CLAIMS

1. A method of joining data records using a computer, the method comprising:

identifying a first plurality of data records and a second plurality of data records;

computing a hash for each data record in the first and second pluralities of data records;

assigning data records of the first plurality of data records to a plurality of first groupings based on the computed hashes, wherein each first grouping is associated with a set of bits that are included in the computed hashes of the data records in the first grouping;

assigning data records of the second plurality of records to a plurality of second groupings based on the computed hashes, wherein each second grouping is associated with a set of bits that are included in the computed hashes of the data records in the second grouping;

assigning each respective first grouping and a respective second grouping to a respective worker thread, wherein the respective first grouping is associated with a same set of bits as the respective second grouping;

determining, by the respective worker thread based on the computed hashes of the data records in the first and second groupings assigned to the respective worker thread, whether to join data records in the first grouping with data records in the second grouping;

responsive to determining to join one or more data records in the first grouping with one or more data records in the second, joining, by the respective worker thread, the data records; and

outputting, by the respective worker thread, the joined data records of the first and second groupings.

2. The method of claim 1, wherein the data records comprise a plurality of fields having values and wherein computing a hash for each data record in the first and second pluralities of data records comprises:

identifying a subset of the plurality of fields of the data records that are common to data records in the first and second pluralities of data records; and

computing hashes of values of the subset of the plurality of fields that are common to data records in the first and second pluralities of data records.

3. The method of claim 1, wherein determining, by the respective worker thread based on the computed hashes of the data records in the first and second groupings assigned to the respective worker thread, whether to join data records in the first grouping with data records in the second grouping comprises:

determining to join a first data record in the first grouping with a second data record in the second grouping responsive to the first and second data records having matching hashes.

4. The method of claim 1, wherein determining whether to join data records in the first grouping with data records in the second grouping comprises:

sorting the data records in the first grouping and the data records in the second grouping based on the hashes for the data records.

5. The method of claim 1, wherein the computer has a number of processor cores, and wherein a number of the first groupings and a number of the second groupings are determined responsive to the number of processor cores.

6. The method of claim 1, wherein different worker threads process the corresponding first and second groupings in parallel.

7. A computer for joining data records, the computer comprising:

at least one computer processor; and

a non-transitory computer-readable memory storing computer program instructions executable by the processor to perform operations comprising:

identifying a first plurality of data records and a second plurality of data records;

computing a hash for each data record in the first and second pluralities of data records;

assigning data records of the first plurality of data records to a plurality of first groupings based on the computed hashes, wherein each first grouping is associated with a set of bits that are included in the computed hashes of the data records in the first grouping;

assigning data records of the second plurality of records to a plurality of second groupings based on the computed hashes, wherein each second grouping is associated with a set of bits that are included in the computed hashes of the data records in the second grouping;

assigning each respective first grouping and a respective second grouping to a respective worker thread, wherein the respective first grouping is associated with a same set of bits as the respective second grouping;

determining, by the respective worker thread based on the computed hashes of the data records in the first and second groupings assigned to the respective worker thread, whether to join data records in the first grouping with data records in the second grouping;

responsive to determining to join one or more data records in the first grouping with one or more data records in the second, joining, by the respective worker thread, the data records; and

outputting, by the respective worker thread, the joined data records of the first and second groupings.

8. The computer of claim 7, wherein the data records comprise a plurality of fields having values and wherein computing a hash for each data record in the first and second pluralities of data records comprises:

identifying a subset of the plurality of fields of the data records that are common to data records in the first and second pluralities of data records; and

computing hashes of values of the subset of the plurality of fields that are common to data records in the first and second pluralities of data records.

9. The computer of claim 7, wherein determining, by the respective worker thread based on the computed hashes of the data records in the first and second groupings assigned to the respective worker thread, whether to join data records in the first grouping with data records in the second grouping comprises:

determining to join a first data record in the first grouping with a second data record in the second grouping responsive to the first and second data records having matching hashes.

10. The computer of claim 7, wherein determining whether to join data records in the first grouping with data records in the second grouping comprises:

sorting the data records in the first grouping and the data records in the second grouping based on the hashes for the data records.

11. The computer of claim 7, wherein the computer processor has a number of processor cores, and wherein a number of the first groupings and a number of the second groupings are determined responsive to the number of processor cores.

12. The computer of claim 7, wherein different worker threads process the corresponding first and second groupings in parallel.

13. A non-transitory computer-readable memory storing computer program instructions executable by a computer processor to perform operations for joining data records, the operations comprising:

identifying a first plurality of data records and a second plurality of data records;

computing a hash for each data record in the first and second pluralities of data records;

assigning data records of the first plurality of data records to a plurality of first groupings based on the computed hashes, wherein each first grouping is associated with a set of bits that are included in the computed hashes of the data records in the first grouping;

assigning data records of the second plurality of records to a plurality of second groupings based on the computed hashes, wherein each second grouping is associated with a set of bits that are included in the computed hashes of the data records in the second grouping;

assigning each respective first grouping and a respective second grouping to a respective worker thread, wherein the respective first grouping is associated with a same set of bits as the respective second grouping;

determining, by the respective worker thread based on the computed hashes of the data records in the first and second groupings assigned to the respective worker thread, whether to join data records in the first grouping with data records in the second grouping;

responsive to determining to join one or more data records in the first grouping with one or more data records in the second, joining, by the respective worker thread, the data records; and

outputting, by the respective worker thread, the joined data records of the first and second groupings.

14. The non-transitory computer-readable memory of claim 13, wherein the data records comprise a plurality of fields having values and wherein computing a hash for each data record in the first and second pluralities of data records comprises:

identifying a subset of the plurality of fields of the data records that are common to data records in the first and second pluralities of data records; and

computing hashes of values of the subset of the plurality of fields that are common to data records in the first and second pluralities of data records.

15. The non-transitory computer-readable memory of claim 13, wherein determining, by the respective worker thread based on the computed hashes of the data records in the first and second groupings assigned to the respective worker thread, whether to join data records in the first grouping with data records in the second grouping comprises:

determining to join a first data record in the first grouping with a second data record in the second grouping responsive to the first and second data records having matching hashes.

16. The non-transitory computer-readable memory of claim 13, wherein the computer processor has a number of processor cores, and wherein a number of the first groupings and a number of the second groupings are determined responsive to the number of processor cores.

17. The non-transitory computer-readable memory of claim 13, wherein different worker threads process the corresponding first and second groupings in parallel.