



(12) **DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION**

(13) **A1**

(86) Date de dépôt PCT/PCT Filing Date: 2020/01/21
 (87) Date publication PCT/PCT Publication Date: 2020/11/19
 (85) Entrée phase nationale/National Entry: 2021/11/11
 (86) N° demande PCT/PCT Application No.: CN 2020/073429
 (87) N° publication PCT/PCT Publication No.: 2020/228378
 (30) Priorité/Priority: 2019/05/14 (CN201910401715.0)

(51) Cl.Int./Int.Cl. *G06F 16/21* (2019.01)
 (71) Demandeurs/Applicants:
 HUAWEI TECHNOLOGIES CO., LTD., CN;
 TSINGHUA UNIVERSITY, CN
 (72) Inventeurs/Inventors:
 LI, GUOLIANG, CN;
 LI, SHIFU, CN;
 ZHOU, XUANHE, CN;
 GAO, BO, CN
 (74) Agent: GOWLING WLG (CANADA) LLP

(54) Titre : PROCÉDE ET DISPOSITIF DE DETERMINATION DE PARAMETRES DE CONFIGURATION DE BASE DE DONNEES
 (54) Title: METHOD AND APPARATUS FOR DETERMINING CONFIGURATION KNOB OF DATABASE

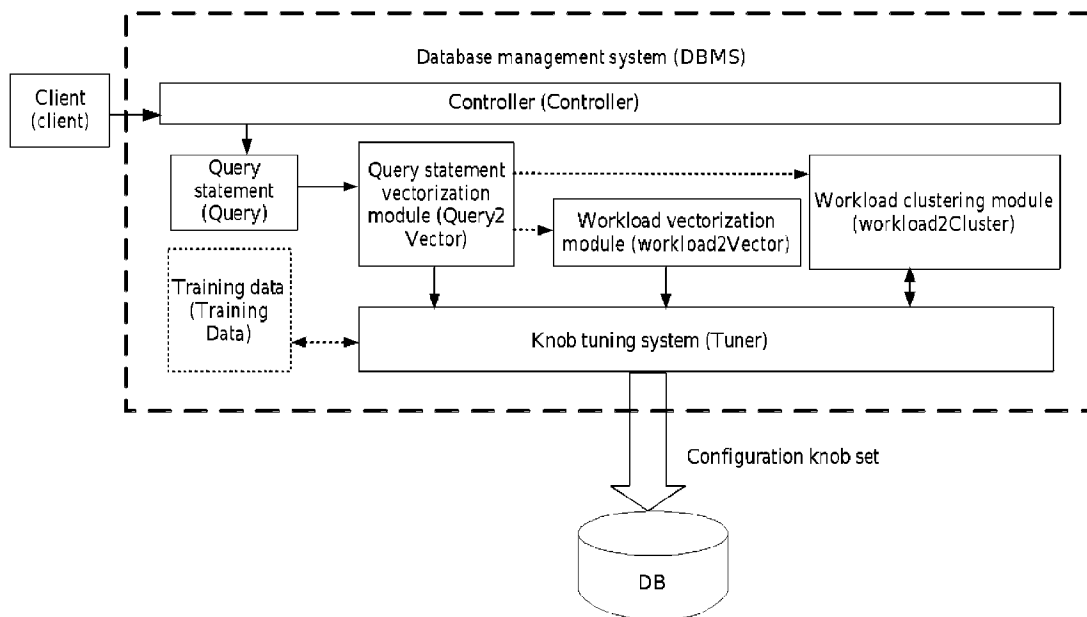


FIG. 8

(57) **Abrégé/Abstract:**

A method for determining database configuration parameters, which can be applied to a database management system. The method comprises: performing vectorization coding on a query sentence to thereby obtain a target feature vector of the query sentence, and then, according to the dual state information of the target feature vector and the current state information of a database (101), determining a configuration parameter set used to determine a database, the configuration parameter set being used to execute N query sentences. The configuration parameters of the database (101) can be dynamically determined according to the query sentence and the current state information the database (101), and it can be ensured that the database (101) performs well under different load requirements, such as performance in latency, throughput, and like areas.

Date Submitted: 2021/11/11

CA App. No.: 3137748

Abstract:

A method for determining a configuration knob of a database is disclosed, and may be applied to a database management system. In this method, vectorization coding is performed on a query statement, to obtain a target feature vector of the query statement, and then a configuration knob set for determining the database is determined based on two aspects of double-state information: the target feature vector and current status information of the database (101). The configuration knob set is used to execute N query statements. The configuration knob of the database (101) may be dynamically determined based on the query statement and the current status information of the database (101), so that good performance of the database (101) in a case of different workload requirements, for example, performance in aspects of a delay and a throughput, can be ensured.

METHOD AND APPARATUS FOR DETERMINING CONFIGURATION KNOB OF DATABASE

[0001] This application claims priority to Chinese Patent Application No. 201910401715.0, filed with the China National Intellectual Property Administration on May 14, 2019 and entitled "METHOD AND APPARATUS FOR DETERMINING CONFIGURATION KNOB OF DATABASE", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] This application relates to the field of database technologies, and specifically, to a method and an apparatus for determining a configuration knob of a database.

BACKGROUND

[0003] A database (database, DB) stores massive data and can provide a data service for a large quantity of clients. In other words, the client is a workload of the database. During working, the database may run based on different configuration knobs and provide the client with data services with different performance. To be specific, if a configuration knob of the database matches a service of the workload, the database responds quickly when providing the service; or if a configuration knob of the database does not match a service of the workload, the database responds slowly when providing the service.

[0004] To better provide a service for the client, during running of the database, a database administrator (database administrator, DBA) determines the configuration knob of the database based on a workload status.

[0005] Currently, the DBA needs to test a to-be-deployed service in a test environment to determine the configuration knob of the database. A large amount of time needs to be taken to determine a configuration knob that is of the database and that is applicable to the service. Once the service changes, the DBA needs to perform a test again. For a service and a quantity of workloads that may change at any time in a network, a manner in which the DBA tunes a knob

cannot meet a current requirement.

SUMMARY

[0006] Embodiments of this application provide a method for determining a configuration knob of a database, to dynamically determine the configuration knob of the database based on a workload status and current status information of the database, so as to ensure good performance of the database in a case of different workload requirements, for example, performance in aspects of a delay and a throughput.

[0007] To achieve the foregoing objective, the following technical solutions are used in the embodiments of this application.

[0008] According to a first aspect, a method for determining a configuration knob of a database is provided, and may include: obtaining N query statements sent by M clients, where M is an integer greater than 0, N is an integer greater than 0, and $N \geq M$; determining N pieces of query information from the N query statements, where the N query statements are in a one-to-one correspondence with the N pieces of query information; performing vectorization coding on the N pieces of query information, to obtain corresponding N target feature vectors, where the N pieces of query information are in a one-to-one correspondence with the N target feature vectors; and determining, based on the N target feature vectors and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements, where the configuration knob set includes at least one target configuration knob group, the target configuration knob group includes at least one configuration knob, and the configuration knob set is used to execute the N query statements.

[0009] In the technical solution provided in the first aspect, after the configuration knob set is determined, a configuration knob currently used for the database may be tuned based on the configuration knob set, and a configuration knob tuning manner may be replacing the configuration knob currently used for the database with a configuration knob in a current configuration knob set. Alternatively, the configuration knob currently used for the database may be tuned by adding an offset value to or reducing an offset value from a value of the configuration knob currently used for the database. In the first aspect, the configuration knob of the database may be dynamically determined based on a workload status and the current status information of

the database, so that good performance of the database in a case of different workload requirements can be ensured. For example, performance in aspects of a delay and a throughput can be met.

[0010] In a possible implementation of the first aspect, the method may further include: determining, in response to a mode selection instruction from a query-level knob tuning mode, a workload-level knob tuning mode, and a cluster-level knob tuning mode, a target knob tuning mode used for the N query statements. In other words, in this possible implementation, knob tuning modes at three granularities are provided: the query-level knob tuning mode, the workload-level knob tuning mode, and the cluster-level knob tuning mode. The mode selection instruction may be triggered by a user or may be triggered by a database administrator. In other words, the user or the database administrator may switch the knob tuning mode in different requirement scenarios, to meet a plurality of knob tuning requirements.

[0011] In a possible implementation of the first aspect, when the target knob tuning mode is the query-level knob tuning mode, $M = 1$, and $N = 1$. To be specific, for the query-level knob tuning mode, one configuration knob set may be determined for one query statement. The configuration knob set includes one target configuration knob group for the query statement, and the target configuration knob group is configured for the database when the query statement is executed.

[0012] In a possible implementation of the first aspect, when the target knob tuning mode is the workload-level knob tuning mode, the method further includes: fusing the N target feature vectors, to obtain a unified vector obtained after fusion; and the determining, based on the N target feature vectors and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements includes: determining, based on the unified vector and the current status information of the database, the configuration knob set that is of the database and that corresponds to the N query statements, where the configuration knob set includes one target configuration knob group, and the one target configuration knob group is used to tune, when the N query statements are executed, a configuration knob currently used for the database. It can be learned from this possible implementation of the first aspect that in the workload-level knob tuning mode, knob tuning may be performed once for the N query statements, to meet a requirement of a scenario in which there is a large throughput.

[0013] In a possible implementation of the first aspect, when the target knob tuning mode is the cluster-level knob tuning mode, the determining, based on the N target feature vectors and

current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements includes: determining N configuration knob groups based on the N target feature vectors and the current status information of the database, where the N configuration knob groups are in a one-to-one correspondence with the N target feature vectors; and performing clustering on the N configuration knob groups, to obtain Q target configuration knob groups, where a first configuration knob group corresponds to at least one target feature vector, the first configuration knob group is any one of the Q target configuration knob groups, Q is an integer greater than 0, $Q < N$, and the first configuration knob group is used to tune, when a query statement corresponding to the at least one target feature vector is executed, a configuration knob currently used for the database. It can be learned from this possible implementation of the first aspect that in the cluster-level knob tuning mode, one target configuration knob group may be determined for a same type of query statement, to ensure a large throughput and meet a requirement for a low delay.

[0014] In a possible implementation of the first aspect, first query information in the N pieces of query information includes a first query type, first table information, and a type and corresponding cost information of an involved operator in a query, and the first query information is any one of the N pieces of query information; the first query type is used to indicate a type of an operation that is requested by a query statement corresponding to the first query information and that is to be performed on the database; the first table information is used to indicate a relationship table involved when the query statement corresponding to the first query information is executed in the database; and the cost information is used to indicate a respective execution cost that is of the involved operator and that is incurred when the involved operator executes the query statement corresponding to the first query information.

[0015] In a possible implementation of the first aspect, a target feature vector corresponding to the first query information includes an identifier of the first query type, an identifier of a second query type, an identifier of the first table information, an identifier of second table information, and a floating-point bit of each type of operator; the identifier of the first query type is represented by using a first value, the identifier of the second query type is represented by using a second value, and the second query type belongs to a query type of the database and is not included in the first query information; the identifier of the first table information is represented by using a third value, the identifier of the second table information is represented by using a fourth value, and the second

table information belongs to table information of the database and is not included in the first query information; and the floating-point bit of each type of operator includes an execution cost of the corresponding type of operator.

5 [0016] In this possible implementation, the identifier may be a bit identifier, or may be a floating-point bit identifier. When the identifier is a bit identifier, 1 may be used to indicate the first query type, and 0 may be used to indicate the second query type; or T may be used to indicate the first query type, and F may be used to indicate the second query type. Alternatively, another bit representation manner may also be applicable to this solution. When the identifier is a floating-point bit identifier, all involved query types may be marked by using different characters. For
10 example, insert (insert) is represented by using 1, select (select) is represented by using 2, update (update) is represented by using 3, and delete (delete) is represented by using 4. Certainly, the four query types are merely used as an example for description herein. The query types are not limited to the four query types, and representations of the query types are not limited to the listed numerical form herein. Different query types may be represented by using another symbol or numerical form.

15 [0017] In a possible implementation of the first aspect, the execution cost of the corresponding type of operator is a normalized execution cost.

[0018] In a possible implementation of the first aspect, the determining, based on the N target feature vectors and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements includes: determining, based on the
20 N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (double-state deep reinforcement learning, DS-DRL) model, the configuration knob set that is of the database and that corresponds to the N query statements. The DS-DRL model can process data in two aspects of a feature vector of a query statement and state information of the database, to obtain a configuration knob that is of the database and that is
25 applicable to the query statement.

[0019] In a possible implementation of the first aspect, the DS-DRL model includes a prediction model, an environment model, and an agent model; the determining, based on the N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (DS-DRL) model, the configuration knob set that is of the database
30 and that corresponds to the N query statements includes: predicting a status information change amount ΔS of the database based on the N target feature vectors by using the prediction model,

where ΔS is an amount of a change from status information that is of the database and that exists before the N query statements are executed to status information that is of the database and that is obtained after the N query statements are executed; determining, by using the environment model and based on ΔS and the status information S that is of the database and that exists before the N query statements are executed, simulated status information S' that is of the database and that is obtained after the N query statements are executed in a simulated manner; determining a recommended configuration knob set based on the simulated status information S' by using the agent model; performing simulated configuration based on the recommended configuration knob set by using the environment model, executing the N target feature vectors, to update S and S', and determining a representation value R obtained after the N target feature vectors are executed; updating a knob tuning policy based on the representation value R by using the agent model, and determining an updated recommended configuration knob set based on the updated simulated status information S'; and if the representation value R meets a preset condition, determining that the updated recommended configuration knob set is the configuration knob set; or if the representation value R does not meet a preset condition, repeatedly performing the foregoing process in which the environment model updates S and S' and the agent model determines the updated recommended configuration knob set, until the representation value R meets the preset condition, where an updated recommended configuration knob set corresponding to a value R that meets the preset condition is the configuration knob set. It can be learned from this possible implementation that the DS-DRL model implements, through cooperation among the prediction model, the environment model, and the agent model, the foregoing process of determining the configuration knob set, to ensure accuracy of the configuration knob set.

[0020] In this possible implementation, meeting the preset condition may include that the value R exceeds a preset threshold, or a pre-specified resource limitation condition is reached. For example, iteration is performed for a predetermined quantity of times, or iteration is performed for preset duration.

[0021] In a possible implementation of the first aspect, when the DS-DRL model is a double-state deep deterministic policy gradient (DS-DDPG) model, the agent model includes an actor (actor) model and a score (critic) model; and the determining a recommended configuration knob set based on the simulated status information S' by using the agent model includes: determining the recommended configuration knob set based on the simulated status information S' by using the

actor model; and determining a score of the recommended configuration knob set based on the simulated status information S' and the recommended configuration knob set by using the critic model.

5 [0022] In a possible implementation of the first aspect, the updating a knob tuning policy based on the representation value R may include: updating a weight of the actor model based on the score by using the actor model.

[0023] In a possible implementation of the first aspect, the updating a knob tuning policy based on the representation value R may include: updating a weight of the critic model based on the representation value R by using the critic model.

10 [0024] According to a second aspect, an apparatus for determining a configuration knob of a database is provided, and is configured to perform the method in any one of the first aspect or the possible implementations of the first aspect. Specifically, the apparatus includes modules or units configured to perform the method in any one of the first aspect or the possible implementations of the first aspect.

15 [0025] According to a third aspect, an apparatus for determining a configuration knob of a database is provided. The apparatus may include at least one processor, a memory, and a communications interface. The processor is coupled to the memory and the communications interface. The memory is configured to store instructions, the processor is configured to execute the instructions, and the communications interface is configured to communicate with another
20 network element under control of the processor. When the instructions are executed by the processor, the processor is enabled to perform the method in any one of the first aspect or the possible implementations of the first aspect.

[0026] According to a fourth aspect, a database system is provided. The database system includes the apparatus for determining a configuration knob of a database according to the second
25 aspect or the third aspect and a database.

[0027] According to a fifth aspect, a computer-readable storage medium is provided. The computer-readable storage medium stores a program, and the program enables a data query apparatus to perform the method for determining a configuration knob of a database in any one of the first aspect and the implementations of the first aspect.

30 [0028] According to a sixth aspect, a computer program product is provided. The computer program product includes computer-executable instructions, and the computer-executable

instructions are stored in a computer-readable storage medium. At least one processor of a device may read the computer-executable instructions from the computer-readable storage medium, and the at least one processor executes the computer-executable instructions, so that the device implements the method for determining a configuration knob of a database in any one of the first aspect or the possible implementations of the first aspect.

[0029] It may be understood that any apparatus for determining a configuration knob of a database, the computer storage medium, or the computer program product provided above are all used to perform the corresponding method for determining a configuration knob of a database provided above. Therefore, for beneficial effects that can be achieved, refer to beneficial effects in the corresponding method provided above. Details are not described herein again.

