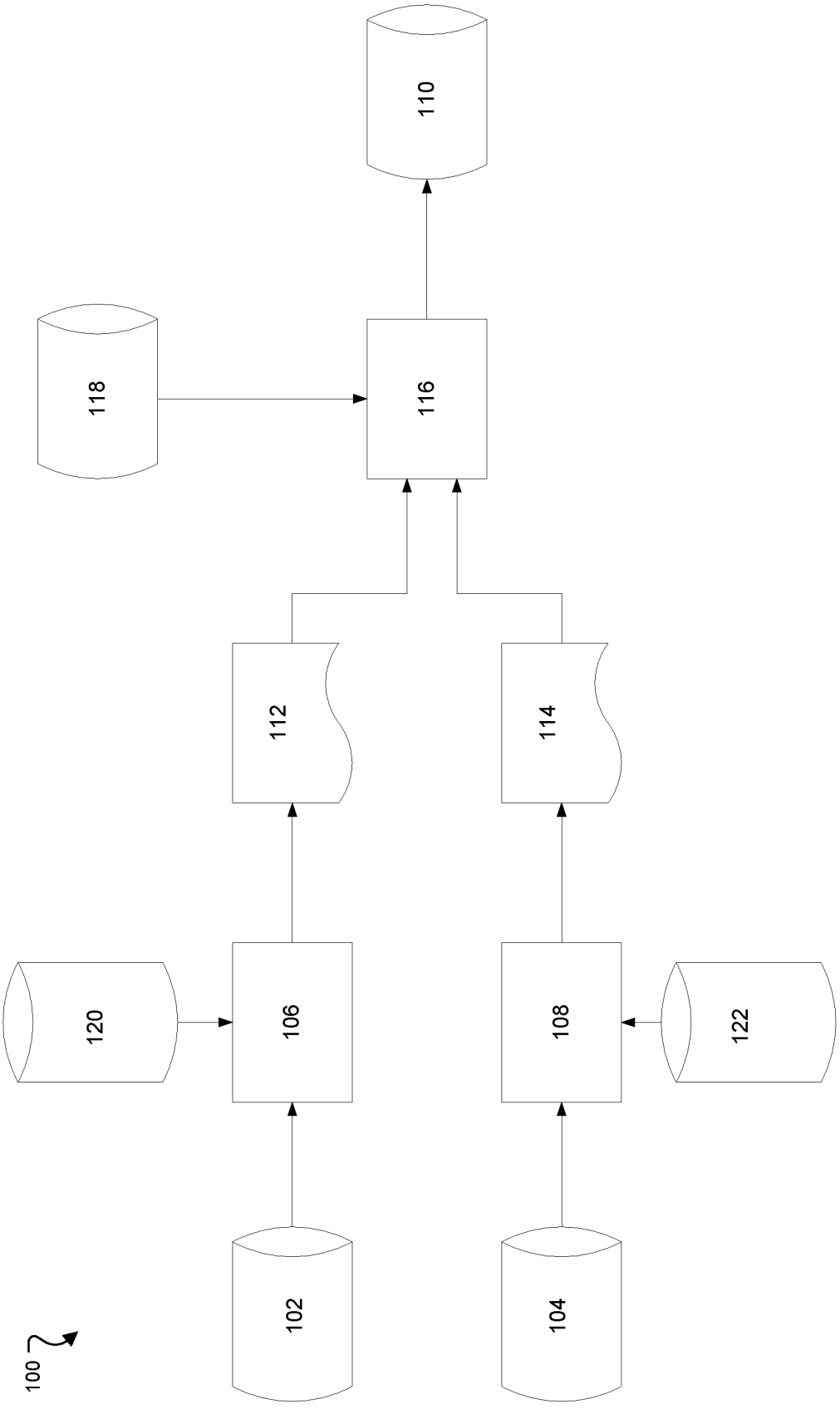

FIG. 1

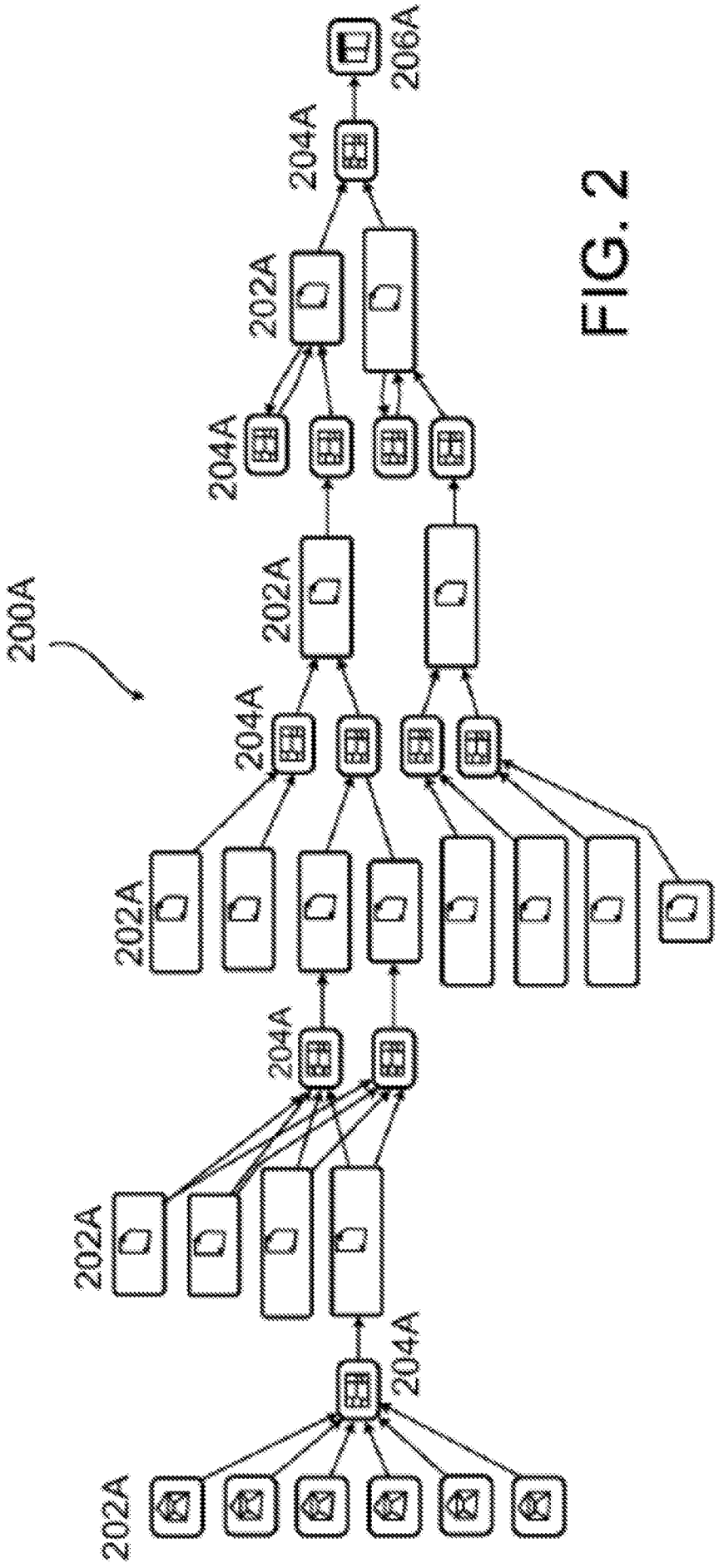

FIG. 2

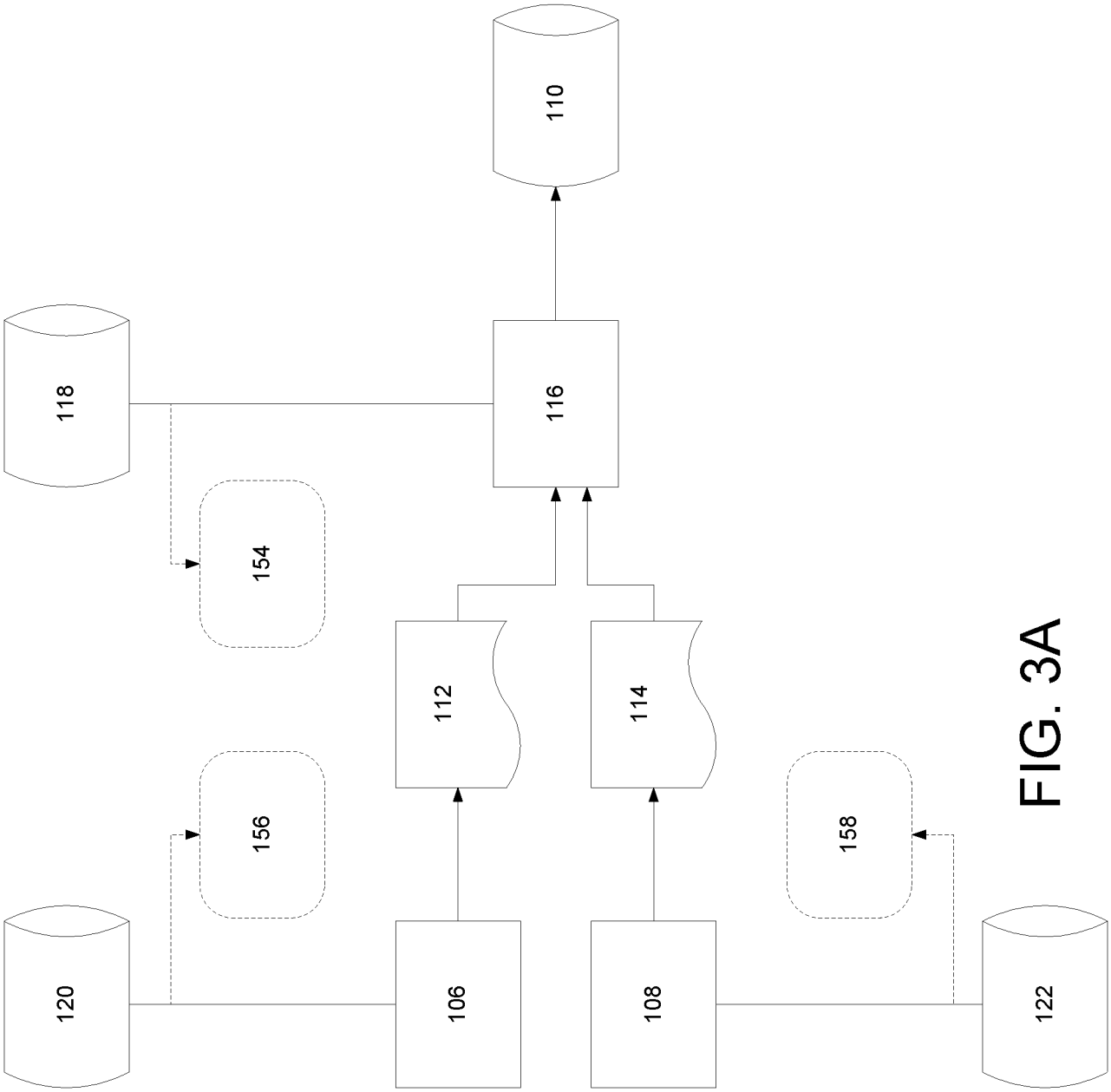

FIG. 3A

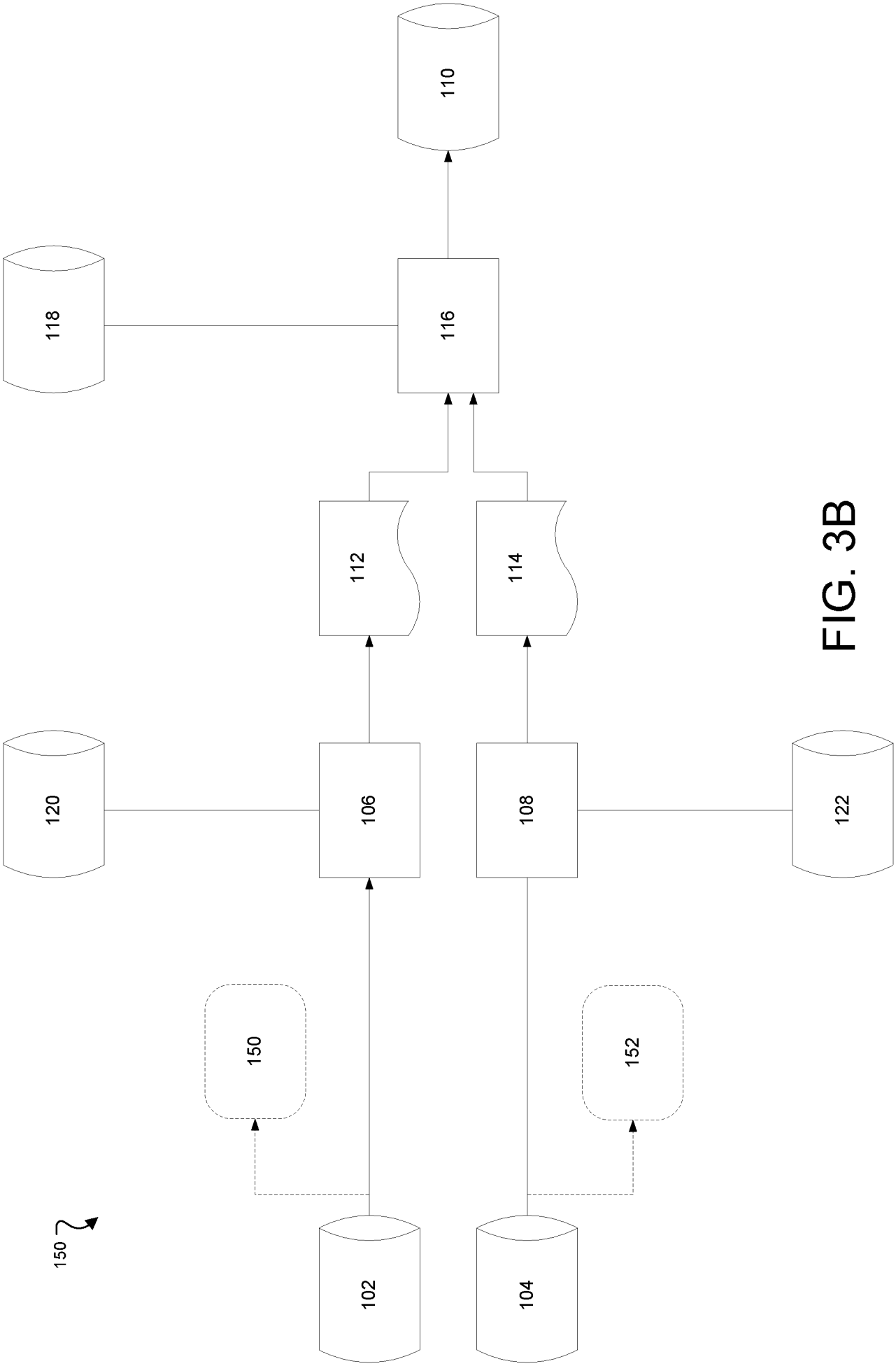

FIG. 3B

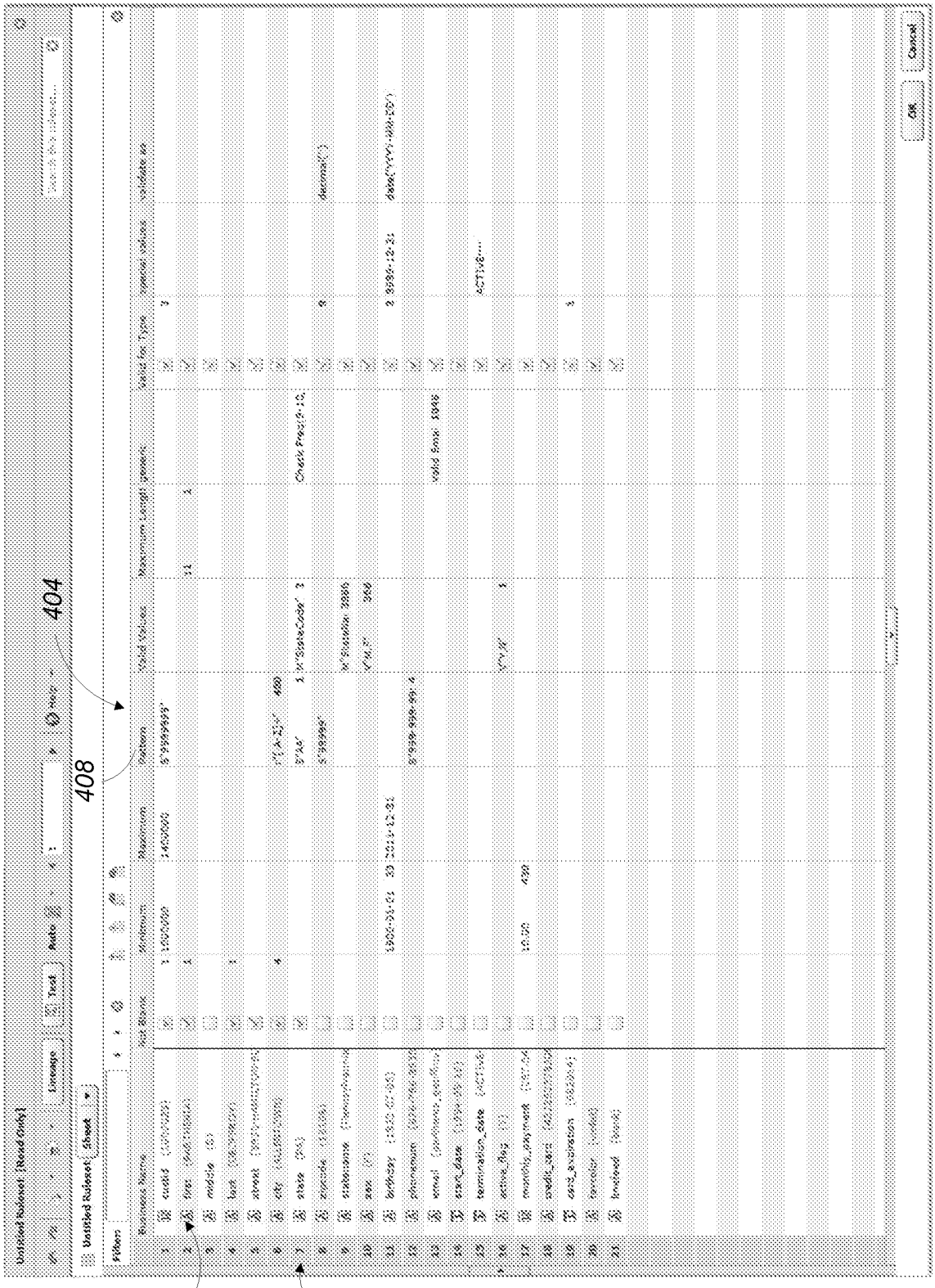

