] The secondary storage computing device(s) 106 can comprise any appropriate type of computing device and can include, without limitation, any of the types of fixed and portable computing devices described above with respect to the client computing devices 102. In some cases, the secondary storage computing device(s) 106 include specialized hardware and/or software componentry for interacting with the secondary storage devices 108.

[0096] To create a secondary copy 116, the client computing device 102 communicates the primary data 112 to be copied (or a processed version thereof) to the designated secondary storage computing device 106, via the communication pathway 114. The secondary storage computing device 106 in turn conveys the received data (or a processed version thereof) to the secondary storage device 108. In some such configurations, the communication pathway 114 between the client computing device 102 and the secondary storage computing device 106 comprises a portion of a LAN, WAN or SAN. In other cases, at least some client computing devices

102 communicate directly with the secondary storage devices 108 (e.g., via Fibre Channel or SCSI connections).

Exemplary Primary Data and an Exemplary Secondary Copy

[0097] FIG. 1B is a detailed view showing some specific examples of primary data stored on the primary storage device(s) 104 and secondary copy data stored on the secondary storage device(s) 108, with other components in the system removed for the purposes of illustration. Stored on the primary storage device(s) 104 are primary data objects including word processing documents 119A-B, spreadsheets 120, presentation documents 122, video files 124, image files 126, email mailboxes 128 (and corresponding email messages 129A-C), html/xml or other types of markup language files 130, databases 132 and corresponding tables 133A-133C).

[0098] Some or all primary data objects are associated with a primary copy of object metadata (e.g., "Meta1-11"), which may be file system metadata and/or application specific metadata. Stored on the secondary storage device(s) 108 are secondary copy objects 134A-C which may include copies of or otherwise represent corresponding primary data objects and metadata.

[0099] As shown, the secondary copy objects 134A-C can individually represent more than one primary data object. For example, secondary copy data object 134A represents three separate primary data objects 133C, 122 and 129C (represented as 133C', 122' and 129C', respectively). Moreover, as indicated by the prime mark ('), a secondary copy object may store a representation of a primary data object or metadata differently than the original format, e.g., in a compressed, encrypted, deduplicated, or other modified format.

Exemplary Information Management System Architecture

[0100] The information management system 100 can incorporate a variety of different hardware and software components, which can in turn be organized with respect to one another in many different configurations, depending on the embodiment. There are critical design choices involved in specifying the functional responsibilities of the components and the role of each component in the information management system 100. For instance, as will be discussed, such design choices can impact performance as well as the adaptability of the information management system 100 to data growth or other changing circumstances.

[0101] FIG. 1C shows an information management system 100 designed according to these considerations and which includes: a central storage or information manager 140 configured to perform certain control functions, one or more data agents 142 executing on the client computing device(s) 102 configured to process primary data 112, and one or more media agents 144 executing on the one or more secondary storage computing devices 106 for performing tasks involving the secondary storage devices 108.

[0102] Storage Manager

[0103] As noted, the number of components in the information management system 100 and the amount of data under management can be quite large. Managing the components and data is therefore a significant task, and a task that can grow in an often unpredictable fashion as the quantity of components and data scale to meet the needs of the organization.

[0104] For these and other reasons, according to certain embodiments, responsibility for controlling the information management system 100, or at least a significant portion of that responsibility, is allocated to the storage manager 140.

[0105] By distributing control functionality in this manner, the storage manager 140 can be adapted independently according to changing circumstances. Moreover, a host computing device can be selected to best suit the functions of the storage manager 140. These and other advantages are described in further detail below with respect to FIG. 1D.

[0106] The storage manager 140 may be a software module or other application. The storage manager generally initiates, coordinates and/or controls storage and other information management operations performed by the information management system 100, e.g., to protect and control the primary data 112 and secondary copies 116 of data and metadata.

[0107] As shown by the dashed, arrowed lines, the storage manager 140 may communicate with and/or control some or all elements of the information management system 100, such as the data agents 142 and media agents 144. Thus, in certain embodiments, control information originates from the storage manager 140, whereas payload data and metadata is generally communicated between the data agents 142 and the media agents 144 (or otherwise between the client computing device(s) 102 and the secondary storage computing device(s) 106), e.g., at the direction of the storage manager 140. In other embodiments, some information management operations are controlled by other components in the information management system 100 (e.g., the media agent(s) 144 or data agent(s) 142), instead of or in combination with the storage manager 140.

[0108] According to certain embodiments, the storage manager provides one or more of the following functions:

[0109] initiating execution of secondary copy operations;

[0110] managing secondary storage devices 108 and inventory/capacity of the same;

[0111] allocating secondary storage devices 108 for secondary storage operations;

[0112] monitoring completion of and providing status reporting related to secondary storage operations;

[0113] tracking age information relating to secondary copies 116, secondary storage devices 108, and comparing the age information against retention guidelines;

[0114] tracking movement of data within the information management system 100;

[0115] tracking logical associations between components in the information management system 100;

[0116] protecting metadata associated with the information management system 100; and

[0117] implementing operations management functionality.

[0118] The storage manager 140 may maintain a database 146 of management-related data and information management policies 148. The database 146 may include a management index 150 or other data structure that stores logical associations between components of the system, user preferences and/or profiles (e.g., preferences regarding encryption, compression, or deduplication of primary or secondary copy data, preferences regarding the scheduling, type, or other aspects of primary or secondary copy or other operations, mappings of particular information management users or user accounts to certain computing devices or other components, etc.), management tasks, media containerization, or other

useful data. For example, the storage manager 140 may use the index 150 to track logical associations between media agents 144 and secondary storage devices 108 and/or movement of data from primary storage devices 104 to secondary storage devices 108.

[0119] Administrators and other employees may be able to manually configure and initiate certain information management operations on an individual basis. But while this may be acceptable for some recovery operations or other relatively less frequent tasks, it is often not workable for implementing on-going organization-wide data protection and management.

[0120] Thus, the information management system 100 may utilize information management policies 148 for specifying and executing information management operations (e.g., on an automated basis). Generally, an information management policy 148 can include a data structure or other information source that specifies a set of parameters (e.g., criteria and rules) associated with storage or other information management operations.

[0121] The storage manager database 146 may maintain the information management policies 148 and associated data, although the information management policies 148 can be stored in any appropriate location. For instance, a storage policy may be stored as metadata in a media agent database 152 or in a secondary storage device 108 (e.g., as an archive copy) for use in restore operations or other information management operations, depending on the embodiment. Information management policies 148 are described further below.

[0122] According to certain embodiments, the storage manager database 146 comprises a relational database (e.g., an SQL database) for tracking metadata, such as metadata associated with secondary copy operations (e.g., what client computing devices 102 and corresponding data were protected). This and other metadata may additionally be stored in other locations, such as at the secondary storage computing devices 106 or on the secondary storage devices 108, allowing data recovery without the use of the storage manager 140.

[0123] As shown, the storage manager 140 may include a jobs agent 156, a user interface 158, and a management agent 154, all of which may be implemented as interconnected software modules or application programs.

[0124] The jobs agent 156 in some embodiments initiates, controls, and/or monitors the status of some or all storage or other information management operations previously performed, currently being performed, or scheduled to be performed by the information management system 100. For instance, the jobs agent 156 may access information management policies 148 to determine when and how to initiate and control secondary copy and other information management operations, as will be discussed further.

[0125] The user interface 158 may include information processing and display software, such as a graphical user interface (“GUI”), an application program interface (“API”), or other interactive interface through which users and system processes can retrieve information about the status of information management operations (e.g., storage operations) or issue instructions to the information management system 100 and its constituent components.

[0126] The storage manager 140 may also track information that permits it to select, designate, or otherwise identify content indices, deduplication databases, or similar databases or resources or data sets within its information management cell (or another cell) to be searched in response to certain

queries. Such queries may be entered by the user via interaction with the user interface 158.

[0127] Via the user interface 158, users may optionally issue instructions to the components in the information management system 100 regarding performance of storage and recovery operations. For example, a user may modify a schedule concerning the number of pending secondary copy operations. As another example, a user may employ the GUI to view the status of pending storage operations or to monitor the status of certain components in the information management system 100 (e.g., the amount of capacity left in a storage device).

[0128] In general, the management agent 154 allows multiple information management systems 100 to communicate with one another. For example, the information management system 100 in some cases may be one information management subsystem or “cell” of a network of multiple cells adjacent to one another or otherwise logically related in a WAN or LAN. With this arrangement, the cells may be connected to one another through respective management agents 154.

[0129] For instance, the management agent 154 can provide the storage manager 140 with the ability to communicate with other components within the information management system 100 (and/or other cells within a larger information management system) via network protocols and application programming interfaces (“APIs”) including, e.g., HTTP, HTTPS, FTP, REST, virtualization software APIs, cloud service provider APIs, and hosted service provider APIs. Inter-cell communication and hierarchy is described in greater detail in U.S. Pat. No. 7,035,880, which is incorporated by reference herein.

[0130] Data Agents

[0131] As discussed, a variety of different types of applications 110 can reside on a given client computing device 102, including operating systems, database applications, e-mail applications, and virtual machines, just to name a few. And, as part of the as part of the process of creating and restoring secondary copies 116, the client computing devices 102 may be tasked with processing and preparing the primary data 112 from these various different applications 110. Moreover, the nature of the processing/preparation can differ across clients and application types, e.g., due to inherent structural and formatting differences between applications 110.

[0132] The one or more data agent(s) 142 are therefore advantageously configured in some embodiments to assist in the performance of information management operations based on the type of data that is being protected, at a client-specific and/or application-specific level.

[0133] The data agent 142 may be a software module or component that is generally responsible for managing, initiating, or otherwise assisting in the performance of information management operations. For instance, the data agent 142 may take part in performing data storage operations such as the copying, archiving, migrating, replicating of primary data 112 stored in the primary storage device(s) 104. The data agent 142 may receive control information from the storage manager 140, such as commands to transfer copies of data objects, metadata, and other payload data to the media agents 144.

[0134] In some embodiments, a data agent 142 may be distributed between the client computing device 102 and storage manager 140 (and any other intermediate components) or may be deployed from a remote location or its functions

approximated by a remote process that performs some or all of the functions of data agent 142. In addition, a data agent 142 may perform some functions provided by a media agent 144, e.g., encryption and deduplication.