BRIEF DESCRIPTION OF DRAWINGS

[0030] FIG. 1 is a schematic diagram of an architecture of a database system according to an embodiment of this application;

[0031] FIG. 2 is a schematic diagram of another architecture of a database system according to an embodiment of this application;

[0032] FIG. 3 is a schematic diagram of still another architecture of a database system according to an embodiment of this application;

[0033] FIG. 4 is a schematic diagram of yet another architecture of a database system according to an embodiment of this application;

[0034] FIG. 5 is a schematic diagram of a structure of a database server according to an embodiment of this application;

[0035] FIG. 6 is a schematic diagram of another structure of a database server according to an embodiment of this application;

[0036] FIG. 7 is a schematic diagram of a neural network model according to an embodiment of this application;

[0037] FIG. 8 is a schematic diagram of still yet another architecture of a database system according to an embodiment of this application;

[0038] FIG. 9 is a schematic diagram of an embodiment of a method for determining a configuration knob of a database according to an embodiment of this application;

[0039] FIG. 10 is a schematic diagram of a further architecture of a database system according to an embodiment of this application;

[0040] FIG. 11A and FIG. 11B are a schematic diagram of another embodiment of a method for determining a configuration knob of a database according to an embodiment of this application;

5 [0041] FIG. 12 is another schematic diagram of a neural network model according to an embodiment of this application;

[0042] FIG. 13 is a schematic diagram of a query statement vectorization coding process according to an embodiment of this application;

10 [0043] FIG. 14 is a schematic diagram of an architecture of DS-DRL according to an embodiment of this application;

[0044] FIG. 15 is a schematic diagram of an architecture of DS-DDPG according to an embodiment of this application;

[0045] FIG. 16 is a schematic diagram of an embodiment of an apparatus for determining a configuration knob of a database according to an embodiment of this application; and

15 [0046] FIG. 17 is a schematic diagram of still another structure of a database server according to an embodiment of this application.

DESCRIPTION OF EMBODIMENTS

20 [0047] The following describes embodiments of this application with reference to accompanying drawings. It is clear that the described embodiments are merely some rather than all of the embodiments of this application. A person of ordinary skill in the art may learn that technical solutions provided in the embodiments of this application are also applicable to similar technical problems with development of technologies and emergence of new scenarios.

[0048] In the specification, claims, and the accompanying drawings of this application, terms "first", "second", and the like are intended to distinguish similar objects but do not necessarily
25 indicate a specific order or sequence. It should be understood that data used in such a way is interchangeable in proper circumstances, so that the embodiments described herein can be implemented in other orders than the order illustrated or described herein. Moreover, terms "include", "have" and any other variants mean to cover the non-exclusive inclusion, for example, a process, method, system, product, or device that includes a list of steps or units is not necessarily

limited to those steps or units, but may include other steps or units not expressly listed or inherent to such a process, method, system, product, or device.

[0049] An architecture of a database system to which the embodiments of this application are applied is shown in FIG. 1. The database system includes a database 101 and a database management system (database management system, DBMS) 102.

[0050] The database 101 is an organized data set stored in a data store (data store) for a long time, in other words, an associated data set that is organized, stored, and used by using a specific data model. For example, the database 101 may include one or more pieces of table data.

[0051] The DBMS 102 is configured to: establish, use, and maintain the database 101, and uniformly manage and control the database 101, to ensure security and integrity of the database 101. A user may access data in the database 101 by using the DBMS 102, and a database administrator (database administrator, DBA) also maintains the database by using the DBMS 102. The DBMS 102 provides various functions, so that a plurality of applications and user equipments can establish, modify, and query the database in different methods at a same moment or at different moments. The application and the user equipment may be collectively referred to as a client. The DBMS 102 may provide the following functions: (1) Data definition function: The DBMS 102 provides a data definition language (data definition language, DDL) to define a database structure. The DDL is used to describe a database framework and may be stored in a data dictionary. (2) Data access function: The DBMS 102 provides a data manipulation language (data manipulation language, DML) to implement a basic access operation on data in the database, for example, retrieval, insertion, modification, and deletion. (3) Database running management function: The DBMS 102 provides a data control function, to be specific, data security, integrity, concurrency control, and the like, to effectively control and manage running of the database, and ensure correct and effective data. (4) Database establishment and maintenance function, including functions such as loading initial data into the database, dumping, restoring, and re-organizing the database, and monitoring and analyzing system performance. (5) Database transmission: The DBMS 102 provides processed data transmission to implement communication between a client and the DBMS 102, usually in coordination with an operating system.

[0052] Specifically, FIG. 2 is a schematic diagram of a single-server database system. The single-server database system includes a database management system and a data store (data store). The database management system is configured to provide a service such as query and

modification of a database, and the database management system stores data in the data store. In the single-server database system, the database management system and the data store are usually located on a single server, for example, a symmetric multi-processor (symmetric multi-processor, SMP) server. The SMP server includes a plurality of processors. All the processors share resources such as a bus, a memory, and an I/O system. A function of the database management system may be implemented by executing a program in the memory by one or more processors.

[0053] FIG. 3 is a schematic diagram of a cluster database system in which a shared-storage (shared-storage) architecture is used. The cluster database system includes a plurality of nodes (nodes 1–N in FIG. 3). A database management system is deployed on each node to separately provide services such as database query and modification for a user. A plurality of database management systems store shared data in a shared data store, and perform a read/write operation on data in the data store by using a switch. The shared data store may be a shared disk array. A node in the cluster database system may be a physical machine, for example, a database server, or may be a virtual machine running on an abstract hardware resource. If the node is a physical machine, the switch is a storage area network (storage area network, SAN) switch, an Ethernet switch, a fibre channel switch, or another physical switch device. If the node is a virtual machine, the switch is a virtual switch.

[0054] FIG. 4 is a schematic diagram of a cluster database system of a shared-nothing (shared-nothing) architecture. Each node has an exclusive hardware resource (for example, a data store), an operating system, and a database, and nodes communicate with each other by using a network. In this system, data is distributed to each node based on a database model and an application characteristic. A query task is divided into several parts, and the parts are executed concurrently on all the nodes. All the nodes perform calculation coordinately and serve as a whole to provide a database service. All communications functions are implemented in a high-bandwidth network interconnection system. Similar to the cluster database system of the shared disk architecture that is described in FIG. 3, the node herein may be a physical machine or may be a virtual machine.

[0055] In all the embodiments of this application, the data store (data store) of the database system includes but is not limited to a solid-state disk (SSD), a disk array, or a non-transitory computer-readable medium of another type. Although no database is shown in FIG. 2 to FIG. 4, it should be understood that the database is stored in the data store. A person skilled in the art may understand that a communications system may include fewer or more components than those

shown in FIG. 2 to FIG. 4, or include components different from those shown in FIG. 2 to FIG. 4. FIG. 2 to FIG. 4 show only components more related to an implementation disclosed in the embodiments of this application. For example, although four nodes are described in FIG. 3 and FIG. 4, a person skilled in the art may understand that the cluster database system may include any quantity of nodes. A function of a database management system of each mode may be implemented by using an appropriate combination of software, hardware, and/or firmware on the node.

[0056] A person skilled in the art may clearly understand according to teachings of the embodiments of the present invention that, a method in the embodiments of the present invention is applied to a database management system. The database management system may be applied to a single-server database system, a cluster database system of a shared-nothing architecture, a cluster database system of a shared-storage architecture, or a database system of another type.

[0057] Further, referring to FIG. 1, when querying the database 101, the DBMS 102 usually needs to perform steps such as syntax analysis, pre-compiling, and optimization on a query statement, obtains through estimation an execution method having a cost that the database system considers to be minimum, and further generates an execution plan with the minimum cost. During operation, an execution structure performs a data operation according to the generated execution plan, so as to improve performance of the database system. When performing cost estimation on the query statement, the DBMS 102 needs to collect statistical information of the query statement, and performs cost estimation based on the collected statistical information. A statistical information collection method may be model information obtained by performing model training through machine learning, or statistical information obtained through data sampling and statistics collection. The model information may alternatively be referred to as statistical information.

[0058] The DBMS 102 may be located in a database server. For example, the database server may be specifically the SMP server in the single-server database system in FIG. 2 or the node in FIG. 3 or FIG. 4. Specifically, as shown in FIG. 5, the database server may include a kernel 1021 and an external trainer 1022 that is independent of the kernel 1021 and that is located in the database server. Alternatively, as shown in FIG. 6, the database server includes a kernel 1021, and an external trainer 1022 is located outside the database server. The kernel 1021 is a core of the database server and may be configured to perform various functions provided by the DBMS 102. The kernel 1021 may include a utility 10211 and an optimizer 10212. When the database server queries the database 101, the utility 10211 may trigger the external trainer 1022 to perform model

training through machine learning, to obtain model information of a training model. The optimizer 10212 may perform cost estimation based on the model information obtained through training performed by the external trainer 1022, to generate the execution plan with the minimum cost, so that the execution structure performs the data operation according to the generated execution plan, 5 so as to improve the performance of the database system.

[0059] Machine learning is a process of obtaining a new inference model by relying on learning or observation of existing data. Machine learning may be implemented by using a plurality of different algorithms, and common machine learning algorithms may include: models such as a neural network (neural network, NN) and a random forest (random forest, RF). For 10 example, the neural network may include a feed forward neural network (feed forward neural network, FFNN) and a recurrent neural network (recurrent neural network, RNN). FIG. 7 is a schematic diagram of a model of a neural network. The model may include an input layer, a hidden layer, and an output layer. The layers may include different quantities of neurons.

[0060] Because a quantity of clients that request a database to provide a service is not fixed, 15 to enable the database to have better performance (for example, performance in aspects of a delay and a throughput) when the database provides a service for a client, an embodiment of this application provides a method for determining a configuration knob of a database. In this method, a double-state-based configuration knob tuning solution is provided. The double states may be a feature vector of a query statement (query vector) and state information of the database (database 20 states). The feature vector of the query statement affects statistical information of each activity of the database. In other words, the query statement or a workload may affect the statistical information of each activity of the database. The state information of the database may include an internal configuration of the database, and the configuration is affected by knob tuning behavior. The double-state-based configuration knob tuning solution provided in this embodiment of this 25 application may be implemented by using a double-state deep reinforcement learning (double-state deep reinforcement learning, DS-DRL) model. The DS-DRL model may process data in two aspects of the feature vector of the query statement and the state information of the database, to obtain a configuration knob that is of the database and that is suitable for the query statement. The DS-DRL model may have a plurality of specific representation forms. For example, one 30 representation form may be a double-state deep deterministic policy gradient (double-state deep deterministic policy gradient, DS-DDPG) model. Specific structures of the DS-DRL model and

the DS-DDPG model and functions in a process of determining a configuration knob of a database are described below. Details are not described herein.

[0061] In this embodiment of this application, the double-state-based configuration knob tuning solution includes knob tuning modes at three granularities, and the knob tuning modes at three granularities are a query-level knob tuning mode (query-level tuning), a workload-level knob tuning mode (workload-level tuning), and a cluster-level knob tuning mode (cluster-level tuning). Specific content of the knob tuning modes at three granularities in the process of determining a configuration knob of a database is described below. Details are not described herein.

[0062] Because the method for determining a configuration knob of a database provided in this embodiment of this application is applied to a DBMS, the DBMS provides, for the DB, a configuration knob set that matches a workload, and further tunes the configuration knob of the DB based on the configuration knob set. Because both the DBMS and the DB belong to a database system, the following describes the process of determining a configuration knob of a database in this embodiment of this application with reference to FIG. 8.

[0063] FIG. 8 is a schematic diagram of still yet another architecture of a database system according to an embodiment of this application.

[0064] As shown in FIG. 8, the database system includes a DBMS and a DB. A function of the DB is basically the same as that described in the embodiment corresponding to FIG. 1, and details are not described herein again. In addition, it should be noted that the DB may tune, based on a configuration knob set provided by the DBMS, a configuration knob used for the DB.

[0065] The DBMS may include a controller (controller), a query statement vectorization module (Query2Vector), a workload vectorization module (Workload2Vector), a workload clustering module (Workload2Cluster), and a knob tuning system (tuner). The DBMS may further include a training knob set storage module. If the training knob set storage module is included, the module stores training data (training data). It should be noted that "2" in (Query2Vector) indicates a meaning of "to" in English, and indicates that a query statement is converted into a vector. "2" in other modules also has a same meaning, and is not listed one by one herein.

[0066] The training data is mainly used to train a DS-DRL model. If the DS-DRL model is trained offline, the database system may alternatively not include the training knob set storage module. In a possible implementation, even if the DS-DRL model is trained offline, the database system may still include the training knob set storage module. In a knob tuning process, new data

generated by the knob tuning system may also be updated to the training data of the training knob set storage module, to optimize the DS-DRL model online.

[0067] A quantity of clients shown in FIG. 8 is not limited. There may be one client, or there may be two or more clients.

5 [0068] When using a database, the client sends a query statement (query). The query statement may be sent in a manner of a query request.

[0069] The DBMS may perform a process of a method for determining a configuration knob of a database in FIG. 9.

[0070] As shown in FIG. 9, an embodiment of a method for determining a configuration knob
10 of a database provided in an embodiment of this application may include the following steps.

[0071] 201: Obtain N query statements sent by M clients.

[0072] M is an integer greater than 0, N is an integer greater than 0, and $N \geq M$.

[0073] In this embodiment of this application, there may be only one client, or there may be two or more clients. One client may send one query statement, or may send two or more query
15 statements.

[0074] 202: Determine N pieces of query information from the N query statements.

[0075] The N query statements are in a one-to-one correspondence with the N pieces of query information.

[0076] 203: Perform vectorization coding on the N pieces of query information, to obtain
20 corresponding N target feature vectors.

[0077] The N pieces of query information are in a one-to-one correspondence with the N target feature vectors.

[0078] 204: Determine, based on the N target feature vectors and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query
25 statements.

[0079] The configuration knob set includes at least one target configuration knob group, and the target configuration knob group includes at least one configuration knob. The configuration knob set is used to execute the N query statements.

[0080] In a possible implementation, after step 204, step 205 may be further performed.

30 [0081] 205: When the N query statements are executed, tune a configuration knob currently used for the database based on the configuration knob set.

[0082] A configuration knob tuning manner may be replacing the configuration knob currently used for the database with a configuration parameter in a current configuration parameter set. Alternatively, the configuration knob currently used for the database may be tuned by adding an offset value to or reducing an offset value from a value of the configuration knob currently used for the database.

[0083] In a possible implementation, the foregoing steps 201 to 205 may also be performed by a part remained after a controller is removed from a DBMS, and the part is used as an apparatus for determining a configuration knob of a database. In this case, the controller provides an interface for the client to access the database, to forward a query statement. Then, the apparatus for determining a configuration knob of a database performs the foregoing steps 201 to 205.

[0084] In this embodiment of this application, the configuration knob of the database may be determined based on a target feature vector of a query statement and the current status information of the database, to ensure good performance of the database in a case of different workload requirements. For example, performance in aspects of a delay and a throughput can be met.

[0085] As described above, this embodiment of this application provides knob tuning modes at three granularities, and processes of performing vectorization coding on the query statement in the three knob tuning modes are as follows:

[0086] 1. In a query-level knob tuning mode (query-level tuning), a query statement vectorization module (Query2Vector) performs vectorization coding on a single query statement, to obtain a target feature vector of the query statement. Then, the target feature vector of the query statement is transferred to a knob tuning system (tuner), and the knob tuning system determines, based on the feature vector of the query statement, a configuration knob set that is of the database and that corresponds to the query statement. In other words, when the query-level knob tuning mode is executed, one configuration knob set is obtained by using the foregoing process, and the configuration knob set includes one target configuration knob group corresponding to the query statement. The one target configuration knob group is used to tune, when the single query statement is executed, the configuration knob currently used for the database. In the query-level knob tuning mode, knob tuning is performed once for one query statement, and a knob tuning delay is very small, so that a personalized knob tuning requirement of different query statements can be met.

[0087] 2. In a workload-level knob tuning mode (workload-level tuning), a query statement

vectorization module (Query2Vector) performs vectorization coding on each of the N query statements, to obtain the N target feature vectors. The N query statements are all query statements that request the database to provide a service. Then, the query statement vectorization module transfers the N target feature vectors to a workload vectorization module (Workload2Vector), the workload vectorization module fuses the N target feature vectors into a unified vector, the workload vectorization module transfers the unified vector to a knob tuning system, and the knob tuning system determines, based on the unified vector, a configuration knob set that is of the database and that corresponds to the N query statements.

[0088] A process in which the workload vectorization module fuses the N target feature vectors into the unified vector may be fusing corresponding parts of the N target feature vectors. For example, each target feature vector includes three subvectors: a query type, table information, and cost information. Vector fusion may be performed as follows: The subvector of the query type and the subvector of the table information may be fused based on logic of "and", and the subvector of the cost information may be fused through addition. Certainly, a fusion manner is not limited to the listed fusion manner herein. Some subvectors may be fused based on the logic of "and", and some subvectors may be fused in a manner of a vector product.

[0089] In the workload-level knob tuning mode, knob tuning may be performed once for the N query statements, to meet a requirement of a scenario in which there is a large throughput.

