

What is claimed is:

1. A method for processing data in a computing system, the method including:
receiving, over an input device or port of the computing system, a plurality of records that each
have one or more values for respective fields of a plurality of fields;
storing, in a storage medium of the computing system, data type information that associates each
of one or more data types with at least one identifier; and
processing, using at least one processor of the computing system, a plurality of data values from
the records, the processing including:
generating a plurality of data units from the records, each data unit including a field
identifier that uniquely identifies one of the fields and a binary value from one of
the records, the binary value extracted from the field of that record identified by
the field identifier;
aggregating information about binary values from a plurality of the data units;
generating a list of entries for each of one or more of the fields, at least some of the
entries each including one of the binary values and information about that binary
value aggregated from a plurality of the data units;
retrieving a data type associated with a first identifier from the data type information, for
type-dependent processing, and associating the retrieved data type with at least
one binary value included in an entry of one of the lists; and
generating profile information for at least one of the fields based at least in part on a
retrieved data type of a particular binary value appearing in the field, after
aggregating information about binary values from a plurality of the data units.

2. The method of claim 1, wherein the binary value extracted from the field of that record identified by the field identifier is extracted as an un-typed sequence of bits; and generating profile information for at least one of the fields based at least in part on a retrieved data type of a particular binary value appearing in the field includes reinterpreting the un-typed sequence of bits as a typed data value that has the retrieved data type.
3. The method of claim 2, wherein the profile information includes type-dependent profiling results that depend on original data types of the plurality of data values from the records.
4. The method of claim 1, wherein aggregating information about binary values from a plurality of the data units includes comparing binary values from the plurality of data units with binary values in the list of entries to determine if there is a match between the binary values.
5. The method of claim 4, wherein the information about that binary value aggregated from a plurality of the data units includes a total count of matched binary values that is incremented each time a match is determined when comparing the binary values.
6. The method of claim 4, wherein a match between a first binary value and a second binary value corresponds a sequence of bits comprising the first binary value being identical to a sequence of bits comprising the second binary value.
7. The method of claim 1, wherein the data type information associates each of one or more data types with at least one of the field identifiers.
8. The method of claim 7, wherein retrieving a data type associated with the first identifier from the data type information includes retrieving a data type associated with a first field identifier.
9. The method of claim 1, wherein each data unit includes: one of the field identifiers, a binary value from one of the records, and a data type identifier that uniquely identifies one of the data types.

10. The method of claim 9, wherein the data type information associates each of one or more data types with at least one of the data type identifiers.

11. The method of claim 10, wherein retrieving a data type associated with the first identifier from the data type information includes retrieving a data type associated with a first data type identifier.

12. The method of claim 1, wherein associating the retrieved data type with at least one binary value included in an entry of one of the lists includes: instantiating a local variable having the retrieved data type and initializing the instantiated variable to a value based on the binary value included in the entry.

13. The method of claim 1, wherein associating the retrieved data type with at least one binary value included in an entry of one of the lists includes: setting a pointer associated with the retrieved data type to a memory location at which the binary value included in the entry is stored.

14. The method of claim 1, wherein each data unit includes: one of the field identifiers, a binary value from one of the records, and an indicator of the length of the binary value.

15. The method of claim 14, wherein the length indicator is stored as a prefix to the binary value.

16. The method of claim 1, further including receiving, over the input device or port of the computing system, record format information associated with the plurality of records.

17. The method of claim 16, wherein the data type information is generated based at least in part on the received record format information.

18. The method of claim 1, wherein the processing further includes: converting binary values from different respective entries of a first of the lists for a first field from a retrieved data type

associated with the binary values to a target data type, and converting binary values from different respective entries of a second of the lists for a second field from a retrieved data type associated with the binary values to the same target data type.