FIG. 4

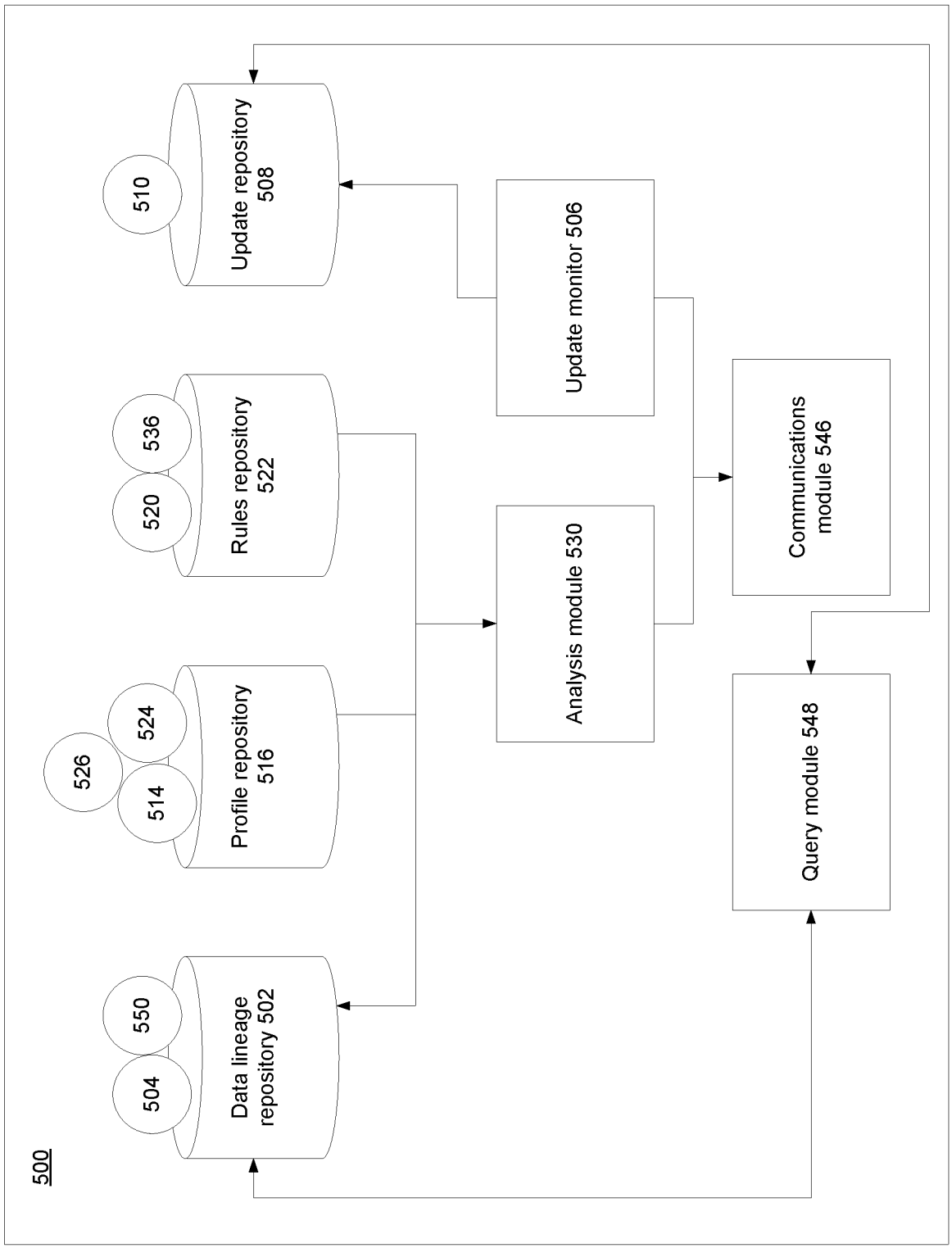
500
Data lineage repository 502
504
550
Profile repository 516
514
524
526
Rules repository 522
520
536
Update repository 508
510
Analysis module 530
Update monitor 506
Communications module 546
Query module 548

FIG. 5

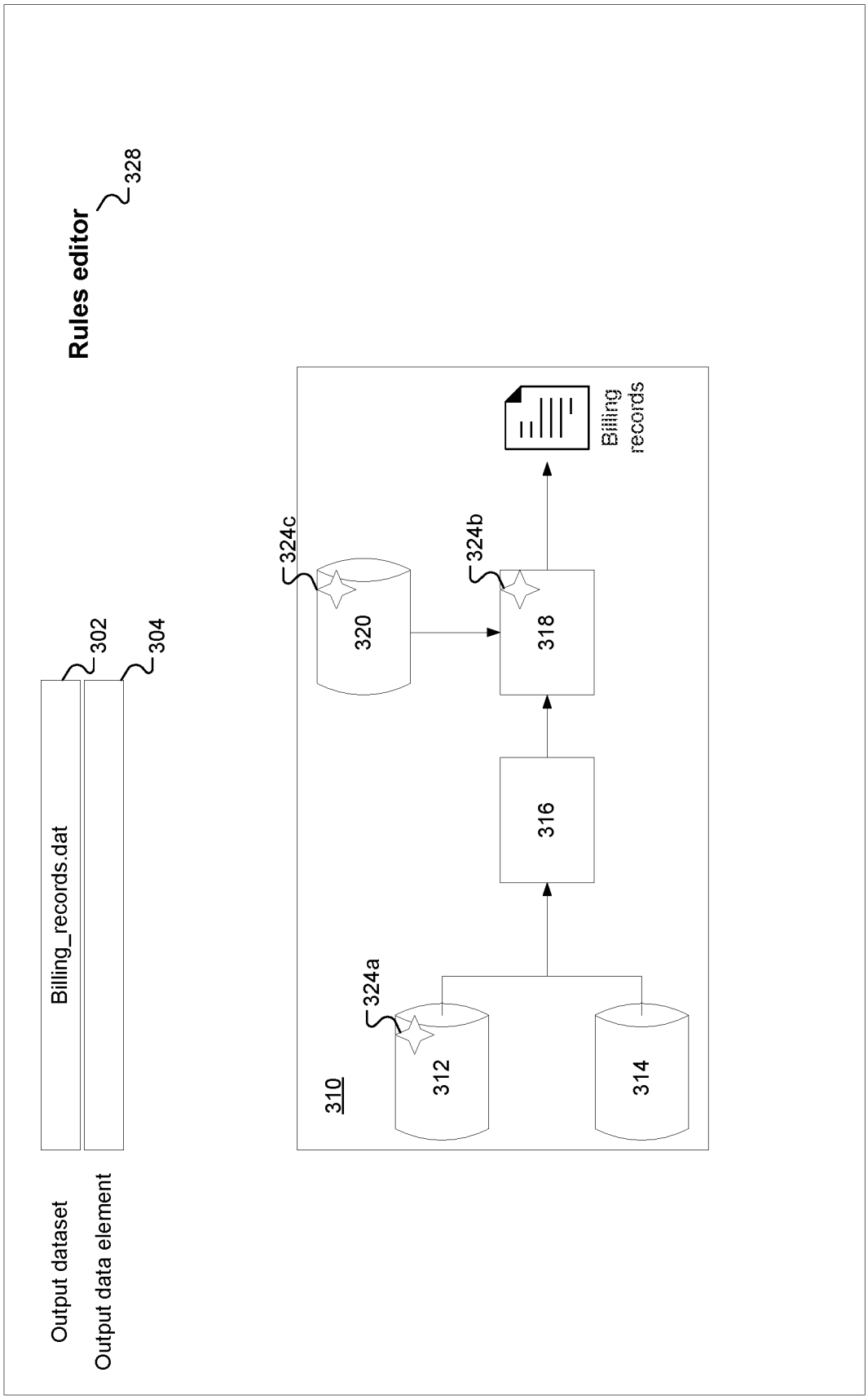
Rules editor
328
Billing_records.dat
302
304
Output dataset
Output data element
310
312
314
316
318
320
324a
324b
324c
Billing records

FIG. 6

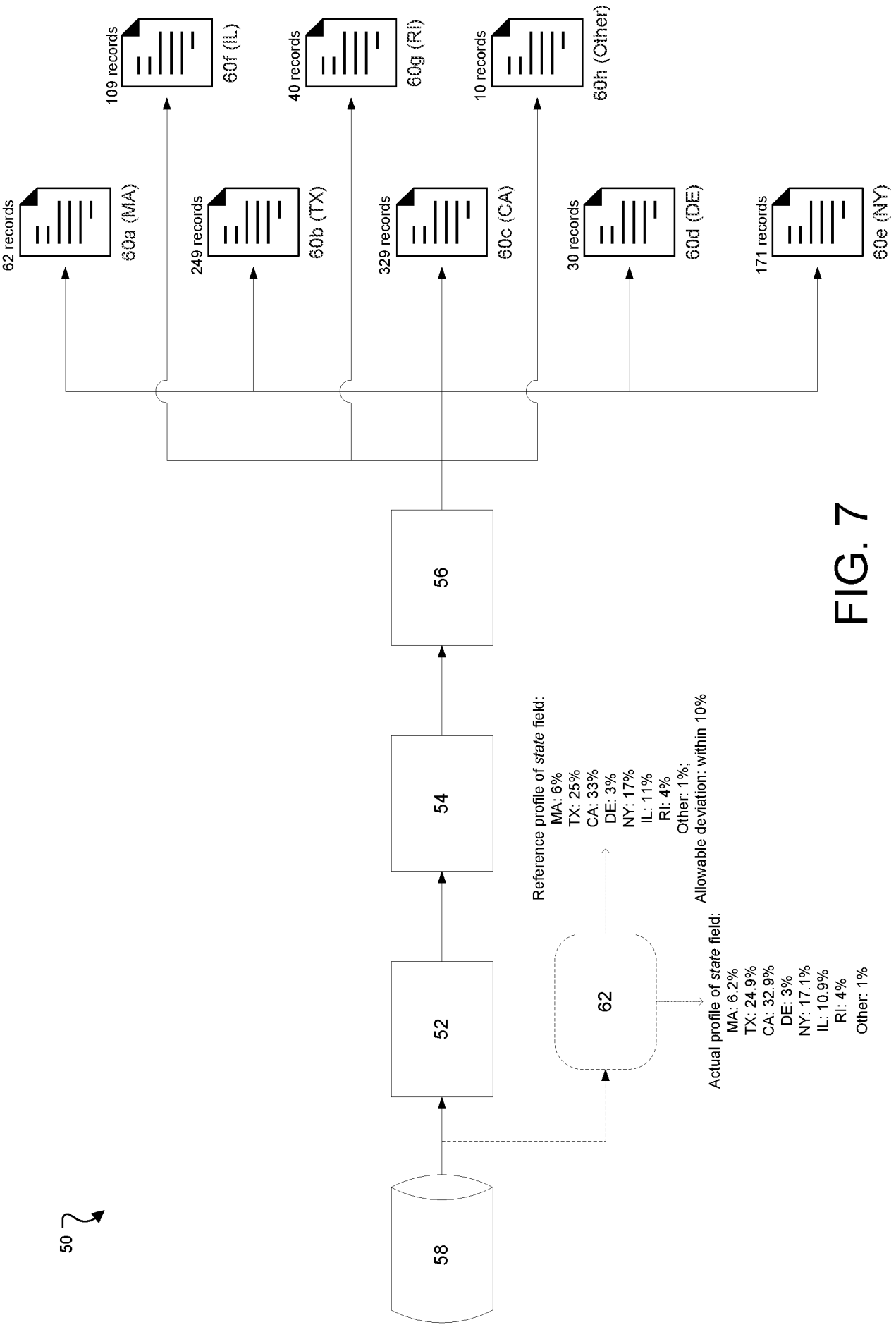
50
58
52
54
56
62
Reference profile of state field:
MA: 6%
TX: 25%
CA: 33%
DE: 3%
NY: 17%
IL: 11%
RI: 4%
Other: 1%;
Allowable deviation: within 10%
Actual profile of state field:
MA: 6.2%
TX: 24.9%
CA: 32.9%
DE: 3%
NY: 17.1%
IL: 10.9%
RI: 4%
Other: 1%
62 records
60a (MA)
249 records
60b (TX)
329 records
60c (CA)
30 records
60d (DE)
171 records
60e (NY)
109 records
60f (IL)
40 records
60g (RI)
10 records
60h (Other)

FIG. 7

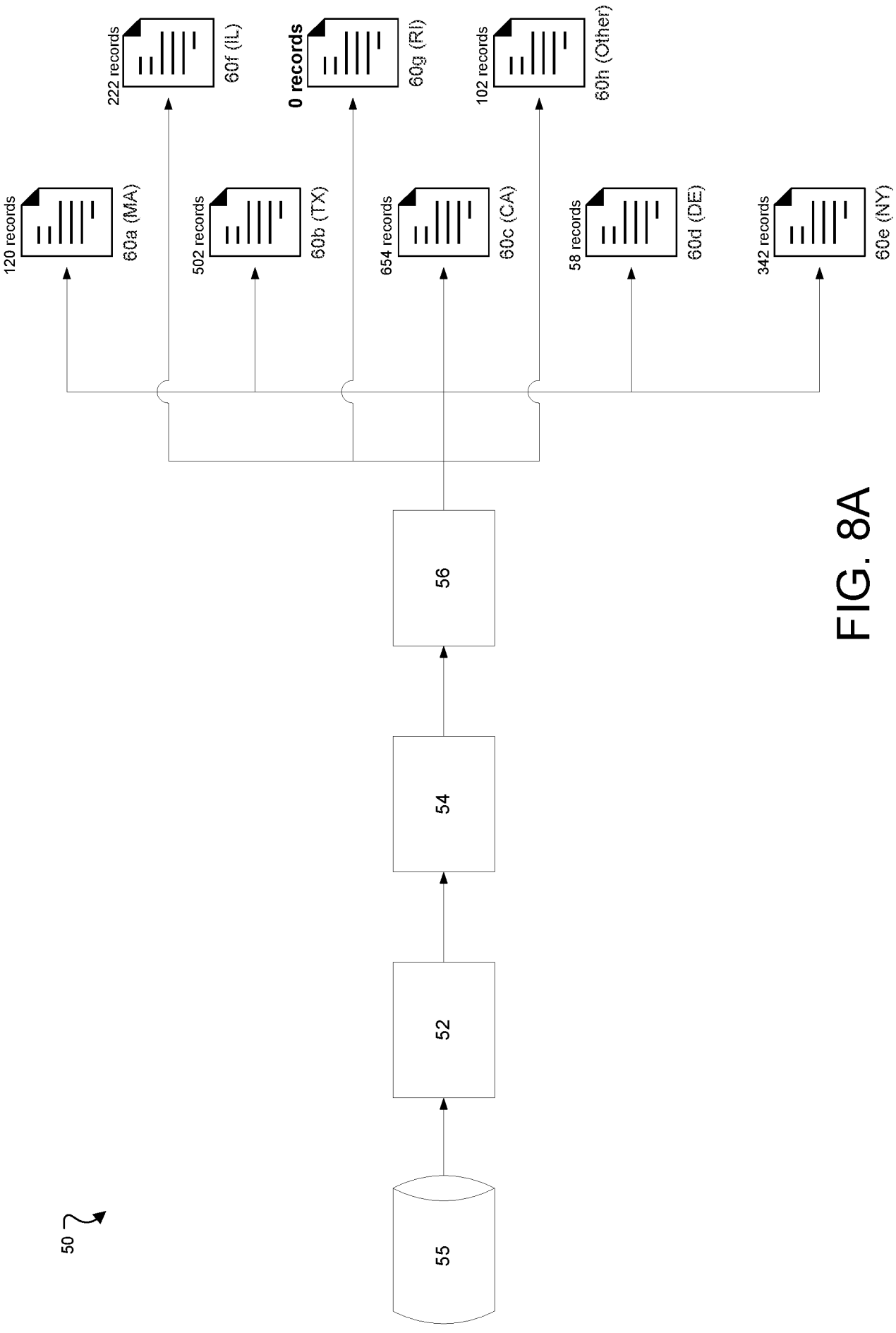
50
55
52
54
56
120 records
60a (MA)
502 records
60b (TX)
654 records
60c (CA)
58 records
60d (DE)
342 records
60e (NY)
222 records
60f (IL)
0 records
60g (RI)
102 records
60h (Other)

FIG. 8A

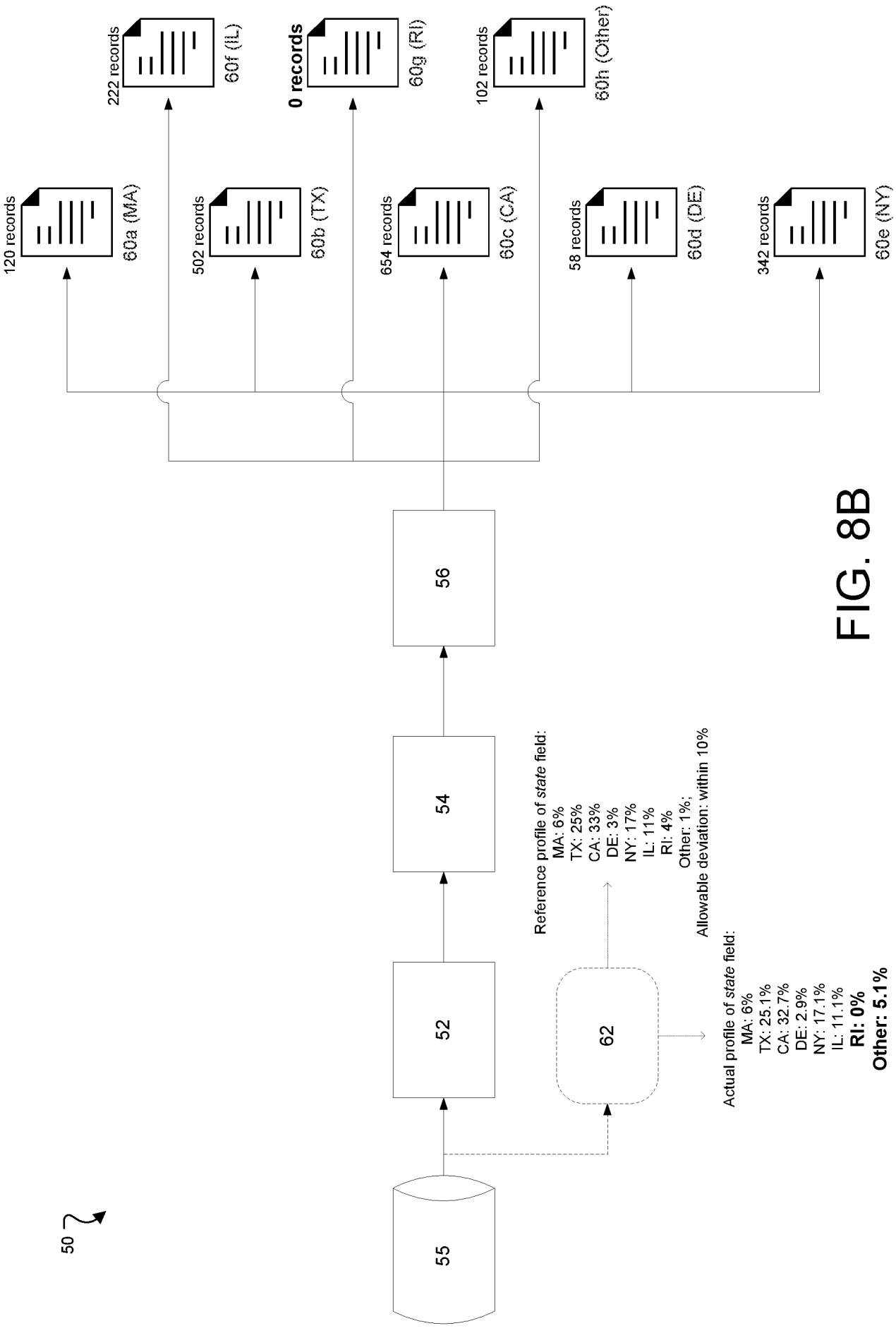
50
55
52
54
56
62
Reference profile of state field:
MA: 6%
TX: 25%
CA: 33%
DE: 3%
NY: 17%
IL: 11%
RI: 4%
Other: 1%;
Allowable deviation: within 10%
Actual profile of state field:
MA: 6%
TX: 25.1%
CA: 32.7%
DE: 2.9%
NY: 17.1%
IL: 11.1%
RI: 0%
Other: 5.1%
120 records
60a (MA)
502 records
60b (TX)
654 records
60c (CA)
58 records
60d (DE)
342 records
60e (NY)
222 records
60f (IL)
0 records
60g (RI)
102 records
60h (Other)