[0090] 3. In a cluster-level knob tuning mode (cluster-level tuning), a query statement vectorization module (Query2Vector) performs vectorization coding on each of the N query statements, to obtain N target feature vectors. The N query statements are all query statements that request the database to provide a service. Then, the query statement vectorization module transfers the N target feature vectors to a knob tuning system, the knob tuning system determines N configuration knob groups for the N target feature vectors, the knob tuning system transfers the N configuration knob groups to a workload clustering module (Workload2Cluster), and the workload clustering module performs clustering on the N configuration knob groups, to obtain Q target configuration knob groups. Then, the workload clustering module transfers the Q target configuration knob groups to the knob tuning system, and the knob tuning system determines, based on the Q target configuration knob groups, the configuration knob set that is of the database and that corresponds to the N query statements. A process in which the workload clustering module performs clustering on the N configuration knob groups to obtain the Q target configuration knob

groups may be performing clustering based on a similarity of configuration knobs between groups. For example, clustering may be performed on a same type of configuration knob in two groups in a manner of a Euclidean shortest path.

[0091] In the cluster-level knob tuning mode, the knob tuning system determines the N configuration knob groups for the N target feature vectors, and then the workload clustering module performs clustering on the N configuration knob groups, to obtain the Q target configuration knob groups. Actually, a manner of obtaining the Q target configuration knob groups is not limited to such a manner in which the Q target configuration knob groups are obtained based on the N target feature vectors. Alternatively, the workload clustering module may determine the N configuration knob groups based on the N target feature vectors, and perform clustering on the N configuration knob groups, to obtain the Q target configuration knob groups. In this implementation, the workload clustering module may include two parts: a vector-to-knob module (Vector2Pattern) and a knob clustering module (Pattern2Cluster). In this case, content related to clustering may be understood with reference to FIG. 10. The workload vectorization module (Workload2Vector) that is unrelated to clustering is not shown in FIG. 10, but it does not mean that the workload vectorization module does not exist.

[0092] In a clustering process shown in FIG. 10, the query statement vectorization module transfers the N target feature vectors to the vector-to-knob module, and the vector-to-knob module determines the N configuration knob groups for the N target feature vectors. Then, the vector-to-knob module transfers the N configuration knob groups to the knob clustering module, and the knob clustering module performs clustering on the N configuration knob groups, to obtain the Q target configuration knob groups. The vector-to-knob module may include a deep learning model (deep learning model, DL Model). The DL model may determine a discrete value as a configuration knob. For example, the configuration knob may be determined in $\{-1, 0, 1\}$ based on a relationship between an estimated value and a default value. Herein, 0 is taken when the estimated value is near the default value; 1 is taken when the estimated value is much greater than the default value; and -1 is taken when the estimated value is much less than the default value. In this solution in which the configuration knob is represented by using the discrete value, a large quantity of calculation overheads can be reduced.

[0093] A specific knob tuning mode to be executed may be selected by a user or an administrator based on an actual requirement. In this case, the method for determining a

configuration knob of a database provided in this embodiment of this application may further include:

determining, in response to a mode selection instruction from a query-level knob tuning mode, a workload-level knob tuning mode, and a cluster-level knob tuning mode, a target knob tuning mode used for the N query statements.

[0094] In other words, when the target knob tuning mode is the query-level knob tuning mode, $M = 1$, and $N = 1$. To be specific, for the query-level knob tuning mode, one configuration knob set may be determined for one query statement. The configuration knob set includes one target configuration knob group for the query statement, and the target configuration knob group is configured for the database when the query statement is executed.

[0095] When the target knob tuning mode is the workload-level knob tuning mode, the method for determining a configuration knob of a database may further include: fusing the N target feature vectors, to obtain a unified vector obtained after fusion. Step 204 may include: determining, based on the unified vector and the current status information of the database, the configuration knob set that is of the database and that corresponds to the N query statements, where the configuration knob set includes a target configuration knob group, and the target configuration knob group is used to tune, when the N query statements are executed, the configuration knob currently used for the database. In the workload-level knob tuning mode, knob tuning may be performed once for the N query statements, to meet a requirement of a scenario in which there is a large throughput.

[0096] When the target knob tuning mode is the cluster-level knob tuning mode, the determining, based on the N target feature vectors and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements may include:

determining N configuration knob groups based on the N target feature vectors and the current status information of the database, where the N configuration knob groups are in a one-to-one correspondence with the N target feature vectors; and

performing clustering on the N configuration knob groups, to obtain Q target configuration knob groups, where a first configuration knob group corresponds to at least one target feature vector, the first configuration knob group is any one of the Q target configuration knob groups, Q is an integer greater than 0, $Q < N$, and the first configuration knob group is used to tune, when a query statement corresponding to the at least one target feature vector is executed,

a configuration knob currently used for the database.

[0097] In this cluster-level knob tuning mode, one target configuration knob group may be determined for a same type of query statement, to ensure a large throughput and meet a requirement for a low delay.

5 [0098] FIG. 8 and FIG. 10 describe a process of determining a configuration knob of a database from a perspective of a modular DBMS. FIG. 11A and FIG. 11B describe, from a perspective of a processing procedure, a process of performing subsequent processing on the N target feature vectors to obtain a configuration knob set, so as to determine the configuration knob of the database in the foregoing embodiment.

10 [0099] FIG. 11A and FIG. 11B show a processing process in the knob tuning modes at three granularities provided in this embodiment of this application, and further extracts a working process of the knob tuning system from the processing process in the knob tuning modes at three granularities and a processing process of vector-to-knob (Vector2Pattern) in the cluster-level knob tuning mode. DS-DDPG is used as an example of a knob tuning model in FIG. 11A and FIG. 11B,
15 but the knob tuning model is not limited to the DS-DDPG.

[00100] In FIG. 11A and FIG. 11B, in the query statement-level (query-level) knob tuning mode, after vectorization coding is performed on any query statement Q_i , a feature vector can be obtained. After the feature vector is input into the DS-DDPG model, a configuration knob group $[a_1, a_2, \dots, a_m]$ for the query statement Q_i may be obtained. In the (query-level) knob tuning mode, a
20 configuration knob set A includes the configuration knob group $[a_1, a_2, \dots, a_m]$. In other words, $A = [a_1, a_2, \dots, a_m]$. When the query statement Q_i is executed, the configuration knob set A may be used to determine the configuration knob of the database.

[00101] In FIG. 11A and FIG. 11B, in the workload-level (workload-level) knob tuning mode, there are N query statements in a universal set, and the N query statements may be represented as
25 $Q = \{Q_1, Q_2, \dots, Q_n\}$. Vectorization coding is separately performed on the N query statements, to obtain N target feature vectors. The N target feature vectors are fused into a unified vector V^* , and V^* is input into the DS-DDPG model, to obtain a configuration knob group $[a_1, a_2, \dots, a_m]$ for the N query statements $Q = \{Q_1, Q_2, \dots, Q_n\}$. In the (workload-level) knob tuning mode, a configuration knob set A includes the configuration knob group $[a_1, a_2, \dots, a_m]$. In other words, A
30 $= [a_1, a_2, \dots, a_m]$. When the N query statements $Q = \{Q_1, Q_2, \dots, Q_n\}$ are executed, the configuration knob set A may be used to determine the configuration knob of the database.

[00102] In FIG. 11A and FIG. 11B, in the cluster-level (cluster-level) knob tuning mode, there are N query statements in a universal set, and the N query statements may be represented as $Q = \{Q_1, Q_2, \dots, Q_n\}$. Vectorization coding is separately performed on the N query statements, to obtain N target feature vectors. N configuration knob groups $[A_1, A_2, \dots, A_n]$ corresponding to the N target feature vectors may be determined by using the DS-DDPG model. Alternatively, N configuration knob groups $[P_1, P_2, \dots, P_n]$ corresponding to the N target feature vectors may be determined by using a DL model of vector-to-knob (Vector2Pattern). Subsequently, clustering is performed on both the N configuration knob groups $[A_1, A_2, \dots, A_n]$ and the N configuration knob groups $[P_1, P_2, \dots, P_n]$, to obtain Q target configuration knob groups $\{U_1, U_2, \dots, U_q\}$. Then, vectorization coding is performed on each target configuration knob group U_j , to obtain vectorization coding V^{**} corresponding to U_j , and the vectorization coding V^{**} corresponding to U_j is input into the DS-DDPG model, to obtain a configuration knob set A corresponding to U_j . When the query statement corresponding to U_j is executed, the configuration knob set A may be used to determine the configuration knob of the database.

[00103] In the knob tuning modes at three granularities, a working principle of the DS-DDPG model is basically the same. The DS-DDPG model predicts a status change amount $\Delta S = [\Delta s_1, \Delta s_2, \dots, \Delta s_m]$ of the database based on a target feature vector of a query statement, determines, with reference to current status information $S = [s_1, s_2, \dots, s_m]$ and based on $S' = \Delta S + S$, that $S' = [S'_1, S'_2, \dots, S'_m]$, and outputs S' to an actor (actor)-score (critic) module by using S' as a performance value (the performance value may also be referred to as an observation value (observation)). The (actor-critic) module may also be referred to as an (actor-critic) model. The (actor-critic) model determines the configuration knob set A. For a specific knob tuning process of the DS-DDPG model, detailed description is provided below. Details are not described herein.

[00104] In the cluster-level (cluster-level) knob tuning mode, in a process of the DL model, the target feature vector of the query statement is used as an input. Then, the target feature configuration knob group $P = [p_1, p_2, \dots, p_n]$ is obtained after conversion at an input layer (input layer), a hidden layer (hidden layer), and an output layer (output layer).

[00105] The DL model may also be understood with reference to FIG. 12. L1 is an input layer, L2, L3, and L4 are hidden layers, L5 is an output layer, and target feature vectors v_1 , and v_2 to v_k are input into the L1 layer. Then, after a relationship between L1, L2, L3, L4, and L5, the target feature configuration knob group $P = [p_1, p_2, \dots, p_n]$ is output.

[00106] It can be learned from the foregoing description that in a knob tuning mode at each granularity, vectorization coding needs to be performed on the query statement. Vectorization coding is performed on the query information in the query statement. The query information in this embodiment of this application includes a query type, table information, and a type and corresponding cost information of an involved operator in a query. For example, the query type may include insert (insert), select (select), update (update), and delete (delete). The table information may include identification information of a table, or may include a data volume and a data structure of the table. The type of the involved operator in the query may include types such as sequential scan (Seq_Scan), hash join (Hash_Join), and an aggregation operation (Aggregate).
5 The cost information of the operator is an execution cost of the operator when the query statement is executed. An execution cost of each operator may be calculated in advance based on a query plan of the query statement by using an optimizer. In other words, the execution cost of each operator may be obtained in advance.

[00107] FIG. 13 shows a process of vectorizing a query statement. In the process, an execution cost of an operator is estimated based on the query plan, and then each part of the query plan is vectorized.
15

[00108] As shown in FIG. 13, a template vector may include three parts: a query type (DML), a table (Tables), and an operator cost (Operation Costs).

[00109] For the DML, the template vector may include insert (insert), select (select), update (update), and delete (delete). If the query statement is a selection operation, a bit of select is filled with 1, and other bits of insert, update, and delete are filled with 0.
20

[00110] For the table part, if the query plan of the query statement includes table information of three tables: tbl1, tbl2, and tbl3, bits of tbl1, tbl2, and tbl3 are filled with 1, and other bits of tbl4 to tbl8 are filled with 0.

[00111] For each type of operator, execution costs of operators of a same type in an operator structure tree may be accumulated, and then a floating-point bit of an operator of a corresponding type is filled with a cost value obtained after accumulation. For example, in FIG. 13, each operator includes a time point of an operation corresponding to a cost estimation, and the time point of the operation corresponding to the cost estimation includes a total time (Total), a start time (Startup), and a child node time (Child). The total time indicates a time point at which all execution results are returned. The start time indicates a time point at which a result starts to be returned. The child
25
30

node time indicates an earliest time at which a node operator starts to be executed. For any operator node in the query plan, (total time—child node time) can be used to represent a maximum execution cost of the operator node. For example, there are three sequential scanning operators in FIG. 13, and a child node time of each sequential scanning operator is 0. Therefore, a sum of execution costs of the three sequential scanning operators is $2.41 + 17.8 + 20.7 = 40.91$. Similarly, an execution cost of the operator Hash_Join is $23.19 + 48.16 - 2.43 = 68.92$, and an execution cost of the operator Aggregate is $48.28 - 23.24 = 25.04$.

[00112] The target feature vector of the query statement is obtained after a value is assigned to each of the three parts of the template vector: the query type, the table information, and the operator cost. To eliminate a feature importance difference caused by different value ranges, normalization processing may be performed on the execution costs of these operators, to obtain a target feature vector obtained after normalization processing, and all features are constrained with a same order of magnitude. For example, in a process of generating a final result, a feature whose value is 10000 has a greater proportion than a feature whose value is -1, but both features may have a proportion of 0.5 after normalization. This means that the two features have a same degree of impact.

[00113] The DML and the table of the target feature vector are represented by using a bit identifier. Actually, a representation manner is not limited to such a manner of representation performed by using a bit identifier in this embodiment of this application, and representation may alternatively be performed in form of a floating-point bit identifier. When the bit identifier is used for representation, a representation manner is not limited to 0 and 1 in the foregoing representation manner. For example, T may be used to replace 1, F may be used to replace 0, or other representations may be used to replace 1 and 0. When the floating-point bit identifier is used for representation, all involved query types may be marked by using different characters. For example, insert (insert) is represented by using 1, select (select) is represented by using 2, update (update) is represented by using 3, and delete (delete) is represented by using 4. Certainly, the four query types are merely used as an example for description herein. The query types are not limited to the four query types, and representations of the query types are not limited to the listed numerical form herein. Different query types may be represented by using another symbol or numerical form.

[00114] Vectorization is described in a process of assigning a value to the vector template. Actually, if any one of the N pieces of query information is selected, and is referred to as first query information. The process may alternatively be described as follows: The first query

information includes a first query type, first table information, and a type and corresponding cost information of an involved operator in a query, the first query type is used to indicate a type of an operation that is requested by a query statement corresponding to the first query information and that is to be performed on the database, the first table information is used to indicate a relationship table involved when the query statement corresponding to the first query information is executed in the database, and the cost information is used to indicate a respective execution cost that is of the involved operator and that is incurred when the involved operator executes the query statement corresponding to the first query information.

[00115] A target feature vector corresponding to the first query information includes an identifier of the first query type, an identifier of a second query type, an identifier of the first table information, an identifier of second table information, and a floating-point bit of each type of operator; the identifier of the first query type is represented by using a first value, the identifier of the second query type is represented by using a second value, and the second query type belongs to a query type of the database and is not included in the first query information; the identifier of the first table information is represented by using a third value, the identifier of the second table information is represented by using a fourth value, and the second table information belongs to table information of the database and is not included in the first query information; and the floating-point bit of each type of operator includes an execution cost of the corresponding type of operator. The execution cost of the corresponding type of operator is a normalized execution cost.

[00116] The foregoing describes a process of performing vectorization coding on a query statement. The following describes a process of tuning a knob based on a DS-DRL model in a knob tuning system (tuner) and a process of tuning a knob based on a DS-DDPG model.

[00117] The following first describes a process of tuning a knob based on a DS-DRL model with reference to FIG. 14.

[00118] The DS-DRL model includes a prediction model (predictor), an environment model (environment), and an agent model (agent). The environment model includes two parts: outer measurement (outer metric) and an inner state (inner state). The process of tuning a knob based on a DS-DRL model may include the following steps.

[00119] 301: A workload (workload) provides a query statement to a query statement vectorization module.

[00120] The query statement vectorization module performs the foregoing described

vectorization coding process for the query statement, to obtain N target feature vectors.

[00121] 302: The query statement vectorization module transfers the N target feature vectors to the prediction model.

[00122] The prediction model predicts a status information change amount ΔS of a database
5 based on the N target feature vectors. Herein, ΔS is an amount of a change from status information that is of the database and that exists before the N query statements are executed to status information that is of the database and that is obtained after the N query statements are executed.

[00123] The prediction model may be a neural network model.

[00124] 303: The prediction model transfers ΔS to the environment model.

10 [00125] The environment model learns of ΔS , and determines, based on the status information S that is of the database, that exists before the N query statements are executed, and that is stored in an outer state (outer metric), simulated status information S' that is of the database and that is obtained after the N query statements are executed in a simulated manner. A formula may be $S' = \Delta S + S$.

15 [00126] The state information of the database may include utilization of a central processing unit (central processing unit, CPU) of the database, network utilization, a quantity of input/outputs (input/output, I/O) of a disk, a busy degree of the disk, a memory page switching amount, and the like.

[00127] 304: The environment model transfers S' to the agent model.

20 [00128] The agent model uses S' as an input, to generate a recommended configuration knob set (action).

[00129] 305: The agent model transfers the recommended configuration knob set to the environment model.

[00130] The environment model performs simulated configuration based on the recommended
25 configuration knob set.

[00131] The environment model further includes the inner state (inner state). The inner state stores a current configuration knob of the database, so that the configuration knob in the inner state can be tuned based on the recommended configuration knob set (action).

[00132] 306: The query statement vectorization module transfers the N target feature vectors to
30 the environment model.

[00133] The environment model executes the N target feature vectors, to update S and S'.

[00134] The environment model determines a performance value (reward, R) obtained after the N target feature vectors are executed.

[00135] A calculation process of R may be understood with reference to the following formula:

$$R = \sum_{i=1}^m w_i r_i$$

5 [00136] Herein, R is the performance value obtained after the N target feature vectors are executed, w_i is a weight of the i^{th} performance indicator in m performance indicators (for example, a delay and a throughput), and r_i is a performance value of the i^{th} performance indicator.

$$\text{Herein, } r_i = \begin{cases} \left((1 + \Delta_{(t-1),t})^2 - 1 \right) |1 + \Delta_{0,t}|, & \Delta_{0,t} > 0 \\ - \left((1 - \Delta_{(t-1),t})^2 - 1 \right) |1 + \Delta_{0,t}|, & \Delta_{0,t} \leq 0 \end{cases}$$

[00137] Herein, $\Delta_{0,t}$ indicates an amount of a change from initial performance to current
10 performance, and $\Delta_{(t-1),t}$ indicates an amount of a change from previous performance to the current performance.

$$\text{Herein, } \Delta_{0,t} = \begin{cases} \frac{m_t - m_0}{m_0}, & \text{when a larger measurement value} \\ & \text{indicates better performance} \\ \frac{m_0 - m_t}{m_0}, & \text{when a smaller measurement value} \\ & \text{indicates better performance} \end{cases}$$

$$\text{Herein, } \Delta_{(t-1),t} = \begin{cases} \frac{m_t - m_{t-1}}{m_{t-1}}, & \text{when a larger measurement value} \\ & \text{indicates better performance} \\ \frac{m_{t-1} - m_t}{m_{t-1}}, & \text{when a smaller measurement value} \\ & \text{indicates better performance} \end{cases}$$

[00138] Herein, m_0 indicates initial performance measurement, m_t indicates current
15 performance measurement, and m_{t-1} indicates previous performance measurement relative to the current performance measurement. The performance measurement may be, for example, a delay, a throughput, or an execution time.

[00139] 307: The environment model transfers the value R to the agent model.

[00140] The agent model updates a knob tuning policy based on the representation value R, and
20 determines an updated recommended configuration knob set based on the updated simulated status

information S'.

[00141] If the representation value R meets a preset condition, the updated recommended configuration knob set is determined as the configuration knob set.

[00142] If the representation value R does not meet a preset condition, the process in which the environment model updates S and S' and the agent model determines the updated recommended configuration knob set in 305 and 306 is repeatedly performed, until the representation value R meets the preset condition. An updated recommended configuration knob set corresponding to a value R that meets the preset condition is the configuration knob set.

[00143] In this possible implementation, meeting the preset condition may include that the value R exceeds a preset threshold, or a pre-specified resource limitation condition is reached. For example, iteration is performed for a predetermined quantity of times, or iteration is performed for preset duration.

[00144] In other words, step 204 may include: determining, based on the N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (DS-DRL) model, the configuration knob set that is of the database and that corresponds to the N query statements.

[00145] When the DS-DRL model includes the prediction model, the environment model, and the agent model, the determining, based on the N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (DS-DRL) model, the configuration knob set that is of the database and that corresponds to the N query statements may include:

predicting a status information change amount ΔS of the database based on the N target feature vectors by using the prediction model, where ΔS is an amount of a change from status information that is of the database and that exists before the N query statements are executed to status information that is of the database and that is obtained after the N query statements are executed;

determining, by using the environment model and based on ΔS and the status information S that is of the database and that exists before the N query statements are executed, simulated status information S' that is of the database and that is obtained after the N query statements are executed in a simulated manner;

determining a recommended configuration knob set based on the simulated status

information S' by using the agent model;

performing simulated configuration based on the recommended configuration knob set by using the environment model, executing the N target feature vectors, to update S and S' , and determining a representation value R obtained after the N target feature vectors are executed;

5 updating a knob tuning policy based on the representation value R by using the agent model, and determining an updated recommended configuration knob set based on the updated simulated status information S' ; and

if the representation value R meets a preset condition, determining that the updated recommended configuration knob set is the configuration knob set; or

10 if the representation value R does not meet a preset condition, repeatedly performing the foregoing process in which the environment model updates S and S' and the agent model determines the updated recommended configuration knob set, until the representation value R meets the preset condition, where an updated recommended configuration knob set corresponding to a value R that meets the preset condition is the configuration knob set.

15 **[00146]** The foregoing describes the process of tuning a knob based on a DS-DRL model. The following describes a process of tuning a knob based on a DS-DDPG model with reference to FIG. 15.

[00147] When the DS-DRL model is a DS-DDPG model, the agent model includes an actor (actor) model and a score critic model. The actor (actor) model and the score (critic) model may
20 be neural network models.

[00148] Steps 401 to 403 in the process of tuning a knob based on a DS-DDPG model are the same as steps 301 to 303 in the foregoing embodiment. Details are not described herein again.

[00149] 404: The environment model transfers S' to the actor model.

[00150] The actor model uses S' as an input, to generate a recommended configuration knob set
25 (action). To be specific, the determining a recommended configuration knob set based on the simulated status information S' by using the agent model may include: determining the recommended configuration knob set based on the simulated status information S' by using the actor model; and determining a score of the recommended configuration knob set based on the simulated status information S' and the recommended configuration knob set by using the critic
30 model.

[00151] Steps 405 and 406 are the same as steps 305 and 306 in the foregoing embodiment.

Details are not described herein again.

[00152] 407: The environment model transfers the value R to the critic model.

[00153] 408: The actor model transfers the configuration knob set to the critic model.

[00154] The critic model determines a score of the recommended configuration knob set based
5 on the simulated status information S' and the recommended configuration knob set.

[00155] 409: The critic model transfers the score to the actor model.

[00156] The updating a knob tuning policy based on the representation value R in 307 may
include the following step:

[00157] The actor model updates a weight of the actor model based on the score. In other words,
10 the weight of the actor model is updated based on the score by using the actor model.

[00158] The critic model updates the weight of the critic model based on the representation
value R. In other words, the weight of the critic model is updated based on the representation value
R by using the critic model.

[00159] In comparison with a default configuration or a configuration recommended by a DBA,
15 there is a shorter delay in the solution for tuning a configuration knob of a database provided in
this embodiment of this application. Table 1 shows delay test results of testing a use case of TPCH
10x and a use case of TPCH 30x in a case of different configurations by using a use case of
transaction processing performance council BenchmarkTMH (transaction processing performance
council BenchmarkTMH, TPC-H) as an example.

20 **Table 1: Total job execution delay (seconds: s) in case of different configurations**

	Default Configuration	Configuration recommended by a DBA	Configuration recommended for optimization
TPCH 10x	293.37	283.12	197.90
TPCH 30x	1083.67	872.45	539.86

[00160] It can be learned from Table 1 that, in both the use case of TPCH 10x and the use case
of TPCH 30x, the delay in the solution of the configuration recommended for optimization
provided in this embodiment of this application is far less than those of the default configuration
25 and the configuration recommended by the DBA.

[00161] The database system and the method for determining a configuration knob of a database

are described in the foregoing plurality of embodiments. The following describes, with reference to accompanying drawings, an apparatus 50 for determining a configuration knob of a database provided in an embodiment of this application.

[00162] As shown in FIG. 16, an embodiment of an apparatus 50 for determining a configuration knob of a database provided in an embodiment of this application includes:

an obtaining unit 501, configured to obtain N query statements sent by M clients, where M is an integer greater than 0, N is an integer greater than 0, and $N \geq M$;

a first determining unit 502, configured to determine N pieces of query information from the N query statements obtained by the obtaining unit 501, where the N query statements are in a one-to-one correspondence with the N pieces of query information;

a vectorization coding unit 503, configured to perform vectorization coding on the N pieces of query information determined by the first determining unit 502, to obtain corresponding N target feature vectors, where the N pieces of query information are in a one-to-one correspondence with the N target feature vectors; and

a second determining unit 504, configured to determine, based on the N target feature vectors obtained by the vectorization coding unit 503 and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements, where the configuration knob set includes at least one target configuration knob group, the target configuration knob group includes at least one configuration knob, and the configuration knob set is used to execute the N query statements.

[00163] In this embodiment of this application, the configuration knob of the database may be determined based on the target feature vector of the query statement and the current status information of the database, to ensure good performance of the database in a case of different workload requirements. For example, performance in aspects of a delay and a throughput can be met.

[00164] In a possible implementation, the apparatus further includes:

a third determining unit 505, configured to determine, in response to a mode selection instruction from a query-level knob tuning mode, a workload-level knob tuning mode, and a cluster-level knob tuning mode, a target knob tuning mode used for the N query statements obtained by the obtaining unit 501.

[00165] In a possible implementation, the vectorization coding unit 503 is further configured

to: when the target knob tuning mode is the workload-level knob tuning mode, fuse the N target feature vectors, to obtain a unified vector obtained through fusion; and

the second determining unit 504 is configured to determine, based on the unified vector and the current status information of the database, the configuration knob set that is of the database and that corresponds to the N query statements, where the configuration knob set includes one target configuration knob group, and the one target configuration knob group is used to tune, when the N query statements are executed, a configuration knob currently used for the database.

[00166] In a possible implementation, the second determining unit 504 is configured to: when the target knob tuning mode is the cluster-level knob tuning mode, determine N configuration knob groups based on the N target feature vectors and the current status information of the database, where the N configuration knob groups are in a one-to-one correspondence with the N target feature vectors; and perform clustering on the N configuration knob groups, to obtain Q target configuration knob groups, where a first configuration knob group corresponds to at least one target feature vector, the first configuration knob group is any one of the Q target configuration knob groups, Q is an integer greater than 0, $Q < N$, and the first configuration knob group is used to tune, when a query statement corresponding to the at least one target feature vector is executed, a configuration knob currently used for the database.

[00167] In a possible implementation, first query information in the N pieces of query information includes a first query type, first table information, and a type and corresponding cost information of an involved operator in a query, and the first query information is any one of the N pieces of query information;

the first query type is used to indicate a type of an operation that is requested by a query statement corresponding to the first query information and that is to be performed on the database;

the first table information is used to indicate a relationship table involved when the query statement corresponding to the first query information is executed in the database; and

the cost information is used to indicate a respective execution cost that is of the involved operator and that is incurred when the involved operator executes the query statement corresponding to the first query information.

[00168] In a possible implementation, a target feature vector corresponding to the first query information includes an identifier of the first query type, an identifier of a second query type, an identifier of the first table information, an identifier of second table information, and a floating-

point bit of each type of operator;

the identifier of the first query type is represented by using a first value, the identifier of the second query type is represented by using a second value, and the second query type belongs to a query type of the database and is not included in the first query information;

5 the identifier of the first table information is represented by using a third value, the identifier of the second table information is represented by using a fourth value, and the second table information belongs to table information of the database and is not included in the first query information; and

10 the floating-point bit of each type of operator includes an execution cost of the corresponding type of operator.

[00169] In a possible implementation, the execution cost of the corresponding type of operator is a normalized execution cost.

15 **[00170]** In a possible implementation, the second determining unit 504 is configured to determine, based on the N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (DS-DRL) model, the configuration knob set that is of the database and that corresponds to the N query statements.

[00171] In a possible implementation, when the DS-DRL model includes a prediction model, an environment model, and an agent model, the second determining unit 504 is configured to:

20 predict a status information change amount ΔS of the database based on the N target feature vectors by using the prediction model, where ΔS is an amount of a change from status information that is of the database and that exists before the N query statements are executed to status information that is of the database and that is obtained after the N query statements are executed;

25 determine, by using the environment model and based on ΔS and the status information S that is of the database and that exists before the N query statements are executed, simulated status information S' that is of the database and that is obtained after the N query statements are executed in a simulated manner;

determine a recommended configuration knob set based on the simulated status information S' by using the agent model;

30 perform simulated configuration based on the recommended configuration knob set by using the environment model, execute the N target feature vectors, to update S and S', and

determine a representation value R obtained after the N target feature vectors are executed;

update a knob tuning policy based on the representation value R by using the agent model, and determine an updated recommended configuration knob set based on the updated simulated status information S'; and

5 if the representation value R meets a preset condition, determine that the updated recommended configuration knob set is the configuration knob set; or

if the representation value R does not meet a preset condition, repeatedly perform the foregoing process in which the environment model updates S and S' and the agent model determines the updated recommended configuration knob set, until the representation value R
10 meets the preset condition, where an updated recommended configuration knob set corresponding to a value R that meets the preset condition is the configuration knob set.

[00172] In a possible implementation, when the DS-DRL model is a double-state deep deterministic policy gradient (DS-DDPG) model, and the agent model includes an actor (actor) model and a score (critic) model, the second determining unit 504 is configured to:

15 determine the recommended configuration knob set based on the simulated status information S' by using the actor model; and

determine a score of the recommended configuration knob set based on the simulated status information S' and the recommended configuration knob set by using the critic model.

[00173] In a possible implementation, the second determining unit 504 is configured to update
20 a weight of the actor model based on the score by using the actor model.

[00174] In a possible implementation, the second determining unit 504 is configured to update a weight of the critic model based on the representation value R by using the critic model.

[00175] Corresponding content of the apparatus 50 for determining a configuration knob of a database may be understood with reference to corresponding content of the embodiments
25 corresponding to FIG. 8 to FIG. 15. Details are not described herein again.

[00176] In a hardware implementation, the apparatus for determining a configuration knob of a database may be a database server. The obtaining unit 501, the first determining unit 502, the vectorization coding unit 503, the second determining unit 504, and the knob tuning unit each may be a processor. Certainly, the obtaining unit 501 may also be a communications interface.

30 **[00177]** FIG. 17 is a schematic diagram of a possible logical structure of a database server 60 involved in the foregoing embodiments according to an embodiment of this application. The

database server 60 includes a processor 601, a communications interface 602, a memory 603, and a bus 604. The processor 601, the communications interface 602, and the memory 603 are connected to each other by using the bus 604. In this embodiment of this application, the processor 601 is configured to control and manage an action of the database server 60. For example, the processor 601 is configured to perform steps 201 to 205 in FIG. 9 and/or another process of a technology described in this specification. The communications interface 602 is configured to support the database server 60 to perform communication. The memory 603 is configured to store program code and data of the database server 60.

[00178] The processor 601 may be a central processing unit, a general-purpose processor, a digital signal processor, an application-specific integrated circuit, a field programmable gate array or another programmable logic device, a transistor logic device, a hardware component, or any combination thereof. The processor 601 may implement or execute various example logical blocks, modules, and circuits described with reference to content disclosed in this application. Alternatively, the processor may be a combination of processors implementing a computing function, for example, a combination of one or more microprocessors, or a combination of a digital signal processor and a microprocessor. The bus 604 may be a peripheral component interconnect (peripheral component interconnect, PCI) bus, an extended industry standard architecture (extended industry standard architecture, EISA) bus, or the like. The bus may be classified into an address bus, a data bus, a control bus, and the like. For ease of representation, only one thick line is used to represent the bus in FIG. 17, but this does not mean that there is only one bus or only one type of bus.

[00179] In another embodiment of this application, a computer-readable storage medium is further provided. The computer-readable storage medium stores computer-executable instructions. When at least one processor of a device executes the computer-executable instructions, the device performs the method for determining a configuration knob of a database described in some embodiments in FIG. 8 to FIG. 15.

[00180] In another embodiment of this application, a computer program product is further provided. The computer program product includes computer-executable instructions, and the computer-executable instructions are stored in a computer-readable storage medium. At least one processor of a device may read the computer-executable instructions from the computer-readable storage medium, and the at least one processor executes the computer-executable instructions, so

that the device performs the method for determining a configuration knob of a database described in the embodiments in FIG. 8 to FIG. 15.

5 [00181] A person of ordinary skill in the art may be aware that, in combination with the examples described in the embodiments disclosed in this specification, units and algorithm steps can be implemented by electronic hardware or a combination of computer software and electronic hardware. Whether the functions are performed by hardware or software depends on particular applications and design constraint conditions of the technical solutions. A person skilled in the art may use different methods to implement the described functions for each particular application, but it should not be considered that the implementation goes beyond the scope of the embodiments
10 of this application.

[00182] It may be clearly understood by a person skilled in the art that, for the purpose of convenient and brief description, for a detailed working process of the foregoing system, apparatus, and unit, refer to a corresponding process in the foregoing method embodiments. Details are not described herein again.

