ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Parallel andOptimization Aggregation in SQL ToOrganize Data sets for Data Mining Analysis

Mr.Fahad Bin Awad Quraishi
P.G Scholar,
Dept of CSE,
Hi Point College of Engineering & Technology.

Ms.Ifath Nazia Ghori
HoD,CS & IT,
Girls Community College,
Abu Arish, Jazan University,KSA.

ABSTRACT:

Preparing a data set for analysis isgenerally the most time consuming task in a data miningproject, requiring many complex SQL queries, joining tables,and aggregating columns. Existing SQL aggregations havelimitations to prepare data sets because they return onecolumn per aggregated group. In general, a significantmanual effort is required to build data sets, where ahorizontal layout is required. We propose simple, yetpowerful, methods to generate SQL code to returnaggregated columns in a horizontal tabular layout, returninga set of numbers instead of one number per row.

This newclass of functions is called horizontal aggregations.Horizontal aggregations build data sets with a horizontaldenormalized layout (e.g., point-dimension, observationvariable, instance-feature), which is the standard layoutrequired by most data mining algorithms. We propose threefundamental methods to evaluate horizontal aggregations:CASE: Exploiting the programming CASE construct; SPJ:Based on standard relational algebra operators (SPJ queries);PIVOT: Using the PIVOT operator, which is offered bysome DBMSs.

KEY WORDS:

Initial data Analysis, Characteristics of data sample, SQL Code generation, Main data Analysis,Properties.

INTRODUCTION:

Horizontal aggregation is new class of function to returnaggregated columns in a horizontal layout. Most algorithmsrequire datasets with horizontal layout as input with severalrecords and one variable or dimensions per columns.

Managing large data sets without DBMS support can be adifficult task. Trying different subsets of data points anddimensions is more flexible, faster and easier to do inside arelational database with SQL queries than outside withalternative tool. Horizontal aggregation can be performing byusing operator, it can easily be implemented inside a queryprocessor, much like a select, project and join.

PIVOToperator on tabular data that exchange rows, enable datatransformations useful in data modelling, data analysis, anddata presentation There are many existing functions andoperators for aggregation in Structured Query Language. Themost commonly used aggregation is the sum of a column andother aggregation operators return the average, maximum,minimum or row count over groups of rows. All operationsfor aggregation have many limitations to build large data setsfor data mining purposes. Database schemas are also highlynormalized for On-Line Transaction Processing (OLTP)systems where data sets that are stored in a relationaldatabase or data warehouse.

SQL CODE GENERATION:

Our main goal is to define a template to generate SQL codecombining aggregation and transposition (pivoting). Asecond goal is to extend the SELECT statement with a clausethat combines transposition with aggregation. Considerthe following GROUP BY query in standard SQL that takesa subset L1; . . . ; Lm from D1; . . .;Dp: SELECT L1; :::;Lm, sum(A)FROM FGROUP BY L1; . . . ; Lm; This aggregation query will produce a wide table with m þ 1columns (automatically determined), with one group for eachunique combination of values L1; . . . ; Lm and oneaggregated value per group (sum(A) in this case).

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

In order toevaluate this query the query optimizer takes three inputparameters: 1) the input table F, 2) the list of groupingcolumns L1; . . . ; Lm, 3) the column to aggregate (A). Thebasic goal of a horizontal aggregation is to transpose (pivot)the aggregated column A by a column subset of L1; . . . ;Lm; for simplicity assume such subset is R1; . . .;Rk where k< m. In other words, we partition the GROUP BY list intotwo sublists: one list to produce each group (j columns L1; . .. ; Lj) and another list (k columns R1; . . .;Rk) to transposeaggregated values, where fL1; . . . ; Ljg \ fR1; . . .;Rkg ¼ ;.

Each distinct combination of fR1; . . .;Rkg will automaticallyproduce an output column. In particular, if k ¼ 1 then thereare j_R1 ðFÞj columns (i.e., each value in R1 becomes acolumn storing one aggregation). Therefore, in a horizontalaggregation there are four input parameters to generate SQLcode: 1. the input table F, 2. the list of GROUP BY columnsL1; . . . ; Lj, 3. the column to aggregate (A), 4. The list oftransposing columns R1; . . .;Rk. Horizontal aggregationspreserve evaluation semantics of standard (vertical) SQLaggregations. The main difference will be returning a tablewith a horizontal layout, possibly having extra nulls. TheSQL code generation aspect isExample.

In the Fig.1 there is a common field K in F1 and F2.In F2,D2 consist of only two distinct values X and Y and is used totranspose the table. The aggregate operation is used in this issum (). The values within D1 are repeated, 1appears 3 times, for row 3, 4 and, and for row 3 & 4 value ofD2 is X & Y. So D2X and D2Y is newly generated columnsin FH.

K	D_1	D_2
1	3	X
2	2	Y
3	1	Y
4	1	Y
5	2	X
6	1	X
7	3	X
8	2	X

F_1

K	A
1	9
2	6
3	10
4	0
5	1
6	Null
7	8
8	7

F_2

D_1	D_2X	D_2Y
1	Null	10
2	8	6
3	17	null

F_H

Fig 1.An example of Horizontal aggregationCommonly using Query Evaluation methods in Horizontalaggregation functions .Data mining (the analysis step of the"Knowledge Discovery in Databases" process, or KDD), afield at the intersection of computer science and statistics, isthe process that attempts to discover patterns in large datasets.

It utilizes methods at the intersection of artificialintelligence, machine learning, statistics, and databasesystems. The overall goal of the data mining process is toextract information from a data set and transform it into anunderstandable structure for further use. Aside from the rawanalysis step, it involves database and data managementaspects, data preprocessing, model and inferenceconsiderations, interestingness metrics, complexityconsiderations, post-processing of discovered structures,visualization, and online updating.The term is a buzzword, and is frequently misused to meanany form of large-scale data or information processing(collection, extraction, warehousing, analysis, and statistics)but is also generalized to any kind of computer decisionsupport system, including artificial intelligence, machinelearning, and business intelligence. In the proper use of theword, the key term is discovery, commonly defined as"detecting something new".]

Even the popular book "Datamining: Practical machine learning tools and techniques withJava" (which covers mostly machine learning material) wasoriginally to be named just "Practical machine learning", andthe term "data mining" was only added for marketingreasons. Often the more general terms "(large scale) dataanalysis", or "analytics" – or when referring to actualmethods, artificial intelligence and machine learning – aremore appropriate.The actual data mining task is the automatic or semiautomaticanalysis of large quantities of data to extractpreviously unknown interesting patterns such as groups ofdata records (cluster analysis), unusual records (anomalydetection) and dependencies (association rule mining).

Thisusually involves using database techniques such as spatialindexes. These patterns can then be seen as a kind ofsummary of the input data, and may be used in further thesecurity expenditures are seen as wasteful because success istoo invisible". However, Schneider assures one that, despitethe lack of visible results, the need to secure information stillexists. Active attacks attempt to modify system resources ornetwork functionality. Examples of these attacks are messagemodification, message replay, impersonation and denial ofservice attacks.for example, in machine learning and predictiveanalytics. For example, the data mining step might identifymultiple groups in the data, which can then be used to obtainmore accurate prediction results by a decision supportsystem.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Neither the data collection, data preparation, norresult interpretation and reporting are part of the data miningstep, but do belong to the overall KDD process as additionalsteps.

Initial Data Analysis:

The most important distinction between the initial dataanalysis phase and the main analysis phase, is that duringinitial data analysis one refrains from any analysis that areaimed at answering the original research question. The initialdata analysis phase is guided by the following fourquestions:

Quality of Data:

The quality of the data should be checked as early aspossible. Data quality can be assessed in several ways, using different types of analyses: frequency counts, descriptivestatistics (mean, standard deviation, median), normality (skewness, kurtosis, frequency histograms, normalprobability plots), associations (correlations, scatter plots).

Other initial data quality checks are:

• Checks on data cleaning: have decisions influencedthe distribution of the variables? The distribution ofthe variables before data cleaning is compared tothe distribution of the variables after data cleaningto see whether data cleaning has had unwantedeffects on the data.

• Analysis of missing observations: are there manymissing values, and are the values missing atrandom? The missing observations in the data areanalyzed to see whether more than 25% of thevalues are missing, whether they are missing atrandom (MAR), and whether some form ofimputation is needed.

Quality of measurements:

The quality of the measurement instruments should only bechecked during the initial data analysis phase when this isnot the focus or research question of the study. One shouldcheck whether structure of measurement instrumentscorresponds to structure reported in the literature.

There are two ways to assess measurement quality:

• Confirmatory factor analysis

• Analysis of homogeneity (internal consistency),which gives an indication of the reliability of ameasurement instrument. During this analysis, oneinspects the variances of the items and the scales,the Cronbach'sα of the scales, and the change in theCronbach's alpha when an item would be deletedfrom a scale.

• Initial transformations

After assessing the quality of the data and of themeasurements, one might decide to impute missing data, orto perform initial transformations of one or more variables,although this can also be done during the main analysisphase.

