CLAIMS:

1. A method, comprising:

performing, by one or more computers:

storing columnar data of a database table in a plurality of physical data blocks in a distributed data storage system on behalf of one or more clients, wherein the distributed data storage system comprises a cluster of one or more nodes, each of which comprises one or more disks on which physical data blocks are stored, and wherein each of the plurality of physical data blocks is associated with a respective unique identifier;

storing a copy of each of the plurality of physical data blocks in a remote key-value durable backup storage system, wherein for each of the plurality of physical data blocks, the respective unique identifier serves as a key to access the data block in the remote key-value durable backup storage system;

detecting a failure in the distributed data storage system affecting at least one of the plurality of physical data blocks in which the columnar data was stored;

in response to said detecting, automatically initiating a restore of the columnar data that was stored in the at least one of the plurality of physical data blocks from the remote key-value durable backup storage system; and

prior to restoring all of the columnar data that was stored in the at least one of the plurality of physical data blocks:

receiving one or more query requests directed to the columnar data of the database table; and

accepting and servicing the one or more query requests, wherein said servicing comprises obtaining at least some of the columnar data of the database table to which the one or more query requests are directed from the remote key-value durable backup storage system

using the respective unique identifiers as keys to access data blocks in the remote key-value durable backup storage system comprising the at least some of the columnar data.

2. The method of claim 1, wherein said storing columnar data of a database table comprises storing a portion of the columnar data as a primary copy of the portion of the columnar data in a respective physical data block on a given disk, and storing the portion of the columnar data as one or more secondary copies of the portion of the columnar data in respective physical data blocks on one or more disks other than the given disk.

3. The method of claim 1, wherein said storing columnar data of a database table comprises storing the unique identifier of each of the physical data blocks stored on disks of a given node in a respective entry in a superblock data structure that stores information about the physical data blocks stored on the given node.

4. The method of claim 1, wherein said storing columnar data of a database table comprises storing information indicating a location at which each of the physical data blocks stored on disks of a given node are stored in a respective entry in a superblock data structure that stores information about the physical data blocks stored on the given node.

5. The method of claim 1, wherein said servicing comprises obtaining at least some of the columnar data of the database table to which the one or more query requests are directed from disks in the distributed data storage system.

6. A method, comprising:

performing, by one or more computers:

maintaining data in one or more physical data blocks of a data storage system on behalf of one or more clients, wherein each physical data block is associated with a unique identifier;

performing a backup operation to store a respective copy of data stored in a given physical data block in a key-value storage system that is distinct from the data storage system;

subsequent to storing the respective copy of the data stored in the given physical data block, restoring the data stored in the given physical data block from the key-value storage system to the data storage system while accepting and servicing queries directed to the data maintained on behalf of the one or more clients, wherein said restoring comprises accessing the respective copy of data in the key-value storage system using the unique identifier associated with the given physical data block as a key in the key-value storage system.

7. The method of claim 6, wherein said restoring is performed as part of an operation to restore data stored in multiple physical data blocks from the key-value storage system to the data storage system, and wherein said accepting and servicing queries is performed prior to restoring all of the data stored in the multiple physical data blocks.

8. The method of claim 6, further comprising receiving data representing entries in a database table, wherein said maintaining data in one or more physical data blocks of a data storage system comprises storing data representing one or more columns of data in the database table in each of the one or more physical data blocks.

9. The method of claim 6, wherein said restoring is performed in response to a failure of a storage device on which the data stored in the given physical data block is stored in the data storage system.

10. The method of claim 6, wherein said restoring is performed in response to a failure of a node comprising a storage device on which the data stored in the given physical data block is stored in the data storage system.

11. The method of claim 6, wherein said restoring is performed in response to a failure of a cluster of one or more nodes comprising a storage device on which the data stored in the given physical data block is stored in the data storage system.

12. The method of claim 6, wherein said restoring is performed in response to an explicit request from one of the one or more clients to perform a restore operation.

13. The method of claim 6,

wherein said maintaining data in one or more physical data blocks of a data storage system comprises maintaining a data structure that stores a mapping between the data maintained on a particular node in a cluster of one or more nodes and a location in the data storage system at which the data is stored in a physical data block on the particular node, wherein the data structure is maintained on the particular node, and wherein each entry in the data structure stores a mapping for a particular physical data block and the unique identifier associated with the particular physical data block; and

wherein the method comprises, prior to said restoring, restoring the data structure on the particular node in response to a failure of the node or a failure of the cluster of nodes.

14. A non-transitory computer-readable storage medium storing program instructions that when executed on one or more computers cause the one or more computers to perform:

storing one or more data blocks in a data warehouse system;

creating a respective entry for each of the one or more data blocks in a data structure that stores information about data blocks stored in the data warehouse system, wherein each of the respective entries for the one or more data blocks comprises a unique identifier for the data block and an indication that the data block has not yet been backed up;

performing a backup operation for a plurality of data blocks stored in the data warehouse system, including the one or more data blocks, wherein said performing comprises:

storing a backup copy of the data structure in a remote key-value storage system;

storing, for each data block stored in the data warehouse system for which a corresponding entry in the data structure indicates that it has not yet been backed up, a backup copy of the data block in the remote key-value storage system; and

updating the entries in the data structure corresponding to each data block that was backed up by the backup operation to indicate that the data block has been backed up.

15. The non-transitory computer-readable storage medium of claim 14, wherein said storing one or more data blocks in a data warehouse system comprises storing two or more copies of each of the one or more data blocks on different storage devices in the data warehouse system.

16. The non-transitory computer-readable storage medium of claim 14, wherein said storing one or more data blocks in a data warehouse system comprises generating the unique identifier for each of the one or more data blocks.

17. The non-transitory computer-readable storage medium of claim 14, wherein when executed on one or more computers, the program instructions further cause the one or more computers to perform:

receiving a query directed to data in one of the one or more data blocks; and

in response to receiving the query, accessing the backup copy of the one of the one or more data blocks in the remote key-value storage system using the unique identifier for the one of the one or more blocks as an access key in the remote key-value storage system.