[0135] As indicated, each data agent 142 may be specialized for a particular application 110, and the system can employ multiple data agents 142, each of which may backup, migrate, and recover data associated with a different application 110. For instance, different individual data agents 142 may be designed to handle Microsoft Exchange data, Lotus Notes data, Microsoft Windows file system data, Microsoft Active Directory Objects data, SQL Server data, SharePoint data, Oracle database data, SAP database data, virtual machines and/or associated data, and other types of data.

[0136] A file system data agent, for example, may handle data files and/or other file system information. If a client computing device 102 has two or more types of data, one data agent 142 may be used for each data type to copy, archive, migrate, and restore the client computing device 102 data. For example, to backup, migrate, and restore all of the data on a Microsoft Exchange server, the client computing device 102 may use one Microsoft Exchange Mailbox data agent 142 to backup the Exchange mailboxes, one Microsoft Exchange Database data agent 142 to backup the Exchange databases, one Microsoft Exchange Public Folder data agent 142 to backup the Exchange Public Folders, and one Microsoft Windows File System data agent 142 to backup the file system of the client computing device 102. In such embodiments, these data agents 142 may be treated as four separate data agents 142 by even though they reside on the same client computing device 102.

[0137] Other embodiments may employ one or more generic data agents 142 that can handle and process data from two or more different applications 110, or that can handle and process multiple data types, instead of or in addition to using specialized data agents 142. For example, one generic data agent 142 may be used to back up, migrate and restore Microsoft Exchange Mailbox data and Microsoft Exchange Database data while another generic data agent may handle Microsoft Exchange Public Folder data and Microsoft Windows File System data.

[0138] Each data agent 142 may be configured to access data and/or metadata stored in the primary storage device(s) 104 associated with the data agent 142 and process the data as appropriate. For example, during a secondary copy operation, the data agent 142 may arrange or assemble the data and metadata into one or more files having a certain format (e.g., a particular backup or archive format) before transferring the file(s) to a media agent 144 or other component. The file(s) may include a list of files or other metadata. Each data agent 142 can also assist in restoring data or metadata to primary storage devices 104 from a secondary copy 116. For instance, the data agent 142 may operate in conjunction with the storage manager 140 and one or more of the media agents 144 to restore data from secondary storage device(s) 108.

[0139] Media Agents

[0140] As indicated above with respect to FIG. 1A, off-loading certain responsibilities from the client computing devices 102 to intermediary components such as the media agent(s) 144 can provide a number of benefits including improved client computing device 102 operation, faster secondary copy operation performance, and enhanced scalability. As one specific example which will be discussed below in further detail, the media agent 144 can act as a local cache of

copied data and/or metadata that it has stored to the secondary storage device(s) 108, providing improved restore capabilities.

[0141] Generally speaking, a media agent 144 may be implemented as a software module that manages, coordinates, and facilitates the transmission of data, as directed by the storage manager 140, between a client computing device 102 and one or more secondary storage devices 108. Whereas the storage manager 140 controls the operation of the information management system 100, the media agent 144 generally provides a portal to secondary storage devices 108.

[0142] Media agents 144 can comprise logically and/or physically separate nodes in the information management system 100 (e.g., separate from the client computing devices 102, storage manager 140, and/or secondary storage devices 108). In addition, each media agent 144 may reside on a dedicated secondary storage computing device 106 in some cases, while in other embodiments a plurality of media agents 144 reside on the same secondary storage computing device 106.

[0143] A media agent 144 (and corresponding media agent database 152) may be considered to be “associated with” a particular secondary storage device 108 if that media agent 144 is capable of one or more of: routing and/or storing data to the particular secondary storage device 108, coordinating the routing and/or storing of data to the particular secondary storage device 108, retrieving data from the particular secondary storage device 108, and coordinating the retrieval of data from a particular secondary storage device 108.

[0144] While media agent(s) 144 are generally associated with one or more secondary storage devices 108, the media agents 144 in certain embodiments are physically separate from the secondary storage devices 108. For instance, the media agents 144 may reside on secondary storage computing devices 106 having different housings or packages than the secondary storage devices 108. In one example, a media agent 144 resides on a first server computer and is in communication with a secondary storage device(s) 108 residing in a separate, rack-mounted RAID-based system.

[0145] In operation, a media agent 144 associated with a particular secondary storage device 108 may instruct the secondary storage device 108 (e.g., a tape library) to use a robotic arm or other retrieval means to load or eject a certain storage media, and to subsequently archive, migrate, or retrieve data to or from that media, e.g., for the purpose of restoring the data to a client computing device 102. The media agent 144 may communicate with a secondary storage device 108 via a suitable communications link, such as a SCSI or Fiber Channel link.

[0146] As shown, each media agent 144 may maintain an associated media agent database 152. The media agent database 152 may be stored in a disk or other storage device (not shown) that is local to the secondary storage computing device 106 on which the media agent 144 resides. In other cases, the media agent database 152 is stored remotely from the secondary storage computing device 106.

[0147] The media agent database 152 can include, among other things, an index 153 including data generated during secondary copy operations and other storage or information management operations. The index 153 provides a media agent 144 or other component with a fast and efficient mechanism for locating secondary copies 116 or other data stored in the secondary storage devices 108. In one configuration, a storage manager index 150 or other data structure may store

data associating a client computing device 102 with a particular media agent 144 and/or secondary storage device 108, as specified in a storage policy. A media agent index 153 or other data structure associated with the particular media agent 144 may in turn include information about the stored data.

[0148] For instance, for each secondary copy 116, the index 153 may include metadata such as a list of the data objects (e.g., files/subdirectories, database objects, mailbox objects, etc.), a path to the secondary copy 116 on the corresponding secondary storage device 108, location information indicating where the data objects are stored in the secondary storage device 108, when the data objects were created or modified, etc. Thus, the index 153 includes metadata associated with the secondary copies 116 that is readily available for use in storage operations and other activities without having to be first retrieved from the secondary storage device 108. In yet further embodiments, some or all of the data in the index 153 may instead or additionally be stored along with the data in a secondary storage device 108, e.g., with a copy of the index 153.

[0149] Because the index 153 maintained in the database 152 may operate as a cache, it can also be referred to as an index cache. In such cases, information stored in the index cache 153 typically comprises data that reflects certain particulars about storage operations that have occurred relatively recently. After some triggering event, such as after a certain period of time elapses, or the index cache 153 reaches a particular size, the index cache 153 may be copied or migrated to a secondary storage device(s) 108. This information may need to be retrieved and uploaded back into the index cache 153 or otherwise restored to a media agent 144 to facilitate retrieval of data from the secondary storage device (s) 108. In some embodiments, the cached information may include format or containerization information related to archives or other files stored on the storage device(s) 108. In this manner, the index cache 153 allows for accelerated restores.

[0150] In some alternative embodiments the media agent 144 generally acts as a coordinator or facilitator of storage operations between client computing devices 102 and corresponding secondary storage devices 108, but does not actually write the data to the secondary storage device 108. For instance, the storage manager 140 (or the media agent 144) may instruct a client computing device 102 and secondary storage device 108 to communicate with one another directly. In such a case the client computing device 102 transmits the data directly to the secondary storage device 108 according to the received instructions, and vice versa. In some such cases, the media agent 144 may still receive, process, and/or maintain metadata related to the storage operations. Moreover, in these embodiments, the payload data can flow through the media agent 144 for the purposes of populating the index cache 153 maintained in the media agent database 152, but not for writing to the secondary storage device 108.

[0151] The media agent 144 and/or other components such as the storage manager 140 may in some cases incorporate additional functionality, such as data classification, content indexing, deduplication, encryption, compression, and the like. Further details regarding these and other functions are described below.

[0152] Distributed, Scalable Architecture

[0153] As described, certain functions of the information management system 100 can be distributed amongst various physical and/or logical components in the system. For

instance, one or more of the storage manager **140**, data agents **142**, and media agents **144** may reside on computing devices that are physically separate from one another. This architecture can provide a number of benefits.

[0154] For instance, hardware and software design choices for each distributed component can be targeted to suit its particular function. The secondary computing devices **106** on which the media agents **144** reside can be tailored for interaction with associated secondary storage devices **108** and provide fast index cache operation, among other specific tasks. Similarly, the client computing device(s) **102** can be selected to effectively service the applications **110** residing thereon, in order to efficiently produce and store primary data **112**.

[0155] Moreover, in some cases, one or more of the individual components in the information management system **100** can be distributed to multiple, separate computing devices. As one example, for large file systems where the amount of data stored in the storage management database **146** is relatively large, the management database **146** may be migrated to or otherwise reside on a specialized database server (e.g., an SQL server) separate from a server that implements the other functions of the storage manager **140**. This configuration can provide added protection because the database **146** can be protected with standard database utilities (e.g., SQL log shipping or database replication) independent from other functions of the storage manager **140**. The database **146** can be efficiently replicated to a remote site for use in the event of a disaster or other data loss incident at the primary site. Or the database **146** can be replicated to another computing device within the same site, such as to a higher performance machine in the event that a storage manager host device can no longer service the needs of a growing information management system **100**.

[0156] The distributed architecture also provides both scalability and efficient component utilization. FIG. 1D shows an embodiment of the information management system **100** including a plurality of client computing devices **102** and associated data agents **142** as well as a plurality of secondary storage computing devices **106** and associated media agents **144**.

[0157] Additional components can be added or subtracted based on the evolving needs of the information management system **100**. For instance, depending on where bottlenecks are identified, administrators can add additional client computing devices **102**, secondary storage devices **106** (and corresponding media agents **144**), and/or secondary storage devices **108**.

[0158] Moreover, each client computing device **102** in some embodiments can communicate with any of the media agents **144**, e.g., as directed by the storage manager **140**. And each media agent **144** may be able to communicate with any of the secondary storage devices **108**, e.g., as directed by the storage manager **140**. Thus, operations can be routed to the secondary storage devices **108** in a dynamic and highly flexible manner. Further examples of scalable systems capable of dynamic storage operations are provided in U.S. Pat. No. 7,246,207, which is incorporated by reference herein.

[0159] In alternative configurations, certain components are not distributed and may instead reside and execute on the same computing device. For example, in some embodiments one or more data agents **142** and the storage manager **140** reside on the same client computing device **102**. In another

embodiment, one or more data agents **142** and one or more media agents **144** reside on a single computing device.

Exemplary Types of Information Management Operations

[0160] In order to protect and leverage stored data, the information management system **100** can be configured to perform a variety of information management operations. As will be described, these operations can generally include secondary copy and other data movement operations, processing and data manipulation operations, and management operations.

[0161] Data Movement Operations

[0162] Data movement operations according to certain embodiments are generally operations that involve the copying or migration of data (e.g., payload data) between different locations in the information management system **100**. For example, data movement operations can include operations in which stored data is copied, migrated, or otherwise transferred from primary storage device(s) **104** to secondary storage device(s) **108**, from secondary storage device(s) **108** to different secondary storage device(s) **108**, or from primary storage device(s) **104** to different primary storage device(s) **104**.

[0163] Data movement operations can include by way of example, backup operations, archive operations, information lifecycle management operations such as hierarchical storage management operations, replication operations (e.g., continuous data replication operations), snapshot operations, deduplication operations, single-instancing operations, auxiliary copy operations, and the like. As will be discussed, some of these operations involve the copying, migration or other movement of data, without actually creating multiple, distinct copies. Nonetheless, some or all of these operations are referred to as “copy” operations for simplicity.

[0164] Backup Operations

[0165] A backup operation creates a copy of primary data **112** at a particular point in time. Each subsequent backup copy may be maintained independently of the first. Further, a backup copy in some embodiments is stored in a backup format. This can be in contrast to the version in primary data **112** from which the backup copy is derived, and which may instead be stored in a native format of the source application (s) **110**. In various cases, backup copies can be stored in a format in which the data is compressed, encrypted, deduplicated, and/or otherwise modified from the original application format. For example, a backup copy may be stored in a backup format that facilitates compression and/or efficient long-term storage.

[0166] Backup copies can have relatively long retention periods as compared to primary data **112**, and may be stored on media with slower retrieval times than primary data **112** and certain other types of secondary copies **116**. On the other hand, backups may have relatively shorter retention periods than some other types of secondary copies **116**, such as archive copies (described below). Backups may sometimes be stored at an offsite location.

[0167] Backup operations can include full, synthetic or incremental backups. A full backup in some embodiments is generally a complete image of the data to be protected. However, because full backup copies can consume a relatively large amount of storage, it can be useful to use a full backup copy as a baseline and only store changes relative to the full backup copy for subsequent backup copies.

[0168] For instance, a differential backup operation (or cumulative incremental backup operation) tracks and stores changes that have occurred since the last full backup. Differential backups can grow quickly in size, but can provide relatively efficient restore times because a restore can be completed in some cases using only the full backup copy and the latest differential copy.

[0169] An incremental backup operation generally tracks and stores changes since the most recent backup copy of any type, which can greatly reduce storage utilization. In some cases, however, restore times can be relatively long in comparison to full or differential backups because completing a restore operation may involve accessing a full backup in addition to multiple incremental backups.

[0170] Any of the above types of backup operations can be at the file-level, e.g., where the information management system **100** generally tracks changes to files at the file-level, and includes copies of files in the backup copy. In other cases, block-level backups are employed, where files are broken into constituent blocks, and changes are tracked at the block-level. Upon restore, the information management system **100** reassembles the blocks into files in a transparent fashion.

[0171] Far less data may actually be transferred and copied to the secondary storage devices **108** during a block-level copy than during a file-level copy, resulting in faster execution times. However, when restoring a block-level copy, the process of locating constituent blocks can sometimes result in longer restore times as compared to file-level backups. Similar to backup operations, the other types of secondary copy operations described herein can also be implemented at either the file-level or the block-level.

[0172] Archive Operations

[0173] Because backup operations generally involve maintaining a version of the copied data in primary data **112** and also maintaining backup copies in secondary storage device (s) **108**, they can consume significant storage capacity. To help reduce storage consumption, an archive operation according to certain embodiments creates a secondary copy **116** by both copying and removing source data. Or, seen another way, archive operations can involve moving some or all of the source data to the archive destination. Thus, data satisfying criteria for removal (e.g., data of a threshold age or size) from the source copy may be removed from source storage. Archive copies are sometimes stored in an archive format or other non-native application format. The source data may be primary data **112** or a secondary copy **116**, depending on the situation. As with backup copies, archive copies can be stored in a format in which the data is compressed, encrypted, deduplicated, and/or otherwise modified from the original application format.

[0174] In addition, archive copies may be retained for relatively long periods of time (e.g., years) and, in some cases, are never deleted. Archive copies are generally retained for longer periods of time than backup copies, for example. In certain embodiments, archive copies may be made and kept for extended periods in order to meet compliance regulations.

[0175] Moreover, when primary data **112** is archived, in some cases the archived primary data **112** or a portion thereof is deleted when creating the archive copy. Thus, archiving can serve the purpose of freeing up space in the primary storage device(s) **104**. Similarly, when a secondary copy **116** is archived, the secondary copy **116** may be deleted, and an archive copy can therefore serve the purpose of freeing up

space in secondary storage device(s) **108**. In contrast, source copies often remain intact when creating backup copies.

[0176] Snapshot Operations

[0177] Snapshot operations can provide a relatively lightweight, efficient mechanism for protecting data. From an end-user viewpoint, a snapshot may be thought of as an “instant” image of the primary data **112** at a given point in time. In one embodiment, a snapshot may generally capture the directory structure of an object in primary data **112** such as a file or volume or other data set at a particular moment in time and may also preserve file attributes and contents. A snapshot in some cases is created relatively quickly, e.g., substantially instantly, using a minimum amount of file space, but may still function as a conventional file system backup.

[0178] A snapshot copy in many cases can be made quickly and without significantly impacting primary computing resources because large amounts of data need not be copied or moved. In some embodiments, a snapshot may exist as a virtual file system, parallel to the actual file system. Users in some cases gain read-only access to the record of files and directories of the snapshot. By electing to restore primary data **112** from a snapshot taken at a given point in time, users may also return the current file system to the state of the file system that existed when the snapshot was taken.

[0179] Some types of snapshots do not actually create another physical copy of all the data as it existed at the particular point in time, but may simply create pointers that are able to map files and directories to specific memory locations (e.g., disk blocks) where the data resides, as it existed at the particular point in time. For example, a snapshot copy may include a set of pointers derived from the file system or an application. Each pointer points to a respective stored data block, so collectively, the set of pointers reflect the storage location and state of the data object (e.g., file(s) or volume(s) or data set(s)) at a particular point in time when the snapshot copy was created.

[0180] In some embodiments, once a snapshot has been taken, subsequent changes to the file system typically do not overwrite the blocks in use at the time of the snapshot. Therefore, the initial snapshot may use only a small amount of disk space needed to record a mapping or other data structure representing or otherwise tracking the blocks that correspond to the current state of the file system. Additional disk space is usually required only when files and directories are actually modified later. Furthermore, when files are modified, typically only the pointers which map to blocks are copied, not the blocks themselves. In some embodiments, for example in the case of “copy-on-write” snapshots, when a block changes in primary storage, the block is copied to secondary storage or cached in primary storage before the block is overwritten in primary storage. The snapshot mapping of file system data is also updated to reflect the changed block(s) at that particular point in time. In some other cases, a snapshot includes a full physical copy of all or substantially all of the data represented by the snapshot. Further examples of snapshot operations are provided in U.S. Pat. No. 7,529,782, which is incorporated by reference herein.

[0181] Replication Operations

[0182] Another type of secondary copy operation is a replication operation. Some types of secondary copies **116** are used to periodically capture images of primary data **112** at particular points in time (e.g., backups, archives, and snapshots). However, it can also be useful for recovery purposes to protect primary data **112** in a more continuous fashion, by

replicating the primary data **112** substantially as changes occur. In some cases a replication copy can be a mirror copy, for instance, where changes made to primary data **112** are mirrored to another location (e.g., to secondary storage device(s) **108**). By copying each write operation to the replication copy, two storage systems are kept synchronized or substantially synchronized so that they are virtually identical at approximately the same time. Where entire disk volumes are mirrored, however, mirroring can require significant amount of storage space and utilizes a large amount of processing resources.

[0183] According to some embodiments storage operations are performed on replicated data that represents a recoverable state, or “known good state” of a particular application running on the source system. For instance, in certain embodiments, known good replication copies may be viewed as copies of primary data **112**. This feature allows the system to directly access, copy, restore, backup or otherwise manipulate the replication copies as if the data was the “live”, primary data **112**. This can reduce access time, storage utilization, and impact on source applications **110**, among other benefits.

[0184] Based on known good state information, the information management system **100** can replicate sections of application data that represent a recoverable state rather than rote copying of blocks of data. Examples of compatible replication operations (e.g., continuous data replication) are provided in U.S. Pat. No. 7,617,262, which is incorporated by reference herein.

[0185] Deduplication/Single-Instancing Operations

[0186] Another type of data movement operation is deduplication, which is useful to reduce the amount of data within the system. For instance, some or all of the above-described secondary storage operations can involve deduplication in some fashion. New data is read, broken down into blocks (e.g., sub-file level blocks) of a selected granularity, compared with blocks that are already stored, and only the new blocks are stored. Blocks that already exist are represented as pointers to the already stored data.

[0187] In order to stream-line the comparison process, the information management system **100** may calculate and/or store signatures (e.g., hashes) corresponding to the individual data blocks and compare the hashes instead of comparing entire data blocks. In some cases, only a single instance of each element is stored, and deduplication operations may therefore be referred to interchangeably as “single-instancing” operations. Depending on the implementation, however, deduplication or single-instancing operations can store more than one instance of certain data blocks, but nonetheless significantly reduce data redundancy. Moreover, single-instancing in some cases is distinguished from deduplication as a process of analyzing and reducing data at the file level, rather than the sub-file level.

[0188] Depending on the embodiment, deduplication blocks can be of fixed or variable length. Using variable length blocks can provide enhanced deduplication by responding to changes in the data stream, but can involve complex processing. In some cases, the information management system **100** utilizes a technique for dynamically aligning deduplication blocks (e.g., fixed-length blocks) based on changing content in the data stream, as described in U.S. Pat. Pub. No. 2012/0084268, which is incorporated by reference herein.

[0189] The information management system **100** can perform deduplication in a variety of manners at a variety of locations in the information management system **100**. For instance, in some embodiments, the information management system **100** implements “target-side” deduplication by deduplicating data (e.g., secondary copies **116**) stored in the secondary storage devices **108**. In some such cases, the media agents **144** are generally configured to manage the deduplication process. For instance, one or more of the media agents **144** maintain a corresponding deduplication database that stores deduplication information (e.g., datablock signatures). Examples of such a configuration are provided in U.S. Pat. Pub. No. 2012/0150826, which is incorporated by reference herein. Deduplication can also be performed on the “source-side” (or “client-side”), e.g., to reduce the amount of traffic between the media agents **144** and the client computing device(s) **102** and/or reduce redundant data stored in the primary storage devices **104**. Examples of such deduplication techniques are provided in U.S. Pat. Pub. No. 2012/0150818, which is incorporated by reference herein.

[0190] Information Lifecycle Management and Hierarchical Storage Management Operations

[0191] In some embodiments, files and other data over their lifetime move from more expensive, quick access storage to less expensive, slower access storage. Operations associated with moving data through various tiers of storage are sometimes referred to as information lifecycle management (ILM) operations.

[0192] One type of ILM operation is a hierarchical storage management (HSM) operation. A HSM operation is generally an operation for automatically moving data between classes of storage devices, such as between high-cost and low-cost storage devices. For instance, an HSM operation may involve movement of data from primary storage devices **104** to secondary storage devices **108**, or between tiers of secondary storage devices **108**. With each tier, the storage devices may be progressively relatively cheaper, have relatively slower access/restore times, etc. For example, movement of data between tiers may occur as data becomes less important over time.

[0193] In some embodiments, an HSM operation is similar to an archive operation in that creating an HSM copy may (though not always) involve deleting some of the source data. For example, an HSM copy may include data from primary data **112** or a secondary copy **116** that is larger than a given size threshold or older than a given age threshold and that is stored in a backup format.

[0194] Often, and unlike some types of archive copies, HSM data that is removed or aged from the source copy is replaced by a logical reference pointer or stub. The reference pointer or stub can be stored in the primary storage device **104** to replace the deleted data in primary data **112** (or other source copy) and to point to or otherwise indicate the new location in a secondary storage device **108**.

[0195] According to one example, files are generally moved between higher and lower cost storage depending on how often the files are accessed. When a user requests access to the HSM data that has been removed or migrated, the information management system **100** uses the stub to locate the data and often make recovery of the data appear transparent, even though the HSM data may be stored at a location different from the remaining source data. The stub may also include some metadata associated with the corresponding data, so that a file system and/or application can provide some

information about the data object and/or a limited-functionality version (e.g., a preview) of the data object.

[0196] An HSM copy may be stored in a format other than the native application format (e.g., where the data is compressed, encrypted, deduplicated, and/or otherwise modified from the original application format). In some cases, copies which involve the removal of data from source storage and the maintenance of stub or other logical reference information on source storage may be referred to generally as “on-line archive copies”. On the other hand, copies which involve the removal of data from source storage without the maintenance of stub or other logical reference information on source storage may be referred to as “off-line archive copies”.

[0197] Auxiliary Copy and Disaster Recovery Operations

[0198] An auxiliary copy is generally a copy operation in which a copy is created of an existing secondary copy **116**. For instance, an initial or “primary” secondary copy **116** may be generated using or otherwise be derived from primary data **112**, whereas an auxiliary copy is generated from the initial secondary copy **116**. Auxiliary copies can be used to create additional standby copies of data and may reside on different secondary storage devices **108** than initial secondary copies **116**. Thus, auxiliary copies can be used for recovery purposes if initial secondary copies **116** become unavailable. Exemplary compatible auxiliary copy techniques are described in further detail in U.S. Pat. No. 8,230,195, which is incorporated by reference herein.

[0199] The information management system **100** may also perform disaster recovery operations that make or retain disaster recovery copies, often as secondary, high-availability disk copies. The information management system **100** may create secondary disk copies and store the copies at disaster recovery locations using auxiliary copy or replication operations, such as continuous data replication technologies. Depending on the particular data protection goals, disaster recovery locations can be remote from the client computing devices **102** and primary storage devices **104**, remote from some or all of the secondary storage devices **108**, or both.

[0200] Data Processing and Manipulation Operations

[0201] As indicated, the information management system **100** can also be configured to implement certain data manipulation operations, which according to certain embodiments are generally operations involving the processing or modification of stored data. Some data manipulation operations include content indexing operations and classification operations can be useful in leveraging the data under management to provide enhanced search and other features. Other data manipulation operations such as compression and encryption can provide data reduction and security benefits, respectively.

[0202] Data manipulation operations can be different than data movement operations in that they do not necessarily involve the copying, migration or other transfer of data (e.g., primary data **112** or secondary copies **116**) between different locations in the system. For instance, data manipulation operations may involve processing (e.g., offline processing) or modification of already stored primary data **112** and/or secondary copies **116**. However, in some embodiments data manipulation operations are performed in conjunction with data movement operations. As one example, the information management system **100** may encrypt data while performing an archive operation.

[0203] Content Indexing

[0204] In some embodiments, the information management system **100** “content indexes” data stored within the

primary data **112** and/or secondary copies **116**, providing enhanced search capabilities for data discovery and other purposes. The content indexing can be used to identify files or other data objects having pre-defined content (e.g., user-defined keywords or phrases), metadata (e.g., email metadata such as “to”, “from”, “cc”, “bcc”, attachment name, received time, etc.).

[0205] The information management system **100** generally organizes and catalogues the results in a content index, which may be stored within the media agent database **152**, for example. The content index can also include the storage locations of (or pointer references to) the indexed data in the primary data **112** or secondary copies **116**, as appropriate. The results may also be stored, in the form of a content index database or otherwise, elsewhere in the information management system **100** (e.g., in the primary storage devices **104**, or in the secondary storage device **108**). Such index data provides the storage manager **140** or another component with an efficient mechanism for locating primary data **112** and/or secondary copies **116** of data objects that match particular criteria.

[0206] For instance, search criteria can be specified by a user through user interface **158** of the storage manager **140**. In some cases, the information management system **100** analyzes data and/or metadata in secondary copies **116** to create an “off-line” content index, without significantly impacting the performance of the client computing devices **102**. Depending on the embodiment, the system can also implement “on-line” content indexing, e.g., of primary data **112**. Examples of compatible content indexing techniques are provided in U.S. Pat. No. 8,170,995, which is incorporated by reference herein.

[0207] Classification Operations—Metabase

[0208] In order to help leverage the data stored in the information management system **100**, one or more components can be configured to scan data and/or associated metadata for classification purposes to populate a metabase of information. Such scanned, classified data and/or metadata may be included in a separate database and/or on a separate storage device from primary data **112** (and/or secondary copies **116**), such that metabase related operations do not significantly impact performance on other components in the information management system **100**.

[0209] In other cases, the metabase(s) may be stored along with primary data **112** and/or secondary copies **116**. Files or other data objects can be associated with user-specified identifiers (e.g., tag entries) in the media agent **144** (or other indices) to facilitate searches of stored data objects. Among a number of other benefits, the metabase can also allow efficient, automatic identification of files or other data objects to associate with secondary copy or other information management operations (e.g., in lieu of scanning an entire file system). Examples of compatible metabases and data classification operations are provided in U.S. Pat. Nos. 8,229,954 and 7,747,579, which are incorporated by reference herein.

[0210] Encryption Operations

[0211] The information management system **100** in some cases is configured to process data (e.g., files or other data objects, secondary copies **116**, etc.), according to an appropriate encryption algorithm (e.g., Blowfish, Advanced Encryption Standard [AES], Triple Data Encryption Standard [3-DES], etc.) to limit access and provide data security in the information management system **100**.

[0212] The information management system **100** in some cases encrypts the data at the client level, such that the client computing devices **102** (e.g., the data agents **142**) encrypt the data prior to forwarding the data to other components, e.g., before sending the data media agents **144** during a secondary copy operation. In such cases, the client computing device **102** may maintain or have access to an encryption key or passphrase for decrypting the data upon restore. Encryption can also occur when creating copies of secondary copies, e.g., when creating auxiliary copies. In yet further embodiments, the secondary storage devices **108** can implement built-in, high performance hardware encryption.

[0213] Management Operations

[0214] Certain embodiments leverage the integrated, ubiquitous nature of the information management system **100** to provide useful system-wide management functions. As two non-limiting examples, the information management system **100** can be configured to implement operations management and e-discovery functions.

[0215] Operations management can generally include monitoring and managing the health and performance of information management system **100** by, without limitation, performing error tracking, generating granular storage/performance metrics (e.g., job success/failure information, deduplication efficiency, etc.), generating storage modeling and costing information, and the like.

[0216] Such information can be provided to users via the user interface **158** in a single, integrated view. For instance, the integrated user interface **158** can include an option to show a “virtual view” of the system that graphically depicts the various components in the system using appropriate icons. The operations management functionality can facilitate planning and decision-making. For example, in some embodiments, a user may view the status of some or all jobs as well as the status of each component of the information management system **100**. Users may then plan and make decisions based on this data. For instance, a user may view high-level information regarding storage operations for the information management system **100**, such as job status, component status, resource status (e.g., network pathways, etc.), and other information. The user may also drill down or use other means to obtain more detailed information regarding a particular component, job, or the like.

[0217] In some cases the information management system **100** alerts a user such as a system administrator when a particular resource is unavailable or congested. For example, a particular primary storage device **104** or secondary storage device **108** might be full or require additional capacity. Or a component may be unavailable due to hardware failure, software problems, or other reasons. In response, the information management system **100** may suggest solutions to such problems when they occur (or provide a warning prior to occurrence). For example, the storage manager **140** may alert the user that a secondary storage device **108** is full or otherwise congested. The storage manager **140** may then suggest, based on job and data storage information contained in its database **146**, an alternate secondary storage device **108**.

[0218] Other types of corrective actions may include suggesting an alternate data path to a particular primary or secondary storage device **104**, **108**, or dividing data to be stored among various available primary or secondary storage devices **104**, **108** as a load balancing measure or to otherwise optimize storage or retrieval time. Such suggestions or corrective actions may be performed automatically, if desired.

Further examples of some compatible operations management techniques and of interfaces providing an integrated view of an information management system are provided in U.S. Pat. No. 7,343,453, which is incorporated by reference herein. In some embodiments, the storage manager **140** implements the operations management functions described herein.

[0219] The information management system **100** can also be configured to perform system-wide e-discovery operations in some embodiments. In general, e-discovery operations provide a unified collection and search capability for data in the system, such as data stored in the secondary storage devices **108** (e.g., backups, archives, or other secondary copies **116**). For example, the information management system **100** may construct and maintain a virtual repository for data stored in the information management system **100** that is integrated across source applications **110**, different storage device types, etc. According to some embodiments, e-discovery utilizes other techniques described herein, such as data classification and/or content indexing.

Information Management Policies

[0220] As indicated previously, an information management policy **148** can include a data structure or other information source that specifies a set of parameters (e.g., criteria and rules) associated with secondary copy or other information management operations.

[0221] One type of information management policy **148** is a storage policy. According to certain embodiments, a storage policy generally comprises a logical container that defines (or includes information sufficient to determine) one or more of the following items: (1) what data will be associated with the storage policy; (2) a destination to which the data will be stored; (3) datapath information specifying how the data will be communicated to the destination; (4) the type of storage operation to be performed; and (5) retention information specifying how long the data will be retained at the destination.

[0222] Data associated with a storage policy can be logically organized into groups, which can be referred to as “sub-clients”. A sub-client may represent static or dynamic associations of portions of a data volume. Sub-clients may represent mutually exclusive portions. Thus, in certain embodiments, a portion of data may be given a label and the association is stored as a static entity in an index, database or other storage location.

[0223] Sub-clients may also be used as an effective administrative scheme of organizing data according to data type, department within the enterprise, storage preferences, or the like. Depending on the configuration, sub-clients can correspond to files, folders, virtual machines, databases, etc. In one exemplary scenario, an administrator may find it preferable to separate e-mail data from financial data using two different sub-clients.

[0224] A storage policy can define where data is stored by specifying a target or destination storage device (or group of storage devices). For instance, where the secondary storage device **108** includes a group of disk libraries, the storage policy may specify a particular disk library for storing the sub-clients associated with the policy. As another example, where the secondary storage devices **108** include one or more tape libraries, the storage policy may specify a particular tape library for storing the sub-clients associated with the storage policy, and may also specify a drive pool and a tape pool

defining a group of tape drives and a group of tapes, respectively, for use in storing the sub-client data.

[0225] Datapath information can also be included in the storage policy. For instance, the storage policy may specify network pathways and components to utilize when moving the data to the destination storage device(s). In some embodiments, the storage policy specifies one or more media agents **144** for conveying data (e.g., one or more sub-clients) associated with the storage policy between the source (e.g., one or more host client computing devices **102**) and destination (e.g., a particular target secondary storage device **108**).

[0226] A storage policy can also specify the type(s) of operations associated with the storage policy, such as a backup, archive, snapshot, auxiliary copy, or the like. Retention information can specify how long the data will be kept, depending on organizational needs (e.g., a number of days, months, years, etc.)

[0227] The information management policies **148** may also include one or more scheduling policies specifying when and how often to perform operations. Scheduling information may specify with what frequency (e.g., hourly, weekly, daily, event-based, etc.) or under what triggering conditions secondary copy or other information management operations will take place. Scheduling policies in some cases are associated with particular components, such as particular sub-clients, client computing device **102**, and the like. In one configuration, a separate scheduling policy is maintained for particular sub-clients on a client computing device **102**. The scheduling policy specifies that those sub-clients are to be moved to secondary storage devices **108** every hour according to storage policies associated with the respective sub-clients.

[0228] When adding a new client computing device **102**, administrators can manually configure information management policies **148** and/or other settings, e.g., via the user interface **158**. However, this can be an involved process resulting in delays, and it may be desirable to begin data protecting operations quickly.

[0229] Thus, in some embodiments, the information management system **100** automatically applies a default configuration to client computing device **102**. As one example, when a data agent(s) **142** is installed on a client computing devices **102**, the installation script may register the client computing device **102** with the storage manager **140**, which in turn applies the default configuration to the new client computing device **102**. In this manner, data protection operations can begin substantially immediately. The default configuration can include a default storage policy, for example, and can specify any appropriate information sufficient to begin data protection operations. This can include a type of data protection operation, scheduling information, a target secondary storage device **108**, data path information (e.g., a particular media agent **144**), and the like.

[0230] Other types of information management policies **148** are possible. For instance, the information management policies **148** can also include one or more audit or security policies. An audit policy is a set of preferences, rules and/or criteria that protect sensitive data in the information management system **100**. For example, an audit policy may define “sensitive objects” as files or objects that contain particular keywords (e.g. “confidential,” or “privileged”) and/or are associated with particular keywords (e.g., in metadata) or particular flags (e.g., in metadata identifying a document or email as personal, confidential, etc.).

[0231] An audit policy may further specify rules for handling sensitive objects. As an example, an audit policy may require that a reviewer approve the transfer of any sensitive objects to a cloud storage site, and that if approval is denied for a particular sensitive object, the sensitive object should be transferred to a local storage device **104** instead. To facilitate this approval, the audit policy may further specify how a secondary storage computing device **106** or other system component should notify a reviewer that a sensitive object is slated for transfer.

[0232] In some implementations, the information management policies **148** may include one or more provisioning policies. A provisioning policy can include a set of preferences, priorities, rules, and/or criteria that specify how clients **102** (or groups thereof) may utilize system resources, such as available storage on cloud storage and/or network bandwidth. A provisioning policy specifies, for example, data quotas for particular client computing devices **102** (e.g. a number of gigabytes that can be stored monthly, quarterly or annually). The storage manager **140** or other components may enforce the provisioning policy. For instance, the media agents **144** may enforce the policy when transferring data to secondary storage devices **108**. If a client computing device **102** exceeds a quota, a budget for the client computing device **102** (or associated department) is adjusted accordingly or an alert may trigger.

[0233] While the above types of information management policies **148** have been described as separate policies, one or more of these can be generally combined into a single information management policy **148**. For instance, a storage policy may also include or otherwise be associated with one or more scheduling, audit, or provisioning policies. Moreover, while storage policies are typically associated with moving and storing data, other policies may be associated with other types of information management operations. The following is a non-exhaustive list of items the information management policies **148** may specify:

[0234] schedules or other timing information, e.g., specifying when and/or how often to perform information management operations;

[0235] the type of secondary copy **116** and/or secondary copy format (e.g., snapshot, backup, archive, HSM, etc.);

[0236] a location or a class or quality of storage for storing secondary copies **116** (e.g., one or more particular secondary storage devices **108**);

[0237] preferences regarding whether and how to encrypt, compress, deduplicate, or otherwise modify or transform secondary copies **116**;

[0238] which system components and/or network pathways (e.g., preferred media agents **144**) should be used to perform secondary storage operations;

[0239] resource allocation between different computing devices or other system components used in performing information management operations (e.g., bandwidth allocation, available storage capacity, etc.);

[0240] whether and how to synchronize or otherwise distribute files or other data objects across multiple computing devices or hosted services; and

[0241] retention information specifying the length of time primary data **112** and/or secondary copies **116** should be retained, e.g., in a particular class or tier of storage devices, or within the information management system **100**.

[0242] Policies can additionally specify or depend on a variety of historical or current criteria that may be used to determine which rules to apply to a particular data object, system component, or information management operation, such as:

- [0243] frequency with which primary data 112 or a secondary copy 116 of a data object or metadata has been or is predicted to be used, accessed, or modified;
- [0244] time-related factors (e.g., aging information such as time since the creation or modification of a data object);
- [0245] deduplication information (e.g., hashes, data blocks, deduplication block size, deduplication efficiency or other metrics);
- [0246] an estimated or historic usage or cost associated with different components (e.g., with secondary storage devices 108);
- [0247] the identity of users, applications 110, client computing devices 102 and/or other computing devices that created, accessed, modified, or otherwise utilized primary data 112 or secondary copies 116;
- [0248] a relative sensitivity (e.g., confidentiality) of a data object, e.g., as determined by its content and/or metadata;
- [0249] the current or historical storage capacity of various storage devices;
- [0250] the current or historical network capacity of network pathways connecting various components within the storage operation cell;
- [0251] access control lists or other security information; and
- [0252] the content of a particular data object (e.g., its textual content) or of metadata associated with the data object.

Exemplary Storage Policy and Secondary Storage Operations

[0253] FIG. 1E shows a data flow data diagram depicting performance of storage operations by an embodiment of an information management system 100, according to an exemplary data storage policy 148A. The information management system 100 includes a storage manager 140, a client computing device 102 having a file system data agent 142A and an email data agent 142B residing thereon, a primary storage device 104, two media agents 144A, 144B, and two secondary storage devices 108A, 108B: a disk library 108A and a tape library 108B. As shown, the primary storage device 104 includes primary data 112A, 112B associated with a file system sub-client and an email sub-client, respectively.

[0254] As indicated by the dashed box, the second media agent 144B and the tape library 108B are “off-site”, and may therefore be remotely located from the other components in the information management system 100 (e.g., in a different city, office building, etc.). In this manner, information stored on the tape library 108B may provide protection in the event of a disaster or other failure.

[0255] The file system sub-client and its associated primary data 112A in certain embodiments generally comprise information generated by the file system and/or operating system of the client computing device 102, and can include, for example, file system data (e.g., regular files, file tables, mount points, etc.), operating system data (e.g., registries, event logs, etc.), and the like. The e-mail sub-client, on the other hand, and its associated primary data 112B, include data generated by an e-mail client application operating on the

client computing device 102, and can include mailbox information, folder information, emails, attachments, associated database information, and the like. As described above, the sub-clients can be logical containers, and the data included in the corresponding primary data 112A, 112B may or may not be stored contiguously.

[0256] The exemplary storage policy 148A includes a backup copy rule set 160, a disaster recovery copy rule set 162, and a compliance copy rule set 164. The backup copy rule set 160 specifies that it is associated with a file system sub-client 166 and an email sub-client 168. Each of these sub-clients 166, 168 are associated with the particular client computing device 102. The backup copy rule set 160 further specifies that the backup operation will be written to the disk library 108A, and designates a particular media agent 144A to convey the data to the disk library 108A. Finally, the backup copy rule set 160 specifies that backup copies created according to the rule set 160 are scheduled to be generated on an hourly basis and to be retained for 30 days. In some other embodiments, scheduling information is not included in the storage policy 148A, and is instead specified by a separate scheduling policy.

[0257] The disaster recovery copy rule set 162 is associated with the same two sub-clients 166, 168. However, the disaster recovery copy rule set 162 is associated with the tape library 108B, unlike the backup copy rule set 160. Moreover, the disaster recovery copy rule set 162 specifies that a different media agent 144B than the media agent 144A associated with the backup copy rule set 160 will be used to convey the data to the tape library 108B. As indicated, disaster recovery copies created according to the rule set 162 will be retained for 60 days, and will be generated on a daily basis. Disaster recovery copies generated according to the disaster recovery copy rule set 162 can provide protection in the event of a disaster or other data-loss event that would affect the backup copy 116A maintained on the disk library 108A.

[0258] The compliance copy rule set 164 is only associated with the email sub-client 166, and not the file system sub-client 168. Compliance copies generated according to the compliance copy rule set 164 will therefore not include primary data 112A from the file system sub-client 166. For instance, the organization may be under an obligation to store maintain copies of email data for a particular period of time (e.g., 10 years) to comply with state or federal regulations, while similar regulations do not apply to the file system data. The compliance copy rule set 164 is associated with the same tape library 108B and media agent 144B as the disaster recovery copy rule set 162, although a different storage device or media agent could be used in other embodiments. Finally, the compliance copy rule set 164 specifies that copies generated under the compliance copy rule set 164 will be retained for 10 years, and will be generated on a quarterly basis.

[0259] At step 1, the storage manager 140 initiates a backup operation according to the backup copy rule set 160. For instance, a scheduling service running on the storage manager 140 accesses scheduling information from the backup copy rule set 160 or a separate scheduling policy associated with the client computing device 102, and initiates a backup copy operation on an hourly basis. Thus, at the scheduled time slot the storage manager 140 sends instructions to the client computing device 102 to begin the backup operation.

[0260] At step 2, the file system data agent 142A and the email data agent 142B residing on the client computing device 102 respond to the instructions received from the

storage manager 140 by accessing and processing the primary data 112A, 112B involved in the copy operation from the primary storage device 104. Because the operation is a backup copy operation, the data agent(s) 142A, 142B may format the data into a backup format or otherwise process the data.

[0261] At step 3, the client computing device 102 communicates the retrieved, processed data to the first media agent 144A, as directed by the storage manager 140, according to the backup copy rule set 160. In some other embodiments, the information management system 100 may implement a load-balancing, availability-based, or other appropriate algorithm to select from the available set of media agents 144A, 144B. Regardless of the manner the media agent 144A is selected, the storage manager 140 may further keep a record in the storage manager database 140 of the association between the selected media agent 144A and the client computing device 102 and/or between the selected media agent 144A and the backup copy 116A.

[0262] The target media agent 144A receives the data from the client computing device 102, and at step 4 conveys the data to the disk library 108A to create the backup copy 116A, again at the direction of the storage manager 140 and according to the backup copy rule set 160. The secondary storage device 108A can be selected in other ways. For instance, the media agent 144A may have a dedicated association with a particular secondary storage device(s), or the storage manager 140 or media agent 144A may select from a plurality of secondary storage devices, e.g., according to availability, using one of the techniques described in U.S. Pat. No. 7,246,207, which is incorporated by reference herein.

[0263] The media agent 144A can also update its index 153 to include data and/or metadata related to the backup copy 116A, such as information indicating where the backup copy 116A resides on the disk library 108A, data and metadata for cache retrieval, etc. After the 30 day retention period expires, the storage manager 140 instructs the media agent 144A to delete the backup copy 116A from the disk library 108A.

[0264] At step 5, the storage manager 140 initiates the creation of a disaster recovery copy 1166 according to the disaster recovery copy rule set 162. For instance, at step 6, based on instructions received from the storage manager 140 at step 5, the specified media agent 144B retrieves the most recent backup copy 116A from the disk library 108A.

[0265] At step 7, again at the direction of the storage manager 140 and as specified in the disaster recovery copy rule set 162, the media agent 144B uses the retrieved data to create a disaster recovery copy 116B on the tape library 108B. In some cases, the disaster recovery copy 1166 is a direct, mirror copy of the backup copy 116A, and remains in the backup format. In other embodiments, the disaster recovery copy 116C may be generated in some other manner, such as by using the primary data 112A, 112B from the storage device 104 as source data. The disaster recovery copy operation is initiated once a day and the disaster recovery copies 116A are deleted after 60 days.

[0266] At step 8, the storage manager 140 initiates the creation of a compliance copy 116C, according to the compliance copy rule set 164. For instance, the storage manager 140 instructs the media agent 144B to create the compliance copy 116C on the tape library 108B at step 9, as specified in the compliance copy rule set 164. In the example, the compliance copy 116C is generated using the disaster recovery copy 116B. In other embodiments, the compliance copy

116C is instead generated using either the primary data 112B corresponding to the email sub-client or using the backup copy 116A from the disk library 108A as source data. As specified, compliance copies 116C are created quarterly, and are deleted after ten years.

[0267] While not shown in FIG. 1E, at some later point in time, a restore operation can be initiated involving one or more of the secondary copies 116A, 1166, 116C. As one example, a user may manually initiate a restore of the backup copy 116A by interacting with the user interface 158 of the storage manager 140. The storage manager 140 then accesses data in its index 150 (and/or the respective storage policy 148A) associated with the selected backup copy 116A to identify the appropriate media agent 144A and/or secondary storage device 116A.

[0268] In other cases, a media agent may be selected for use in the restore operation based on a load balancing algorithm, an availability based algorithm, or other criteria. The selected media agent 144A retrieves the data from the disk library 108A. For instance, the media agent 144A may access its index 153 to identify a location of the backup copy 116A on the disk library 108A, or may access location information residing on the disk 108A itself.

[0269] When the backup copy 116A was recently created or accessed, the media agent 144A accesses a cached version of the backup copy 116A residing in the media agent index 153, without having to access the disk library 108A for some or all of the data. Once it has retrieved the backup copy 116A, the media agent 144A communicates the data to the source client computing device 102. Upon receipt, the file system data agent 142A and the email data agent 142B may unpackage (e.g., restore from a backup format to the native application format) the data in the backup copy 116A and restore the unpackaged data to the primary storage device 104.

Exemplary Secondary Copy Formatting

[0270] The formatting and structure of secondary copies 116 can vary, depending on the embodiment. In some cases, secondary copies 116 are formatted as a series of logical data units or “chunks” (e.g., 512 MB, 1 GB, 2 GB, 4 GB, or 8 GB chunks). This can facilitate efficient communication and writing to secondary storage devices 108, e.g., according to resource availability. For example, a single secondary copy 116 may be written on a chunk-by-chunk basis to a single secondary storage device 108 or across multiple secondary storage devices 108. In some cases, users can select different chunk sizes, e.g., to improve throughput to tape storage devices.

[0271] Generally, each chunk can include a header and a payload. The payload can include files (or other data units) or subsets thereof included in the chunk, whereas the chunk header generally includes metadata relating to the chunk, some or all of which may be derived from the payload. For example, during a secondary copy operation, the media agent 144, storage manager 140, or other component may divide the associated files into chunks and generate headers for each chunk by processing the constituent files.

[0272] The headers can include a variety of information such as file identifier(s), volume(s), offset(s), or other information associated with the payload data items, a chunk sequence number, etc. Importantly, in addition to being stored with the secondary copy 116 on the secondary storage device 108, the chunk headers 300 can also be stored to the index 153 of the associated media agent(s) 144 and/or the storage man-

ager index **150**. This is useful in some cases for providing faster processing of secondary copies **116** during restores or other operations. In some cases, once a chunk is successfully transferred to a secondary storage device **108**, the secondary storage device **108** returns an indication of receipt, e.g., to the media agent **144** and/or storage manager **140**, which may update their respective indexes **150**, **153** accordingly.

[0273] During restore, chunks may be processed (e.g., by the media agent **144**) according to the information in the chunk header to reassemble the files. Additional information relating to chunks can be found in U.S. Pat. No. 8,156,086, which is incorporated by reference herein.

System Overview

[0274] The systems and methods described with respect to FIGS. 1A-1E can be used for protecting secondary copy data. For instance, the system of FIG. 1D applies backup policies and backs up the data from client computing devices **102** in the data storage system. In other embodiments, data storage systems include virtual computing devices. As indicated above, when virtual machines are created, network administrators will sometimes fail to set the protection schemes to ensure that the virtual machines will be backed up. Systems and methods are described herein to automatically discover and protect unprotected virtual machines. Further examples of systems and methods for 1) identifying virtual machine backup policies; 2) automatically categorizing unprotected virtual machines; 3) identifying uncategorized virtual machines as unprotected; and 4) protecting unprotected virtual machines are described below with respect to FIGS. 2-7.

Auto Discovery

[0275] FIG. 2 is a block diagram illustrating an arrangement of resources that form an example information management cell **750**. According to certain embodiments, some or all of the components of the information management cell **750** of FIG. 2 may have the same or similar structure and/or functionality as the similarly named components of the information management cell of FIG. 1D.

[0276] As shown, the information management cell **750** can include virtual machine management software **740**, a storage or information manager **701**, one or more virtual machines **785**, one or more primary data storage mediums **290**, one or more media agents **205**, and one or more secondary storage mediums **215**.

[0277] The virtual machine management software **740**, in an embodiment, configures, provisions, and manages virtualized environments, and stores information about the physical servers, resource pools and virtual machines. Examples of virtual machine management software **740** are VMware® VirtualCenter, Microsoft® System Center Virtual Machine Manager, and the like. The virtual machine management software **740** interfaces with the virtual machines **785** to retrieve information about the virtual machine **785**, which is stored in a database associated with the virtual machine management software **740**. Examples of the stored information are the number of virtual machines in the information management cell **750**, the physical server associated with each virtual machine **785**, and the like.

[0278] The storage or information manager **701** may be a software module or other application that coordinates and/or controls storage operations performed by one or more information management cells **750**, similar to that described above

for the storage or information manager **140** and the information management cell **100** of FIGS. 1C and 1D. In this manner, the storage or information manager **701** may act as a generally central control component with respect to the other components in the cell **750**. As shown by the dashed lines, the storage or information manager **701** may communicate with and/or control some or all elements of the information management cell **750**, such as the media agents **205** and virtual computing devices **785**, to initiate, coordinate, and/or manage secondary copy operations.

[0279] The storage or information manager **701** comprises an auto discovery module **710** which, when executed, interfaces with the virtual machine management software **740** to retrieve information about the virtual machines **785** in the information management cell **750**. Moreover, as will be described below, the auto discovery module **710**, when executed, can also identify virtual machine backup policies, and automatically categorize virtual machines without designated backup policies, identify uncategorized virtual machines as unprotected, and implement a default backup policy for these unprotected virtual machines.

[0280] While the auto discovery module is shown as residing on the storage manager **701**, in some embodiments auto discovery functionality is advantageously distributed amongst other components in the system. For instance, depending on the embodiment, auto discovery agents can execute on or form a part of one or more of the clients **785**, one or more of the data agents **295**, or one or more of the media agents **205**.

[0281] As described above with respect to storage or information manager **140** of FIG. 1D, the storage or information manager **701** may maintain the database **260** of management-related data and policies. As shown, the database **260** may include a management index or other data structure that stores logical associations between components of the system, user preferences and/or profiles, such as preferences regarding the scheduling, type, or other aspects of secondary copy operations, management tasks, media containerization, or other useful data.

[0282] The storage manager database **260** may maintain various "information management policies," and associated data, which in some embodiments are stored in the management database **260** of the storage manager **701**, although the policies can be stored in any appropriate location. A storage or other information management policy may be stored in the media agent database **226** or in secondary storage media **215** (e.g., as an archive copy) as metadata for use in information management operations, depending on the embodiment.

[0283] Generally, a policy can include a data structure or other information source that specifies a set of parameters (e.g., criteria and rules) associated with performing one or more information management operations. For example, a policy can contain information sufficient for the information management system **750** to set up and execute information management operations corresponding to the policy.

[0284] According to some embodiments, a policy comprises a data structure that defines at least timing information and type(s) sufficient for the system to perform corresponding information management operations, such as secondary copy operations (e.g. data backup operations). Timing information can include scheduling information, and may specify when (e.g., under what triggering conditions) and/or with what frequency (e.g., hourly, weekly, daily, event-based, etc.) operations will take place.

[0285] Specifically with respect to secondary copy operations, in some embodiments a policy can be a data structure that defines at least timing information, the type(s) of secondary copy, and information sufficient to determine a secondary storage destination (e.g., a location or class or quality of storage media) for storing the secondary copy. A storage policy may include a schedule policy specifying when and how often to perform secondary storage operations.

[0286] Policies can additionally specify a variety of historical or current criteria that may be used to determine which rules to apply to a particular data object, system component, or information management operation. In an embodiment, the policies can specify criteria that may be used to determine which backup rules apply to virtual machines that do not have a designated backup policy. These criteria or auto-discovery rules can be used to ensure that virtual machines are protected automatically without manual intervention.

[0287] The virtual machine 785 is a software implementation of a computing environment in which an operating system or program can be installed and run. The virtual machine 785 typically emulates a physical computing environment, but requests for CPU, memory, hard disk, network and other hardware resources are managed by a virtualization layer which translates these requests to the underlying physical hardware.

[0288] The virtual machines 785 are created within the virtualization layer, such as a hypervisor or a virtualization platform that runs on top of a client or server operating system. This operating system is often known as the host operating system. The virtualization layer can be used to create many individual, isolated virtual machine environments.

[0289] Typically, guest operating systems and programs are not aware that they are running on a virtual platform and, as long as the virtual machine's virtual platform is supported, this software can be installed in the same way it would be deployed to physical server hardware. For example, the guest operating system might appear to have a physical hard disk attached to it, but actual I/O requests are translated by the virtualization layer so they actually occur against a file that is accessible by the host operating system.

[0290] As indicated above, the virtual machines 785 interface with and provide information to the virtual machine management software 740. In another embodiment, the virtual machine 785 interfaces with the storage or information manager 701. By executing the auto discovery module 710, the information or storage manager 701 queries each computing device 102, 785 to confirm whether the device 102, 785 is a virtual machine and to determine whether the confirmed virtual machines 785 are associated with a backup policy. Each virtual machine 785 provides information, such as its existence in the information management cell 750, its associated physical server, its associated backup policy or lack of backup policy, and the like, to the storage or information manager 701. In an embodiment, the information is stored in the management database 260 for access by the storage or information manager 701.

[0291] With further reference to FIG. 2, the interaction between the various components of the example data storage system will now be described in greater detail.

[0292] FIG. 3 illustrates a flow chart of an exemplary embodiment of a process 800 to identify virtual machine backup policies usable by the information management system 750. At block 802, the information or storage manager

701 queries the virtual machine management software 740 for a list of virtual machines 785 in the information management system 750.

[0293] At block 804, the information or storage manager 701 receives a list of the virtual machines 785 in the information management system 750 from the virtual machine management software 740.

[0294] At block 806, the process 800 retrieves a list of machines (virtual machines 785 and client computing devices 102) in the information management system 750 that have designated backup policies. In an embodiment, the list of virtual machines 785 and client computing devices 102 is stored in the management database 260.

[0295] At block 808, the process 800 compares the list of virtual machines 785 in the information management system 750 with the list of machines in the information management system 750 having designated backup policies.

[0296] At block 810, the process 800 identifies which virtual machines 785 in the information management system 750 do not have a backup policy. The virtual machines 785 without an associated backup policy are unprotected. In an embodiment, an indication of the virtual machines 785 without designated backup policies is stored in the management database 260.

[0297] At block 812, the process 800 notifies the system or network administrator of the virtual machines 785 without an associated backup policy or the unprotected virtual machines 785. In an embodiment, the process 800 notifies the administrator through a graphical user interface.

[0298] FIG. 4 illustrates a flow chart of an exemplary embodiment of a process 900 to automatically categorize unprotected virtual machines usable by the information management system 750.

[0299] At block 902, the process 900 identifies the virtual machine backup policies of the virtual machines 785 in the information management system 750. In an embodiment, the process 900 identifies which virtual machines 785 in the information management system 750 are not assigned a backup policy by comparing the list of virtual machines in the system 750 with the list of machines in the system 750 having designated backup policies, as is described above with respect to FIG. 3.

[0300] At block 904, the process 900 automatically retrieves backup characteristics associated with the virtual machines 785 not assigned backup policies. In an embodiment, the backup characteristics are stored in the management database 260. In another embodiment, the information or storage manager 701 queries each virtual machine 785 not assigned a backup policy and receives backup characteristics from the queried virtual machines 785. In a further embodiment, the process 900 retrieves the backup characteristics associated with the virtual machines 785 not assigned a backup policy from the virtual machine management software 740. Examples of backup characteristics are the type of client, processor platform, operating system, type and version of software being run, the user, the location of the user, the department to which the virtual machine 785 is assigned, name of the virtual machine 785, guest DNS hostname, guest operating system, location in the virtual machine management software 740 of one or more of a folder, resource pool, datacenter, and vApp, location of the virtual machine datastore 290, attributes in the virtual machine management software 740, state of the virtual machine 785, such as power on, and the like.

[0301] At block 906, the process 900 automatically determines whether the virtual machine's backup characteristics meet the criteria for pre-defined categories or auto-discovery rules associated with backup policies. Each category is associated with criteria and a backup policy. In an embodiment, the process 900 automatically compares the backup characteristics associated with the virtual server 785 to the criteria associated with the categories of auto-discovery rules.

[0302] In an embodiment, backup policies are fine tuned to apply the timing, frequency, storage location, and the like, appropriate to the client device being backed up. For example, client devices associated with an accounting department may be backed up every 2 hours, while client devices associated with an engineering department may be backed up every week.

[0303] FIG. 5 is an exemplary screen shot 1000 of available auto-discovery rules usable by the information management system 750. Exemplary screen shot 1000 shows auto-discovery rules in the following categories: cluster affinity, data center affinity, data store affinity, ESX server affinity, matching guest host names by regular expression, operating system group, resource pool affinity, and virtual application affinity. In other embodiments, the categories of auto-discovery rules may be different from the exemplary categories of FIG. 5.

[0304] As described above, each category is associated with criteria and a backup policy. For example, the criteria for the cluster affinity category may be any client device connected to the administration department cluster and the backup policy associated with the cluster affinity category may be a backup schedule of once a day at 11 p.m.

[0305] Referring to block 906 of FIG. 4, when the backup characteristics of the virtual machine 785 meet the criteria associated with a particular auto-discovery rule, the process 900 moves to block 908. At block 908, the process 900 automatically applies the backup policy associated with the auto-discovery rules to the virtual device 785. Thus, the virtual device 785 is automatically protected by the process 900. In an embodiment, the process 900 associates the backup policy with the virtual device 785 and stores the association in the management database 260.

[0306] Referring to block 906, when the backup characteristics of the virtual machine 785 do not fit the criteria associated with any of the auto-discovery rules, the process 900 moves to block 910. At block 910, the process 900 identifies the virtual machine 785 as uncategorized. The virtual machine 785 is uncategorized when it does not have a designated backup policy and it cannot be automatically assigned a backup policy associated with the auto-discovery rules because it does not fit into any of the categories of auto-discovery rules.

[0307] FIG. 6 illustrates a flow chart of an exemplary embodiment of a process 1100 to identify an unprotected category of virtual machines 785 usable by the information management system 750.

[0308] At block 1102, the process 1100 identifies the virtual machine backup policies of the virtual machines 785 in the information management system 750. In an embodiment, the process 1100 identifies which virtual machines 785 in the information management system 750 are not assigned a backup policy by comparing the list of virtual machines in the system 750 with the list of machines in the system 750 having designated backup policies, as described above with respect to FIG. 3.

[0309] At block 1104, the process 1100 automatically categorizes virtual machines 785 not associated with a backup policy. In an embodiment, the process 1100 automatically compares the backup characteristics of the virtual machines 785 without designated backup policies to the criteria associated with categories of auto-discovery rules such that when there is a match, the virtual machine is assigned the backup policy associated with the matching category and when there is no match, the virtual machine 785 is identified as uncategorized, as described above with respect to FIG. 4.