15 [00183] In the several embodiments provided in this application, it should be understood that the disclosed system, apparatus, and method may be implemented in other manners. For example, the described apparatus embodiment is merely an example. For example, division into the units is merely logical function division, and there may be another division manner in an actual implementation. For example, a plurality of units or components may be combined or integrated
20 into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual couplings or direct couplings or communication connections may be implemented through some interfaces. The indirect couplings or communication connections between the apparatuses or units may be implemented in electronic, mechanical, or other forms.

[00184] The units described as separate parts may or may not be physically separate, and parts
25 displayed as units may or may not be physical units, may be located in one location, or may be distributed on a plurality of network units. Some or all of the units may be selected based on an actual requirement to achieve the objectives of the solutions of the embodiments.

[00185] In addition, function units in the embodiments of this application may be integrated into one processing unit, or each of the units may exist alone physically, or two or more units are
30 integrated into one unit.

[00186] When the functions are implemented in the form of a software function unit and sold

or used as an independent product, the functions may be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of this application essentially, or the part contributing to the prior art, or some of the technical solutions may be implemented in a form of a software product. The computer software product is stored in a storage medium, and
5 includes several instructions for instructing a computer device (which may be a personal computer, a server, or a network device) to perform all or some of the steps of the methods described in the embodiments of this application. The foregoing storage medium includes various media that can store program code, such as a USB flash drive, a removable hard disk, a read-only memory (read-only memory, ROM), a random access memory (random access memory, RAM), a magnetic disk,
10 or an optical disc.

[00187] The foregoing descriptions are merely specific implementations of this application, but are not intended to limit the protection scope of this application. Any variation or replacement readily figured out by a person skilled in the art within the technical scope disclosed in this application shall fall within the protection scope of this application. Therefore, the protection scope
15 of the embodiments of this application shall be subject to the protection scope of the claims.

CLAIMS

What is claimed is:

1. A method for determining a configuration knob of a database, comprising:

obtaining N query statements sent by M clients, wherein M is an integer greater than 0, N is
5 an integer greater than 0, and $N \geq M$;

determining N pieces of query information from the N query statements, wherein the N query
statements are in a one-to-one correspondence with the N pieces of query information;

performing vectorization coding on the N pieces of query information, to obtain
corresponding N target feature vectors, wherein the N pieces of query information are in a one-to-
10 one correspondence with the N target feature vectors; and

determining, based on the N target feature vectors and current status information of the
database, a configuration knob set that is of the database and that corresponds to the N query
statements, wherein the configuration knob set comprises at least one target configuration knob
group, the target configuration knob group comprises at least one configuration knob, and the
15 configuration knob set is used to execute the N query statements.

2. The method according to claim 1, wherein the method further comprises:

determining, in response to a mode selection instruction from a query-level knob tuning mode,
a workload-level knob tuning mode, and a cluster-level knob tuning mode, a target knob tuning
mode used for the N query statements.

3. The method according to claim 2, wherein when the target knob tuning mode is the
workload-level knob tuning mode, the method further comprises:

fusing the N target feature vectors, to obtain a unified vector obtained after fusion; and

the determining, based on the N target feature vectors and current status information of the
database, a configuration knob set that is of the database and that corresponds to the N query
statements comprises:
25

determining, based on the unified vector and the current status information of the database,
the configuration knob set that is of the database and that corresponds to the N query statements,
wherein the configuration knob set comprises one target configuration knob group, and the one
target configuration knob group is used to tune, when the N query statements are executed, a
30 configuration knob currently used for the database.

4. The method according to claim 2, wherein when the target knob tuning mode is the cluster-level knob tuning mode, the determining, based on the N target feature vectors and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements comprises:

5 determining N configuration knob groups based on the N target feature vectors and the current status information of the database, wherein the N configuration knob groups are in a one-to-one correspondence with the N target feature vectors; and

performing clustering on the N configuration knob groups, to obtain Q target configuration knob groups, wherein a first configuration knob group corresponds to at least one target feature vector, the first configuration knob group is any one of the Q target configuration knob groups, Q is an integer greater than 0, $Q < N$, and the first configuration knob group is used to tune, when a query statement corresponding to the at least one target feature vector is executed, a configuration knob currently used for the database.

5. The method according to any one of claims 1 to 4, wherein first query information in the N pieces of query information comprises a first query type, first table information, and a type and corresponding cost information of an involved operator in a query, and the first query information is any one of the N pieces of query information;

the first query type is used to indicate a type of an operation that is requested by a query statement corresponding to the first query information and that is to be performed on the database;

20 the first table information is used to indicate a relationship table involved when the query statement corresponding to the first query information is executed in the database; and

the cost information is used to indicate a respective execution cost that is of the involved operator and that is incurred when the involved operator executes the query statement corresponding to the first query information.

25 6. The method according to claim 5, wherein

a target feature vector corresponding to the first query information comprises an identifier of the first query type, an identifier of a second query type, an identifier of the first table information, an identifier of second table information, and a floating-point bit of each type of operator;

30 the identifier of the first query type is represented by using a first value, the identifier of the second query type is represented by using a second value, and the second query type belongs to a query type of the database and is not comprised in the first query information;

the identifier of the first table information is represented by using a third value, the identifier of the second table information is represented by using a fourth value, and the second table information belongs to table information of the database and is not comprised in the first query information; and

5 the floating-point bit of each type of operator comprises an execution cost of the corresponding type of operator.

7. The method according to claim 6, wherein

the execution cost of the corresponding type of operator is a normalized execution cost.

8. The method according to any one of claims 1 to 7, wherein the determining, based on the
10 N target feature vectors and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements comprises:

determining, based on the N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (DS-DRL) model, the configuration knob set that is of the database and that corresponds to the N query statements.

15 9. The method according to claim 8, wherein the DS-DRL model comprises a prediction model, an environment model, and an agent model;

the determining, based on the N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (DS-DRL) model, the configuration knob set that is of the database and that corresponds to the N query statements
20 comprises:

predicting a status information change amount ΔS of the database based on the N target feature vectors by using the prediction model, wherein ΔS is an amount of a change from status information that is of the database and that exists before the N query statements are executed to status information that is of the database and that is obtained after the N query statements are
25 executed;

determining, by using the environment model and based on ΔS and the status information S that is of the database and that exists before the N query statements are executed, simulated status information S' that is of the database and that is obtained after the N query statements are executed in a simulated manner;

30 determining a recommended configuration knob set based on the simulated status information S' by using the agent model;

performing simulated configuration based on the recommended configuration knob set by using the environment model, executing the N target feature vectors, to update S and S', and determining a representation value R obtained after the N target feature vectors are executed;

5 updating a knob tuning policy based on the representation value R by using the agent model, and determining an updated recommended configuration knob set based on the updated simulated status information S'; and

if the representation value R meets a preset condition, determining that the updated recommended configuration knob set is the configuration knob set; or

10 if the representation value R does not meet a preset condition, repeatedly performing the foregoing process in which the environment model updates S and S' and the agent model determines the updated recommended configuration knob set, until the representation value R meets the preset condition, wherein an updated recommended configuration knob set corresponding to a value R that meets the preset condition is the configuration knob set.

15 10. The method according to claim 9, wherein when the DS-DRL model is a double-state deep deterministic policy gradient (DS-DDPG) model, the agent model comprises an actor (actor) model and a score (critic) model; and

the determining a recommended configuration knob set based on the simulated status information S' by using the agent model comprises:

20 determining the recommended configuration knob set based on the simulated status information S' by using the actor model; and

determining a score of the recommended configuration knob set based on the simulated status information S' and the recommended configuration knob set by using the critic model.

11. The method according to claim 10, wherein the updating a knob tuning policy based on the representation value R comprises:

25 updating a weight of the actor model based on the score by using the actor model.

12. The method according to claim 10 or 11, wherein the updating a knob tuning policy based on the representation value R comprises:

updating a weight of the critic model based on the representation value R by using the critic model.

30 13. An apparatus for determining a configuration knob of a database, comprising:

an obtaining unit, configured to obtain N query statements sent by M clients, wherein M is

an integer greater than 0, N is an integer greater than 0, and $N \geq M$;

a first determining unit, configured to determine N pieces of query information from the N query statements obtained by the obtaining unit, wherein the N query statements are in a one-to-one correspondence with the N pieces of query information;

5 a vectorization coding unit, configured to perform vectorization coding on the N pieces of query information determined by the first determining unit, to obtain corresponding N target feature vectors, wherein the N pieces of query information are in a one-to-one correspondence with the N target feature vectors; and

a second determining unit, configured to determine, based on the N target feature vectors
10 obtained by the vectorization coding unit and current status information of the database, a configuration knob set that is of the database and that corresponds to the N query statements, wherein the configuration knob set comprises at least one target configuration knob group, the target configuration knob group comprises at least one configuration knob, and the configuration knob set is used to execute the N query statements.

15 14. The apparatus according to claim 13, wherein the apparatus further comprises:

a third determining unit, configured to determine, in response to a mode selection instruction from a query-level knob tuning mode, a workload-level knob tuning mode, and a cluster-level knob tuning mode, a target knob tuning mode used for the N query statements obtained by the obtaining unit.

20 15. The apparatus according to claim 14, wherein

the vectorization coding unit is further configured to: when the target knob tuning mode is the workload-level knob tuning mode, fuse the N target feature vectors, to obtain a unified vector obtained through fusion; and

the second determining unit is configured to determine, based on the unified vector and the
25 current status information of the database, the configuration knob set that is of the database and that corresponds to the N query statements, wherein the configuration knob set comprises one target configuration knob group, and the one target configuration knob group is used to tune, when the N query statements are executed, a configuration knob currently used for the database.

16. The apparatus according to claim 14, wherein

30 the second determining unit is configured to: when the target knob tuning mode is the cluster-level knob tuning mode, determine N configuration knob groups based on the N target feature

vectors and the current status information of the database, wherein the N configuration knob groups are in a one-to-one correspondence with the N target feature vectors; and perform clustering on the N configuration knob groups, to obtain Q target configuration knob groups, wherein a first configuration knob group corresponds to at least one target feature vector, the first configuration knob group is any one of the Q target configuration knob groups, Q is an integer greater than 0, $Q < N$, and the first configuration knob group is used to tune, when a query statement corresponding to the at least one target feature vector is executed, a configuration knob currently used for the database.

17. The apparatus according to any one of claims 13 to 16, wherein first query information in the N pieces of query information comprises a first query type, first table information, and a type and corresponding cost information of an involved operator in a query, and the first query information is any one of the N pieces of query information;

the first query type is used to indicate a type of an operation that is requested by a query statement corresponding to the first query information and that is to be performed on the database;

the first table information is used to indicate a relationship table involved when the query statement corresponding to the first query information is executed in the database; and

the cost information is used to indicate a respective execution cost that is of the involved operator and that is incurred when the involved operator executes the query statement corresponding to the first query information.

18. The apparatus according to claim 17, wherein a target feature vector corresponding to the first query information comprises an identifier of the first query type, an identifier of a second query type, an identifier of the first table information, an identifier of second table information, and a floating-point bit of each type of operator;

the identifier of the first query type is represented by using a first value, the identifier of the second query type is represented by using a second value, and the second query type belongs to a query type of the database and is not comprised in the first query information;

the identifier of the first table information is represented by using a third value, the identifier of the second table information is represented by using a fourth value, and the second table information belongs to table information of the database and is not comprised in the first query information; and

the floating-point bit of each type of operator comprises an execution cost of the

corresponding type of operator.

19. The apparatus according to claim 18, wherein
the execution cost of the corresponding type of operator is a normalized execution cost.

20. The apparatus according to any one of claims 13 to 19, wherein:

5 the second determining unit is configured to determine, based on the N target feature vectors and the current status information of the database by using a double-state deep reinforcement learning (DS-DRL) model, the configuration knob set that is of the database and that corresponds to the N query statements.

21. The apparatus according to claim 20, wherein

10 when the DS-DRL model comprises a prediction model, an environment model, and an agent model, the second determining unit is configured to:

predict a status information change amount ΔS of the database based on the N target feature vectors by using the prediction model, wherein ΔS is an amount of a change from status information that is of the database and that exists before the N query statements are executed to status information that is of the database and that is obtained after the N query statements are executed;

15 determine, by using the environment model and based on ΔS and the status information S that is of the database and that exists before the N query statements are executed, simulated status information S' that is of the database and that is obtained after the N query statements are executed in a simulated manner;

determine a recommended configuration knob set based on the simulated status information S' by using the agent model;

20 perform simulated configuration based on the recommended configuration knob set by using the environment model, execute the N target feature vectors, to update S and S', and determine a representation value R obtained after the N target feature vectors are executed;

update a knob tuning policy based on the representation value R by using the agent model, and determine an updated recommended configuration knob set based on the updated simulated status information S'; and

30 if the representation value R meets a preset condition, determine that the updated recommended configuration knob set is the configuration knob set; or

if the representation value R does not meet a preset condition, repeatedly perform the

foregoing process in which the environment model updates S and S' and the agent model determines the updated recommended configuration knob set, until the representation value R meets the preset condition, wherein an updated recommended configuration knob set corresponding to a value R that meets the preset condition is the configuration knob set.

5 22. The apparatus according to claim 21, wherein

when the DS-DRL model is a double-state deep deterministic policy gradient (DS-DDPG) model, and the agent model comprises an actor (actor) model and a score (critic) model, the second determining unit is configured to:

10 determine the recommended configuration knob set based on the simulated status information S' by using the actor model; and

determine a score of the recommended configuration knob set based on the simulated status information S' and the recommended configuration knob set by using the critic model.

23. The apparatus according to claim 22, wherein

15 the second determining unit is configured to update a weight of the actor model based on the score by using the actor model.

24. The apparatus according to claim 22 or 23, wherein

the second determining unit is configured to update a weight of the critic model based on the representation value R by using the critic model.

20 25. An apparatus for determining a configuration knob of a database, wherein the apparatus comprises at least one processor, a memory, and instructions that are stored in the memory and that may be executed by the at least one processor, and the at least one processor executes the instructions, to implement steps of the method according to any one of claims 1 to 12.

25 26. A computer-readable storage medium, wherein the computer-readable medium stores a computer program; and when the program is executed by a processor, steps of the method according to any one of claims 1 to 12 are implemented.

27. A database system, comprising the apparatus for determining a configuration knob of a database according to any one of claims 13 to 26 and the database.

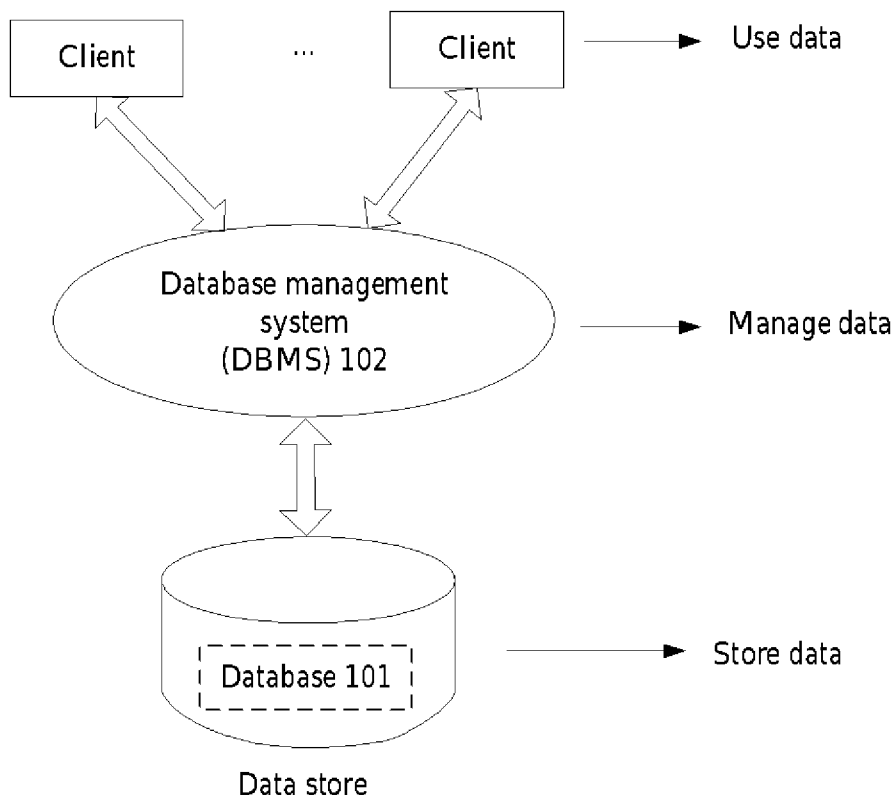


FIG. 1

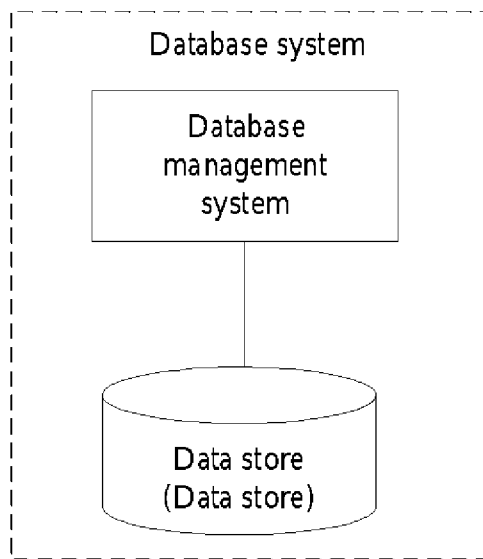


FIG. 2

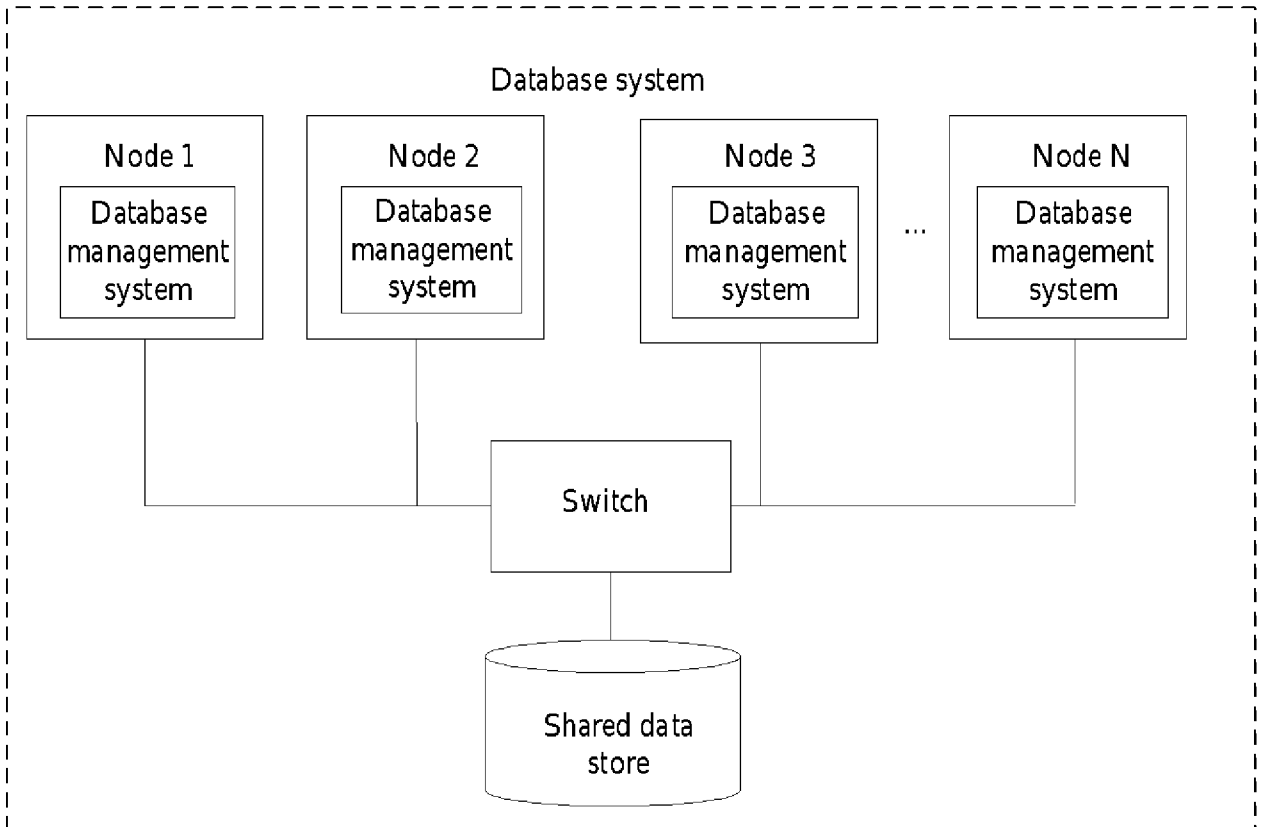


FIG. 3

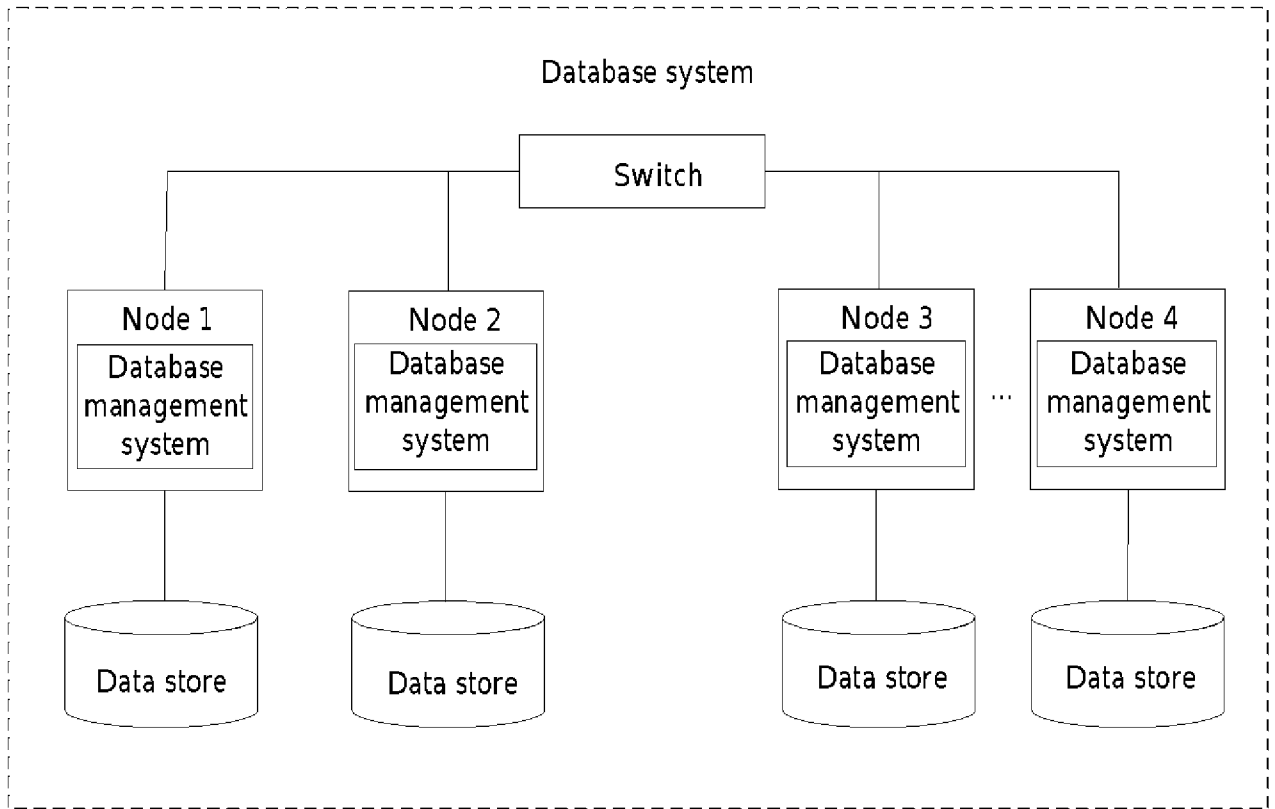


FIG. 4

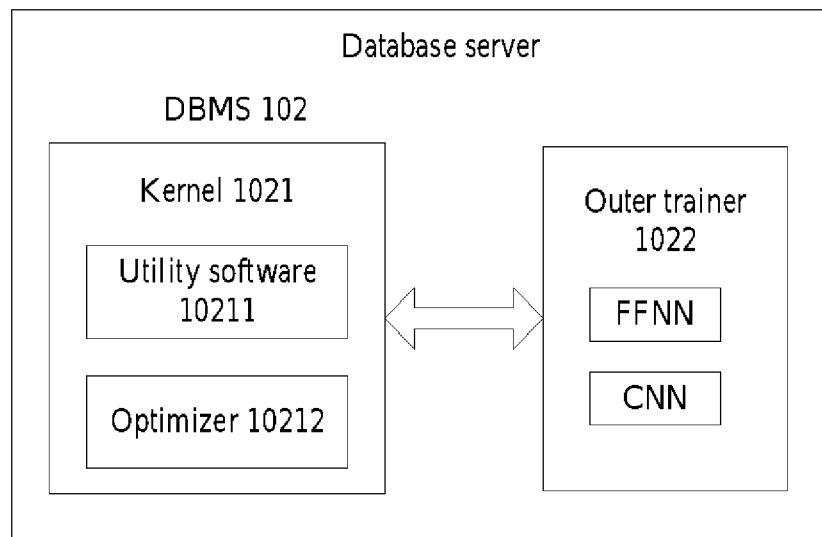


FIG. 5

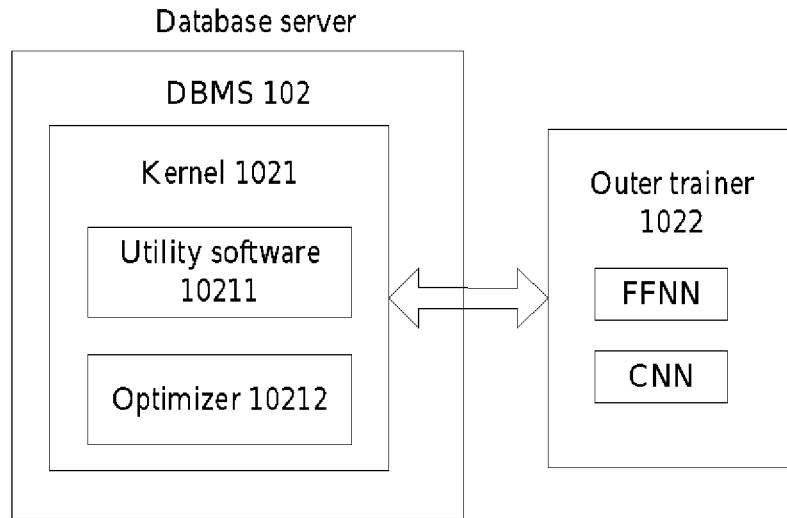


FIG. 6

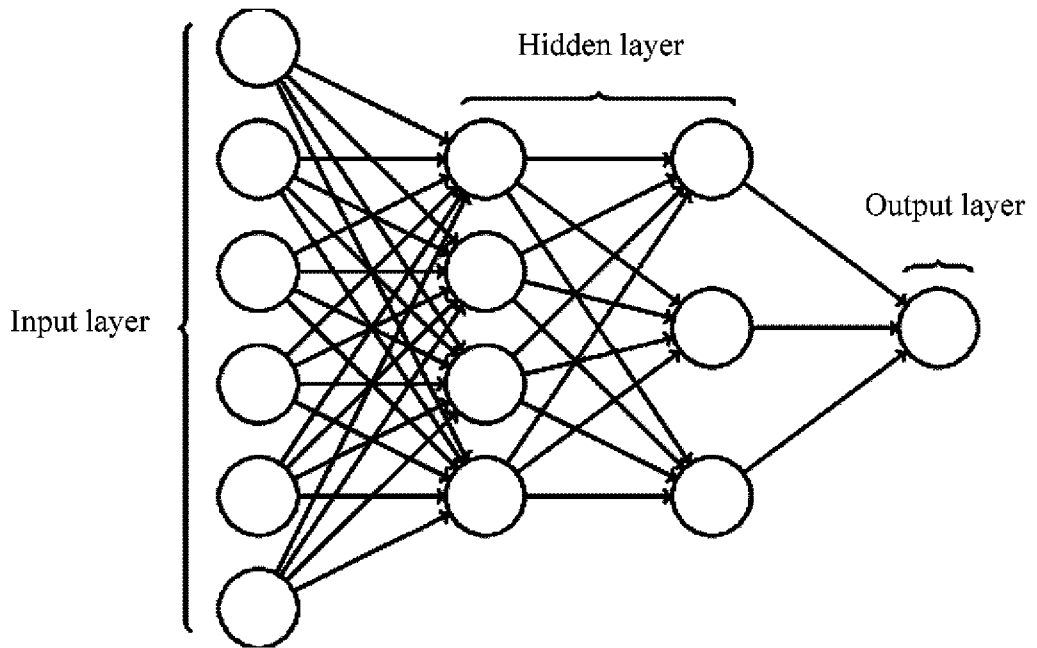


FIG. 7

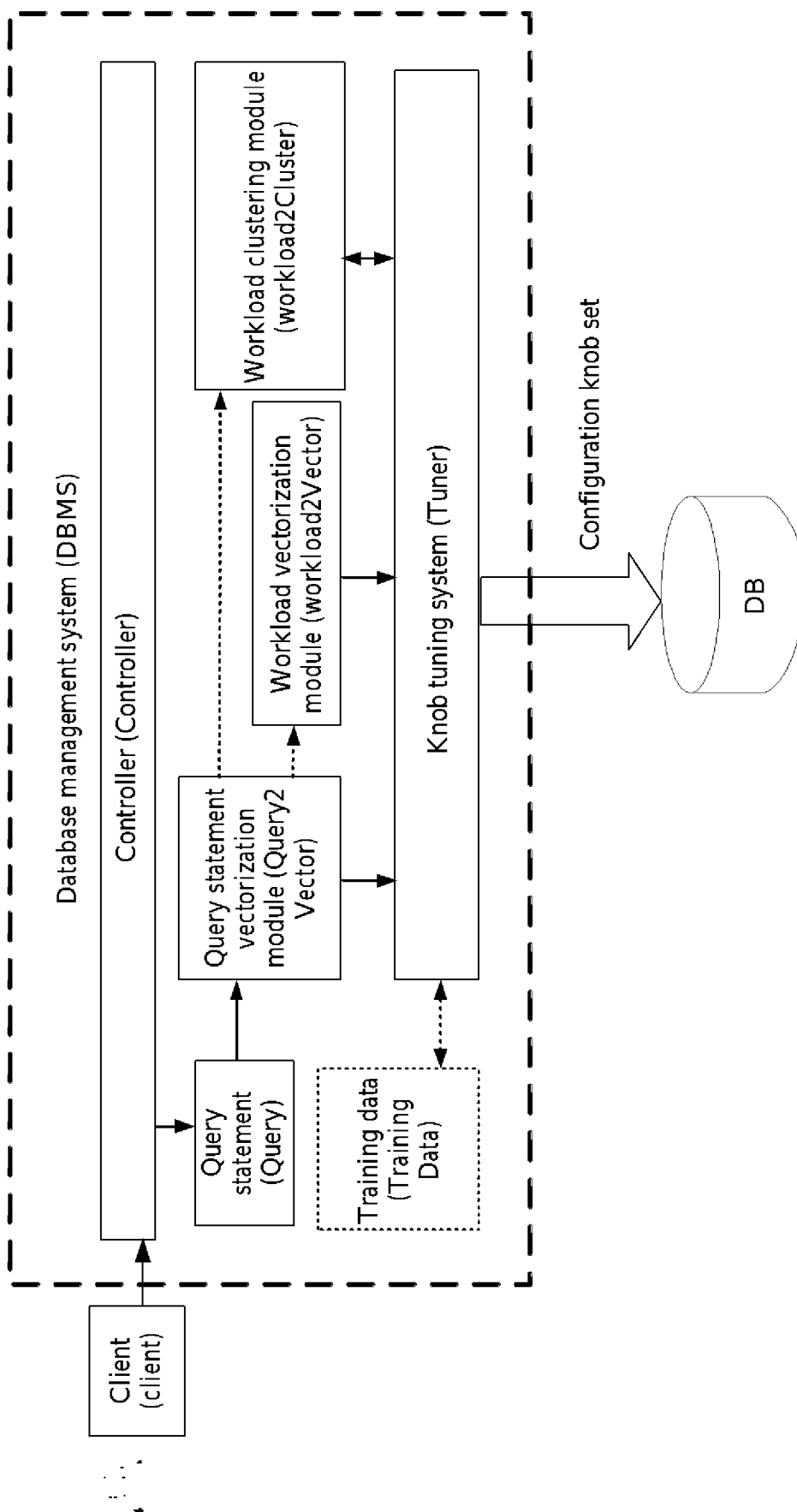


FIG. 8

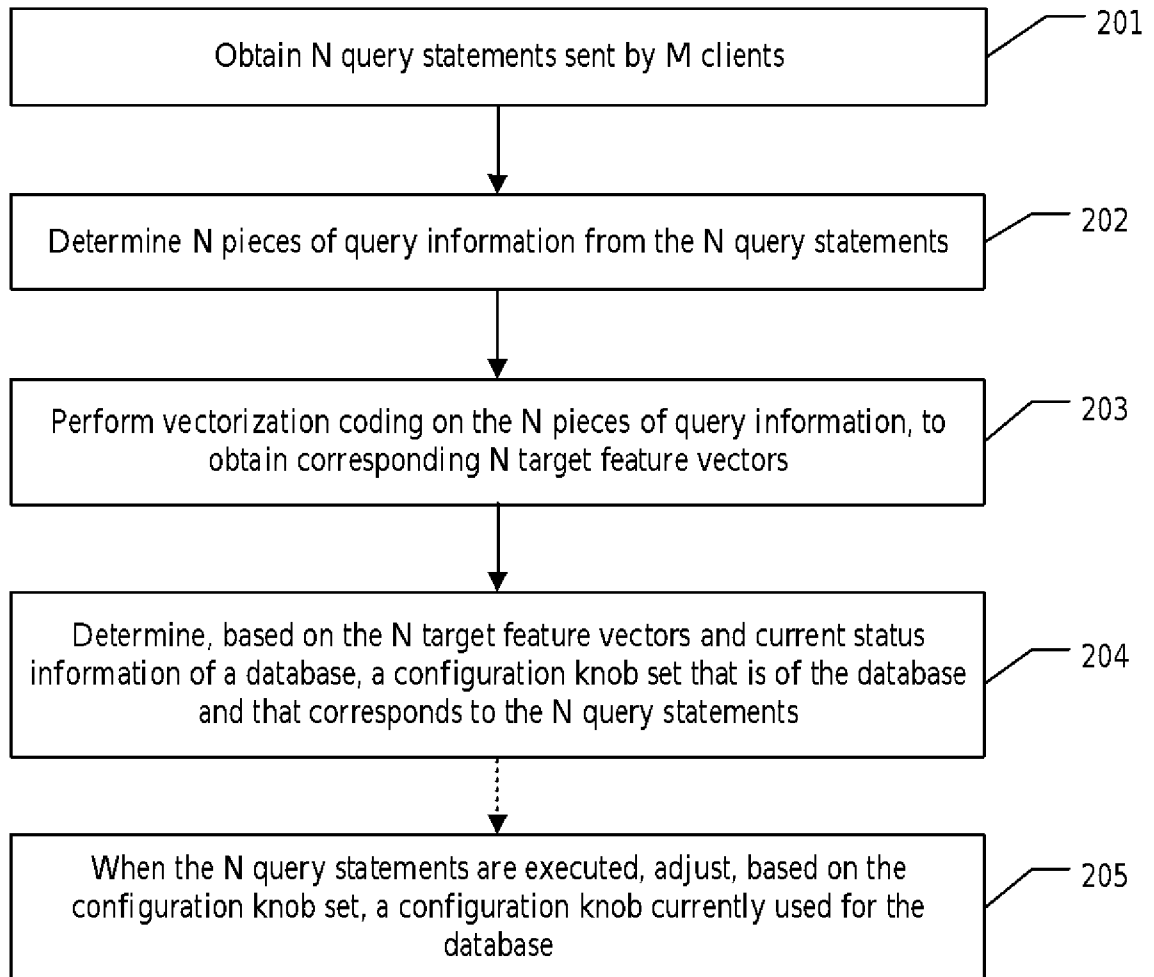


FIG. 9

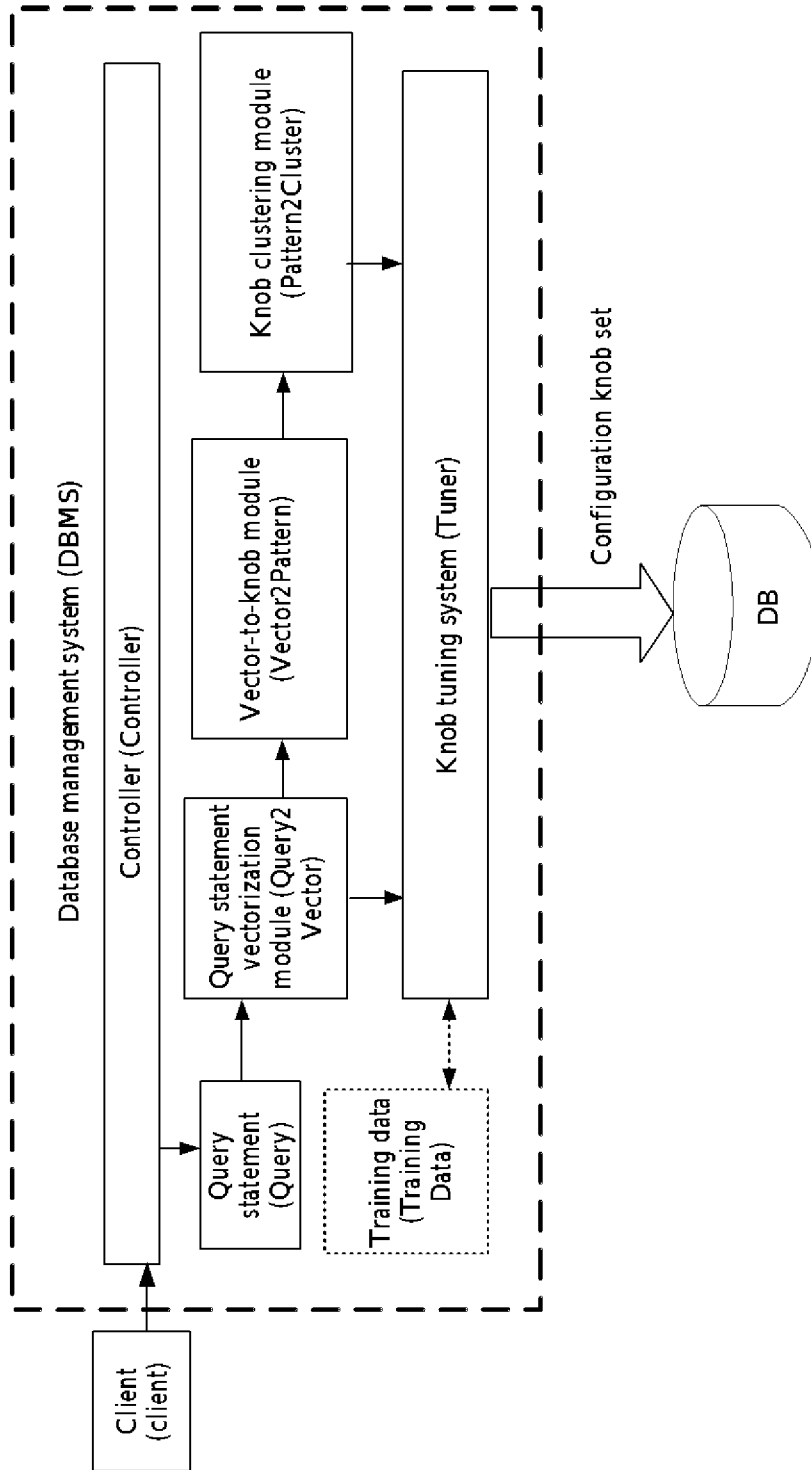
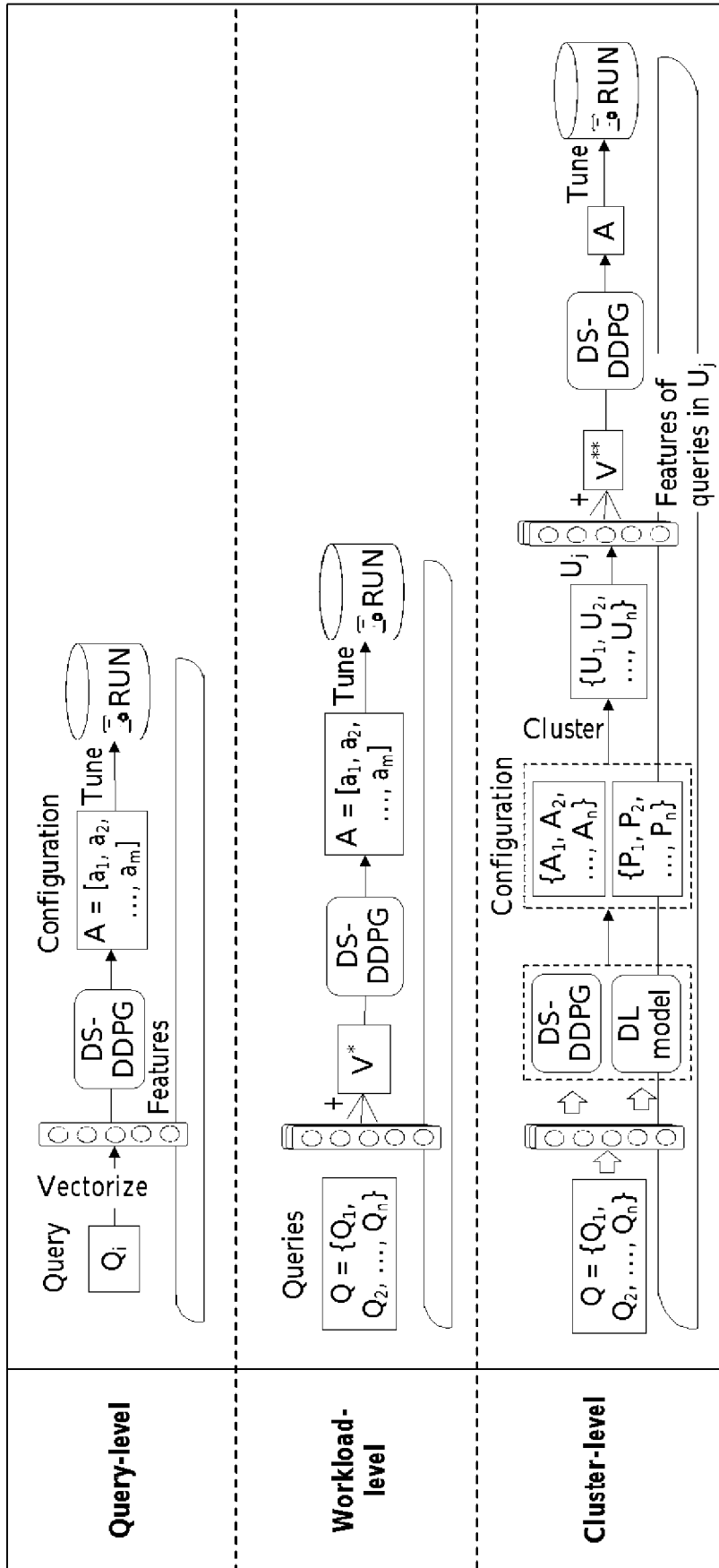


FIG. 10



TO
FIG. 11B

FIG. 11A

CONT.
FROM
FIG. 11A