FIG. 8B

Cust_ID	State	Purchase_price
12575	CA	127.51
35468	MA	547.77
26842	NY	29.80
32458	TX	245.01
54893	IL	55.76
05786	TX	18.76
65420	CA	442.10
12475	CA	54.30
54978	NY	22.99
97845	IR	65.17
65432	IL	18.19
54983	CA	25.44
32877	TX	509.60
54446	TX	63.12
23158	CA	45.18
35458	NY	56.04
87653	IR	17.10
65487	NY	152.60
32455	CA	146.12
54877	IL	33.32
78361	TX	423.01
12756	DE	18.76
45886	TX	87.45
32115	CA	108.69
68575	CA	146.08

FIG. 8C

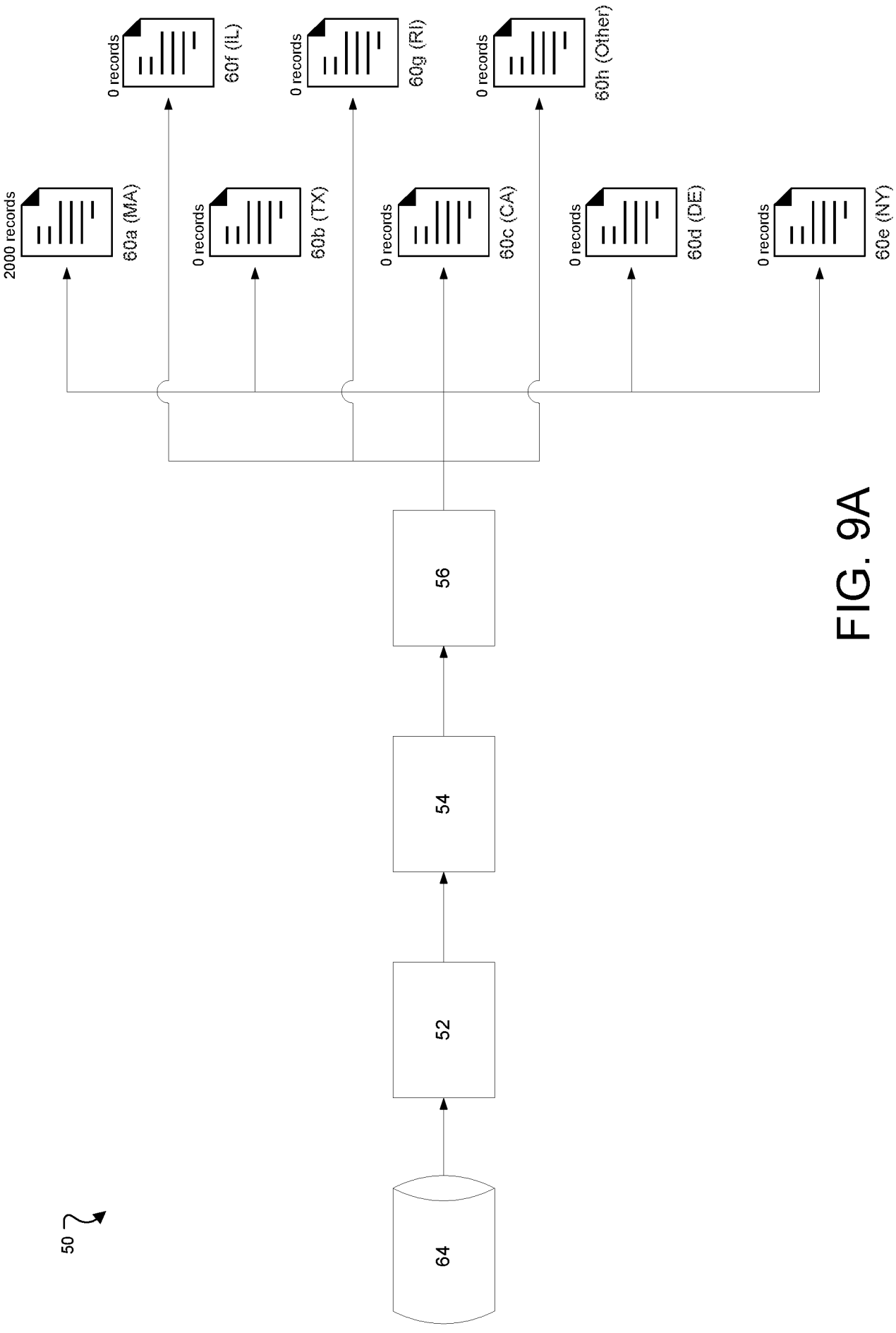
50
64
52
54
56
2000 records
60a (MA)
0 records
60b (TX)
0 records
60c (CA)
0 records
60d (DE)
0 records
60e (NY)
0 records
60f (IL)
0 records
60g (RI)
0 records
60h (Other)

FIG. 9A

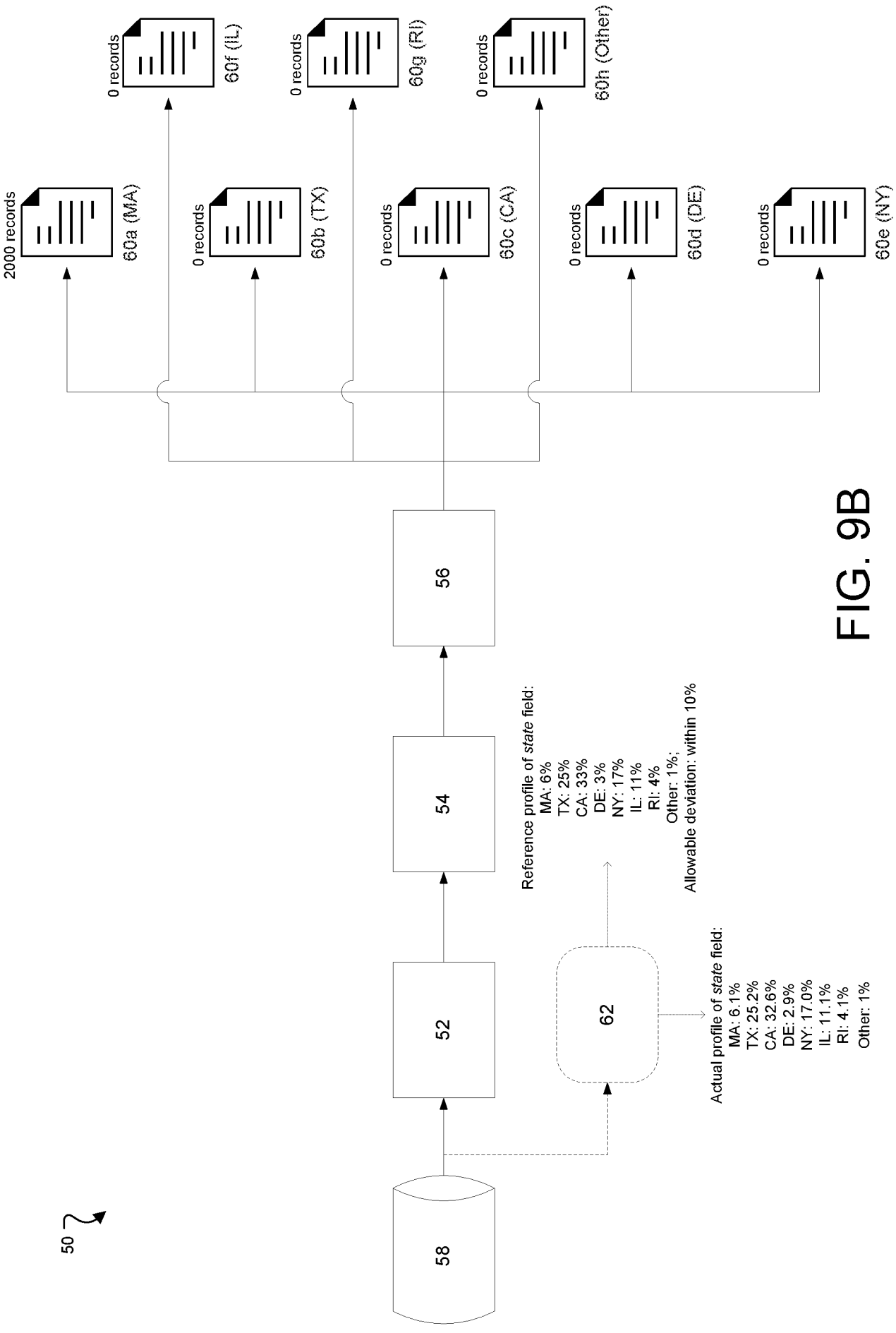
50
58
52
54
56
62
Reference profile of state field:
MA: 6%
TX: 25%
CA: 33%
DE: 3%
NY: 17%
IL: 11%
RI: 4%
Other: 1%;
Allowable deviation: within 10%
Actual profile of state field:
MA: 6.1%
TX: 25.2%
CA: 32.6%
DE: 2.9%
NY: 17.0%
IL: 11.1%
RI: 4.1%
Other: 1%
2000 records
60a (MA)
0 records
60b (TX)
0 records
60c (CA)
0 records
60d (DE)
0 records
60e (NY)
0 records
60f (IL)
0 records
60g (RI)
0 records
60h (Other)

FIG. 9B

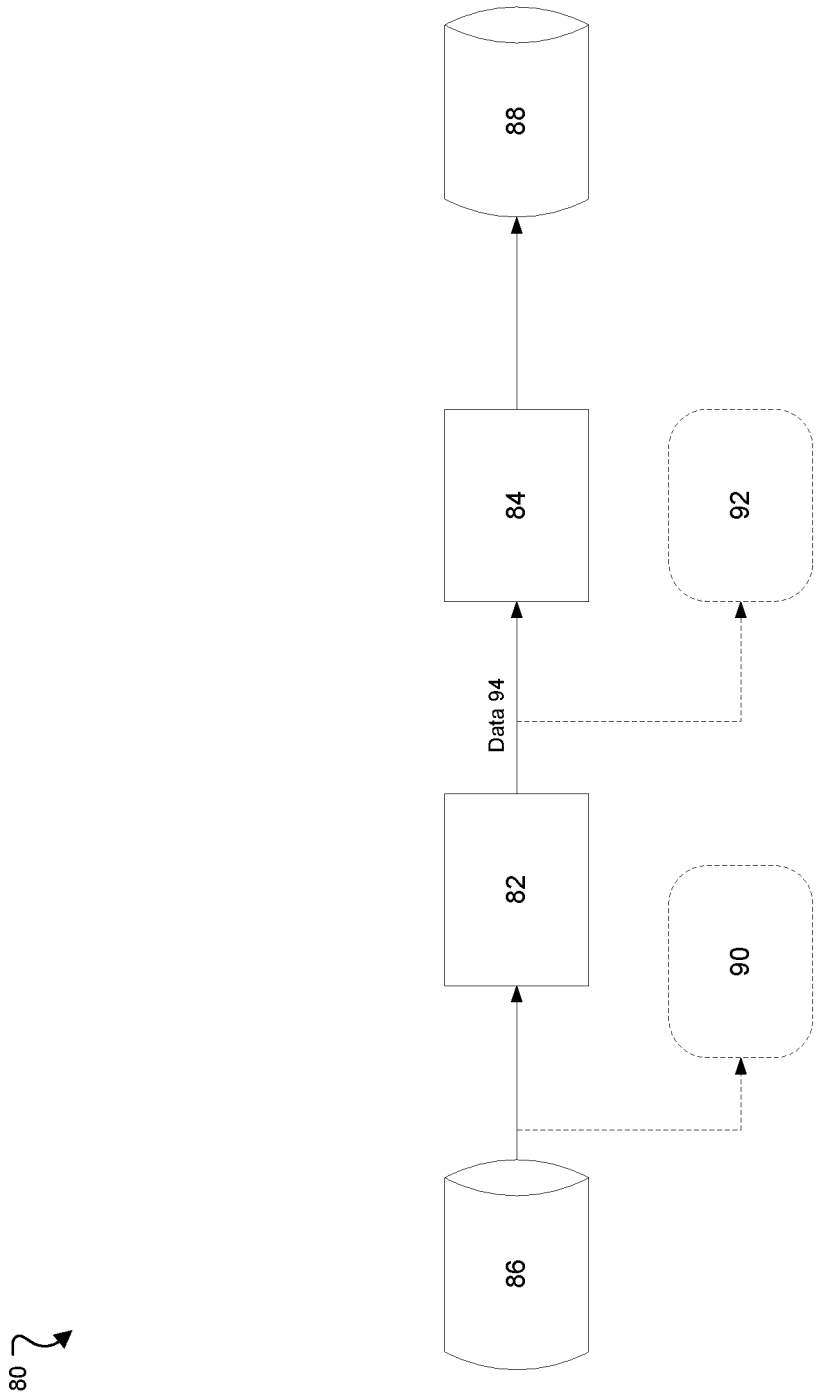

FIG. 10A

94

Phone_number	Start_time	Duration
6508001245	12:45:00 PM	0:30
4082604525	12:45:08 PM	1:15
212508412E	12:45:12 PM	10:12
6174438771	12:45:18 PM	8:07
3024023548	12:45:22 PM	0:02
21445893546	12:45:25 PM	12:17
972450546S	12:45:30 PM	8:32
917508459E	12:45:30 PM	0:09
7738645947	12:45:32 PM	1:12
2105782156	12:45:35 PM	4:01
4085647895	12:45:38 PM	0:14
6174546824	12:45:40 PM	2:08
214568778F	12:45:41 PM	14:16
7735684456	12:45:42 PM	9:12
6175443215	12:45:45 PM	58:17
212565788E	12:45:46 PM	7:26
3025488865	12:45:49 PM	0:16
9175453256	12:45:50 PM	1:38

FIG. 10B

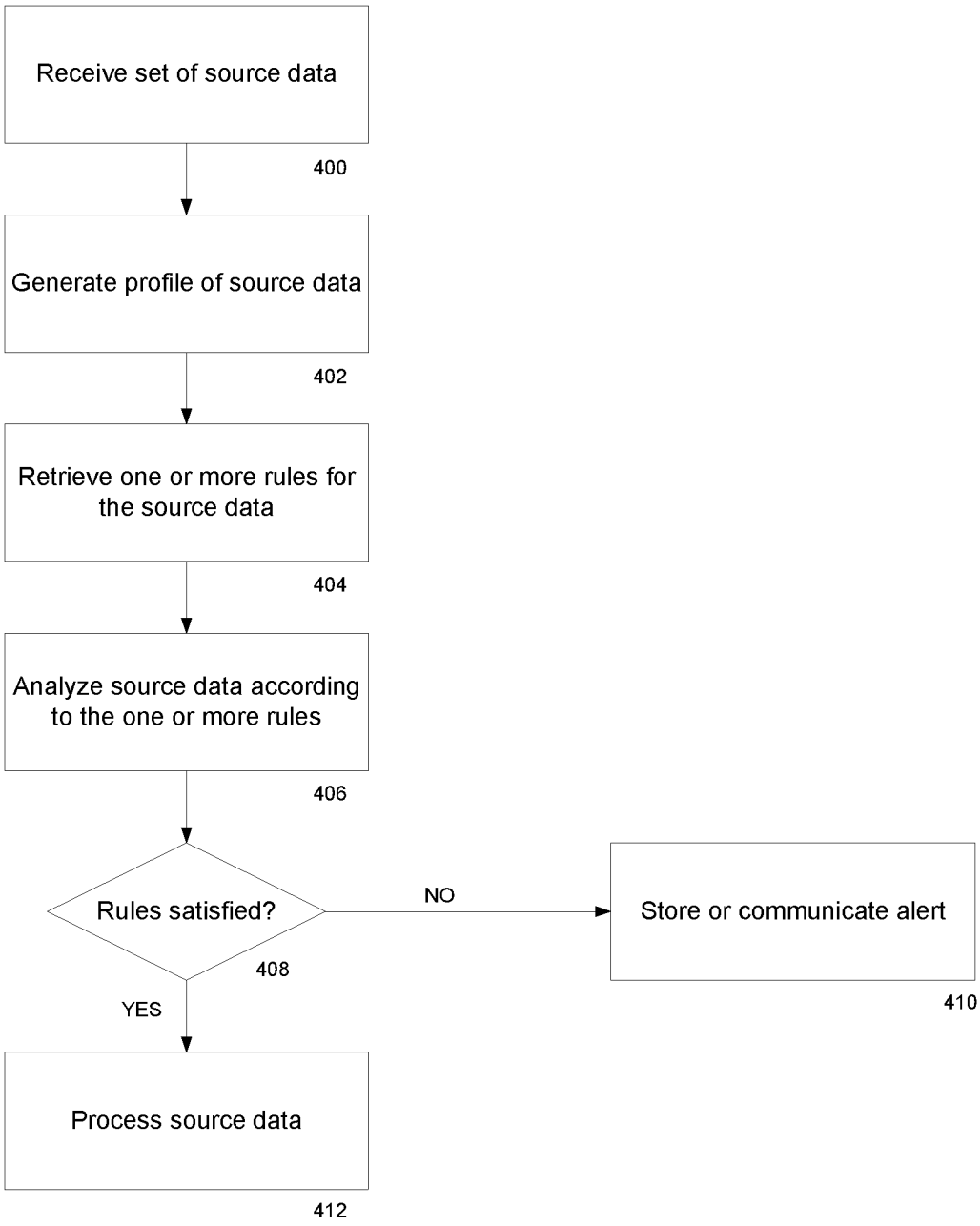
Receive set of source data
400
Generate profile of source data
402
Retrieve one or more rules for the source data
404
Analyze source data according to the one or more rules
406
Rules satisfied?
408
NO
Store or communicate alert
410
YES
Process source data
412

FIG. 11

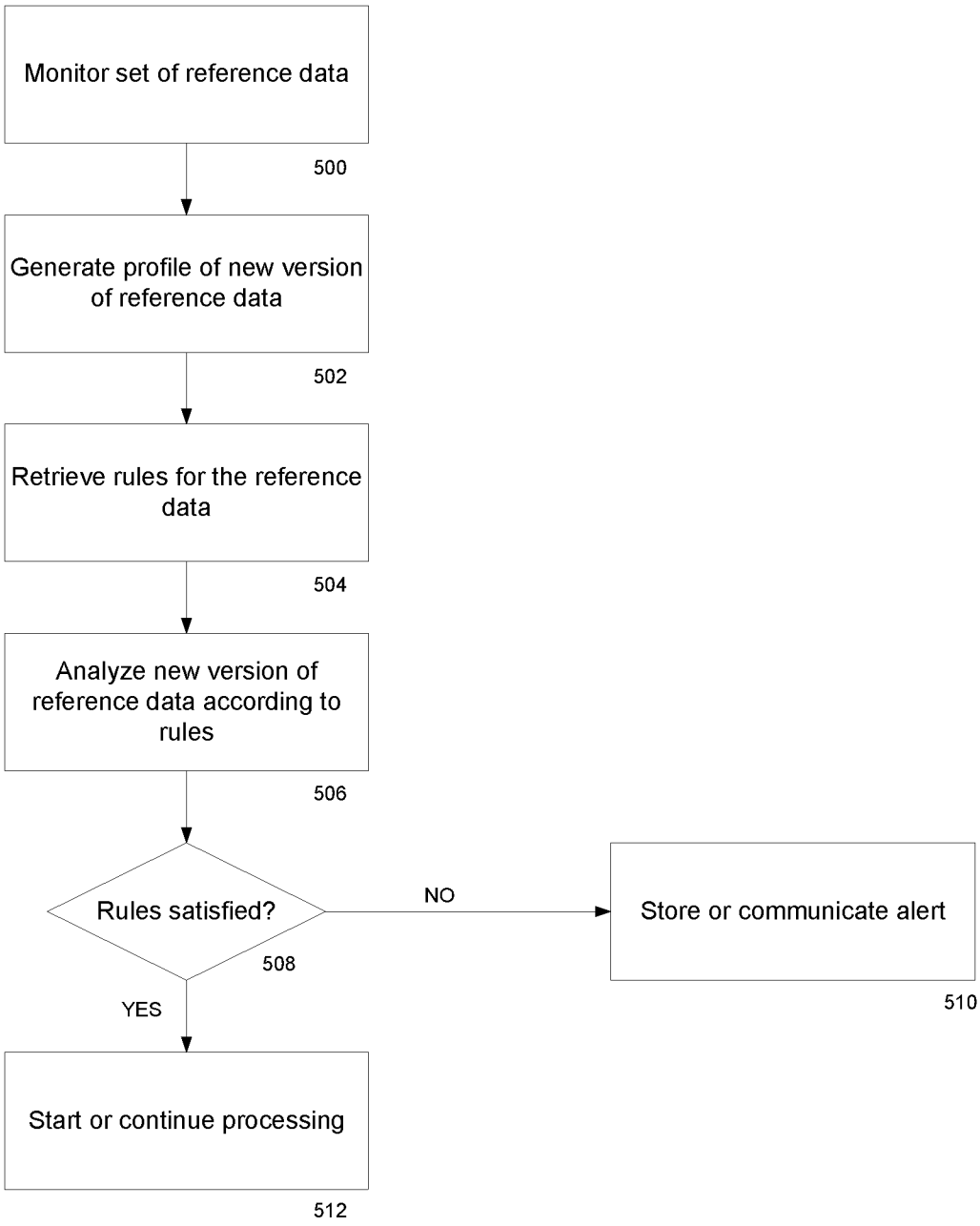
Monitor set of reference data
500
Generate profile of new version of reference data
502
Retrieve rules for the reference data
504
Analyze new version of reference data according to rules
506
Rules satisfied?
508
NO
Store or communicate alert
510
YES
Start or continue processing
512

FIG. 12

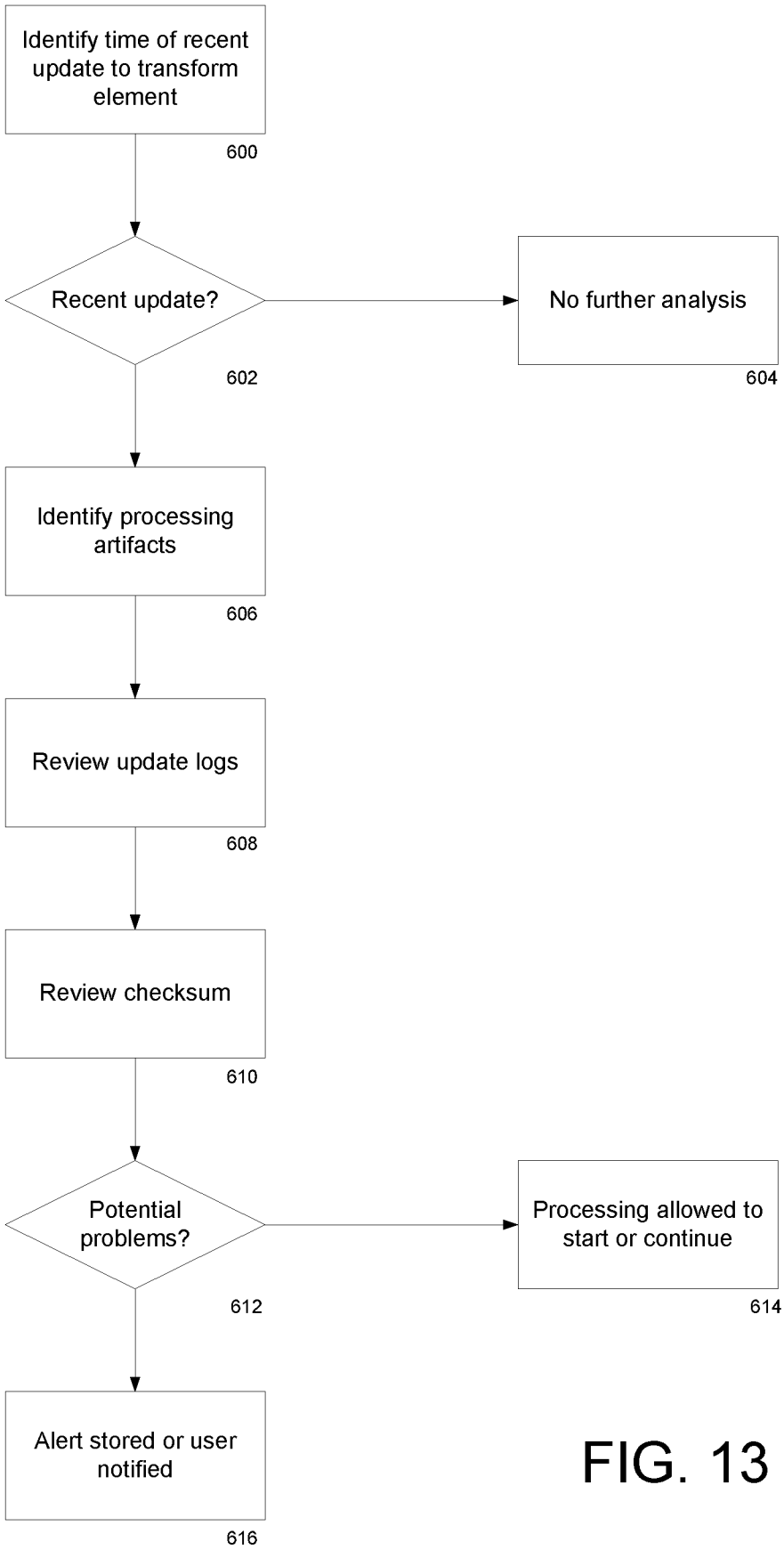
Identify time of recent update to transform element
600
Recent update?
602
No further analysis
604
Identify processing artifacts
606
Review update logs
608
Review checksum
610
Potential problems?
612
Processing allowed to start or continue
614
Alert stored or user notified
616

