Distribution of Matrix Values

Issue

Skim matrices by their very nature contain an exponential amount of information. For the TIM 2.3 zonal system, there are 3,574 zones which equates to 12.8 million (12,773,476) zone pairs. Further increasing the amount of data are multiple skim types (roughly 48) and a temporal component (4 time periods). In summation, to store just the skim matrix data for a completed model run inside of a database, an estimated 2.5 billion (2,452,507,392) rows need to be inserted.

Each record at a minimum requires fields for origin and destination zone identifiers and the corresponding skim value. To make the table usable, two zonal indices are added since nearly all queries will have at least one zonal |WHERE| clause.

Originally, the table was set up with a |bigserial| counter field, |integer| zone identifiers, and |double precision| values. The required disk space to store a complete skim matrix given the table definition is 636 MB. With the added primary key and zonal indices, the grand total is 1457 MB. For a full model run the disk space required balloons to an estimated 279.7 GB.

Spacing Saving Strategies

1. Redefine Table  
   Both the zone identifiers can be set as |smallint| data types which are 2 bytes in size. The skim value could be set as |real| (at a cost of 4 bytes) but a higher degree of precision is desired; skim values are saved as 8 byte |double precision| data types. While redefining the data types for existing fields helps, dropping the boilerplate |bigserial| primary key yields the most significant disk savings.
2. Omit Certain Values  
   By selecting a common value to simply omit, significant disk savings can be achieved – this omission can be accounted for in subsequent queries and analysis. Not only is disk space saved due to the missing data but the associated indices will also see significant savings.

Skim values are always zero or positive. At the extrema, the upper bound of values tends to be 999,999 – this is the default ‘high’ value the TIM 2.3 model uses to denote impassible conditions. However due to impedance averaging/internal processes this value may be some fraction of the default high value although this value is still typically larger than 100,000.

Analyzing the distribution of values in the base model run for TIM 2.3, almost all matrices have significantly more high values than zero values. The few exceptions were matrices that were either blank or ‘continuous’ in distribution (such as PrT Skims where all of the zones are accessible).

Results

By changing the definition of the table, a 25% reduction is achieved in disk utilization. Looking at the split between table and index sizes, it is apparent that the |bigserial| primary key was using a large amount of space. Further significant savings in both initial insert run time and disk utilization can be achieved by omitting high values. In total, a 50.43% reduction in database size is accomplished (from 31.29 GB to 15.51 GB using an abbreviated set of skim matrices)

Table - Distribution of Matrix Values (ranged fields contain counts of values)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number | Code | 0-1 | 0 | >0-1 | 1-10 | 10-100 | 100-1,000 | 1,000-10,000 | 10,000-100,000 | 100,000+ |
| 210 | IMP | 3329 | 122 | 3207 | 263231 | 11833902 | 648009 | 3570 | 21435 | 0 |
| 220 | IVT | 3347 | 122 | 3225 | 268851 | 11941514 | 534759 | 3570 | 0 | 21435 |
| 250 | OVT | 21643 | 174 | 21469 | 12332607 | 397794 | 0 | 0 | 0 | 21432 |
| 260 | TOL | 6207472 | 5794393 | 413079 | 6537188 | 7381 | 0 | 0 | 0 | 21435 |
| 270 | DIS | 26050 | 122 | 25928 | 1161113 | 11475867 | 89011 | 0 | 0 | 21435 |
| 290 | TTC | 3347 | 122 | 3225 | 268851 | 11941514 | 534759 | 3570 | 0 | 21435 |
| 291 | UDS | 21643 | 174 | 21469 | 12332607 | 397794 | 0 | 0 | 0 | 21432 |
| 400 | IPD | 12773476 | 12773476 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 420 | IVT | 4024 | 184 | 3840 | 98591 | 4840816 | 1103065 | 0 | 0 | 6726980 |
| 421 | IVTT(RR) | 2419916 | 1858377 | 561539 | 699225 | 2867586 | 59769 | 0 | 0 | 6726980 |
| 422 | IVTT(Sub) | 3093295 | 2501534 | 591761 | 1348193 | 1605008 | 0 | 0 | 0 | 6726980 |
| 423 | IVTT(Pat) | 4788132 | 4606900 | 181232 | 764083 | 494281 | 0 | 0 | 0 | 6726980 |
| 424 | IVTT(LRT) | 4891532 | 4660090 | 231442 | 363614 | 791350 | 0 | 0 | 0 | 6726980 |
| 426 | IVTT(BRT) | 6046496 | 6046496 | 0 | 0 | 0 | 0 | 0 | 0 | 6726980 |
| 428 | IVTT(Bus) | 619034 | 389838 | 229196 | 795122 | 4444028 | 188312 | 0 | 0 | 6726980 |
| 429 | IVTT(Trl) | 6004273 | 5965782 | 38491 | 34916 | 7307 | 0 | 0 | 0 | 6726980 |
| 450 | OVT | 0 | 0 | 0 | 45778 | 5323883 | 676835 | 0 | 0 | 6726980 |
| 451 | OWTA | 91948 | 0 | 91948 | 2053005 | 3901543 | 0 | 0 | 0 | 6726980 |
| 460 | FAR | 61501 | 34922 | 26579 | 5735993 | 229631 | 0 | 0 | 0 | 6745213 |
| 480 | NTR | 803354 | 216101 | 587253 | 5243142 | 0 | 0 | 0 | 0 | 6726980 |
| 481 | XIMP | 6046496 | 6046496 | 0 | 0 | 0 | 0 | 0 | 0 | 6726980 |
| 490 | JRT | 0 | 0 | 0 | 8103 | 3042441 | 2995952 | 0 | 0 | 6726980 |

Table - Disk Usage by Table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table Name | Row Estimate | Total Size | Index Size | Table Size | Insert Time | Size Reduc. | Insert Reduc. |
| mtx\_420\_nt | 12773500 | 1457 MB | 821 MB | 636 MB | 164.16 s | 100% | 100% |
| mtx\_420\_nt\_small | 12773500 | 1087 MB | 547 MB | 539 MB | 156.07 s | 75% | 95% |
| mtx\_420\_nt\_small\_omit | 6046500 | 514 MB | 259 MB | 255 MB | 52.95 s | 35% | 32% |