DWM Practical's data Handwritten on file pages

Experiment 2

Aim: Implementation of all dimension table and fact table based on experiment 1 case study

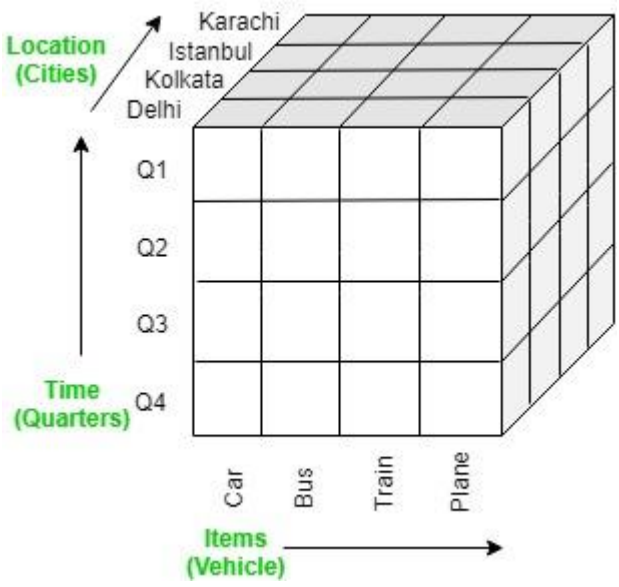
THEORY : Fact tables and dimension tables play different but important roles in a data warehouse. **Fact tables contain numerical data, while dimension tables provide context and background information**. Both types of tables are necessary for effective data analysis and decision-making.

Conclusion: The conclusion of implementing dimension and fact tables is that they are essential components of a data warehouse, providing the structure for data analysis and decision-making.

Experiment 3

Aim: Implementation of OLAP operation: Slice, Dice, Rollup, Drilldown and pivot Based on experiment 1 Case study

Theory: OLAP stands for ***Online Analytical Processing*** Server. It is a software technology that allows users to analyse information from multiple database systems at the same time. It is based on multidimensional data model and allows the user to query on multi-dimensional data (e.g. Delhi -> 2018 -> Sales data). OLAP databases are divided into one or more cubes and these cubes are known as *Hyper-cubes*.

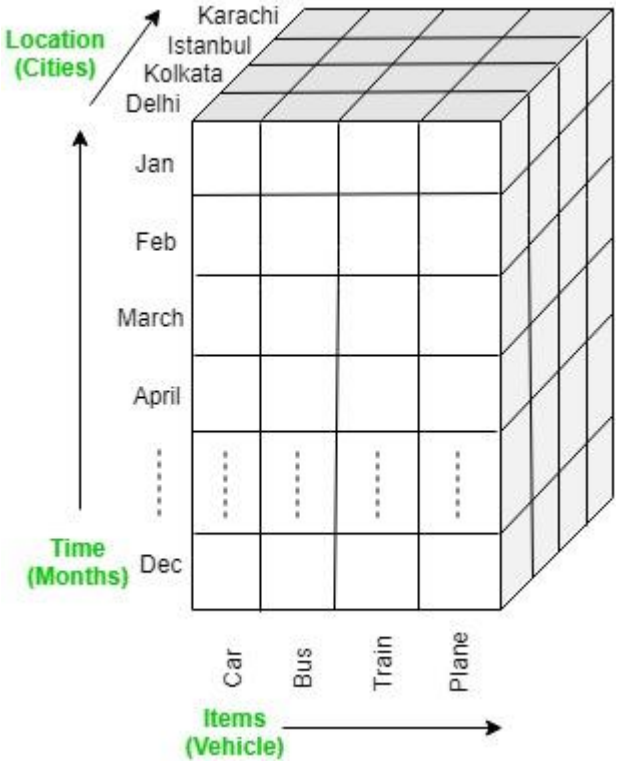


OLAP operations:

There are five basic analytical operations that can be performed on an OLAP cube:

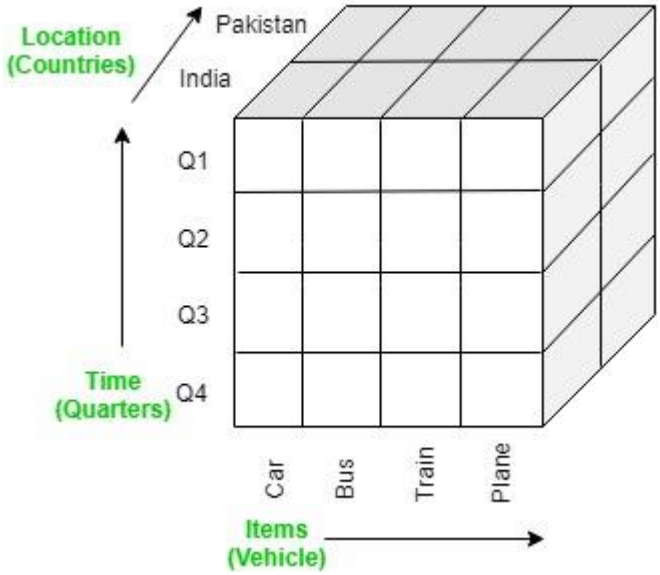
1. **Drill down:** In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:
    - Moving down in the concept hierarchy
    - Adding a new dimension

    In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).

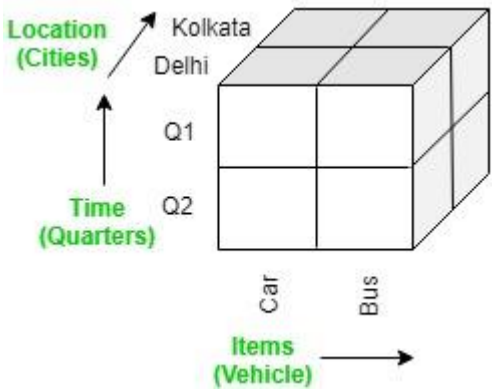


2. **Roll up:** It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:
    - Climbing up in the concept hierarchy
    - Reducing the dimensions

In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).

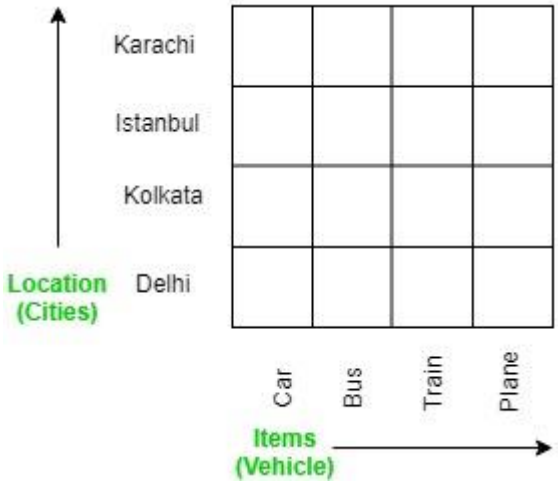


3. **Dice:** It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:
    - Location = "Delhi" or "Kolkata"
    - Time = "Q1" or "Q2"
    - Item = "Car" or "Bus"

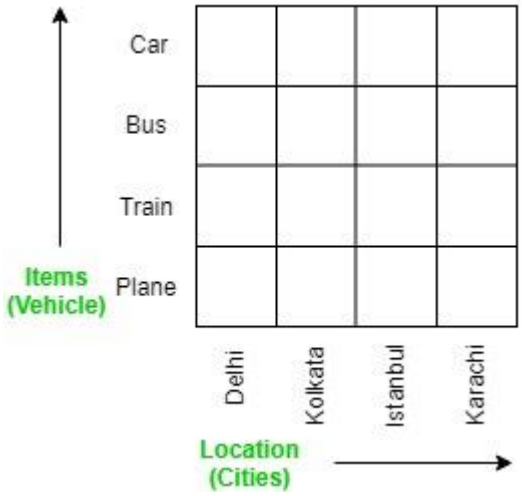


4. **Slice:** It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the

dimension Time = “Q1”.