18. A computing system, comprising:

one or more computing nodes, each of which comprises at least one processor and a memory, wherein the one or more computing nodes are configured to collectively implement a database service; and

an interface to a remote key-value storage system;

wherein the database service is configured to maintain data on behalf of one or more subscribers to the database service;

wherein the one or more computing nodes are configured to store the data maintained on behalf of the one or more subscribers in a plurality of physical data blocks on one or more storage devices, wherein each of the plurality of physical data blocks is associated with a unique identifier;

wherein the database service is configured to perform a backup operation for the data maintained on behalf of the one or more subscribers, wherein to perform the backup operation, the database service is configured to send to the remote key-value storage system, via the interface, a copy of each of the plurality of physical data blocks for storage in the remote key-value storage system, and the unique identifiers associated with each of the plurality of physical data blocks to be used as access keys for copies of the plurality of physical data blocks in the remote key-value storage system.

19. The computing system of claim 18,

wherein the one or more computing nodes comprise a leader node that is configured to maintain one or more query maps and a mapping between the data maintained on behalf of the one or more subscribers and the locations at which the data is stored by the database service on the one or more computing nodes.

20. The computing system of claim 18,

wherein subsequent to performing the backup operation, the database service is configured to perform a restore operation;

wherein to perform the restore operation, the database service is configured to retrieve from the remote key-value storage system, via the interface, the copies of each of the plurality of physical data blocks, using the unique identifiers associated with each of the plurality of physical data blocks as access keys for the copies of the plurality of physical data blocks in the remote key-value storage system; and

wherein during performance of the restore operation, the database service is further configured to accept and service query requests directed to the data maintained on behalf of the one or more subscribers.

21. The computing system of claim 20, wherein to service query requests directed to the data maintained on behalf of the one or more subscribers, the database service is configured to retrieve at least some of the data targeted by the query requests from the remote key-value storage system prior to completion of the restore operation.

22. The computing system of claim 18, wherein the remote key-value storage system is configured to apply replication, parity, erasure coding, or another error correction technique to the copies of the plurality of physical data blocks in the remote key-value storage system.

23. A method, comprising:

performing, by one or more computers:

storing columnar data of a database table in a plurality of physical data blocks in a distributed data storage system on behalf of one or more clients, wherein the distributed data storage system comprises a cluster of one or more nodes, each of which comprises one or more disks on which physical data blocks are stored, and wherein each of the plurality of physical data blocks is associated with a respective unique identifier;

storing a copy of each of the plurality of physical data blocks in a remote key-value durable backup storage system, wherein for each of the plurality of physical data blocks, the respective unique identifier serves as a key to

access the data block in the remote key-value durable backup storage system;

detecting a failure in the distributed data storage system affecting two or more of the plurality of physical data blocks in which the columnar data was stored;

in response to said detecting, automatically restoring the two or more physical data blocks, wherein said restoring comprises:

determining a priority order in which to restore the two or more physical data blocks based, at least in part, on the relative likelihood that each of the two or more physical data blocks will be accessed in the near future;

retrieving a copy of the one of the two or more physical data blocks having the highest priority from the remote key-value durable backup storage system, wherein said retrieving comprises using the respective unique identifier associated with the one of the two or more physical data blocks as a key to access the copy of the one of the two or more physical data blocks in the key-value durable backup storage system;

writing a primary copy of the retrieved copy of the physical data block to a given disk on a given node in the distributed data storage system; and

initiating replication of the retrieved copy of the physical data block on one or more disks in the distributed data storage system other than the given disk.

24. The method of claim 23, wherein said determining a priority order in which to restore the two or more physical data blocks is based, at least in part, on a respective count value associated with each of the two or more physical data blocks indicating a number of recent accesses targeting the physical data block.

25. The method of claim 23, wherein said restoring further comprises updating an entry in a superblock data structure that stores information about physical data blocks stored on the given node to indicate that the one of the two or more physical data blocks has been restored.

26. The method of claim 23, further comprising:

retrieving a copy of each of the two or more physical data blocks other than the one of the two or more physical data blocks from the remote key-value durable backup storage system in the determined priority order;

writing a primary copy of each of the retrieved copies of the two or more physical data blocks other than the one of the two or more physical data blocks to a disk in the distributed data storage system; and

initiating replication of each of the retrieved copies of the two or more physical data blocks other than the one of the two or more physical data blocks on one or more disks in the distributed data storage system other than the disk on which its primary copy was written.

27. A method, comprising:

performing, by one or more computers:

maintaining data in a plurality of physical data blocks of a data storage system on behalf of one or more clients, wherein each physical data block is associated with a unique identifier;

beginning an operation to restore data that was stored in two or more of the plurality of physical data blocks in the data storage system from copies of the two or more physical data blocks stored in a key-value storage system that is distinct from the data storage system;

determining a highest priority data block to be restored based, at least in part, on a determination of the relative likelihood that data in each of the two or more physical data blocks will be accessed in the near future;

restoring the copy of the highest priority data block to the data storage system from the key-value storage system prior to restoring copies of the two or more physical data blocks other than the highest priority data block.

28. The method of claim 27,

wherein the method further comprises detecting a failure in the data storage system affecting the two or more of the plurality of physical data blocks; and

wherein said beginning an operation to restore data is performed in response to said detecting.

29. The method of claim 27,

wherein the method further comprises receiving a request from a client to restore the two or more of the plurality of physical data blocks; and

wherein said beginning an operation to restore data is performed in response to said receiving.

30. The method of claim 27,

wherein the data storage system stores columnar data of one or more database tables in the plurality of physical data blocks on behalf of one or more clients; and

wherein said determining a highest priority data block to be restored comprises determining which of the one or more database tables was most recently loaded into the data storage system.

31. The method of claim 27,

wherein the data storage system stores columnar data of one or more database tables in the plurality of physical data blocks on behalf of one or more clients; and

wherein said determining a highest priority data block to be restored comprises determining which of the one or more database tables or columns thereof were most frequently queried in a recent time period.

32. The method of claim 27,

wherein the data storage system stores columnar data of one or more database tables in the plurality of physical data blocks on behalf of one or more clients; and

wherein said determining a highest priority data block to be restored comprises identifying a physical data block that stores data for a same table or column thereof as a copy of a physical data block that was recently retrieved from the key-value storage system in response to receiving a query request targeting data in the physical data block that was no longer available in the data storage system.

33. The method of claim 27, wherein said restoring the copy of the highest priority data block comprises:

retrieving a copy of the highest priority data block from the remote key-value storage system, wherein said retrieving comprises using the unique identifier associated with the highest priority data block as a key to access the copy of the highest priority data block in the key-value storage system;

writing a primary copy of the retrieved copy of the highest priority data block to a storage device in the data storage system.

34. The method of claim 33, further comprising:

prior to said writing:

receiving a query that targets data maintained in the highest priority data block; and

satisfying the query, wherein satisfying the query comprises returning at least a portion of the data in the retrieved copy of the highest priority data block.

35. The method of claim 33, wherein said restoring the copy of the highest priority data block further comprises:

initiating replication of the highest priority data block on one or more storage devices in the data storage system other than the storage device on which the primary copy was written.

36. The method of claim 27,

wherein said restoring the copy of the highest priority data block comprises restoring the data stored in the highest priority data block from the key-value storage system to the data storage system while accepting and servicing queries directed to the data maintained on behalf of the one or more clients.

37. The method of claim 27, wherein said determining a highest priority data block to be restored is based, at least in part, on a historical pattern of accesses by previous queries that target the plurality of physical data blocks.

38. The method of claim 27,

wherein said restoring is performed by a background process; and

wherein the method further comprises, while the background process performs said restoring:

a foreground process receiving one or more queries directed to the data maintained on behalf of the one or more clients; and

the foreground process servicing the one or more queries.

39. The method of claim 27,

wherein the method further comprises receiving a query that targets data previously maintained in two or more physical data blocks that are no longer available in the data storage system; and

wherein said beginning an operation to restore data is performed in response to said receiving.

40. A computing system, comprising:

one or more computing nodes, each of which comprises at least one processor and a memory, wherein the one or more computing nodes are configured to collectively implement a database service; and

an interface to a remote key-value storage system;

wherein the database service is configured to:

store data on behalf of one or more subscribers to the database service in a plurality of physical data blocks on one or more storage devices, wherein each of the plurality of physical data blocks is associated with a unique identifier;

back up the data stored on behalf of the one or more subscribers, wherein to back up the data, the database service is configured to send to the remote key-value storage system, via the interface, a copy of each of the plurality of physical data blocks for storage in the remote key-value storage system, and the unique identifiers associated with each of the plurality of physical data blocks to be used as access keys for copies of the plurality of physical data blocks in the remote key-value storage system;

detect a condition or event that triggers a restore operation for two or more of the plurality of physical data blocks;

determine an order in which to restore the two or more of the plurality of physical data blocks based, at least in part, on the relative likelihood that each of the two or more of the plurality of physical data blocks will be accessed in the near future; and

restore the two or more of the plurality of physical data blocks in the determined order.

41. The computing system of claim 40, wherein to determine the order in which to restore the two or more of the plurality of physical data blocks, the database service is configured

to determine how recently each of the two or more of the plurality of physical data blocks was a target of a query.

42. The computing system of claim 40, wherein to determine the order in which to restore the two or more of the plurality of physical data blocks, the database service is configured to determine how recently each of the two or more of the plurality of physical data blocks was written to.

43. The computing system of claim 40, wherein to determine the order in which to restore the two or more of the plurality of physical data blocks, the database service is configured to determine how recently each of the two or more of the plurality of physical data blocks was backed up.

44. A non-transitory computer-readable storage medium storing program instructions that when executed on one or more computers cause the one or more computers to perform:

maintaining data in a plurality of physical data blocks of a data storage system on behalf of one or more clients, wherein each physical data block is associated with a unique identifier;

during a pre-determined time period:

receiving a query targeting data in a given one of the plurality of physical data blocks;

servicing the query, wherein servicing the query comprises accessing the given one of the plurality of physical data blocks; and

incrementing a counter associated with the given one of the plurality of physical data blocks, wherein the value of the counter indicates the number of times the given one of the plurality of physical data blocks was accessed within the pre-determined time period;

beginning an operation to restore two or more of the plurality of physical data blocks including the given one of the plurality of physical data blocks;

determining an order in which to restore the two or more of the plurality of physical data blocks based, at least in part, on the value of the counter associated with the given one of the plurality of physical data blocks and the value of a respective counter associated with each of the two or more of the plurality of physical data blocks other than the given one of the plurality of physical data blocks.

45. The non-transitory computer-readable storage medium of claim 44,

wherein said determining an order in which to restore the two or more of the plurality of physical data blocks is further dependent on the value of another respective counter associated with each of the plurality of physical data blocks, wherein the value of the other respective counter associated with each of the plurality of physical data blocks indicates the number of times that the physical data block was assessed within a previous pre-determined time period.

46. The non-transitory computer-readable storage medium of claim 45, wherein said determining an order in which to restore the two or more of the plurality of physical data blocks comprises computing a value representing a logical or mathematical combination of the values of the respective counter and the other respective counter associated with each of the two or more of the plurality of physical data blocks.