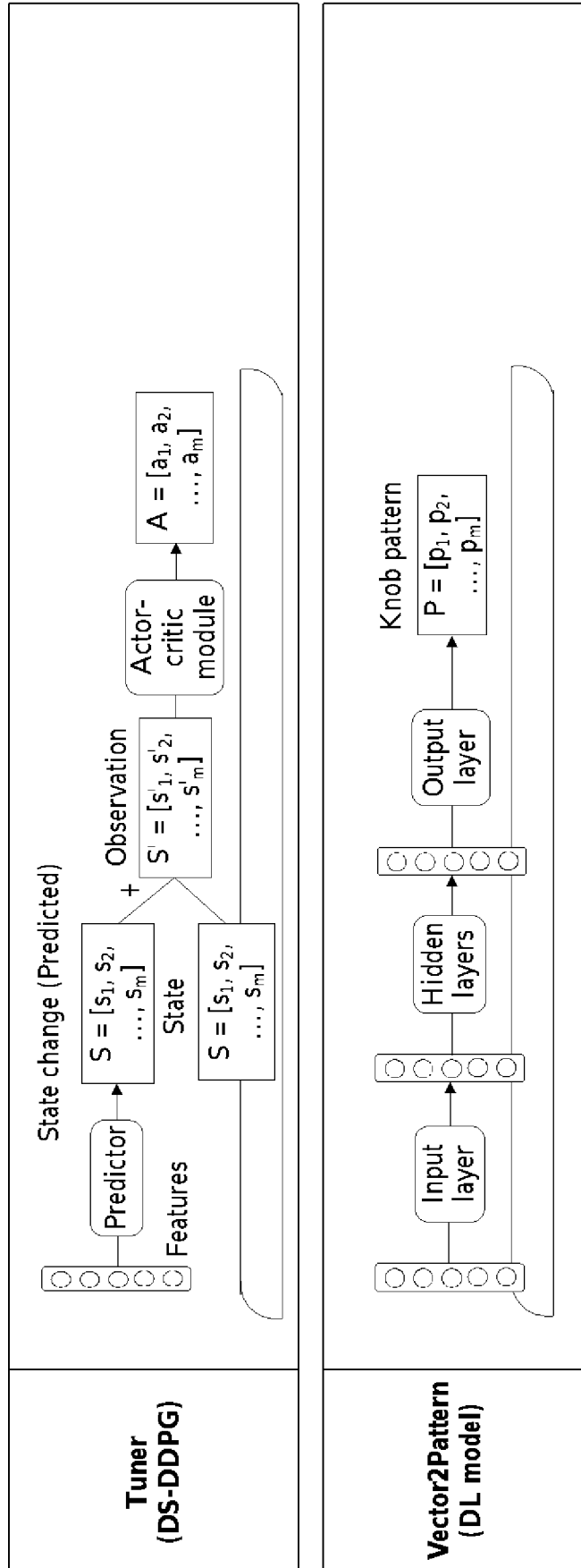


FIG. 11B

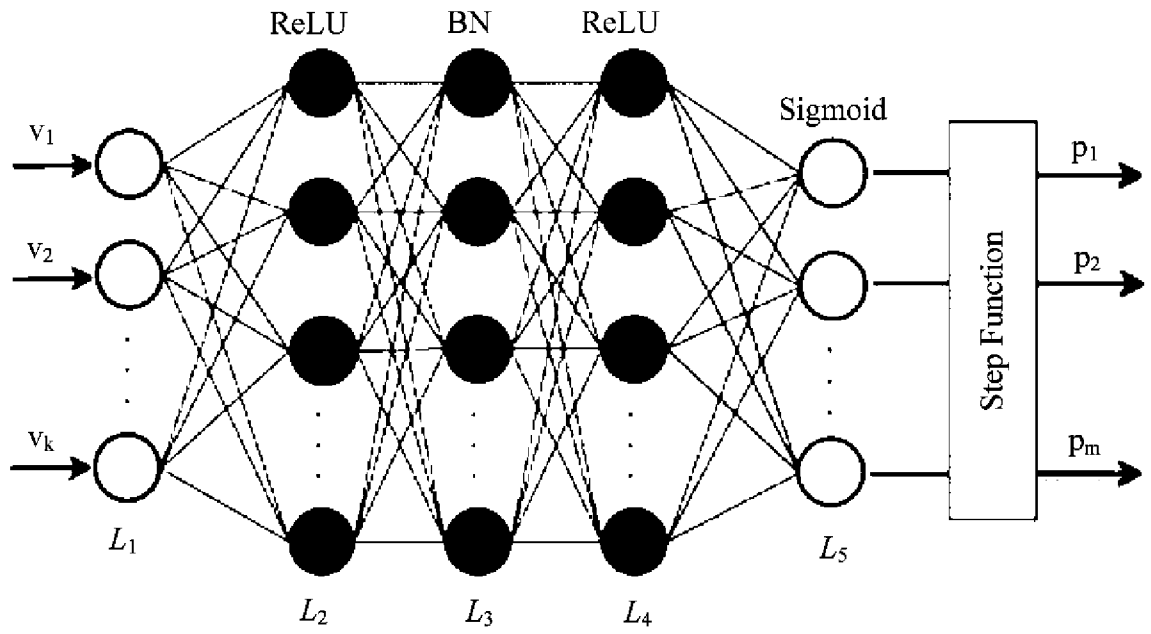
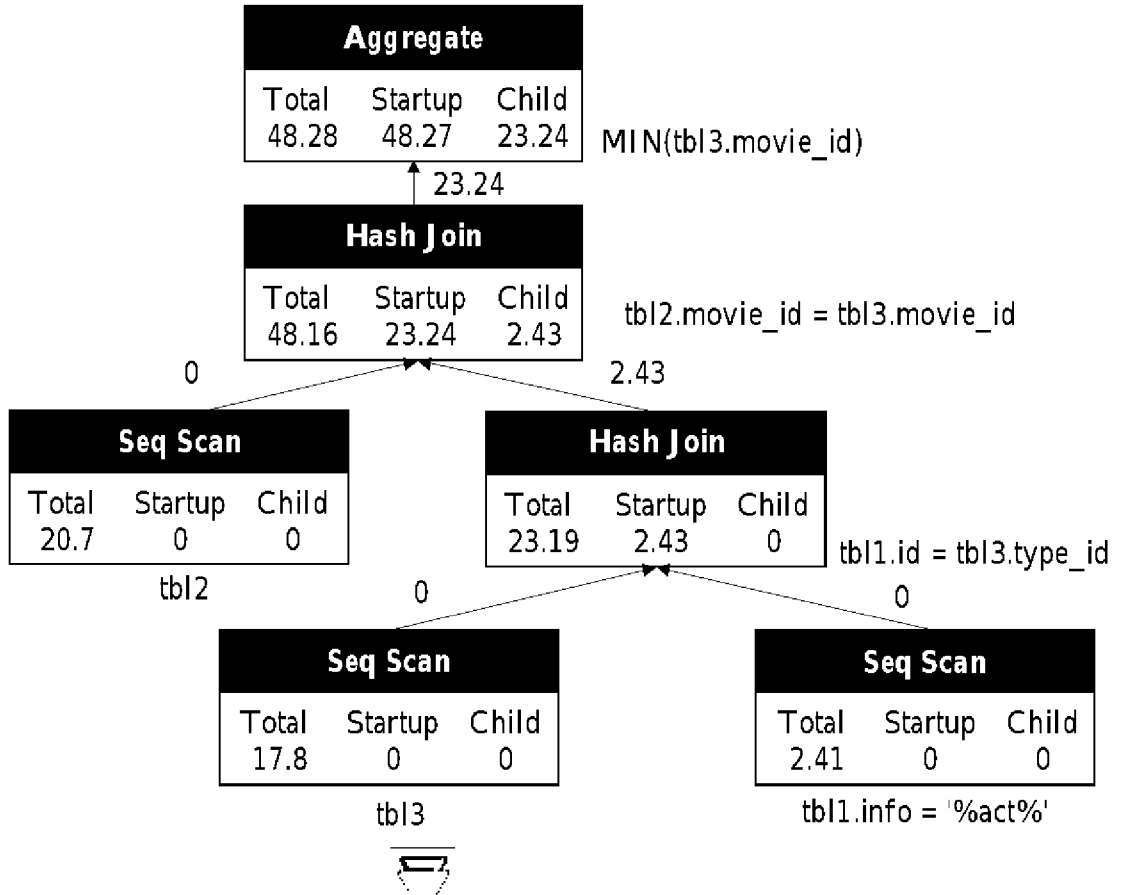


FIG. 12

```

SELECT
FROM
WHERE
AND
AND
MIN(tbl3.movie_id)
tbl1, tbl2, tbl3
tbl1.info = '%act%'
tbl1.id = tbl3.type_id
tbl2.movie_id = tbl3.movie_id
    
```



Insert	Delete	Update	Select	tbl1	tbl2	tbl3	...	tbl8	Hash_Join	Seq_Scan	Aggregate	...
[0	0	0	1	1	1	1	...	0	68.92	40.91	25.04	...
(1) DML			(2) Tables						(3) Operation Costs			

Insert	Delete	Update	Select	tbl1	tbl2	tbl3	...	tbl8	Hash_Join	Seq_Scan	Aggregate	...
[0	0	0	1	1	1	1	...	0	0.1401	-0.166	-0.2423	...

