(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 November 2012 (22.11.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/158718 A1

(51) International Patent Classification:
G06F 9/50 (2006.01)

(21) International Application Number:
PCT/US2012/037997

(22) International Filing Date:
15 May 2012 (15.05.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/486,701 16 May 2011 (16.05.2011) US
13/239,253 21 September 2011 (21.09.2011) US

(71) Applicant *(for all designated States except US)*: **ORACLE INTERNATIONAL CORPORATION** [US/US]; 500 Oracle Parkway, M/S 5op7, Redwood Shores, CA 94065 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LEE, Robert, H.** [US/US]; 928 Sunset Drive, San Carlos, California 94070 (US). **GLEYZER, Gene** [US/US]; 1 Joseph Comee Road, Lexington, Massachusetts 02421 (US). **FALCO, Mark** [US/US]; 74 Wilmington Road, Burlington, Massachusetts 01803 (US). **PURDY, Cameron** [US/US]; 35 Network Drive, Burlington, Massachusetts 01803 (US).

(74) Agents: **MEYER, Sheldon, R.** et al.; Fliesler Meyer LLP, 650 California Street, Fourteenth Floor, San Francisco, California 94108 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: EXTENSIBLE CENTRALIZED DYNAMIC RESOURCE DISTRIBUTION IN A CLUSTERED DATA GRID

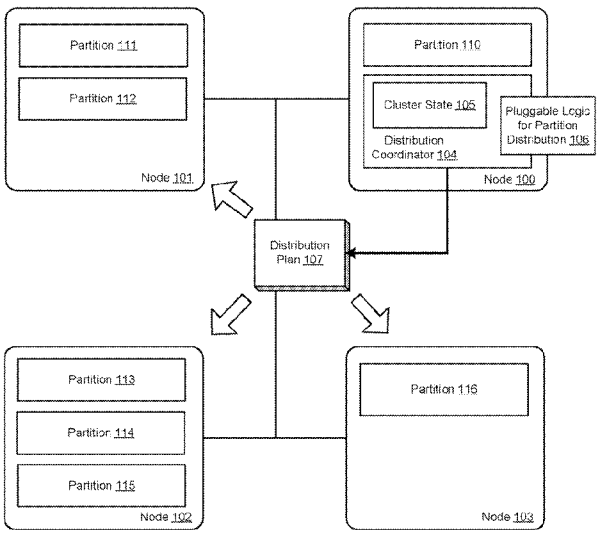

FIGURE 1

(57) **Abstract**: A centralized resource distribution is described where the decision portion of partitioning data among cluster nodes is made centralized while the actual mechanics to implement the partitioning remain a distributed algorithm. A central distribution coordinator is used to create an extensible central strategy that controls how the data will be partitioned across the cluster. The work to implement this strategy is performed by all of the members individually and asynchronously, in accordance with a distributed algorithm. The central strategy can be communicated to all cluster members and each member can perform the partitioning as it relates to itself. For example, in accordance with the distributed algorithm, one node may decide that it needs to obtain a particular partition in light of the central strategy and carry out the necessary steps to obtain that data, while other nodes may be asynchronously performing other individual partition transfers relevant to those particular nodes.

WO 2012/158718 A1

EXTENSIBLE CENTRALIZED DYNAMIC RESOURCE DISTRIBUTION IN A CLUSTERED DATA GRID

COPYRIGHT NOTICE

FIELD OF THE INVENTION

[0002] The current invention relates to data caching techniques in distributed computing environments and in particular to partitioning data among members of a cluster network.

BACKGROUND

[0003] Distributed computing and distributed algorithms have become prevalent in a wide variety of contexts, for reasons of increased performance and load capacity, high availability and failover and faster access to data. Distributed computing typically involves a number of autonomous computers (also called nodes) in communication with one another to solve a task, such as execute an application, solve complex computational problems or provide users access to various services. Each of the computer nodes typically includes its own processor(s), memory and a communication link to other nodes. The computers can be located within a particular location (e.g. cluster network) or can be connected over a large area network (LAN) such as the Internet. In most cases, distributed computers use messages to communicate with one another and to coordinate task processing and data management.

[0004] Data management is a significant issue in distributed computing. In the context of a cluster network, large data sets can be partitioned among the various nodes of the cluster. Each node usually stores a number of such partitions (subparts of the entire data set) and performs transactions on the partitions. In many cases, partitions include primary and backup copies of data distributed among the members for purposes of failover. The distribution of data in this partitioned manner can improve manageability, performance and availability of information.

[0005] There exist a number of constraints and concerns that influence the ideal distribution of data within this context. For example, moving data from one server to another incurs expenses of time and/or processor capacity. For high availability reasons, it is often preferable to locate the primary and backup copy of data on physically distinct machines. Additionally, for performance, scalability, and capacity reasons, it is often preferable to balance the distribution of data somewhat equally among available storage servers and to adjust that distribution when nodes are added or removed from the cluster.

[0006] In some use-cases, further application-specific preferences may be desired. For example, specifying that a particular set of data should be located on a specific server can be useful under certain circumstances. Furthermore, it may be desirable to specify that the distribution should use runtime feedback and inputs to place data among the node members. In light of the foregoing, what is needed is a simple and efficient way to address all or many of data distribution concerns and to optimize the balancing of partitions among the distributed cluster members.

BRIEF SUMMARY

[0007] In accordance with various embodiments of the invention, the decision portion of partitioning data among cluster nodes is made centralized while the actual mechanics to implement the partitioning remain a distributed algorithm. A central distribution coordinator can execute a centralized logic or algorithm (strategy) that generates a distribution plan. The distribution plan controls how data will be partitioned across the entire cluster. The work to implement this distribution plan is performed by all of the members individually and asynchronously, in accordance with a distributed algorithm. The distribution plan can be communicated to all of the members of the cluster and then each member performs the partitioning as it relates to itself only. For example, in accordance with the distributed algorithm, one node may decide that it needs to obtain a particular partition in light of the distribution plan and carry out the necessary steps to obtain that data, while other nodes may be asynchronously performing other individual partition transfers relevant to those particular nodes. In this manner, bottlenecks caused by a single point of coordination can be prevented, while simultaneously enabling centralized management and control for the data distribution.

BRIEF DESCRIPTION OF THE FIGURES

[0008] FIGURE 1 is an illustration of extensible centralized resource distribution being implemented in a cluster, in accordance with various embodiments of the invention.

[0009] FIGURE 2 is an illustration of the distribution plan being implemented by the nodes in the cluster, in accordance with various embodiments of the invention.

[0010] FIGURE 3 is a flow chart illustration of a process for providing centralized resource distribution, in accordance with various embodiments of the invention.

[0011] FIGURE 4 is a flow chart illustration of a process performed by the distribution coordinator, in accordance with various embodiments of the invention.

[0012] FIGURE 5 is a flow chart illustration of a process performed by the nodes in the cluster, in accordance with various embodiments of the invention.