FIG. 13

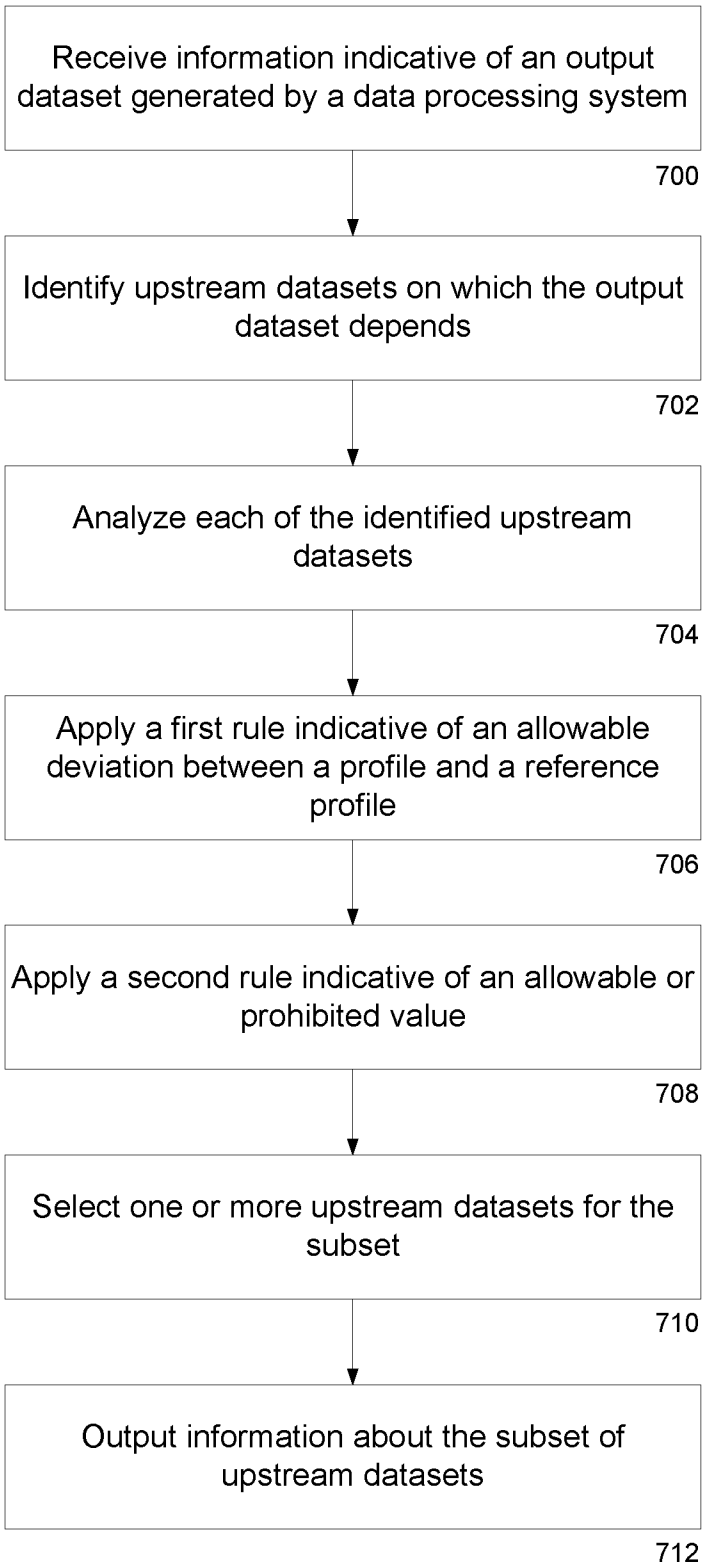
Receive information indicative of an output dataset generated by a data processing system
700
Identify upstream datasets on which the output dataset depends
702
Analyze each of the identified upstream datasets
704
Apply a first rule indicative of an allowable deviation between a profile and a reference profile
706
Apply a second rule indicative of an allowable or prohibited value
708
Select one or more upstream datasets for the subset
710
Output information about the subset of upstream datasets
712

FIG. 14

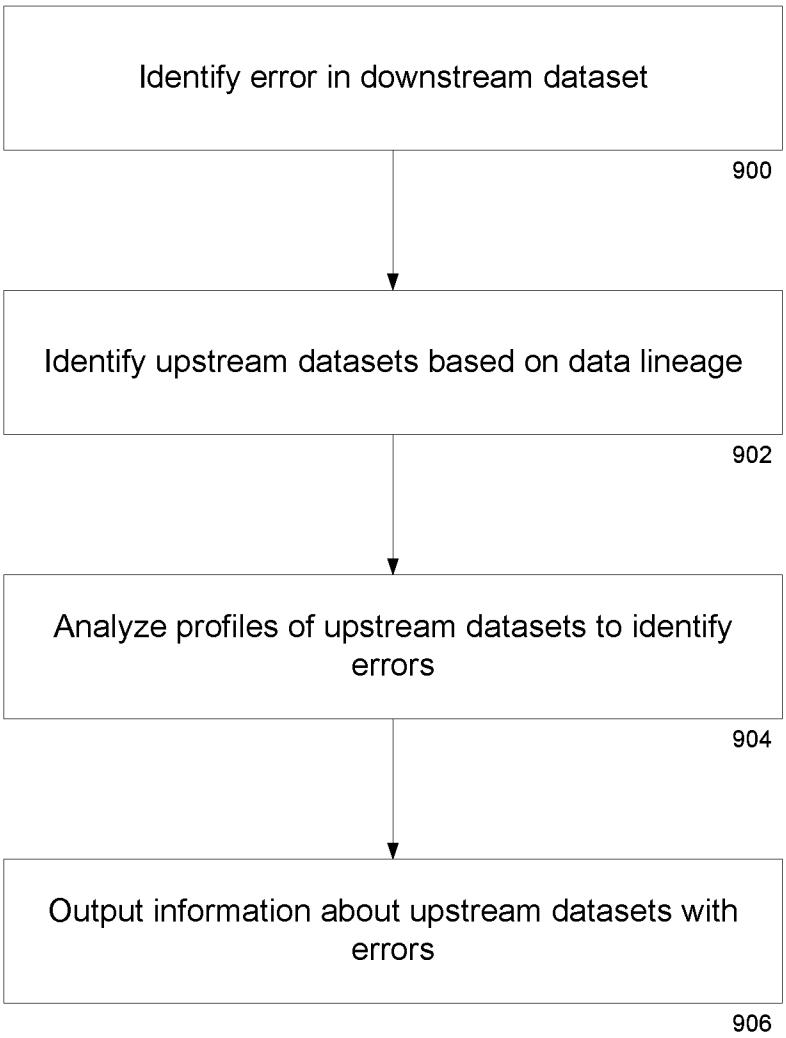
Identify error in downstream dataset
900
Identify upstream datasets based on data lineage
902
Analyze profiles of upstream datasets to identify errors
904
Output information about upstream datasets with errors
906

FIG. 15

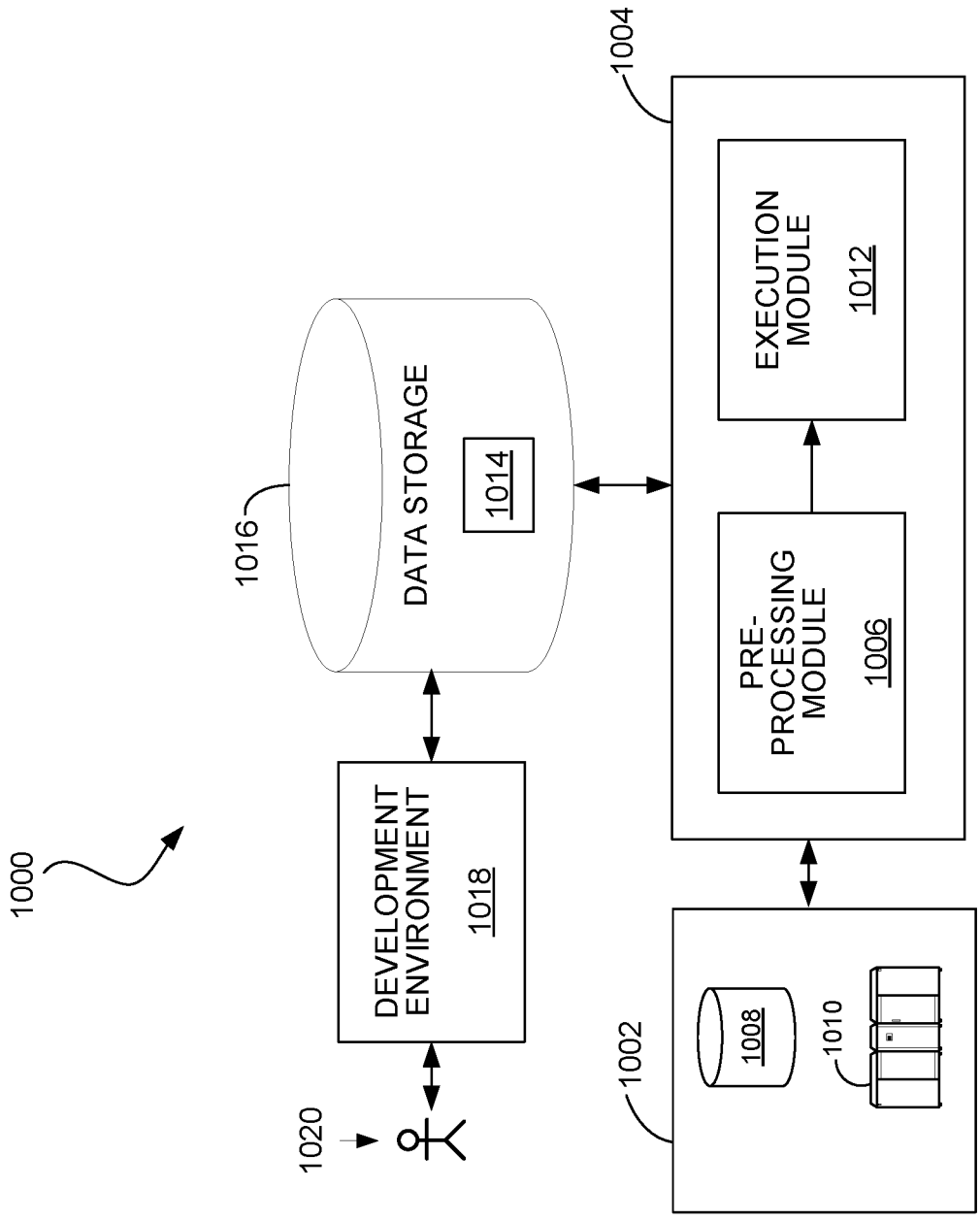
1000
1020
DEVELOPMENT ENVIRONMENT
1018
1016
DATA STORAGE
1014
1004
EXECUTION MODULE
1012
PRE-PROCESSING MODULE
1006
1002
1008
1010

FIG. 16