47. A method, comprising:

performing, by one or more computers:

storing columnar data of a database table in a plurality of physical data blocks in a distributed data storage system on behalf of one or more clients, wherein the distributed data storage system comprises a cluster of one or more nodes, each of which comprises one or more disks on which physical data blocks are stored, wherein each of the plurality of physical data blocks is associated with a respective unique identifier, and wherein said storing columnar data comprises storing two or more copies of each portion of the

columnar data in different physical data blocks in the distributed data storage system;

storing a backup copy of each of the plurality of physical data blocks in a remote key-value durable backup storage system, wherein for each of the plurality of physical data blocks, the respective unique identifier serves as a key to access the data block in the remote key-value durable backup storage system;

receiving, from a client, a query directed to a portion of the columnar data stored in the distributed data storage system;

in response to said receiving:

accessing one of the plurality of physical data blocks in the distributed data storage system in which a copy of the portion of the columnar data is stored;

applying a consistency check to the one of the plurality of physical data blocks;

determining, based, at least in part, on said applying, that data on the one of the plurality of physical data blocks is corrupted or inconsistent with other ones of the plurality of physical data blocks;

automatically retrieving the backup copy of the one of the plurality of physical data blocks stored in the remote key-value durable backup storage system, wherein said retrieving comprises using the unique identifier for the one of the one or more blocks as a key to access the backup copy of the one of the plurality of physical data blocks in the remote key-value durable backup storage system; and

satisfying the query, wherein said satisfying the query comprises returning at least a portion of the data in the retrieved backup copy of the one of the plurality of physical data blocks to the client.

48. The method of claim 47,

wherein said storing columnar data comprises storing one primary copy and one or more secondary copies of each portion of the columnar data; and

wherein said accessing one of the plurality of physical data blocks comprises accessing a data block in which a secondary copy of the portion of the columnar data is stored in response to determining that a data block in which the primary copy of the portion of the columnar data is stored is corrupted or inconsistent with other ones of the plurality of physical data blocks.

49. The method of claim 47, wherein said retrieving the backup copy of the one of the plurality of physical data blocks comprises copying the data stored in the backup copy of the one of the plurality of physical data blocks from the key-value durable backup storage system to system memory in the distributed data storage system while a separate background process copies one or more backup copies of physical data blocks from the key-value durable backup storage system to one or more disks of the cluster of nodes in the distributed data storage system.

50. The method of claim 47, further comprising:

subsequent to said retrieving the backup copy of the one of the plurality of physical data blocks, initiating an operation to restore the one of the plurality of physical data blocks in the distributed data storage system, wherein restoring the one of the plurality of physical data blocks comprises storing one primary copy and one or more secondary copies of the portion of the columnar data in the backup copy of the one of the plurality of physical data blocks to one or more disks of the cluster of nodes in the distributed data storage system.

51. A method, comprising:

performing, by one or more computers:

maintaining data in a plurality of physical data blocks of a data storage system on behalf of one or more clients;

receiving a read or write request directed to a portion of the data;

in response to said receiving:

determining that a consistent and uncorrupted copy of the portion of the data is not available in the data storage system;

in response to determining that a consistent and uncorrupted copy of the portion of the data is not available in the data storage system, automatically retrieving a backup copy of the portion of the data from a key-value storage system that is separate and distinct from the data storage system; and

returning a response to the read or write request comprising data in the retrieved backup copy of the portion of the data.

52. The method of claim 51, wherein said determining comprises accessing a data structure that stores information indicating one or more locations at which a respective copy of each portion of the data maintained on behalf of the one or more clients is stored in a physical data block to locate a primary copy or a secondary copy of the portion of the data.

53. The method of claim 51, wherein said determining that a consistent and uncorrupted copy of the portion of the data is not available in the data storage system comprises determining that a copy of the portion of the data was previously corrupted and has not yet been restored.

54. The method of claim 51, wherein said determining that a consistent and uncorrupted copy of the portion of the data is not available in the data storage system comprises determining that a node or disk on which a copy of the portion of the data was stored has failed and that data stored on the failed node or disk has not yet been restored.

55. The method of claim 51, wherein said determining comprises applying a consistency check to a physical data block in which a primary copy of the portion of the data is stored.

56. The method of claim 55, wherein said determining comprises:

determining, based, at least in part, on said applying, that the primary copy of the portion of the data has been corrupted or is inconsistent with other ones of the plurality of physical data blocks; and

applying a consistency check to a physical data block in which a secondary copy of the portion of the data is stored.

57. The method of claim 51,

wherein each physical data block in which data is maintained in the data storage system on behalf of the one or more clients is associated with a unique identifier; and

wherein automatically retrieving a backup copy of the portion of the data comprises using the respective unique identifier associated with a physical data block in which the portion of the data was stored in the data storage system as a key to access the backup copy of the physical data block in which the portion of the data was stored in the data storage system in the key-value storage system.

58. The method of claim 51, further comprising:

prior to said receiving a read or write request, backing up at least some of the plurality of physical data blocks in which data is maintained in the data storage system on behalf of the one or more clients, wherein said backing up comprises, for each of the at least some of the plurality of physical data blocks:

sending to the key-value storage system a copy of the physical data block and a unique identifier associated with the physical data block to be used as an access key for retrieving a backup copy of the physical data block in the remote key-value storage system.

59. The method of claim 51, further comprising, for each of the plurality of physical data blocks in which data is maintained in the data storage system on behalf of the one or more clients:

generating a unique identifier to be associated with the physical data block when data is first written to the physical data block.

60. The method of claim 51, further comprising:

receiving another read or write request directed to another portion of the data maintained in the data storage system on behalf of the one or more clients;

in response to receiving the other read or write request:

determining that a consistent and uncorrupted copy of the other portion of the data is available in the data storage system;

in response to determining that a consistent and uncorrupted copy of the portion of the data is available in the data storage system, returning a response to the other read or write request comprising data in the consistent and uncorrupted copy of the other portion of the data in the data storage system.

61. The method of claim 51, further comprising:

receiving another read or write request directed to another portion of the data maintained in the data storage system on behalf of the one or more clients;

in response to receiving the other read or write request:

determining that a primary copy of the other portion of the data in the data storage system is consistent and uncorrupted;

in response to determining that the primary copy of the other portion of the data in the data storage system is corrupted or inconsistent with other ones of the plurality of physical data blocks:

accessing a secondary copy of the other portion of the data in the data storage system;

determining that the secondary copy of the other portion of the data in the data storage system is consistent and uncorrupted; and

in response to determining that the secondary copy of the other portion of the data in the data storage system is consistent and uncorrupted, returning a response to the other read or write request comprising data in the secondary copy of the other portion of the data in the data storage system.