DETAILED DESCRIPTION

[0013] One of the functions in a clustered data-grid is to maintain and effect the distribution of the data storage partitions among various cluster members (nodes). In this context, distribution can be considered the placement of both primary and backup copies of the data on a given cluster member server providing data storage.

[0014] One way to implement data partitioning across the cluster is by implementing a distributed algorithm where every storage server independently makes a decision whether to move a partition that it owns to another server or whether to obtain a partition from another server. By way of example, the data set of the cluster may initially comprise 256 primary partitions (with corresponding 256 backup partitions) which are evenly distributed across two cluster members (nodes), each member storing 128 primary and 128 backup partitions. If a third member were to join the cluster, the members could re-distribute the partitions amongst one another such that each node would store 85 partitions with one of the nodes storing 86. For example, the newly joined member node can request that each of the other two nodes in the cluster transfer 43 partitions to the new node. Alternatively, the other two nodes may determine that a new member has joined the cluster and independently transfer the partitions to the new node.

[0015] This autonomous and disconnected approach of using a distributed algorithm to allocate the partitions provides a number of advantages, including better scalability, eliminating single points of failure and the like. However, it also becomes significantly more difficult to implement complex distribution strategies because the member nodes are making

independent decisions regarding the partition transfers. For example, it may be desirable to distribute the partitions according to more dynamic and complex factors, such as how often the particular partitions are likely to be accessed, the current processing load on a particular member node and the CPU/memory capacity of each node. Moreover, it may be desirable to coordinate the arrangement of primary and backup partitions in a more complex and optimized manner for reasons of high availability and failover.

[0016] In accordance with various embodiments, a solution is described wherein individual partition transfers are performed point-to-point (as a result of direct but asynchronous communication between the sender node and the recipient node), but allow for a single point of coordination (the distribution coordinator). This single-point of coordination provides a global state of the cluster that includes the current distribution (data owners), as well as dynamic runtime feedback, such as processing load on each node in the cluster at a given point in time. The ability to have a global-view of the system at a single point allows for much more expressive distribution logic than the alternative of a distributed algorithm.

[0017] In accordance with an embodiment, each server node in the cluster communicates distribution state to the distribution coordinator on an infrequent basis. The distribution coordinator collects this information and periodically (or as a result of membership change) calls the configured distribution strategy to analyze the distribution. Additionally, an interface to the pluggable logic (strategy) can be supplied, which provides the strategy access to the current distribution state and is used by the strategy to suggest a new distribution (arrangement) of the partitions among the available server nodes.

[0018] In accordance with an embodiment, the distribution coordinator collects the suggestions made by the distribution strategy into an overall distribution plan or goal. The plan is communicated to all nodes in the cluster. Thereafter, each server initiates individual partition transfers (data movements) to approach the new distribution plan.

[0019] In accordance with various embodiments, the extensible centralized distribution can also form the basis for various additional features. For example, adaptive distribution allows the cluster to adapt dynamically to fluctuations in application load or data distribution. Furthermore, the centralized distribution can be useful for WAN-safe clustering.

[0020] FIGURE 1 is an illustration of extensible centralized resource distribution being implemented in a cluster, in accordance with various embodiments of the invention. Although this diagram depicts components as logically separate, such depiction is merely for illustrative purposes. It will be apparent to those skilled in the art that the components portrayed in this

figure can be combined or divided into separate software, firmware and/or hardware. Furthermore, it will also be apparent to those skilled in the art that such components, regardless of how they are combined or divided, can execute on the same computing device or can be distributed among different computing devices connected by one or more networks or other suitable communication means.

[0021] As illustrated, the cluster is comprised of a number of nodes (100, 101, 102, 103) that store data partitions (110, 111, 112, 113, 114, 115, 116) distributed throughout. One of the nodes 100 in the cluster is designated to be the central distribution coordinator 104. In accordance with an embodiment, the distribution coordinator periodically reevaluates the distribution (arrangement) of partitions across the cluster based on a number of factors and determines whether changes should be made to it. For example, the distribution coordinator can evaluate the partition placement according to the current request load on each node, how often a particular partition is likely to be accessed and/or the memory and CPU capacity of the node. Similarly, the central distribution coordinator 104 can be used to perform more complex distribution of primary versus backup partitions. For example, for all the primary partitions on a particular node, the distribution coordinator can ensure that the backup partitions associated with those primary partitions are not distributed across a large number of nodes. It is normally preferable that the backup partition is placed on a different physical node than the primary partition associated with it for failover purposes. However, when a primary partition is updated, the backup partitions on it will also need to be updated. This can cause a significant number of node jumps, adding to network traffic across the cluster. For this reason, it can be advantageous to limit the number of nodes that the backup partitions are located on. The central distribution coordinator can enforce this strategy, as well as any other partition arrangement strategies.

[0022] In accordance with an embodiment, the distribution coordinator 104 maintains a view of the global cluster state 105. The global state can include information including (but not limited to) the location of the partitions among the nodes, the processing load on each node, the likely demand for the data stored in each partition, the CPU and/or memory capacity of each node, and the like. In accordance with an embodiment, the distribution coordinator employs the global state to periodically (or in response to node member changes) reevaluate the partition distribution across the cluster. It should be noted that the global cluster state need not necessarily be stored on the distribution coordinator and can alternatively be stored on other members of the cluster, as well as remotely on other machines.

[0023] In accordance with an embodiment, the distribution coordinator 104 can invoke a pluggable logic component 106 in order to evaluate the partition distribution. The pluggable logic 106 can specify a particular custom distribution strategy that should be used for the cluster. The distribution coordinator can evaluate the partition distribution according to that strategy and determine whether changes should be made to it.

[0024] If the distribution coordinator 104 determines that changes should be made to the partition distribution, it can generate a distribution plan 107. This distribution plan can then be made available to each node in the cluster. In accordance with an embodiment, the distribution plan 107 can specify which partitions should be located on which node. Once the new distribution plan is made available, the various nodes can go about making the appropriate partition transfers in a distributed fashion, as will be described below.

[0025] FIGURE 2 is an illustration of the distribution plan being implemented by the nodes in the cluster, in accordance with various embodiments of the invention. Although this diagram depicts components as logically separate, such depiction is merely for illustrative purposes. It will be apparent to those skilled in the art that the components portrayed in this figure can be combined or divided into separate software, firmware and/or hardware. Furthermore, it will also be apparent to those skilled in the art that such components, regardless of how they are combined or divided, can execute on the same computing device or can be distributed among different computing devices connected by one or more networks or other suitable communication means.

[0026] In accordance with the illustrated embodiment, once the distribution coordinator generates the distribution plan, it can make the plan available to each node in the cluster. The nodes can then independently perform the decisions and steps necessary to transfer the partitions to the right nodes according to the distribution plan. In this manner, the mechanics of distributing the partitions remains a distributed algorithm, while the decision and strategy of partition arrangement is made centralized.

[0027] As illustrated, the new distribution plan 107 specifies that node 100 should store partitions 110 and 112; node 101 should store partition 111; node 102 should store partitions 113 and 114; and node 103 should store partitions 115 and 116. Because partition 112 was currently located on node 101, node 100 can inspect the new distribution plan and request partition 112 from node 101. Similarly, node 103 can request partition 115 from node 102. In accordance with alternative embodiments, nodes 101 and 102 can transfer the necessary

partitions to nodes 100 and 103 after receiving the distribution plan, without waiting for any requests from nodes 100 and 103.

[0028] FIGURE 3 is a flow chart illustration of a process for providing centralized resource distribution, in accordance with various embodiments of the invention. Although this figure depicts functional steps in a particular sequence for purposes of illustration, the process is not necessarily limited to this particular order or steps. One skilled in the art will appreciate that the various steps portrayed in this figure can be changed, rearranged, performed in parallel or adapted in various ways. Furthermore, it is to be understood that certain steps or sequences of steps can be added to or omitted from this process, without departing from the spirit and scope of the invention.

[0029] As illustrated in step 300, a cluster of computer nodes store a data set as a number of partitions. These partitions are distributed across the nodes in the cluster. In step 301, one of the computer nodes is designated to be a central distribution coordinator. In accordance with an embodiment, the distribution coordinator collects data that indicates the global state of the cluster and provides access to the global state. As further shown in step 302, the distribution coordinator periodically analyzes the global state of the cluster and determines whether changes should be made to distribution of the partitions among the nodes in the cluster. Alternatively, the distribution coordinator may reevaluate the partition distribution in response to membership changes in the cluster rather than periodically.

[0030] If the distribution coordinator determines that changes should be made, it generates a new distribution plan based on those changes and provides access to the distribution plan to all of the nodes in the cluster, as shown in step 303. The individual partition transfers can then be performed using a distributed algorithm, as shown in step 304. In other words, each node can independently determine how to perform individual partition transfers involving it in order to optimally implement the new distribution plan.

[0031] FIGURE 4 is a flow chart illustration of a process performed by the distribution coordinator, in accordance with various embodiments of the invention. Although this figure depicts functional steps in a particular sequence for purposes of illustration, the process is not necessarily limited to this particular order or steps. One skilled in the art will appreciate that the various steps portrayed in this figure can be changed, rearranged, performed in parallel or adapted in various ways. Furthermore, it is to be understood that certain steps or sequences of steps can be added to or omitted from this process, without departing from the spirit and scope of the invention.

[0032] In accordance with the illustrated embodiment, the process begins in step 400. Once initiated, the distribution coordinator continuously receives runtime feedback and other data from each node in the cluster in order to compile the global state of the cluster (step 401). The distribution coordinator can use this global state to periodically analyze the distribution of partitions across the cluster, as shown in step 402. If no changes are needed (step 403), the distribution coordinator can perform no action until the next time that it needs to evaluate the partition distribution. If, on the other hand, the distribution coordinator determines that changes should be made to the distribution, it can generate a new distribution plan that includes those changes, as shown in step 404. In step 405, the distribution coordinator can provide the distribution plan to all of the nodes in the cluster.

[0033] FIGURE 5 is a flow chart illustration of a process performed by the nodes in the cluster, in accordance with various embodiments of the invention. Although this figure depicts functional steps in a particular sequence for purposes of illustration, the process is not necessarily limited to this particular order or steps. One skilled in the art will appreciate that the various steps portrayed in this figure can be changed, rearranged, performed in parallel or adapted in various ways. Furthermore, it is to be understood that certain steps or sequences of steps can be added to or omitted from this process, without departing from the spirit and scope of the invention.

[0034] In accordance with the illustrated embodiment, the process begins in step 500. Once initiated, each node can periodically transmit runtime feedback, load statistics and other data to the distribution coordinator, as shown in step 501. Step 501 further includes an arrow back onto itself, representing a possibility that no new plan is generated, and the member simply continues to run, periodically gathering and transmitting statistics.

[0035] In step 502, the node may receive a new distribution plan from the distribution coordinator. At this point, the node may inspect the distribution plan and determine whether the plan specifies changes that are relevant to this particular node (step 503). If the new plan does not involve the node, the node can perform no transfers and can continue to periodically send runtime feedback to the distribution coordinator. If, on the other hand, the distribution plan includes partition changes that involve the node, the node can go about obtaining the necessary partitions from other nodes according to the distribution plan and/or provide the necessary partitions to other nodes (step 504).

[0036] Throughout the various contexts described in this disclosure, the embodiments of the invention further encompass computer apparatus, computing systems and machine-

readable media configured to carry out the foregoing systems and methods. In addition to an embodiment consisting of specifically designed integrated circuits or other electronics, the present invention may be conveniently implemented using a conventional general purpose or a specialized digital computer or microprocessor programmed according to the teachings of the present disclosure, as will be apparent to those skilled in the computer art.

[0037] In general, the invention relates to a system for providing extensible centralized dynamic resource distribution, said system comprising: means for storing a plurality of partitions of a data set distributed across a cluster of computer nodes; means for designating one of said computer nodes to be a central distribution coordinator that collects information indicating a global state of said cluster and provides access to said global state; means for periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to distribution of said partitions among said computer nodes; means for generating, by said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all of the cluster of computer nodes; and means for employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

[0038] In said system, said means for periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to the distribution of said partitions among said computer nodes further comprises: means for invoking a pluggable distribution logic module by said distribution coordinator, wherein said pluggable distribution logic module can be switched at runtime to adjust an algorithm used to distribute said partitions among the cluster of computer nodes.

[0039] In said system, said means for employing the distributed algorithm further comprises: means for performing individual partition transfers point-to-point between two computer nodes as a result of direct asynchronous communication between a sender node and a recipient node, wherein the distribution coordinator does not participate in directing said partition transfers.

[0040] In said system, the global state of said cluster includes information that indicates which of said partitions are assigned to each computer node in the cluster.

[0041] In said system, the global state of said cluster includes information that indicates a current processing load on each computer node in the cluster, wherein the current processing

load is determined by said each computer node periodically transmitting runtime feedback statistics to the distribution coordinator.

[0042] In said system, the global state of the cluster includes information that indicates memory capacity and processor capacity of each computer node in the cluster.

[0043] In said system, said plurality of partitions further include a set of primary partitions and a set of backup partitions, wherein the distribution coordinator ensures that each primary partition is located on a different physical node than the backup partition associated with said primary partition.

[0044] In said system, the new distribution plan generated by the distribution coordinator imposes a following restriction: for each given computer node, limit a number of computer nodes that are allowed to contain the backup partitions associated with the primary partitions located on said given computer node.

[0045] In said system, the distribution coordinator provides a single point of coordination for distribution of said partitions among the computer nodes in the cluster.

[0046] In said system, the new distribution plan specifies that a particular partition should be located on a designated computer node in the cluster.

[0047] Appropriate software coding can readily be prepared by skilled programmers based on the teachings of the present disclosure, as will be apparent to those skilled in the software art. The invention may also be implemented by the preparation of application specific integrated circuits or by interconnecting an appropriate network of conventional component circuits, as will be readily apparent to those skilled in the art.

[0048] The various embodiments include a computer program product which is a storage medium (media) having instructions stored thereon/in which can be used to program a general purpose or specialized computing processor(s)/device(s) to perform any of the features presented herein. The storage medium can include, but is not limited to, one or more of the following: any type of physical media including floppy disks, optical discs, DVDs, CD-ROMs, microdrives, magneto-optical disks, holographic storage, ROMs, RAMs, PRAMS, EPROMs, EEPROMs, DRAMs, VRAMs, flash memory devices, magnetic or optical cards, nanosystems (including molecular memory ICs); paper or paper-based media; and any type of media or device suitable for storing instructions and/or information. The computer program product can be transmitted in whole or in parts and over one or more public and/or private networks wherein the transmission includes instructions which can be used by one or more processors to perform any of the features presented herein. The transmission may include a

plurality of separate transmissions. In accordance with certain embodiments, however, the computer storage medium containing the instructions is non-transitory (i.e. not in the process of being transmitted) but rather is persisted on a physical device.

[0049] 1. A method for providing extensible centralized dynamic resource distribution, said method comprising:

storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

designating one of said computer nodes to be a central distribution coordinator that collects information indicating a global state of said cluster and provides access to said global state;

periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to distribution of said partitions among said computer nodes;

generating, by said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all of the cluster of computer nodes; and

employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

[0050] 2. The method of claim 1, wherein periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to the distribution of said partitions among said computer nodes further comprises:

invoking a pluggable distribution logic module by said distribution coordinator, wherein said pluggable distribution logic module can be switched at runtime to adjust an algorithm used to distribute said partitions among the cluster of computer nodes.

[0051] 3. The method of claim 1, wherein employing the distributed algorithm further comprises:

performing individual partition transfers point-to-point between two computer nodes as a result of direct asynchronous communication between a sender node and a recipient node, wherein the distribution coordinator does not participate in directing said partition transfers

[0052] 4. The method of claim 1, wherein the global state of said cluster includes information that indicates which of said partitions are assigned to each computer node in the cluster.

[0053] 5. The method of claim 1, wherein the global state of said cluster includes information that indicates a current processing load on each computer node in the cluster, wherein the current processing load is determined by said each computer node periodically transmitting runtime feedback statistics to the distribution coordinator.

[0054] 6. The method of claim 1, wherein the global state of the cluster includes information that indicates memory capacity and processor capacity of each computer node in the cluster.

[0055] 7. The method of claim 1, wherein said plurality of partitions further include a set of primary partitions and a set of backup partitions, wherein the distribution coordinator ensures that each primary partition is located on a different physical node than the backup partition associated with said primary partition.

[0056] 8. The method of claim 7, wherein the new distribution plan generated by the distribution coordinator imposes a following restriction:

for each given computer node, limit a number of computer nodes that are allowed to contain the backup partitions associated with the primary partitions located on said given computer node.

[0057] 9. The method of claim 1, wherein the distribution coordinator provides a single point of coordination for distribution of said partitions among the computer nodes in the cluster.

[0058] 10. The method of claim 1, wherein the new distribution plan specifies that a particular partition should be located on a designated computer node in the cluster.

[0059] 11. A system for providing extensible centralized dynamic resource distribution, said system comprising:

a cluster of computer nodes storing a plurality of partitions of a data set, said partitions being distributed across said cluster of computer nodes; and

a distribution coordinator selected from the computer nodes, said distribution coordinator collecting information that indicates a global state of said cluster, periodically analyzing the global state in order to determine whether changes should be made to distribution of said partitions among said computer nodes, generating a new distribution plan based on said changes, and providing access to said distribution plan to all of the cluster of computer nodes;

wherein said cluster of computer nodes employ a distributed algorithm to independently determine by each node in the cluster how to perform individual partition

transfers associated with said node in order to implement the new distribution plan generated by the distribution coordinator.

[0060] 12. The system of claim 11, wherein the distribution coordinator invokes a pluggable distribution logic module to determine whether said changes should be made, wherein said pluggable distribution logic module can be switched at runtime to adjust an algorithm used to distribute said partitions among the cluster of computer nodes.

[0061] 13. The system of claim 11, wherein employing the distributed algorithm by the computer nodes further comprises:

performing individual partition transfers point-to-point between two computer nodes as a result of direct asynchronous communication between a sender node and a recipient node, wherein the distribution coordinator does not participate in directing said partition transfers.

[0062] 14. The system of claim 11, wherein the global state of said cluster includes information that indicates which of said partitions are assigned to each computer node in the cluster.

[0063] 15. The system of claim 11, wherein the global state of said cluster includes information that indicates a current processing load on each computer node in the cluster, wherein the current processing load is determined by said each computer node periodically transmitting runtime feedback statistics to the distribution coordinator.

[0064] 16. The system of claim 11, wherein the global state of the cluster includes information that indicates memory capacity and processor capacity of each computer node in the cluster.

[0065] 17. The system of claim 11, wherein said plurality of partitions further include a set of primary partitions and a set of backup partitions, wherein the distribution coordinator ensures that each primary partition is located on a different physical node than the backup partition associated with said primary partition.

[0066] 18. The system of claim 17, wherein the new distribution plan generated by the distribution coordinator imposes a following restriction:

for each given computer node, limit a number of computer nodes that are allowed to contain the backup partitions associated with the primary partitions located on said given computer node.

[0067] 19. The system of claim 11, wherein the new distribution plan specifies that a particular partition should be located on a designated computer node in the cluster.

[0068] 20. A non-transitory computer readable storage medium storing one or more sequences of instructions executable by one or more processors to perform a set of steps comprising:

storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

designating one of said computer nodes to be a central distribution coordinator that collects information indicating a global state of said cluster and provides access to said global state;

periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to distribution of said partitions among said computer nodes;

generating, said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all of the cluster of computer nodes; and

employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

[0069] 21. A system for providing extensible centralized dynamic resource distribution, said system comprising:

means for storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

means for designating one of said computer nodes to be a central distribution coordinator that collects information indicating a global state of said cluster and provides access to said global state;

means for periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to distribution of said partitions among said computer nodes;

means for generating, by said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all of the cluster of computer nodes; and

means for employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

[0070] 22. The system of claim 21, wherein means for periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to the distribution of said partitions among said computer nodes further comprises:

means for invoking a pluggable distribution logic module by said distribution coordinator, wherein said pluggable distribution logic module can be switched at runtime to adjust an algorithm used to distribute said partitions among the cluster of computer nodes

[0071] 23. The system of claim 21, wherein means for employing the distributed algorithm further comprises:

means for performing individual partition transfers point-to-point between two computer nodes as a result of direct asynchronous communication between a sender node and a recipient node, wherein the distribution coordinator does not participate in directing said partition transfers.