19. Software stored in a non-transitory form on a computer-readable medium, the software including instructions for causing a computing system to:

receive, over an input device or port of the computing system, a plurality of records that

each have one or more values for respective fields of a plurality of fields;

store, in a storage medium of the computing system, data type information that associates

each of one or more data types with at least one identifier; and

process, using at least one processor of the computing system, a plurality of data values

from the records, the processing including:

generating a plurality of data units from the records, each data unit including a

field identifier that uniquely identifies one of the fields and a binary value

from one of the records, the binary value extracted from the field of that

record identified by the field identifier;

aggregating information about binary values from a plurality of the data units;

generating a list of entries for each of one or more of the fields, at least some of

the entries each including one of the binary values and information about

that binary value aggregated from a plurality of the data units;

retrieving a data type associated with a first identifier from the data type

information, for type-dependent processing, and associating the retrieved

data type with at least one binary value included in an entry of one of the

lists; and

generating profile information for at least one of the fields based at least in part on

a retrieved data type of a particular binary value appearing in the field,

after aggregating information about binary values from a plurality of the

data units.

20. The software of claim 19, wherein aggregating information about binary values from a plurality of the data units includes comparing binary values from the plurality of data units with binary values in the list of entries to determine if there is a match between the binary values.

21. The software of claim 20, wherein the information about that binary value aggregated from a plurality of the data units includes a total count of matched binary values that is incremented each time a match is determined when comparing the binary values.

22. The software of claim 20, wherein a match between a first binary value and a second binary value corresponds a sequence of bits comprising the first binary value being identical to a sequence of bits comprising the second binary value.

23. The software of claim 19, wherein the data type information associates each of one or more data types with at least one of the field identifiers.

24. The software of claim 23, wherein retrieving a data type associated with the first identifier from the data type information includes retrieving a data type associated with a first field identifier.

25. The software of claim 19, wherein each data unit includes: one of the field identifiers, a binary value from one of the records, and a data type identifier that uniquely identifies one of the data types.

26. The software of claim 25, wherein the data type information associates each of one or more data types with at least one of the data type identifiers.

27. The software of claim 26, wherein retrieving a data type associated with the first identifier from the data type information includes retrieving a data type associated with a first data type identifier.

28. The software of claim 19, wherein associating the retrieved data type with at least one binary value included in an entry of one of the lists includes: instantiating a local variable having the retrieved data type and initializing the instantiated variable to a value based on the binary value included in the entry.

29. The software of claim 19, wherein associating the retrieved data type with at least one binary value included in an entry of one of the lists includes: setting a pointer associated with the retrieved data type to a memory location at which the binary value included in the entry is stored.

30. The software of claim 19, wherein each data unit includes: one of the field identifiers, a binary value from one of the records, and an indicator of the length of the binary value.

31. The software of claim 30, wherein the length indicator is stored as a prefix to the binary value.

32. The software of claim 19, further including instructions for causing the computing system to receive, over the input device or port, record format information associated with the plurality of records.

33. The software of claim 32, wherein the data type information is generated based at least in part on the received record format information.

34. The software of claim 19, wherein the processing further includes: converting binary values from different respective entries of a first of the lists for a first field from a retrieved data type associated with the binary values to a target data type, and converting binary values from different respective entries of a second of the lists for a second field from a retrieved data type associated with the binary values to the same target data type.

35. A computing system including:

an input device or port of the computing system configured to receive a plurality of records that

each have one or more values for respective fields of a plurality of fields;
a storage medium of the computing system configured to store data type information that
associates each of one or more data types with at least one identifier;
and at least one processor of the computing system configured to process a plurality of data
values from the records, the processing including:
generating a plurality of data units from the records, each data unit including a field
identifier that uniquely identifies one of the fields and a binary value from one of
the records, the binary value extracted from the field of that record identified by
the field identifier;
aggregating information about binary values from a plurality of the data units;
generating a list of entries for each of one or more of the fields, at least some of the
entries each including one of the binary values and information about that binary
value aggregated from a plurality of the data units;
retrieving a data type associated with a first identifier from the data type information, for
type-dependent processing, and associating the retrieved data type with at least
one binary value included in an entry of one of the lists; and
generating profile information for at least one of the fields based at least in part on a
retrieved data type of a particular binary value appearing in the field, after
aggregating information about binary values from a plurality of the data units.

36. The computing system of claim 35, wherein the binary value extracted from the field of that
record identified by the field identifier is extracted as an un-typed sequence of bits; and
generating profile information for at least one of the fields based at least in part on a retrieved
data type of a particular binary value appearing in the field includes reinterpreting the un-typed
sequence of bits as a typed data value that has the retrieved data type.

37. The computing system of claim 36, wherein the profile information includes type-dependent profiling results that depend on original data types of the plurality of data values from the records.

38. The computing system of claim 35, wherein aggregating information about binary values from a plurality of the data units includes comparing binary values from the plurality of data units with binary values in the list of entries to determine if there is a match between the binary values.

39. The computing system of claim 38, wherein the information about that binary value aggregated from a plurality of the data units includes a total count of matched binary values that is incremented each time a match is determined when comparing the binary values.

40. The computing system of claim 38, wherein a match between a first binary value and a second binary value corresponds a sequence of bits comprising the first binary value being identical to a sequence of bits comprising the second binary value.

41. The computing system of claim 35, wherein the data type information associates each of one or more data types with at least one of the field identifiers.

42. The computing system of claim 41, wherein retrieving a data type associated with the first identifier from the data type information includes retrieving a data type associated with a first field identifier.

43. The computing system of claim 35, wherein each data unit includes: one of the field identifiers, a binary value from one of the records, and a data type identifier that uniquely identifies one of the data types.

44. The computing system of claim 43, wherein the data type information associates each of one or more data types with at least one of the data type identifiers.

45. The computing system of claim 44, wherein retrieving a data type associated with the first identifier from the data type information includes retrieving a data type associated with a first data type identifier.
46. The computing system of claim 35, wherein associating the retrieved data type with at least one binary value included in an entry of one of the lists includes: instantiating a local variable having the retrieved data type and initializing the instantiated variable to a value based on the binary value included in the entry.
47. The computing system of claim 35, wherein associating the retrieved data type with at least one binary value included in an entry of one of the lists includes: setting a pointer associated with the retrieved data type to a memory location at which the binary value included in the entry is stored.
48. The computing system of claim 35, wherein each data unit includes: one of the field identifiers, a binary value from one of the records, and an indicator of the length of the binary value.
49. The computing system of claim 48, wherein the length indicator is stored as a prefix to the binary value.
50. The computing system of claim 35, wherein the input device or port is further configured to receive record format information associated with the plurality of records.
51. The computing system of claim 50, wherein the data type information is generated based at least in part on the received record format information.
52. The computing system of claim 35, wherein the processing further includes: converting binary values from different respective entries of a first of the lists for a first field from a retrieved data type associated with the binary values to a target data type, and converting binary values from different respective entries of a second of the lists for a second field from a retrieved

data type associated with the binary values to the same target data type.

53. A computing system including:

means for receiving a plurality of records that each have one or more values for respective fields of a plurality of fields;

means for storing data type information that associates each of one or more data types with at least one identifier; and

means for processing a plurality of data values from the records, the processing including:

generating a plurality of data units from the records, each data unit including a field identifier that uniquely identifies one of the fields and a binary value from one of the records, the binary value extracted from the field of that record identified by the field identifier;

aggregating information about binary values from a plurality of the data units;

generating a list of entries for each of one or more of the fields, at least some of the entries each including one of the binary values and information about that binary value aggregated from a plurality of the data units;

retrieving a data type associated with a first identifier from the data type information, for type-dependent processing, and associating the retrieved data type with at least one binary value included in an entry of one of the lists; and

generating profile information for at least one of the fields based at least in part on a retrieved data type of a particular binary value appearing in the field, after aggregating information about binary values from a plurality of the data units.

54. The software of claim 19, wherein the binary value extracted from the field of that record identified by the field identifier is extracted as an un-typed sequence of bits; and generating

profile information for at least one of the fields based at least in part on a retrieved data type of a particular binary value appearing in the field includes reinterpreting the un-typed sequence of bits as a typed data value that has the retrieved data type.

55. The software of claim 54, wherein the profile information includes type-dependent profiling results that depend on original data types of the plurality of data values from the records.