Possible transformations of variables are:

• Square root transformation (if the distributiondiffers moderately from normal)

• Log-transformation (if the distribution differssubstantially from normal)

• Inverse transformation (if the distribution differsseverely from normal)

• Make categorical (ordinal / dichotomous) (if thedistribution differs severely from normal, and notransformations help)

Characteristics of Data Sample:

In any report or article, the structure of the sample must beaccurately described. It is especially important to exactlydetermine the structure of the sample (and specifically thesize of the subgroups) when subgroup analyses will beperformed during the main analysis phase.

The characteristics of the data sample can be assessed bylooking at:

• Basic statistics of important variables.

• Scatter plots.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

• Correlations

• Cross-tabulations

Final stage of the initial data analysis
During the final stage, the findings of the initial data analysisare documented, and necessary, preferable, and possiblecorrective actions are taken.Also, the original plan for the main data analyses can andshould be specified in more detail and/or rewritten.In order to do this, several decisions about the main dataanalyses can and should be made:

• In the case of non-normals: should one transformvariables; make variables categorical(ordinal/dichotomous); adapt the analysis method?

• In the case of missing data: should one neglect orimpute the missing data; which imputation technique should be used?

• In the case of outliers: should one use robustanalysis techniques?

• In case items do not fit the scale: should one adaptthe measurement instrument by omitting items, or rather ensure comparability with other (uses of the)measurement instrument(s)?

Analysis phase:

Several analyses can be used during the initial data analysis

• Univariate statistics

• Bivariate associations (correlations)

• Graphical techniques (scatter plots)
It is important to take the measurement levels of thevariables into account for the analyses, as special statistical techniques are available for each level:

• Nominal and ordinal variables

o Frequency counts (numbers andpercentages)

o Associations

circumambulations(crosstabulations)
hierarchicalloglinear analysis(restricted to a maximum of 8variables)

loglinear analysis (to identifyrelevant/important variables andpossible confounders)

o Exact tests or bootstrapping (in casesubgroups are small)

o Computation of new variables

• Continuous variables

o Distribution

Statistics (M, SD, variance,skewness, kurtosis)
Stem-and-leaf displays
Box plots

Main Data Analysis:

In the main analysis phase analyses aimed at answering theresearch question are performed as well as any other relevantanalysis needed to write the first draft of the research report.Exploratory and confirmatory approachesIn the main analysis phase either an exploratory orconfirmatory approach can be adopted. Usually the approachis decided before data is collected. In an exploratory analysisno clear hypothesis is stated before analyzing the data, andthe data is searched for models that describe the data well. Ina confirmatory analysis clear hypotheses about the data aretested. Exploratory data analysis should be interpreted carefully.

When testing multiple models at once there is a high chanceon finding at least one of them to be significant, but this canbe due to a type 1 error. It is important to always adjust thesignificance level when testing multiple models with, forexample, a bonferroni correction. Also, one should notfollow up an exploratory analysis with a confirmatoryanalysis in the same dataset. An exploratory analysis is usedto find ideas for a theory, but not to test that theory as well. When a model is found exploratory in a dataset, thenfollowing up that analysis with a confirmatory analysis in thesame dataset could simply mean that the results of theconfirmatory analysis are due to the same type 1 error thatresulted in the exploratory model in the first place.

Theconfirmatory analysis therefore will not be more informativethan the original exploratory analysis.A data set (or dataset) is a collection of data, usuallypresented in tabular form. Each column represents aparticular variable. Each row corresponds to a given memberof the data set in question. It lists values for each of thevariables, such as height and weight of an object. Each valueis known as a datum. The data set may comprise data for oneor more members, corresponding to the number of rows.Nontabular data sets can take the form of marked up stringsof characters, such as an XML file.

Properties: A data set has several characteristics which define itsstructure and properties. These include the number and typesof the attributes or variables and the various statisticalmeasures which may be applied to them such as standarddeviation and kurtosis.In the simplest case, there is only one variable, and then thedata set consists of a single column of values, oftenrepresented as a list. In spite of the name, such a univariatedata set is not a set in the usual mathematical sense, since agiven value may occur multiple times. Usually the order doesnot matter, and then the collection of values may beconsidered to be a multiset rather than an (ordered) list.

Thevalues may be numbers, such as real numbers or integers, forexample representing a person's height in centimeters, butmay also be nominal data (i.e., not consisting of numericalvalues), for example representing a person's ethnicity. Moregenerally, values may be of any of the kinds described as alevel of measurement. For each variable, the values willnormally all be of the same kind. However, there may alsobe "missing values", which need to be indicated in someway.In statistics, data sets usually come from actualobservations obtained by sampling a statistical population,and each row corresponds to the observations on one elementof that population.