[0072] 24. The system of claim 21, wherein the global state of said cluster includes information that indicates which of said partitions are assigned to each computer node in the cluster.

[0073] 25. The system of claim 21, wherein the global state of said cluster includes information that indicates a current processing load on each computer node in the cluster, wherein the current processing load is determined by said each computer node periodically transmitting runtime feedback statistics to the distribution coordinator.

[0074] 26. The system of claim 21, wherein the global state of the cluster includes information that indicates memory capacity and processor capacity of each computer node in the cluster.

[0075] 27. The system of claim 21, wherein said plurality of partitions further include a set of primary partitions and a set of backup partitions, wherein the distribution coordinator ensures that each primary partition is located on a different physical node than the backup partition associated with said primary partition.

[0076] 28. The system of claim 27, wherein the new distribution plan generated by the distribution coordinator imposes a following restriction:

for each given computer node, limit a number of computer nodes that are allowed to contain the backup partitions associated with the primary partitions located on said given computer node.

[0077] 29. The system of claim 21, wherein the distribution coordinator provides a single point of coordination for distribution of said partitions among the computer nodes in the cluster.

[0078] 30. The system of claim 21, wherein the new distribution plan specifies that a particular partition should be located on a designated computer node in the cluster.

[0079] The foregoing description of the preferred embodiments of the present invention has been provided for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations can be apparent to the practitioner skilled in the art. Embodiments were chosen and described in order to best explain the principles of the invention and its practical application, thereby enabling others skilled in the relevant art to understand the invention. It is intended that the scope of the invention be defined by the following claims and their equivalents.

CLAIMS

What is claimed is:

1. A computer-implemented method for providing extensible centralized dynamic resource distribution, said method comprising:

storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

designating one of said computer nodes to be a central distribution coordinator that collects information indicating a global state of said cluster and provides access to said global state;

periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to the distribution of said partitions among said computer nodes;

generating, by said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all computer nodes in the cluster of computer nodes; and

employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

2. The method of claim 1, wherein periodically analyzing the global state by the distribution coordinator in order to determine whether changes should be made to the distribution of said partitions further comprises:

invoking a pluggable distribution logic module by said distribution coordinator, wherein said pluggable distribution logic module can be switched at runtime to adjust an algorithm used to distribute said partitions among the cluster of computer nodes.

3. The method of claim 1 or 2, wherein employing the distributed algorithm further comprises:

performing individual partition transfers point-to-point between two computer nodes in the cluster of computer nodes as a result of direct asynchronous communication between a sender node and a recipient node, wherein the distribution coordinator does not participate in directing said partition transfers.

4. The method of any preceding claim, wherein the global state of said cluster includes information that indicates which of said partitions are assigned to each computer node in the cluster.

5. The method of any preceding claim, wherein the global state of said cluster includes information that indicates a current processing load on each computer node in the cluster, wherein the current processing load is determined by said each computer node periodically transmitting runtime feedback statistics to the distribution coordinator.

6. The method of any preceding claim, wherein the global state of the cluster includes information that indicates memory capacity and processor capacity of each computer node in the cluster.

7. The method of any preceding claim, wherein said plurality of partitions further include a set of primary partitions and a set of backup partitions, wherein the distribution coordinator ensures that each primary partition is located on a different physical node than the backup partition associated with said primary partition.

8. The method of claim 7, wherein the new distribution plan generated by the distribution coordinator imposes the restriction that, for each given computer node, a limited number of computer nodes are allowed to contain the backup partitions associated with the primary partitions located on said given computer node.

9. The method of any preceding claim, wherein the distribution coordinator provides a single point of coordination for distribution of said partitions among the computer nodes in the cluster.

10. The method of any preceding claim, wherein the new distribution plan specifies that a particular partition should be located on a designated computer node in the cluster.

11. A computer program comprising instructions that when executed by one or more computers cause the one or more computers to perform the method of any preceding claim.

12. A computer readable medium storing the computer program of claim 11.

13. Apparatus comprising one or more computers configured to perform all the steps of the method of any of claims 1 to 10.

14. A system for providing extensible centralized dynamic resource distribution, said system comprising:

a cluster of computer nodes storing a plurality of partitions of a data set, said partitions being distributed across said cluster of computer nodes; and

a distribution coordinator selected from the computer nodes, said distribution coordinator collecting information that indicates a global state of said cluster, periodically analyzing the global state in order to determine whether changes should be made to the distribution of said partitions among said computer nodes, generating a new distribution plan based on said changes, and providing access to said distribution plan to all the computer nodes in the cluster of computer nodes;

wherein said cluster of computer nodes employ a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan generated by the distribution coordinator.

15. A non-transitory computer readable storage medium storing one or more sequences of instructions executable by one or more processors to perform a set of steps comprising:

storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

designating one of said computer nodes to be a central distribution coordinator that collects information indicating a global state of said cluster and provides access to said global state;

periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to the distribution of said partitions among said computer nodes;

generating, by said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all the computer nodes in the cluster of computer nodes; and

employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

16. A method for providing extensible centralized dynamic resource distribution, said method comprising:

storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

periodically analyzing the global state of the cluster by a distribution coordinator in order to determine whether changes should be made to distribution of said partitions among said computer nodes, the distribution coordinator being one of said computer nodes;

generating, by said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all of the cluster of computer nodes; and

employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

17. A non-volatile computer readable storage medium storing one or more sequences of instructions executable by one or more processors to perform a set of steps comprising:

storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

designating one of said computer nodes to be a central distribution coordinator that collects information indicating a global state of said cluster and provides access to said global state;

periodically analyzing the global state of the cluster by the distribution coordinator in order to determine whether changes should be made to distribution of said partitions among said computer nodes;

generating, said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all of the cluster of computer nodes; and

employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

18. A program causing a processor to perform a set of steps comprising:

storing a plurality of partitions of a data set distributed across a cluster of computer nodes;

periodically analyzing the global state of the cluster by a distribution coordinator in order to determine whether changes should be made to distribution of said partitions among said computer nodes, the distribution coordinator being one of said computer nodes;

generating, by said distribution coordinator, a new distribution plan based on said changes to the distribution of said partitions, and providing access to said distribution plan to all of the cluster of computer nodes; and

employing a distributed algorithm to independently determine by each node in the cluster how to perform individual partition transfers associated with said node in order to implement the new distribution plan.

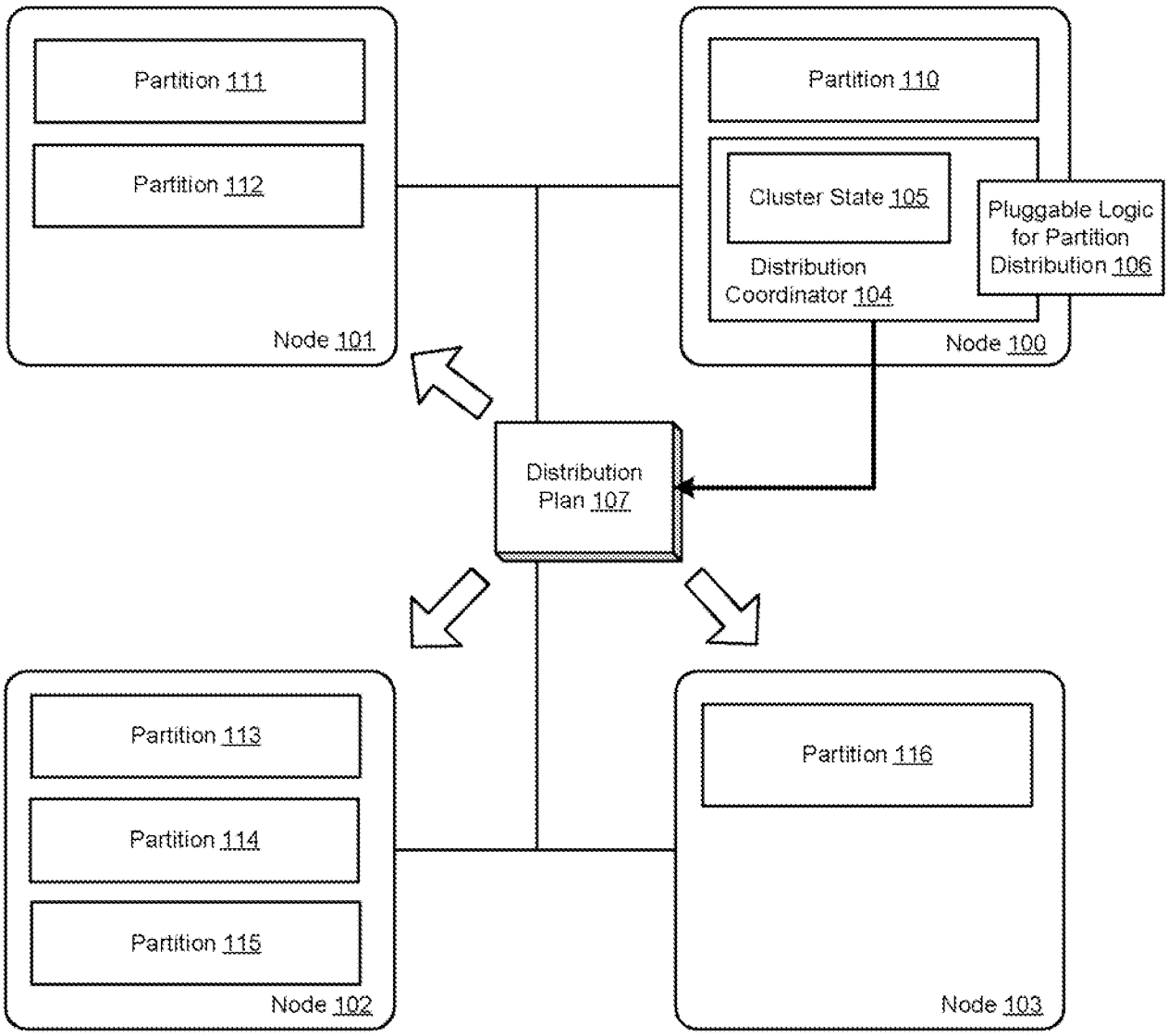
Partition 111
Partition 112
Node 101
Partition 110
Cluster State 105
Distribution Coordinator 104
Pluggable Logic for Partition Distribution 106
Node 100
Distribution Plan 107
Partition 113
Partition 114
Partition 115
Node 102
Partition 116
Node 103

FIGURE 1

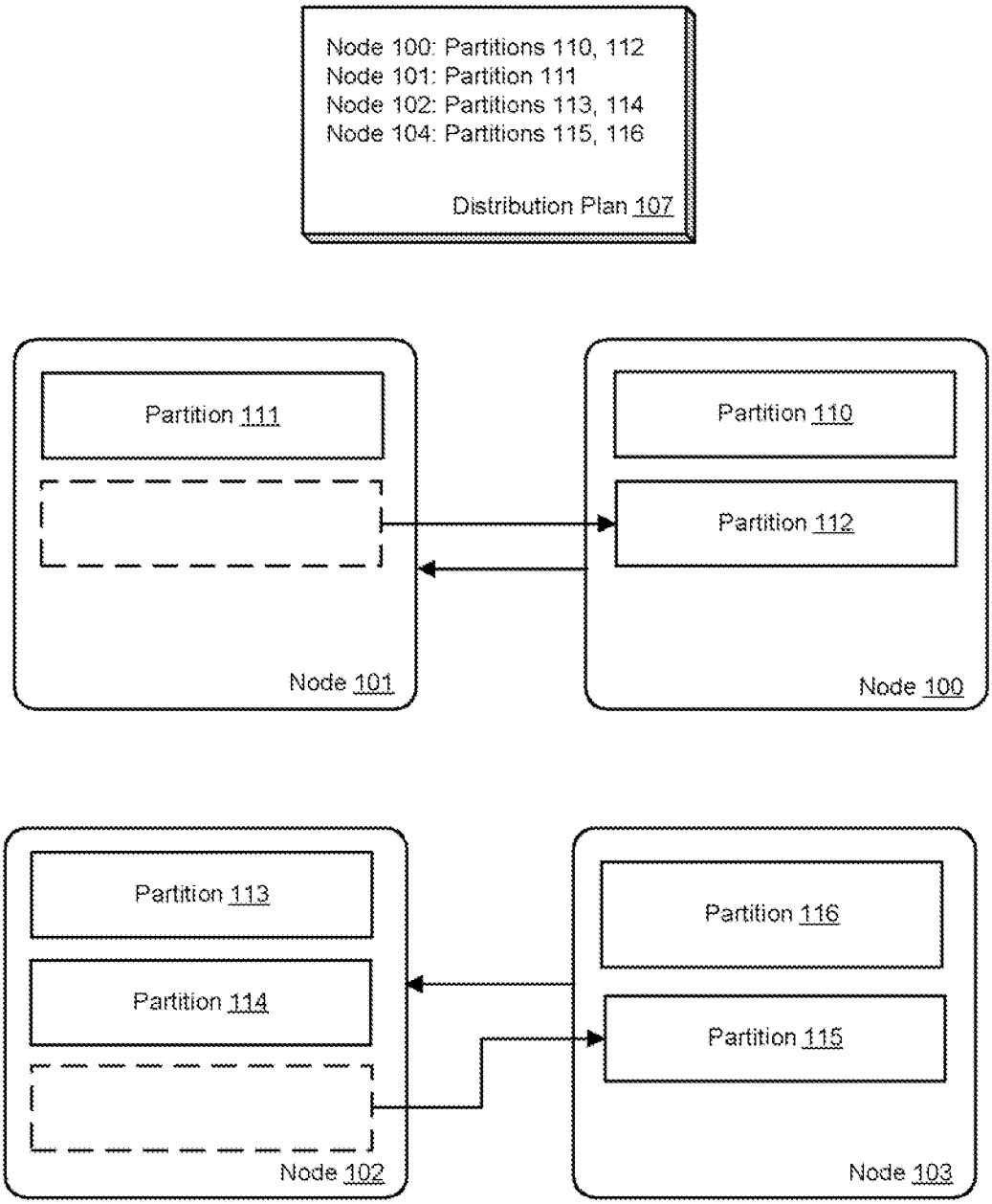
Node 100: Partitions 110, 112
Node 101: Partition 111
Node 102: Partitions 113, 114
Node 104: Partitions 115, 116
Distribution Plan 107
Partition 111
Partition 110
Partition 112
Node 101
Node 100
Partition 113
Partition 116
Partition 114
Partition 115
Node 102
Node 103

FIGURE 2

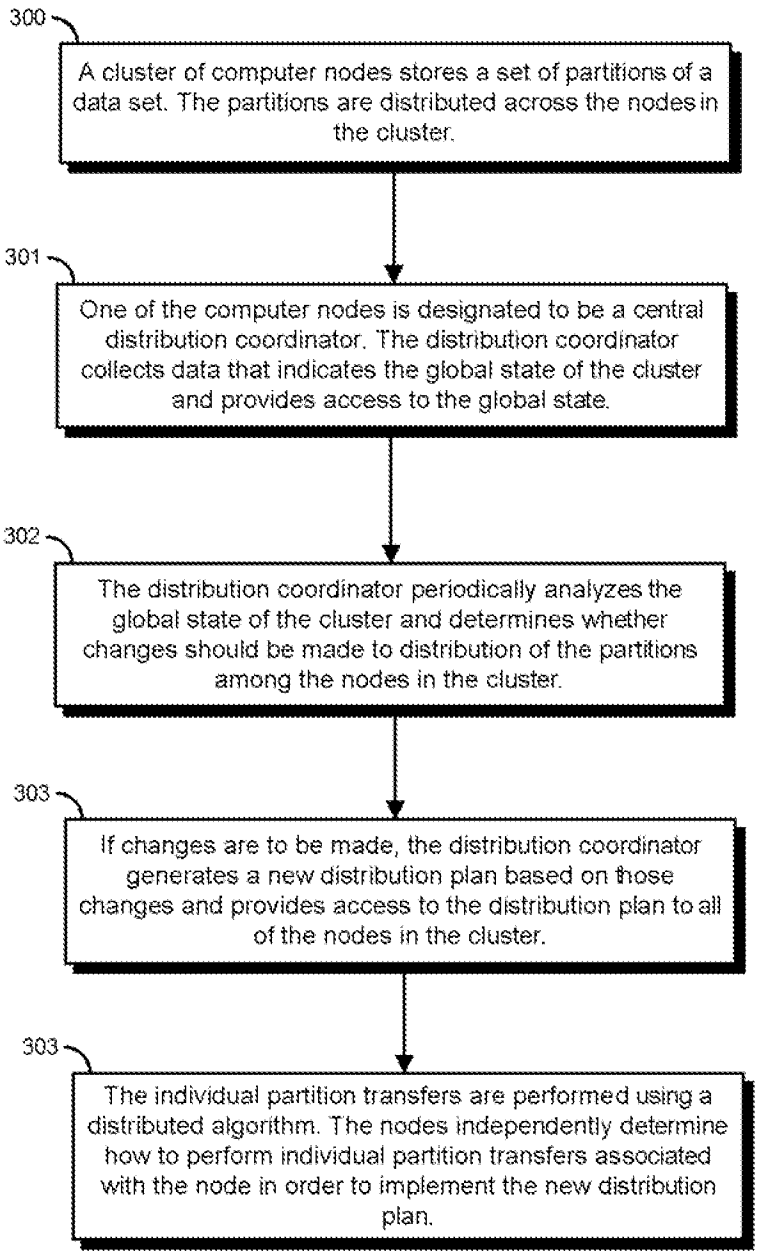
300
A cluster of computer nodes stores a set of partitions of a data set. The partitions are distributed across the nodes in the cluster.
301
One of the computer nodes is designated to be a central distribution coordinator. The distribution coordinator collects data that indicates the global state of the cluster and provides access to the global state.
302
The distribution coordinator periodically analyzes the global state of the cluster and determines whether changes should be made to distribution of the partitions among the nodes in the cluster.
303
If changes are to be made, the distribution coordinator generates a new distribution plan based on those changes and provides access to the distribution plan to all of the nodes in the cluster.
303
The individual partition transfers are performed using a distributed algorithm. The nodes independently determine how to perform individual partition transfers associated with the node in order to implement the new distribution plan.

FIGURE 3

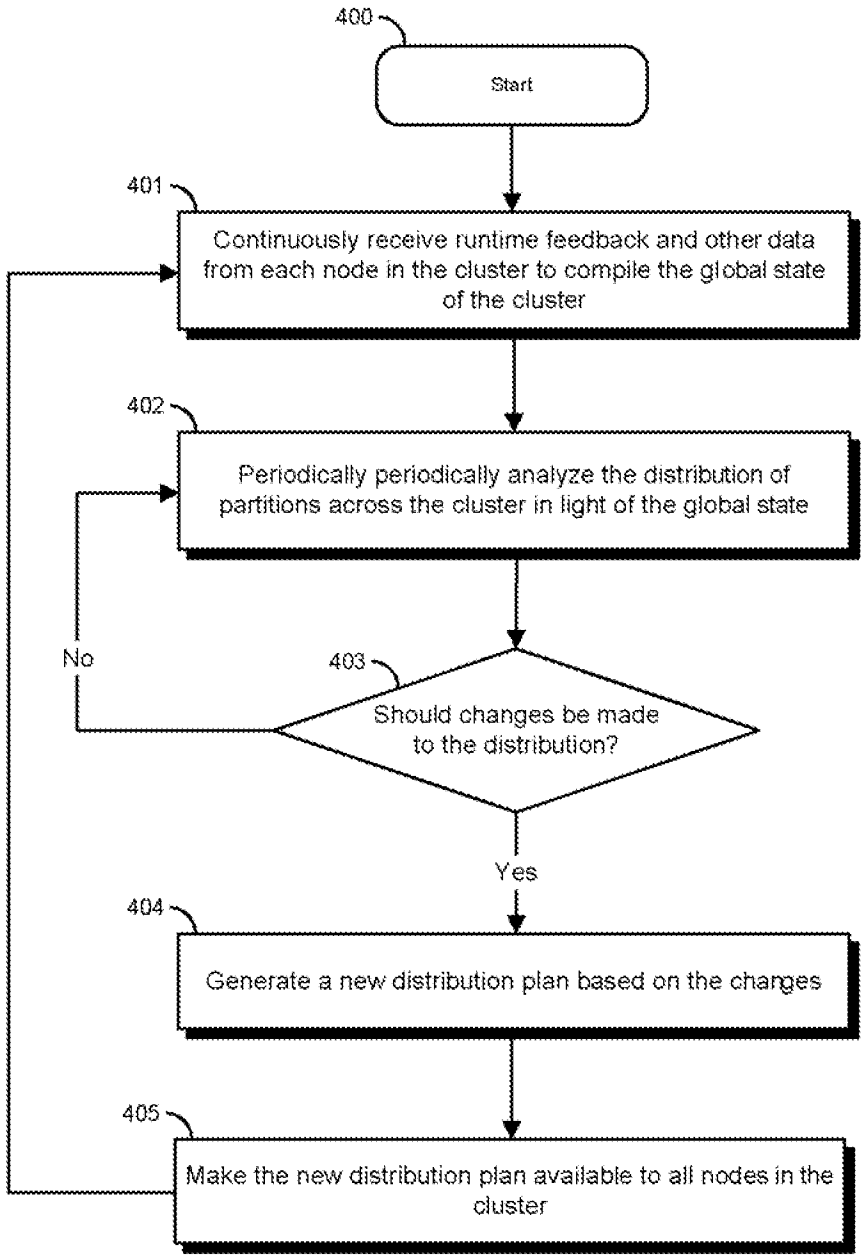
400
Start
401
Continuously receive runtime feedback and other data from each node in the cluster to compile the global state of the cluster
402
Periodically periodically analyze the distribution of partitions across the cluster in light of the global state
403
Should changes be made to the distribution?
No
Yes
404
Generate a new distribution plan based on the changes
405
Make the new distribution plan available to all nodes in the cluster

FIGURE 4

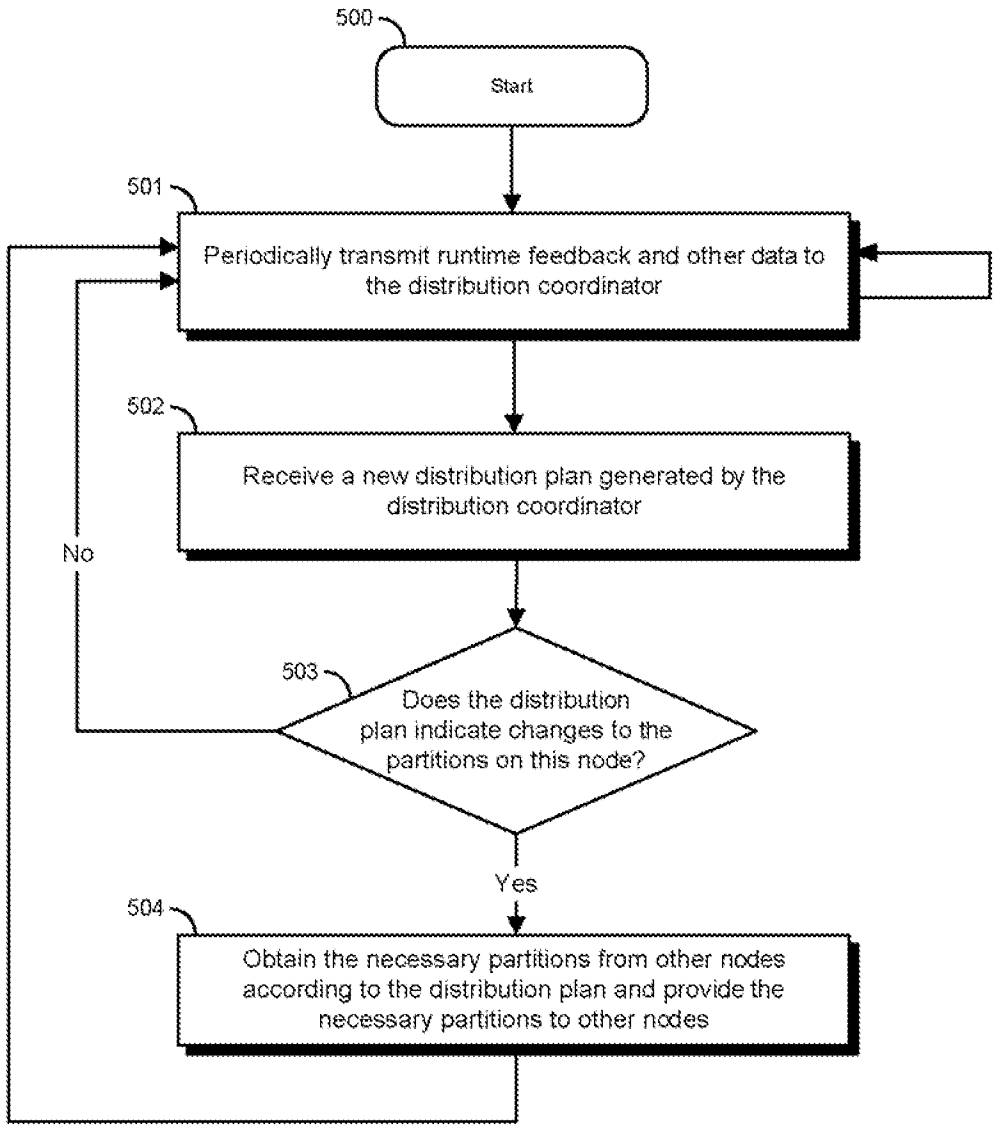
500
Start
501
Periodically transmit runtime feedback and other data to the distribution coordinator
502
Receive a new distribution plan generated by the distribution coordinator
503
Does the distribution plan indicate changes to the partitions on this node?
No
Yes
504
Obtain the necessary partitions from other nodes according to the distribution plan and provide the necessary partitions to other nodes

FIGURE 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2012/037997

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F9/50
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, IBM-TDB, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TENGJIAO WANG ET AL: "Dynamic Data Migration Policies for Query-Intensive Distributed Data Environments", 2 April 2009 (2009-04-02), ADVANCES IN DATA AND WEB MANAGEMENT, SPRINGER BERLIN HEIDELBERG, BERLIN, HEIDELBERG, PAGE(S) 63 - 75, XP019115708, ISBN: 978-3-642-00671-5 page 64, paragraph 2 - paragraph 3 page 65, paragraph 2 - page 66, paragraph 1 page 67, paragraph 1 - page 68, paragraph 3 page 69, paragraph 2 - paragraph 4 -----	1-18

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 July 2012	24/07/2012

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Milasinovic, Goran