[0310] At block 1106, the process 1100 identifies the uncategorized virtual machines 785 as unprotected. Since the virtual machines 785 with no assigned backup policy do not fit into any of the criteria associated with pre-defined backup policies, they are not assigned a backup policy associated with any of the auto-discovery rules. Consequently, the data from these virtual machines is not backed up by the information management system 750.

[0311] At block 1108, the process 1100 notifies the system or network administrator of the unprotected virtual machines 785. In an embodiment, the process 1100 notifies the administrator through a graphical user interface.

[0312] FIG. 7 illustrates a flow chart of an exemplary embodiment of a process 1200 to protect unprotected virtual machines 785 usable by the information management system 750. At block 1202, the process 1200 identifies the virtual machine backup policies of the virtual machines 785 in the information management system 750. In an embodiment, the process 1200 identifies which virtual machines 785 in the information management system 750 are not assigned a backup policy by comparing the list of virtual machines in the system 750 with the list of machines in the system 750 having designated backup policies, as described above with respect to FIG. 3.

[0313] At block 1204, the process 1200 automatically categorizes unprotected virtual machines 785. In an embodiment, the process 1200 automatically compares the backup characteristics of the virtual machines 785 without designated backup policies to the criteria associated with categories of auto-discovery rules such that when there is a match, the virtual machine is assigned the backup policy associated with the matching category and when there is no match, the virtual machine 785 is identified as uncategorized, as described above with respect to FIG. 4.

[0314] At block 1206, the process 1200 identifies the virtual machines with no designated backup policy and which do not fit into any of the categories of auto-discovery rules as belonging to the unprotected category, as described above with respect to FIG. 6.

[0315] At block 1208, the process 1200 retrieves a default backup policy. In an embodiment, the process 1200 retrieves a default backup policy from the policies stored in the management database 260. Since the system 750 does not have the information to automatically assign a backup policy tailored to the usage, location, and the like of the unprotected virtual machine 785, the process 1200 uses a previously defined catch-all backup policy. For example, a default backup policy may specify which volume or secondary storage device 215 to store the backup data, a minimal bi-weekly recovery period, and the like.

[0316] At block 1210, the process 1200 automatically applies the backup default policy to the unprotected virtual machines 785. In an embodiment, the process 1200 associates

the default backup policy with the virtual device **785** and stores an indication of the association in the management database **260**.

[**0317**] At block **1212**, the process **1200** automatically backs up the virtual machines **785** with the default backup policy. Thus, virtual machines **785** which were never assigned a backup policy by system administrators and do not meet any of the criteria for pre-defined backup policies are automatically protected.

TERMINOLOGY

[**0318**] Conditional language, such as, among others, “can,” “could,” “might,” or “may,” unless specifically stated otherwise, or otherwise understood within the context as used, is generally intended to convey that certain embodiments include, while other embodiments do not include, certain features, elements and/or steps. Thus, such conditional language is not generally intended to imply that features, elements and/or steps are in any way required for one or more embodiments or that one or more embodiments necessarily include logic for deciding, with or without user input or prompting, whether these features, elements and/or steps are included or are to be performed in any particular embodiment.

[**0319**] Depending on the embodiment, certain acts, events, or functions of any of the algorithms described herein can be performed in a different sequence, can be added, merged, or left out all together (e.g., not all described acts or events are necessary for the practice of the algorithms). Moreover, in certain embodiments, acts or events can be performed concurrently, e.g., through multi-threaded processing, interrupt processing, or multiple processors or processor cores or on other parallel architectures, rather than sequentially.

[**0320**] Systems and modules described herein may comprise software, firmware, hardware, or any combination(s) of software, firmware, or hardware suitable for the purposes described herein. Software and other modules may reside on servers, workstations, personal computers, computerized tablets, PDAs, and other devices suitable for the purposes described herein. Software and other modules may be accessible via local memory, via a network, via a browser, or via other means suitable for the purposes described herein. Data structures described herein may comprise computer files, variables, programming arrays, programming structures, or any electronic information storage schemes or methods, or any combinations thereof, suitable for the purposes described herein. User interface elements described herein may comprise elements from graphical user interfaces, command line interfaces, and other suitable interfaces.

[**0321**] Further, the processing of the various components of the illustrated systems can be distributed across multiple machines, networks, and other computing resources. In addition, two or more components of a system can be combined into fewer components. Various components of the illustrated systems can be implemented in one or more virtual machines, rather than in dedicated computer hardware systems. Likewise, the data repositories shown can represent physical and/or logical data storage, including, for example, storage area networks or other distributed storage systems. Moreover, in some embodiments the connections between the components shown represent possible paths of data flow, rather than actual connections between hardware. While some examples of possible connections are shown, any of the subset of the compo-

nents shown can communicate with any other subset of components in various implementations.

[**0322**] Embodiments are also described above with reference to flow chart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products. Each block of the flow chart illustrations and/or block diagrams, and combinations of blocks in the flow chart illustrations and/or block diagrams, may be implemented by computer program instructions. Such instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the acts specified in the flow chart and/or block diagram block or blocks.

[**0323**] These computer program instructions may also be stored in a computer-readable memory that can direct a computer or other programmable data processing apparatus to operate in a particular manner, such that the instructions stored in the computer-readable memory produce an article of manufacture including instruction means which implement the acts specified in the flow chart and/or block diagram block or blocks. The computer program instructions may also be loaded onto a computer or other programmable data processing apparatus to cause a series of operations to be performed on the computer or other programmable apparatus to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide steps for implementing the acts specified in the flow chart and/or block diagram block or blocks.

[**0324**] While certain embodiments have been described, these embodiments have been presented by way of example only, and are not intended to limit the scope of the disclosure. Indeed, the novel methods and systems described herein may be embodied in a variety of other forms; furthermore, various omissions, substitutions and changes in the form of the described methods and systems may be made without departing from the spirit of the disclosure. The accompanying claims and their equivalents are intended to cover such forms or modifications as would fall within the scope and spirit of the disclosure.

What is claimed is:

1. A method to categorize unprotected virtual machines in a data management system, the method comprising:

automatically identifying with one or more computer processors one or more virtual machines not associated with a backup policy in a data management system;

automatically retrieving with one or more computer processors backup characteristics of each virtual machine not associated with a backup policy;

automatically retrieving with one or more computer processors one or more categories, each category associated with criteria and a backup policy;

automatically comparing with one or more computer processors the backup characteristics of each virtual machine not associated with a backup policy with the criteria associated with each category; and

when the backup characteristics of the virtual machine match the criteria associated with the category, automatically applying with one or more computer processors the associated backup policy to the virtual machine.

2. The method of claim 1 further comprising automatically identifying virtual machines not associated with a backup policy and not matching the criteria of any category as uncategorized virtual machines.

3. The method of claim 1 further comprising comparing a list of virtual machines in the data management system with a list of computing devices in the data management system, each computing device associated with a backup policy, the list of virtual machines stored in a database associated with virtual machine management software and the list of computing devices stored in a database associated with an information manager for the data management system.

4. The method of claim 1 wherein the backup characteristics are stored in a database associated with an information manager.

5. The method of claim 1 further comprising querying each virtual machine not associated with a backup policy and receiving the backup characteristics from each queried virtual machine.

6. The method of claim 5 wherein an information manager queries each virtual machine not associated with a backup policy.

7. The method of claim 1 wherein the backup characteristics are retrieved from a database associated with virtual machine management software.

8. The method of claim 1 wherein the backup characteristics comprise one or more of a type of client, a processor platform, an operating system, a type of software, a version of software, a user, a location of the user, a department assigned to the virtual machine, a name of the virtual machine, a guest domain name system (DNS) hostname, a guest operating system, and a location of one or more of a folder, a resource pool, a datacenter, and a virtual machine datastore.

9. The method of claim 1 further comprising automatically assigning the associated backup policy to the virtual machine and storing the assignment.

10. The method of claim 11 wherein the assignment is stored in a database associated with an information manager.

11. A system to categorize unprotected virtual machines in an information management cell, the system comprising:

computer hardware including one or more computer processors; and

computer-readable storage comprising computer-readable instructions that, when executed by the one or more processors, cause the computer hardware to perform operations defined by the computer-readable instructions, the computer-readable instructions configured to: automatically identify one or more virtual machines not associated with a backup policy in an information management cell;

automatically retrieve backup characteristics of each virtual machine not associated with a backup policy;

automatically retrieve one or more categories, each category associated with criteria and a backup policy; automatically compare the backup characteristics of each virtual machine not associated with a backup policy with the criteria associated with each category; and

when the backup characteristics of the virtual machine match the criteria associated with the category, automatically apply the associated backup policy to the virtual machine.

12. The system of claim 11 wherein the computer-readable instructions are further configured to automatically identify virtual machines not associated with a backup policy and not matching the criteria of any category as uncategorized virtual machines.

13. The system of claim 11 wherein the computer-readable instructions are further configured to compare a list of virtual machines in the information management cell with a list of computing devices in the information management cell, each computing device associated with a backup policy, the list of virtual machines stored in a database associated with virtual machine management software and the list of computing devices stored in a database associated with an information manager for the information management cell.

14. The system of claim 11 wherein the backup characteristics are stored in a database associated with an information manager.

15. The system of claim 11 wherein the computer-readable instructions are further configured to query each virtual machine not associated with a backup policy and to receive the backup characteristics from each queried virtual machine.

16. The system of claim 15 wherein an information manager queries each virtual machine not associated with a backup policy.

17. The system of claim 11 wherein the backup characteristics are retrieved from a database associated with virtual machine management software.

18. The system of claim 11 wherein the backup characteristics comprise one or more of a type of client, a processor platform, an operating system, a type of software, a version of software, a user, a location of the user, a department assigned to the virtual machine, a name of the virtual machine, a guest domain name system (DNS) hostname, a guest operating system, and a location of one or more of a folder, a resource pool, a datacenter, and a virtual machine datastore.

19. The system of claim 11 wherein the computer-readable instructions are further configured to automatically assign the applied backup policy to the virtual device and store the assignment.

20. The system of claim 11 wherein the categories comprise categories of auto-discovery rules.

* * * * *