62. A computing system, comprising:

one or more computing nodes, each of which comprises at least one processor and a memory, wherein the one or more computing nodes are configured to collectively implement a database service; and

an interface to a remote key-value storage system;

wherein the database service is configured to:

store data on behalf of one or more subscribers to the database service in a plurality of physical data blocks on one or more storage devices, wherein each of the plurality of physical data blocks is associated with a unique identifier;

back up the data stored on behalf of the one or more subscribers, wherein to back up the data, the database service is configured to send to the remote key-value storage system, via the interface, a copy of each of the plurality of physical data blocks for storage in the remote key-value storage system and the unique identifiers associated with each of the plurality of physical data blocks to be used as access keys for retrieving backup copies of the plurality of physical data blocks in the remote key-value storage system;

receive a query directed to a portion of the data;

in response to said receiving:

determine that a consistent and uncorrupted copy of the portion of the data is not available on the one or more computing nodes;

in response to determining that a consistent and uncorrupted copy of the portion of the data is not available on the one or more computing nodes, automatically retrieving a backup copy of the portion of the data from the key-value storage system using the unique identifier for one of the plurality of physical data blocks in which the portion of the data was stored on the one or more computing nodes as a key to access the backup copy of the one of the plurality of physical data blocks in the key-value storage system; and

returning a query response comprising data in the retrieved backup copy of the portion of the data.

63. The computing system of claim 62, wherein determining that a consistent and uncorrupted copy of the portion of the data is not available on the one or more computing nodes comprises determining which of the one or more computing nodes stores a primary copy of the portion of the data.

64. The computing system of claim 62, wherein determining that a consistent and uncorrupted copy of the portion of the data is not available on the one or more computing nodes comprises determining which of the one or more computing nodes stores a secondary copy of the portion of the data.

65. A non-transitory computer-readable storage medium storing program instructions that when executed on one or more computers cause the one or more computers to perform:

receiving a query from a client that targets a given one of a plurality of data blocks that were previously stored in a cluster of one or more computing nodes;

determining that a consistent and uncorrupted copy of the given data block is not available in the cluster;

in response to determining that a consistent and uncorrupted copy of the given data block is not available in the cluster:

automatically initiating an operation to restore the given data block, wherein said initiating comprises retrieving a copy of the given data block from a key-value storage system, wherein the copy of the given data block is indexed in the key-value storage system by a unique identifier associated with the given data block; and

returning at least a portion of the data in the copy of the given data block to the client.

66. The non-transitory computer-readable storage medium of claim 65, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

subsequent to said retrieving the copy of the given data block from the key-value storage system, copying the data in the retrieved copy of the given data block to one of the two or more computing nodes as a primary copy of the given data block.

67. The non-transitory computer-readable storage medium of claim 66, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

subsequent to said retrieving the copy of the given data block from the key-value storage system, copying the data in the retrieved copy of the given data block to another one of the two or more computing nodes as a secondary copy of the given data block.

68. The non-transitory computer-readable storage medium of claim 66, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

subsequent to said copying the data in the retrieved copy of the given data block to one of the two or more computing nodes as a primary copy of the given data block, updating a data structure that stores information about physical data blocks stored

on the one of the two or more computing nodes to indicate that the primary copy of the given data block has been written to the one of the two or more computing nodes.

69. The non-transitory computer-readable storage medium of claim 65,

wherein said automatically initiating a restore operation is performed by a foreground process that receives and services query requests targeting data blocks stored in the cluster; and

wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

initiating a restore operation for two or more other ones of the plurality of data blocks that were previously stored in the cluster as a background process.

70. The non-transitory computer-readable storage medium of claim 65, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

logging, in a file or data structure, information associated with the restore operation, wherein the information comprises: an indication of a failure of one of the two or more computing nodes that resulted in a consistent and uncorrupted copy of the given data block not being available in the cluster, an indication of a condition or event that triggered the initiation of the restore operation, or an indication that the restore operation was performed; and

subsequent to said logging:

accessing the file or data structure to retrieve information associated with a plurality of restore operations; and

analyzing the information associated with the plurality of restore operations as part of a failure analysis operation, a trend analysis operation, or a maintenance operation.

## CLAIMS

1. A method, comprising:

performing, by one or more computers:

storing columnar data of a database table in a plurality of physical data blocks in a distributed data storage system on behalf of one or more clients, wherein the distributed data storage system comprises a cluster of one or more nodes, each of which comprises one or more disks on which physical data blocks are stored, and wherein each of the plurality of physical data blocks is associated with a respective unique identifier;

storing a copy of individual ones of the plurality of physical data blocks in a remote key-value durable backup storage system, wherein for the individual ones of the plurality of physical data blocks, the respective unique identifier serves as a key to access the data block in the remote key-value durable backup storage system;

detecting a failure in the distributed data storage system affecting at least one of the plurality of physical data blocks in which the columnar data was stored;

in response to said detecting, automatically initiating a restore of the columnar data that was stored in the at least one of the plurality of physical data blocks from the remote key-value durable backup storage system;

receiving one or more query requests directed to the columnar data of the database table; and

servicing the one or more query requests, wherein said servicing comprises:

at least partly during the restore, streaming at least some of the columnar data from the key-value durable backup storage system into system memory of the distributed data storage system; and

at least partly during the restore, accessing the columnar data from the system memory in response to the one or more query requests.

2. The method of claim 1, wherein said storing the columnar data of the database table comprises storing a portion of the columnar data as a primary copy of the portion of the columnar data in a respective physical data block on a given disk, and storing the portion of the columnar data as one or more secondary copies of the portion of the columnar data in respective physical data blocks on one or more disks other than the given disk.

3. The method of claim 1, wherein said storing the columnar data of the database table comprises storing the unique identifier of the individual ones of the plurality of physical data blocks in a respective entry in a superblock data structure that stores information about the physical data blocks stored on a given node.