FIG. 13

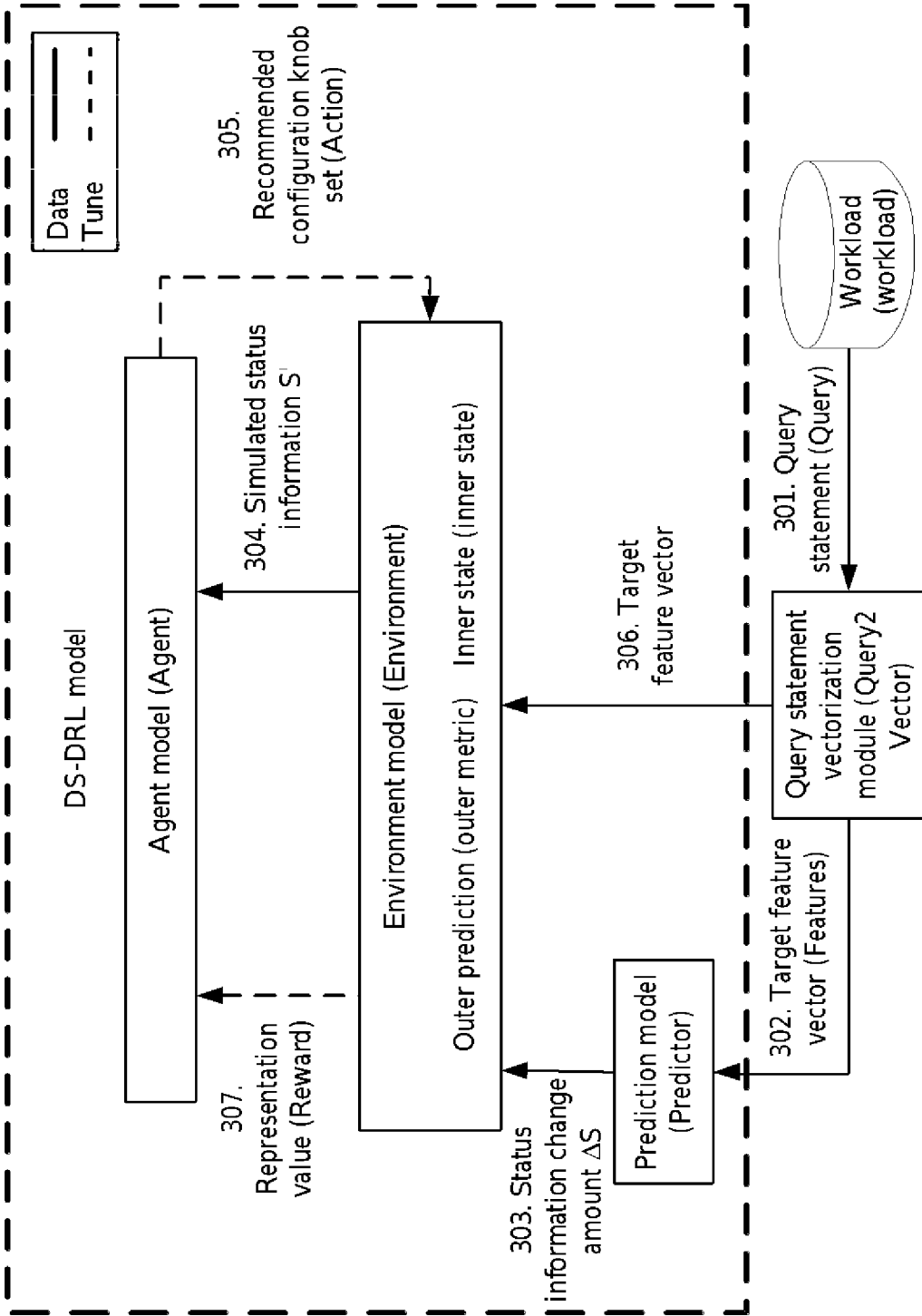


FIG. 14

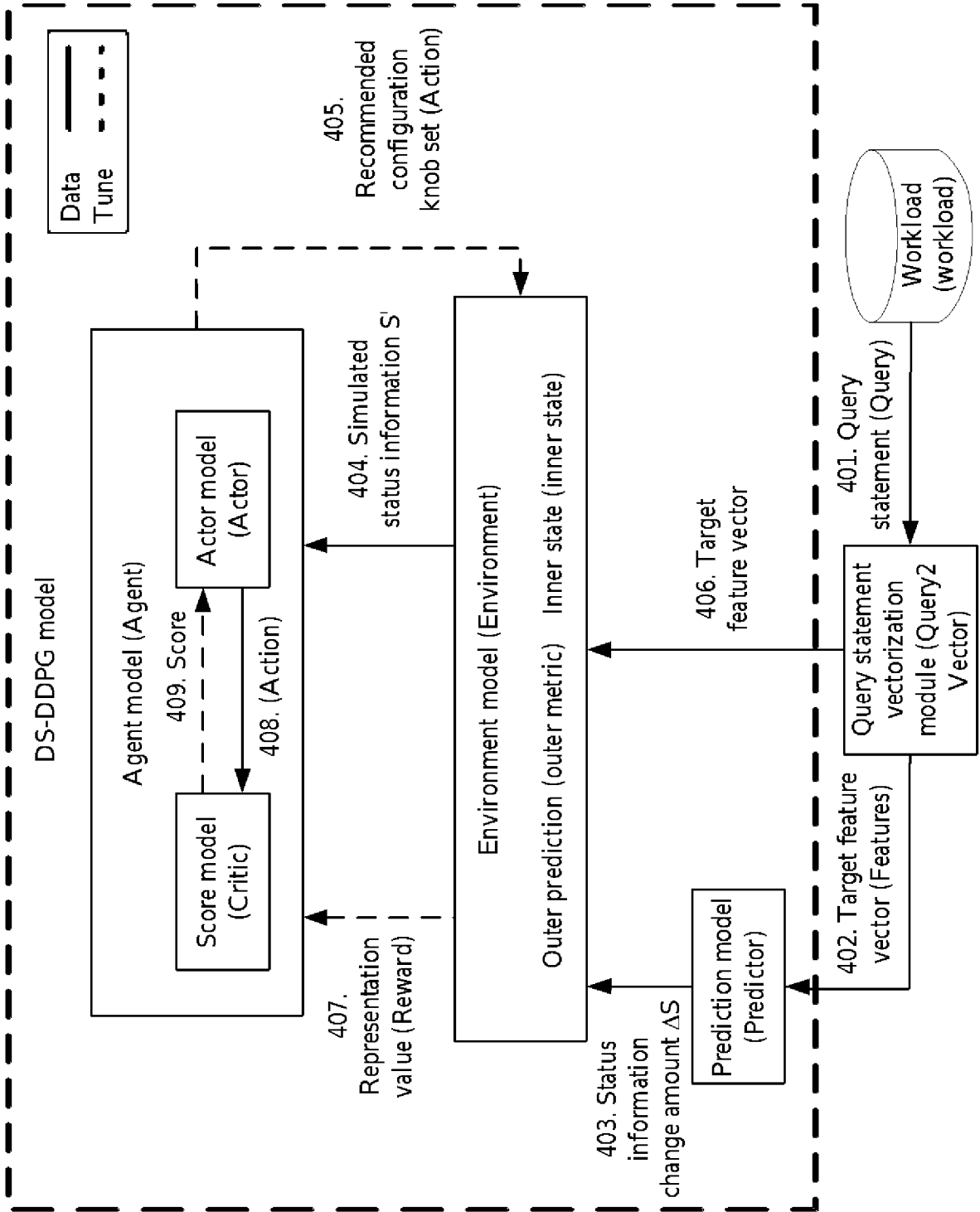


FIG. 15

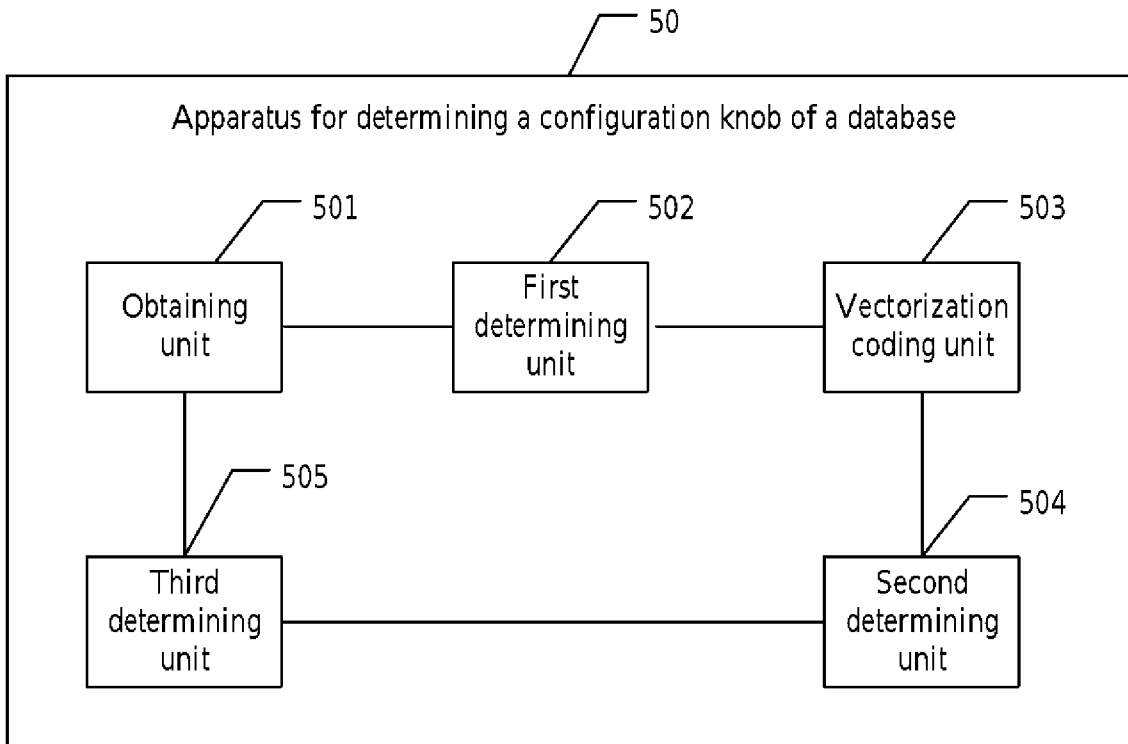


FIG. 16

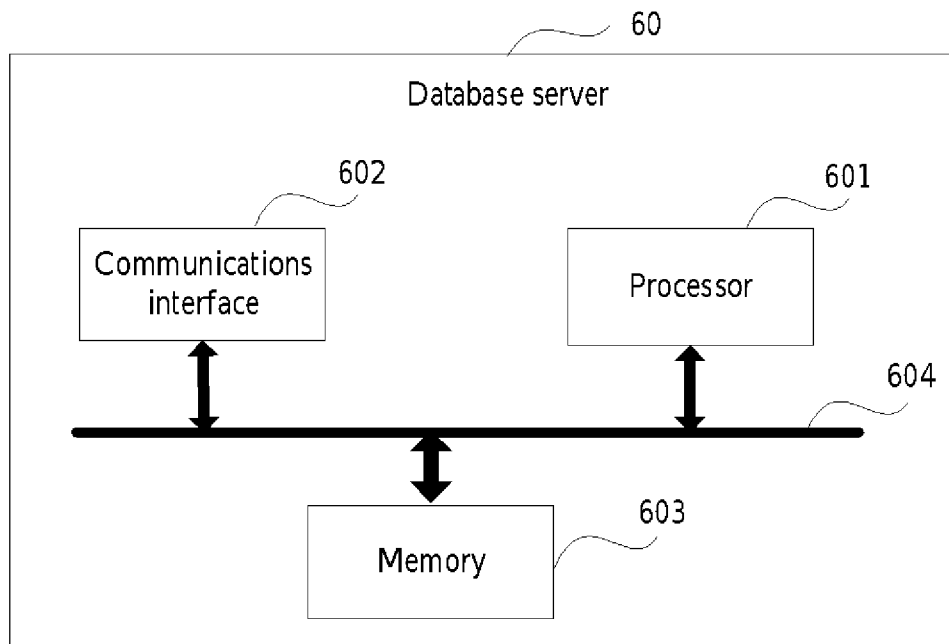


FIG. 17

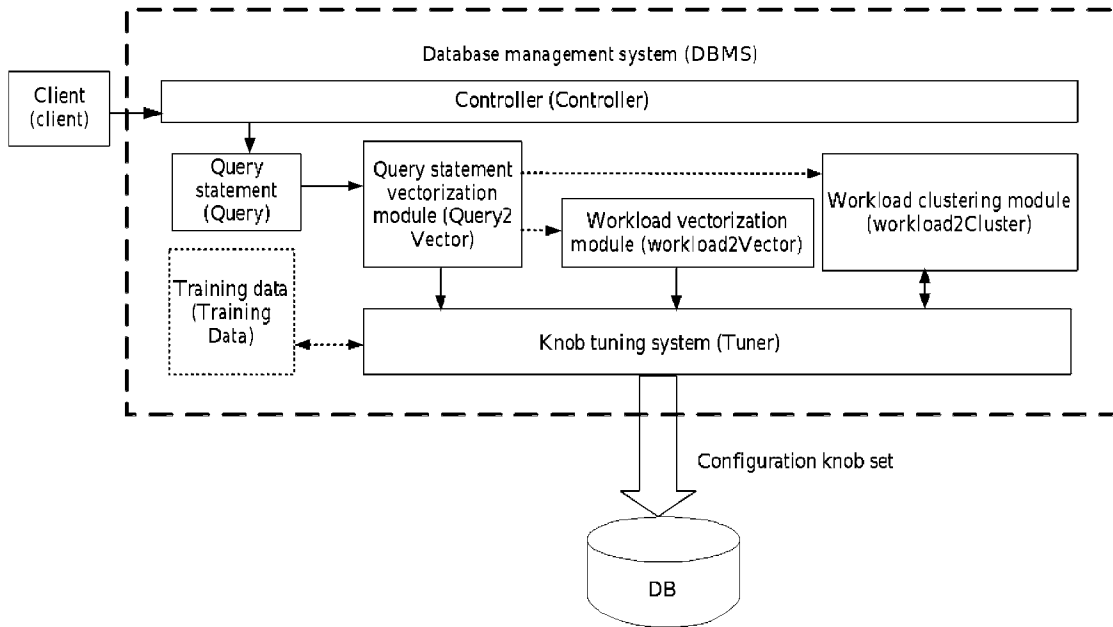


FIG. 8