Data sets may further be generated byalgorithms for the purpose of testing certain kinds ofsoftware. Some modern statistical analysis software such asPSPP still present their data in the classical data set fashionClassic data setsSeveral classic data sets have been used extensively in thestatistical literature:

• Iris flower data set - multivariate data set introducedby Ronald Fisher (1936)

• Categorical data analysis - Data sets used in thebook, An Introduction to Categorical DataAnalysis, by Agrestiare rovided on-line byStatLib.

• Robust statistics - Data sets used in RobustRegression and Outlier Detection (Rousseeuw andLeroy, 1986). rovided on-line at the University ofCologne.

• Time series - Data used in Chatfield's book, TheAnalysis of Time Series, are provided on-line byStatLib.

• Extreme values - Data used in the book, AnIntroduction to the Statistical Modeling of ExtremeValues are provided on-line by Stuart Coles, thebook's author.[Dead link]

• Bayesian Data Analysis - Data used in the book areprovided on-line by Andrew Gelman, one of thebook's authors.

• The Bupa liver data, used in several papers in themachine learning (data mining) literature.

• Anscombe's quartet Small dataset illustrating theimportance of graphing the data to avoid statisticalfallacies

CONCLUSION:

This system extended the horizontal aggregationswith k-means algorithm to cluster the aggregatedcolumn which help preparing datasets for datamining related work. Optimized k-means issignificantly faster because of small data set runclustering outside the DBMS.Input to the system isdata from multiple tables rather than single tableused in traditional horizontal aggregation-IncludeEuclidean distance computation, pivoting a table tohave one dimension value per row. Datamanipulating operator Pivot is easy to compute forwide set of values. Pivot is an extension of GroupBy with unique restrictions and optimizationopportunities, and this makes it easy to introduceincrementally on top of existing groupingimplementation.

REFERENCES:

[1] J. Gray, A. Bosworth, A. Layman, and H. Pirshesh. Datacube: A relational aggregation operator generalizing groupby,cross-tab and subtotal. In ICDE Conference, pages 152–159,1996.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

[2] A. Witkowski, S. Bellakonda, T. Bozkaya, G.Dorman, N. Folkert, A. Gupta, L. Sheng, and S.Subramanian, “Spreadsheetsin RDBMS for OLAP,” Proc. ACM SIGMOD Int’l Conf.Management of Data (SIGMOD ’03), pp. 52-63, 2003.

[3] S.Saraswati, S. Thomas, and R. Agarwal.Integratingassociation rule mining with relational database systems:alternatives and implications. In Proc. ACM SIGMODConference, pages 343–354, 1998.

[4] C. Ordonez. Integrating K-meansclustering with arela.tional DBMS using SQL.IEEE Transactions onKnowledge and Data Engineering (TKDE), 18(2):188–201,2006.

[5] G. Graefe, U. Fayyad, and S.Chaudhuri. On the efficientgathering of sufficient statistics for classification from largeSQL databases. In Proc. ACM KDD Conference, pages 204–208, 1998.

[6] G. Luo, J.F. Naughton, C.J. Elmann, and M. Watske. Locking protocols for materializedaggregate join views. IEEE Transactions on Knowledge andData Engineering (TKDE), 17(6):796–807, 2005.

[7] C. Ordonez, “Integrating K-Means Clustering with aRelational DBMS Using SQL,” IEEE Trans. Knowledgeand Data Eng., vol. 18,no. 2, pp. 188-201, Feb. 2006.

[8] C. Ordonez, “Statistical Model Computation with UDFs,”IEEE Trans. Knowledge and Data Eng., vol. 22, no. 12, pp.1752-1765, Dec.2010.

[9] C. Ordonez, “Data Set Preprocessing andTransformation in a Database System,” Intelligent DataAnalysis, vol. 15, no. 4, pp. 613-631, 2011.

[10] C. Ordonez and S. Pitchamalai, “Bayesian ClassifiersProgrammed in SQL,” IEEE Trans.nowledge and DataEng., vol. 22,no. 1, pp. 139-144, Jan. 2010.

[11] S. Sarawagi, S. Thomas, and R. Agrawal, “IntegratingAssociationRule Mining with Relational Database Systems:Alternatives andImplications,” Proc. ACM SIGMOD Int’l Conf. Management of Data(SIGMOD ’98), pp. 343-354, 1998.

[12] H. Wang, C. Zaniolo, and C.R. Luo, “ATLAS: A Small-But Complete SQL Extension for Data Mining and DataStreams,”Proc. 29th Int’l Conf. Very Large Data Bases (VLDB ’03),
pp.1113-1116,2003.