4. The method of claim 1, wherein said storing columnar data of the database table comprises storing information indicating a location at which the individual ones of the plurality of physical data blocks are stored on disks of a given node, the location indicated in a respective entry in a superblock data structure that stores information about the physical data blocks stored on the given node.

5. The method of claim 1, wherein said servicing further comprises obtaining at least some of the columnar data of the database table to which the one or more query requests are directed from disks in the distributed data storage system.

6. A method, comprising:

performing, by one or more computers:

maintaining data in one or more physical data blocks of a data storage system on behalf of one or more clients, wherein each physical data block is associated with a unique identifier;

performing a backup operation to store a respective copy of data stored in a given physical data block in a key-value storage system that is distinct from the data storage system;

subsequent to storing the respective copy of the data stored in the given physical data block, restoring the data stored in the given physical data block from the key-value storage system to the data storage system; and

servicing one or more queries directed to the data maintained on behalf of the one or more clients, wherein said servicing comprises:

at least partly during the restoring the data, streaming at least some of the data from the key-value storage system into system memory of the data storage system; and

at least partly during the restoring the data, accessing the data from the system memory in response to the one or more queries.

7. The method of claim 6, wherein said restoring is performed as part of an operation to restore the data stored in multiple physical data blocks from the key-value storage system to the data storage system, and wherein said servicing the one or more queries is performed prior to restoring all of the data stored in the multiple physical data blocks.

8. The method of claim 6, further comprising receiving the data representing entries in a database table, wherein said maintaining the data in the one or more physical data blocks of the data storage system comprises storing the data in one or more columns of the database table in the one or more physical data blocks.

9. The method of claim 6, wherein said restoring is performed in response to a failure of a storage device on which the data is stored in the data storage system.

10. The method of claim 6, wherein said restoring is performed in response to a failure of a node comprising a storage device on which the data is stored in the data storage system.

11. The method of claim 6, wherein said restoring is performed in response to a failure of a cluster of one or more nodes comprising a storage device on which the data is stored in the data storage system.

12. The method of claim 6, wherein said restoring is performed in response to a request from one of the one or more clients to perform a restore operation.

13. The method of claim 6,

wherein said maintaining the data in the one or more physical data blocks of the data storage system comprises maintaining a data structure that stores a mapping between the data maintained on a particular node in a cluster of one or more nodes and a location in the data storage system at which the data is stored in a physical data block on the particular node, wherein the data structure is maintained on the particular node, and wherein each entry in the data structure stores a mapping for a particular physical data block and the unique identifier associated with the particular physical data block; and

wherein the method comprises, prior to said restoring, restoring the data structure on the particular node in response to a failure of the node or a failure of the cluster of nodes.

14. A non-transitory computer-readable storage medium storing program instructions that when executed on one or more computers cause the one or more computers to perform:

storing one or more data blocks in a data warehouse system;

creating an entry for each of the one or more data blocks in a data structure that stores information about the one or more data blocks, wherein the entry comprises:

a unique identifier for a data block;

at least one counter indicating a number of accesses to the data block in a given period, wherein the given period corresponds to one or more events comprising backup or restore operations; and

an indication that the data block has not yet been backed up; and

performing a backup operation for a plurality of data blocks stored in the data warehouse system, including the one or more data blocks, wherein said performing comprises:

storing a backup copy of the data structure in a remote key-value storage system;

storing, for each data block stored in the data warehouse system for which a corresponding entry in the data structure indicates that it has not yet been backed up, a backup copy of the data block in the remote key-value storage system; and

updating the entry in the data structure corresponding to each data block that was backed up by the backup operation to indicate that the data block has been backed up.

15. The non-transitory computer-readable storage medium of claim 14, wherein said storing the one or more data blocks in the data warehouse system comprises storing two or more copies of each of the one or more data blocks on different storage devices in the data warehouse system.

16. The non-transitory computer-readable storage medium of claim 14, wherein said storing the one or more data blocks in the data warehouse system comprises generating the unique identifier for each of the one or more data blocks.

17. The non-transitory computer-readable storage medium of claim 14, wherein when executed on one or more computers, the program instructions further cause the one or more computers to perform:

receiving a query directed to the data in one of the one or more data blocks; and

in response to receiving the query, accessing the backup copy of the one of the one or more data blocks in the remote key-value storage system using the unique identifier for the one of the one or more data blocks as an access key in the remote key-value storage system.

18. A computing system, comprising:

one or more computing nodes, each of which comprises at least one processor and a memory, wherein the one or more computing nodes are configured to collectively implement a database service; and

an interface to a remote key-value storage system;

wherein the database service is configured to maintain data on behalf of one or more subscribers to the database service;

wherein the one or more computing nodes are configured to store the data maintained on behalf of the one or more subscribers in a plurality of physical data blocks on one or more storage devices, wherein each of the plurality of physical data blocks is associated with a unique identifier;

wherein the database service is configured to maintain a data structure including an entry describing each of the plurality of physical data blocks, the entry comprising:

the unique identifier for a physical data block; and

at least one counter indicating a number of accesses to the physical data block; and

wherein the database service is configured to perform a backup operation for the data, wherein to perform the backup operation, the database service is configured to:

send to the remote key-value storage system, via the interface, a copy of each of the plurality of physical data blocks for storage in the remote key-value storage system, wherein the unique identifiers associated with each of the plurality of physical data blocks are employable as access keys to access copies of the plurality of physical data blocks in the remote key-value storage system; and

send a backup copy of the data structure for storage in the remote key-value storage system.

19. The computing system of claim 18,

wherein the one or more computing nodes comprise a leader node that is configured to maintain one or more query maps and a mapping between the data maintained on behalf of the one or more subscribers and the locations at which the data is stored by the database service on the one or more computing nodes.

20. The computing system of claim 18,

wherein subsequent to performing the backup operation, the database service is configured to perform a restore operation;

wherein to perform the restore operation, the database service is configured to retrieve from the remote key-value storage system, via the interface, the copies of each of the plurality of physical data blocks, using the unique identifiers associated with each of the plurality of physical data blocks as access keys for the copies of the plurality of physical data blocks in the remote key-value storage system; and

wherein during performance of the restore operation, the database service is further configured to accept and service query requests directed to the data maintained on behalf of the one or more subscribers.

21. The computing system of claim 20, wherein to service the query requests directed to the data maintained on behalf of the one or more subscribers, the database service is configured to retrieve at least some of the data targeted by the query requests from the remote key-value storage system prior to completion of the restore operation.

22. The computing system of claim 18, wherein the remote key-value storage system is configured to apply one or more error correction techniques to the copies of the plurality of

physical data blocks in the remote key-value storage system, the one or more error correction techniques including one or more of replication, parity, or erasure coding.

23. A method, comprising:

performing, by one or more computers:

storing columnar data of a database table in a plurality of physical data blocks in a distributed data storage system on behalf of one or more clients, wherein the distributed data storage system comprises a cluster of nodes, each of which comprises one or more disks on which a plurality of physical data blocks are stored, wherein each of the plurality of physical data blocks is associated with a respective unique identifier, and wherein said storing columnar data comprises storing two or more copies of each portion of the columnar data in different physical data blocks in the distributed data storage system;

storing a backup copy of each of the plurality of physical data blocks in a remote key-value durable backup storage system, wherein for each of the plurality of physical data blocks, the respective unique identifier serves as a key to access a data block that includes the backup copy in the remote key-value durable backup storage system, and wherein the remote key-value durable backup storage system employs a different data storage format than the distributed data storage system;

receiving, from a client, a query to access a portion of the columnar data stored in the distributed data storage system;

in response to said receiving:

accessing one of the plurality of physical data blocks in the distributed data storage system in which a copy of the portion of the columnar data is stored;

applying a consistency check to the one of the plurality of physical data blocks;

determining, based, at least in part, on said applying, that data on the one of the plurality of physical data blocks is corrupted or inconsistent with other ones of the plurality of physical data blocks;

in response to said determining that the data on the one of the plurality of physical data blocks is corrupted or inconsistent with the other ones of the plurality of physical data blocks, automatically retrieving the backup copy of the one of the plurality of physical data blocks stored in the remote key-value durable backup storage system that employs the different data storage format than the distributed data storage system, wherein said retrieving comprises using the unique identifier for the one of the plurality of data blocks as the key to access the backup copy of the one of the plurality of physical data blocks in the remote key-value durable backup storage system; and

satisfying the query, wherein said satisfying the query comprises returning at least a portion of the data in the retrieved backup copy of the one of the plurality of physical data blocks to the client during an operation to restore the one of the plurality of physical data blocks.

24. The method of claim 23,

wherein said storing columnar data comprises storing one primary copy and one or more secondary copies of each portion of the columnar data; and

wherein said accessing one of the plurality of physical data blocks comprises accessing a second data block in which a secondary copy of the portion of the columnar data is stored in response to determining that a first data block in which the primary copy of the portion of the columnar data is stored is corrupted or inconsistent with other ones of the plurality of physical data blocks.

25. The method of claim 23, wherein said retrieving the backup copy of the one of the plurality of physical data blocks comprises copying the data stored in the backup copy of the one of the plurality of physical data blocks from the key-value durable backup storage system to system memory in the distributed data storage system while a separate background process copies one or more backup copies of physical data blocks from the key-value durable backup storage system to one or more disks of the cluster of nodes in the distributed data storage system.

26. The method of claim 23, further comprising:

subsequent to said retrieving the backup copy of the one of the plurality of physical data blocks, initiating the operation to restore the one of the plurality of physical data blocks in the distributed data storage system, wherein restoring the one of the plurality of physical data blocks comprises storing one primary copy and one or more secondary copies of the portion of the columnar data in the backup copy of the one of the plurality of physical data blocks to one or more disks of the cluster of nodes in the distributed data storage system.

27. A method, comprising:

performing, by one or more computers:

maintaining data in a plurality of physical data blocks of a data storage system on behalf of one or more clients;

receiving a request to access a portion of the data;

in response to said receiving:

determining that an uncorrupted copy of the portion of the data is not available in the data storage system;

in response to determining that the uncorrupted copy of the portion of the data is not available in the data storage system, retrieving a backup copy of the portion of the data from a key-value storage system that is separate from the data storage system, wherein the key-value

storage system employs a different data storage format than the data storage system; and

returning a response to the request during an operation to restore a corrupted copy of the portion of the data, the response comprising the retrieved backup copy of the portion of the data.

28. The method of claim 27, wherein said determining comprises:

accessing a data structure that stores information indicating one or more locations at which a respective copy of each portion of the data maintained on behalf of the one or more clients is stored in a physical data block; and

employing the data structure to locate at least one of a primary copy or a secondary copy of the portion of the data.

29. The method of claim 27, wherein said determining that the uncorrupted copy of the portion of the data is not available in the data storage system comprises determining that a copy of the portion of the data was previously corrupted and has not yet been restored.

30. The method of claim 27, wherein said determining that the uncorrupted copy of the portion of the data is not available in the data storage system comprises determining that a node or a disk on which a copy of the portion of the data was stored has failed and that the data stored on the node or the disk has not yet been restored.

31. The method of claim 27, wherein said determining comprises applying a consistency check to a physical data block in which a primary copy of the portion of the data is stored.

32. The method of claim 31, wherein said determining further comprises:

determining, based, at least in part, on said applying, that the primary copy of the portion of the data has been corrupted or is inconsistent with one or more other physical data blocks; and

applying a consistency check to a physical data block in which a secondary copy of the portion of the data is stored.

33. The method of claim 27,

wherein each physical data block in which data is maintained is associated with a unique identifier; and

wherein retrieving the backup copy of the portion of the data comprises using the respective unique identifier associated with a physical data block in which the portion of the data was stored in the data storage system as a key to access the backup copy of the physical data block in which the portion of the data was stored in the data storage system in the key-value storage system.

34. The method of claim 27, further comprising:

prior to receiving the request, backing up at least some of the plurality of physical data blocks in which the data is maintained in the data storage system on behalf of the one or more clients, wherein said backing up comprises, for each of the at least some of the plurality of physical data blocks:

sending to the key-value storage system a copy of a physical data block and a unique identifier associated with the physical data block to be used as an access key for retrieving a backup copy of the physical data block in the remote key-value storage system.

35. The method of claim 27, further comprising, for each of the plurality of physical data blocks in which the data is maintained in the data storage system on behalf of the one or more clients:

generating a unique identifier to be associated with a physical data block, the unique identifier generated when data is written to the physical data block.

36. The method of claim 27, further comprising:

receiving another request to access another portion of the data maintained in the data storage system on behalf of the one or more clients;

in response to receiving the other request:

determining that an uncorrupted copy of the other portion of the data is available in the data storage system;

in response to determining that the uncorrupted copy of the other portion of the data is available in the data storage system, returning a response to the other request, the response to the other request comprising the uncorrupted copy of the other portion of the data in the data storage system.

37. The method of claim 27, further comprising:

receiving another request to access another portion of the data maintained in the data storage system on behalf of the one or more clients;

in response to receiving the other request:

in response to determining that a primary copy of the other portion of the data in the data storage system is corrupted or inconsistent with other ones of the plurality of physical data blocks:

accessing a secondary copy of the other portion of the data in the data storage system; and

in response to determining that the secondary copy of the other portion of the data in the data storage system is uncorrupted, returning a response to the other request, the response to the other request comprising the secondary copy of the other portion of the data in the data storage system.

38. A computing system, comprising:

one or more computing nodes, each of which comprises at least one processor and a memory, wherein the one or more computing nodes are configured to collectively implement a database service; and

an interface to a remote key-value storage system;

wherein the database service is configured to:

store data on behalf of one or more subscribers to the database service in a plurality of physical data blocks on one or more storage devices, wherein each of the plurality of physical data blocks is associated with a unique identifier, wherein the remote key-value storage system employs a different data storage format than the one or more storage devices;

back up the data stored on behalf of the one or more subscribers, wherein to back up the data, the database service is configured to send to the remote key-value storage system, via the interface, a copy of each of the plurality of physical data blocks for storage in the remote key-value storage system, wherein the unique identifiers associated with the plurality of physical data blocks are access keys for retrieving backup copies of the plurality of physical data blocks in the remote key-value storage system;

receive a query directed to a portion of the data;

in response to said receiving:

determine that an uncorrupted copy of the portion of the data is not available on the one or more storage devices;

in response to determining that the uncorrupted copy of the portion of the data is not available on the one or more storage devices, retrieving a backup copy of the portion of the data from the key-value storage system that employs the different data storage format than the one or more storage devices, wherein the retrieving uses the unique identifier for one of the plurality of physical data blocks as a key to access the backup copy of the one of the plurality of physical data blocks in the key-value storage system; and

returning, during an operation to restore a corrupted copy of the portion of the data, a query response comprising the retrieved backup copy of the portion of the data.

39. The computing system of claim 38, wherein determining that the uncorrupted copy of the portion of the data is not available on the one or more storage devices comprises determining which of the one or more storage devices stores a primary copy of the portion of the data.

40. The computing system of claim 38, wherein determining that the uncorrupted copy of the portion of the data is not available on the one or more storage devices comprises determining which of the one or more storage devices stores a secondary copy of the portion of the data.

41. A non-transitory computer-readable storage medium storing program instructions that when executed on one or more computers cause the one or more computers to perform:

receiving a query from a client, the query targeting a data block of a plurality of data blocks that were previously stored in a cluster of computing nodes;

determining that a consistent and uncorrupted copy of the data block is not available in the cluster;

in response to determining that the consistent and uncorrupted copy of the data block is not available in the cluster:

initiating an operation to restore the data block to the cluster by retrieving a copy of the data block from a key-value storage system, wherein the cluster supports a different data storage format than the key-value storage system, and wherein the copy of the data block is indexed in the key-value storage system by a unique identifier associated with the data block; and

returning, during the operation to restore the data block, at least a portion of the copy of the data block to the client.

42. The non-transitory computer-readable storage medium of claim 41, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

subsequent to said retrieving the copy of the data block from the key-value storage system, copying data in the retrieved copy of the data block to one of the cluster of computing nodes as a primary copy of the data block.

43. The non-transitory computer-readable storage medium of claim 42, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

subsequent to said retrieving the copy of the data block from the key-value storage system, copying the data in the retrieved copy of the data block to another one of the cluster of computing nodes as a secondary copy of the data block.

44. The non-transitory computer-readable storage medium of claim 42, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

subsequent to said copying the data in the retrieved copy of the data block to one of the cluster of computing nodes as the primary copy of the data block, updating a data structure that stores information about physical data blocks stored on the one of the cluster of computing nodes to indicate that the primary copy of the data block has been written to the one of the cluster of computing nodes.

45. The non-transitory computer-readable storage medium of claim 41,

wherein said initiating the operation to restore the data block is performed by a foreground process that receives and services query requests targeting data blocks stored in the cluster; and

wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

initiating, as a background process, a restore operation for two or more other ones of the plurality of data blocks that were previously stored in the cluster.

46. The non-transitory computer-readable storage medium of claim 41, wherein when executed on the one or more computers, the program instructions further cause the one or more computers to perform:

logging, in a file or a data structure, information associated with the operation, wherein the information comprises: an indication of a failure of one of the cluster of computing nodes that resulted in the consistent and uncorrupted copy of the data block not being available in the cluster, an indication of a condition or event that triggered the initiation of the operation, or an indication that the operation was performed; and

subsequent to said logging:

accessing the file or the data structure to retrieve information associated with a plurality of restore operations; and

analyzing the information associated with the plurality of restore operations as part of a failure analysis operation, a trend analysis operation, or a maintenance operation.