5. **Pivot:** It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.



Conclusion: The implementation of Online Analytical Processing (OLAP) can be challenging, but it can be simplified by preparing the data and team in advance. The result is a highly efficient system for data analysis, reporting, storage, and forecasting.

Experiment 4

Aim: Implementation of Bayesian algorithm

Theory: Naive Bayes Classifier is a very popular supervised machine learning algorithm based on Bayes’ theorem. It is simple but very powerful algorithm which works well with large datasets and sparse matrices, like pre-processed text data which creates thousands of vectors depending on the number of words in a dictionary. It works really well with text data

projects like sentiment data analysis, performs good with document categorization projects, and also it is great in predicting categorical data in projects such as email spam classification.

It is used to solve many different problem statements, and it is quite fast in training a model since Naive Bayes classifier completely works on probability, so the conversion happens quickly.

Conclusion: Naive Bayes Classifier algorithm implementation is far from ideal, it requires many improvements and modifications to make a better prediction especially for text data

Experiment 5

Aim: Implementation of Data Discretization and visualization.

Theory: Discretization is **the process of converting continuous data into a set of discrete intervals or categories**. This technique can be used for data reduction, simplification, or to make the data more suitable for analysis and it typically applied to very large datasets. Data visualization is the graphical representation of information and data. **By using visual elements like charts, graphs, and maps**, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Conclusion: **Data discretization is an important process in data transformation**. It involves grouping together raw data points into categories, or bins, which are defined by a range of values. This can be done to simplify complex datasets and make them more manageable for analysis.

Experiment 6

Aim: Data pre-processing using WEKA, Classification, Clustering, Association Rule mining on data sets using WEKA

Theory: It provides you a visualization tool to inspect the data. The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose. Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

Conclusion: **WEKA is a powerful tool for developing machine learning models**. It provides implementation of several most widely used ML algorithms. Before these algorithms are applied to your dataset, it also allows you to preprocess the data.

Experiment 7

Aim: Implementation of Clustering algorithm (K-means /K- medoids)

Theory: K means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters. It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the cluster based on its distance from centroid of the cluster

Conclusion: In conclusion, **K-means clustering is a powerful unsupervised machine learning algorithm for grouping unlabeled datasets**. Its objective is to divide data into clusters, making similar data points part of the same group.

Experiment 8

Aim: Implementation of any one Hierarchical Clustering method.

Theory: Hierarchical clustering is a statistical analysis technique that groups similar data into clusters. It's used in many applications, including customer segmentation, crime analysis, and image-based text recognition

Conclusion: Hierarchical cluster analysis is **a versatile and powerful technique for uncovering the structure within complex data sets**

Experiment 9

Aim: Implementation of Association Rule Mining algorithm

Theory: Association rule mining is **a technique used to uncover hidden relationships between variables in large datasets**. It is a popular method in data mining and machine learning and has a wide range of applications in various fields, such as market basket analysis, customer segmentation, and fraud detection. The Apriori algorithm generates association rules by **iterating over the list of frequent item sets and dividing each item set into two or more non-empty subsets**. The algorithm then calculates the support and confidence of each possible rule and keeps only those that meet a minimum confidence threshold.

Conclusion: The Apriori algorithm is a crucial tool in data mining, commonly used for frequent item set mining and association rule learning. It proves particularly valuable in

tasks like market basket analysis for creating recommendation systems, emphasizing interpretability.

Experiment 10

Aim: Implementation of Page rank /HITS Algorithm

Theory: PageRank is a metric for determining how important a website's pages are. **PageRank calculates a rough estimate of the importance of a website by measuring the quantity and quality of links to that page**. The basic premise is that more important websites are more likely to gain links from other sites **Hyperlink Induced Topic Search** (HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.
HITS uses hubs and authorities to define a recursive relationship between webpages. Before understanding the HITS Algorithm, we first need to know about Hubs and Authorities.

Conclusion: PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites