

BACHELOR OF COMPUTER SCIENCE (HONS)

BCS3222 – DATA MINING AND DATA WAREHOUSING

**Individual Assignment 1**

**Instructions:**

* Answer **ALL** questions
* Marks will be awarded for good presentation and thoroughness in your approach
* **NO** marks will be awarded for the entire assignment if any part of it is found to be copied directly from printed materials or from another student.
* Submit your assignment in **hardcopy** (printed) and **softcopy** (*upload it on e-QIU*)
* Complete this cover sheet (page 1 and 2) and attach it to your assignment.

**Student’s Particulars:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Matric No** | **:** | 202007004078 | **Intake / Semester** | **:** | JULY 2022 |
| **Student’s Name** | **:** | SAVITHA SAN MARKAN | |  | |
| **Lecturer’s Name** | **:** | DR ANBUSELVAN SANGODIAH | |  | |
| **Due Date** | **:** | 19/8/2022 | **Submission Date** | **:** | 19/8/2022 |
| **Total number of pages including this cover page** | | | | **:** |  |

**Official Use:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Examiner’s Comment** | | |
|  |  |  | | |
| **Marks Awarded** | **:** | **/ 100** | **Overall Marks (20%):** |  |

**Student declaration:**

*I declare that:*

* *I understand what is meant by plagiarism*
* *The implication of plagiarism has been explained to me by my institution*
* *This assignment is all my own work and I have acknowledged any use of the published and unpublished works of other people.*

|  |  |
| --- | --- |
| Student’s signature: SAVITHA SAN MARKAN | Date: 19/8/2022 |

**Marks Sheet:**

|  |  |  |
| --- | --- | --- |
| **Marks Table** | |  |
| Questions | Marks Allocated | Marks Awarded |
| OLAP  Application | 80 |  |
| Report | 20 |  |
| **Total** | **100** |  |

|  |
| --- |
| **/100** |

**Report on Data Mining Assignment**

Chapter 1: Introduction

Introduction

The OLAP database is a significant and well-liked front-end tool for business analysis and decision assistance. Business information systems, such as ERP systems, refer to multi-dimensional tables, which are used in OLAP databases as "data cubes," as the structure of business data. The business decision-maker can benefit greatly from the manipulation and displaying of such information using interactive multi-dimensional tables and diagrams.

Currently, what-if analysis is not well supported by multi-dimensional business databases. What-if analysis is the study of how different sources of variation in the model's input can be statistically or qualitatively linked to variations in the output of a mathematical model. The OLAP analyst would be able to experiment with "What if...?" inquiries with the use of such analytical functionality. You may ask something like, "If I modify the value of this data cell in this cube by so much, what happens to an aggregated value in higher level cubes?" as an illustration. Therefore, the main inquiry in this work is how what-if analysis might be used to improve the capability of OLAP databases.

In this report, we go into greater detail on a proposed operator that helps analysts respond to these common analytical queries in OLAP databases. The first time that such an operator was stated was in (Caron and Daniels, 2008; Caron and Daniels, 2009), but in this article, we go into greater depth about it and use it in a case study. We propose a new notation for crucial OLAP database concepts, including dimensions, cells, cubes, navigational operators, lattices, upset, and additive measures, for this aim. These ideas help us build the what-if operator. The OLAP database maintaining mathematical consistency throughout the analysis is a crucial concern for the implementation of this operator. It is challenging to maintain consistency in an OLAP database since changing one variable can lead to inconsistencies in the equations that make up a given measure. Determining the prerequisites for consistency and solvability in OLAP databases is crucial, for this reason.

Problem Statement

Large volumes of data are kept or have previously been gathered by businesses all over the world during their operations in corporate databases. These "data sets" provide the most significant potential opportunities for generating fresh, analytical data to support the formulation of corporate strategies to expose market evolution tendencies and to identify new solutions requiring successful development under competitive circumstances. It is clear that some businesses include such analyses into their routine operations, but most of them are just beginning to do so.

The conventional method entails using developed OLTP-systems to support approved administrative solutions, i.e., making an effort to create report sets "under" the corporate database and using reports that have been received (after being interpreted) for approval of strategic business solutions. One drawback of such a way of decision-making is that it's necessary to check at least the following boxes:

* Small amount of data used to support important business solutions;
* Similar processes take a very long time since they involve complex writing and interpreting of queries. When the primary answer may need to be made immediately, it takes many days.
* Difficulty in interpreting reports (lack of visualisation);

Relational database management system expert E.F. Codd stated that despite RDBMS systems being user-accessible, they were never thought of as a resource providing strong capabilities on synthesis, analysis, and consolidation (also known as multidimensional data analysis). Information synthesis, the transformation of operating system data into information, and even quality standards are in dispute.

Objectives

The objective of using OLAP tools is to interactively evaluate multidimensional data from many perspectives. Three fundamental analytical processes make up OLAP: consolidation (roll-up), drill-down, and slicing and dicing. Data that can be amassed and computed in one or more dimensions are combined during consolidation. For instance, all sales offices are gathered in one place at the sales division or sales department to forecast sales trends. The drill-down method, in contrast, enables visitors to browse through the specifics. Users can, for example, see the sales of certain products that make up a region's sales. A feature known as "slicing and dicing" allows users to remove (slice) a specified set of data from an OLAP cube and view (dice) the slices from multiple viewpoints. These viewpoints are frequently referred to as dimensions (such as looking at the same sales by salesperson, or by date, or by customer, or by product, or by region, etc.).

Scope

OLAP mining is a method that combines on-line analytical processing (OLAP) and data mining so that users can do mining at various levels of abstraction and in various areas of databases or data warehouses. It is potential to build OLAP mining algorithms given the OLAP and data warehouse technologies' rapid development in the database sector.

An OLAP-based data mining system has been developed after years of data mining research, and it uses OLAP mining for association, classification, prediction, clustering, and sequencing in addition to data characterisation. Users can find desired knowledge with the aid of such integration, which boosts mining's versatility. In this report. describe how to implement OLAP mining in a data mining system and introduce the idea of OLAP mining using basic analytical operations: roll-up, drill-down, slicing and dicing, drill-across, and pivoting.

Chapter 2: Literature Review

Literature on Data Warehousing

Data Warehouse is a decision support database that is kept apart from the operational database of the company. It supports information processing by giving an analysis-ready framework of consolidated historical data.

“A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision-making process.”—W. H. Inmon

Data Warehousing is the process of constructing and using data warehouses.

Data Warehouse (Integrated)

* Built by combining several, disparate data sources, including relational databases, flat files, and online transaction records
* By establishing consistency in name conventions, encoding structures, attribute measures, etc. among various data sources, data cleaning and integration techniques are used.
* Data is transformed before being transferred to the warehouse.

Data Warehouse (Time Variant)

Comparing to operational systems, the data warehouse has a far longer time horizon.

* Operational database: data on the current value
* Provide information with a historical perspective using data from a data warehouse (e.g., past 5-10 years)

The data warehouse's whole fundamental structure

* Has a time component, either explicitly or implicitly, while the operational data's key may or may not have a time component.

Data Warehouse (Non-volatile)

* A physically separate store of data transformed from the operational environment

The data warehouse environment does not involve operational data updating.

* not requiring systems for transaction processing, recovery, and concurrency control

Only needs two operations to retrieve data

* Initial loading of data and access of data

Types of Data Warehouse

Enterprise Warehouse

* A centralised warehouse is the Enterprise Data Warehouse (EDW). It offers enterprise-wide decision support services. It provides a consistent method for gathering and displaying data. It gathers all the data on topics that concern the entire organisation. Additionally, it offers the capability of categorising data by subject and granting access in accordance with such divisions.

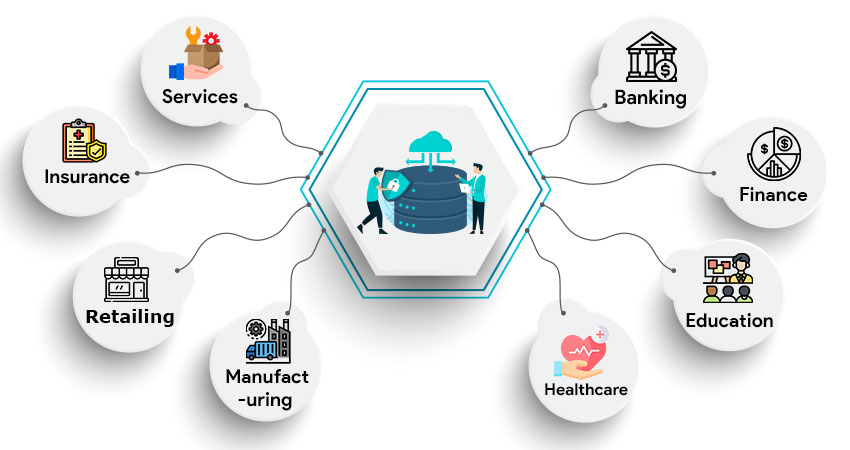
Data Mart

* A subset of the data warehouse is called a data mart. It is especially made for a specific industry, such sales, finance, sales, or finance. Data can be gathered directly from sources and stored in a separate data mart.

Virtual Warehouse

* A collection of database views for active databases. Only a few of the potential summary views might come to pass.

Applications of Data Warehouse



Banking:

It is frequently used in the banking industry to efficiently manage the resources on the desk. A few banks are also used for operations, product performance analysis, and market research.

Finance:

The use of data warehousing in the financial sector is similar to that in the banking business. The proper solution aids in the analysis of customer spending, allowing the financing sector to develop more effective plans to increase profits on both ends.

Education

Data warehousing is necessary for the educational industry to have a complete understanding of their faculty members' and students' data. In order for educational institutions to make wise and informed judgments, it gives them access to real-time data flows.

Healthcare:

Data warehouses were also used by the healthcare sector to plan and forecast outcomes, provide patient treatment reports, and communicate data with affiliated insurance firms, medical aid services, etc.

Services

Data warehousing is used in the services industry to store client information, financial records, and resources to analyse patterns and improve decision-making for successful outcomes.

Insurance

Data warehousing is necessary in the insurance industry to preserve records of current clients and analyse those records to spot client patterns and draw in new clients.

Retailing

Because retailers act as a bridge between wholesalers and final consumers, it is important that they keep accurate data for both parties. The use of data warehousing enters the picture to assist them in managing their data storage.

Manufacturing & Distribution

Manufacturing and distribution companies may consolidate all of their data under one roof, forecast market changes, examine current patterns, identify potential growth areas, and ultimately make decisions that will have a positive impact.

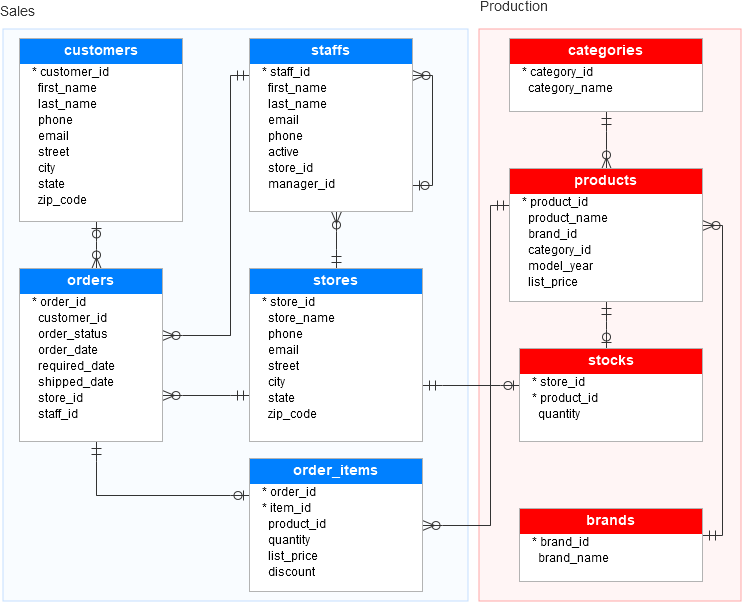
Literature on the selected framework and development platform

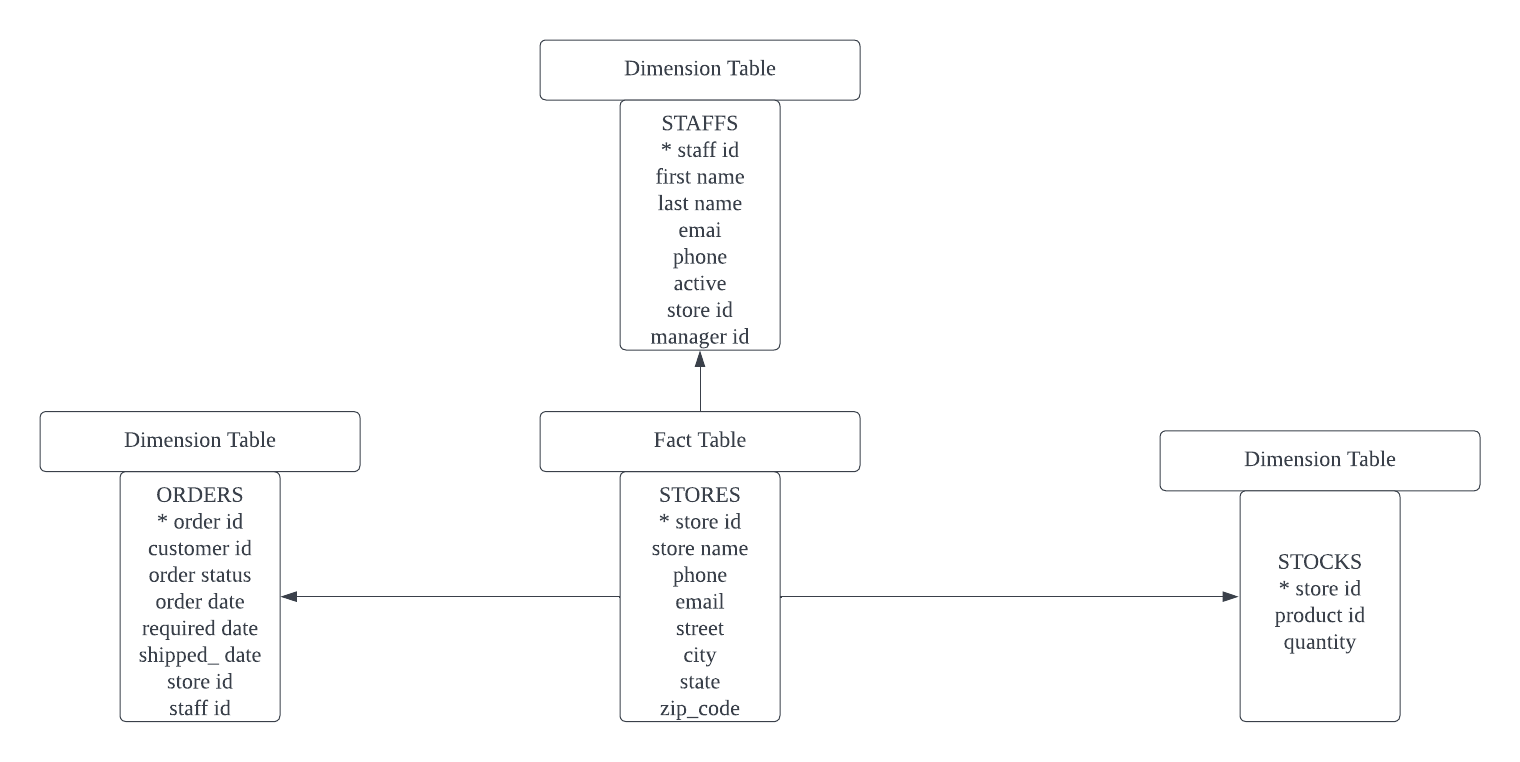
Framework used is OLAP which consist roll-up, drill-down, slicing and dicing, drill-across, and pivoting. The purpose of choosing SQL Server to execute the SQL commands is that, this tool offers so many features which is easier to used compared to other SQL apps. For instance, when we create relational Database, it automatically creates it ERD in the diagram section. Also, SQL Server can be connected to any cloud database or we can make website and use our database there. Using SQL server is safer than MySQL. It prevents any process from accessing and changing the database files while it is running. By executing an instance, users can alter files or execute specified activities. This stops hackers from immediately accessing or modifying the data.

Star schema was used to for modelling the data warehouse. Star schema is where the fact table in the middle connected to a set of dimension tables. As an example, the fact table for my dataset is ‘store’ and it is connected to other dimensions’ table such as ‘customer, orders, staffs, order items, categories, products, stocks and brands’.

Chapter 3: Methodology

Database schema and ERD of the constructed database

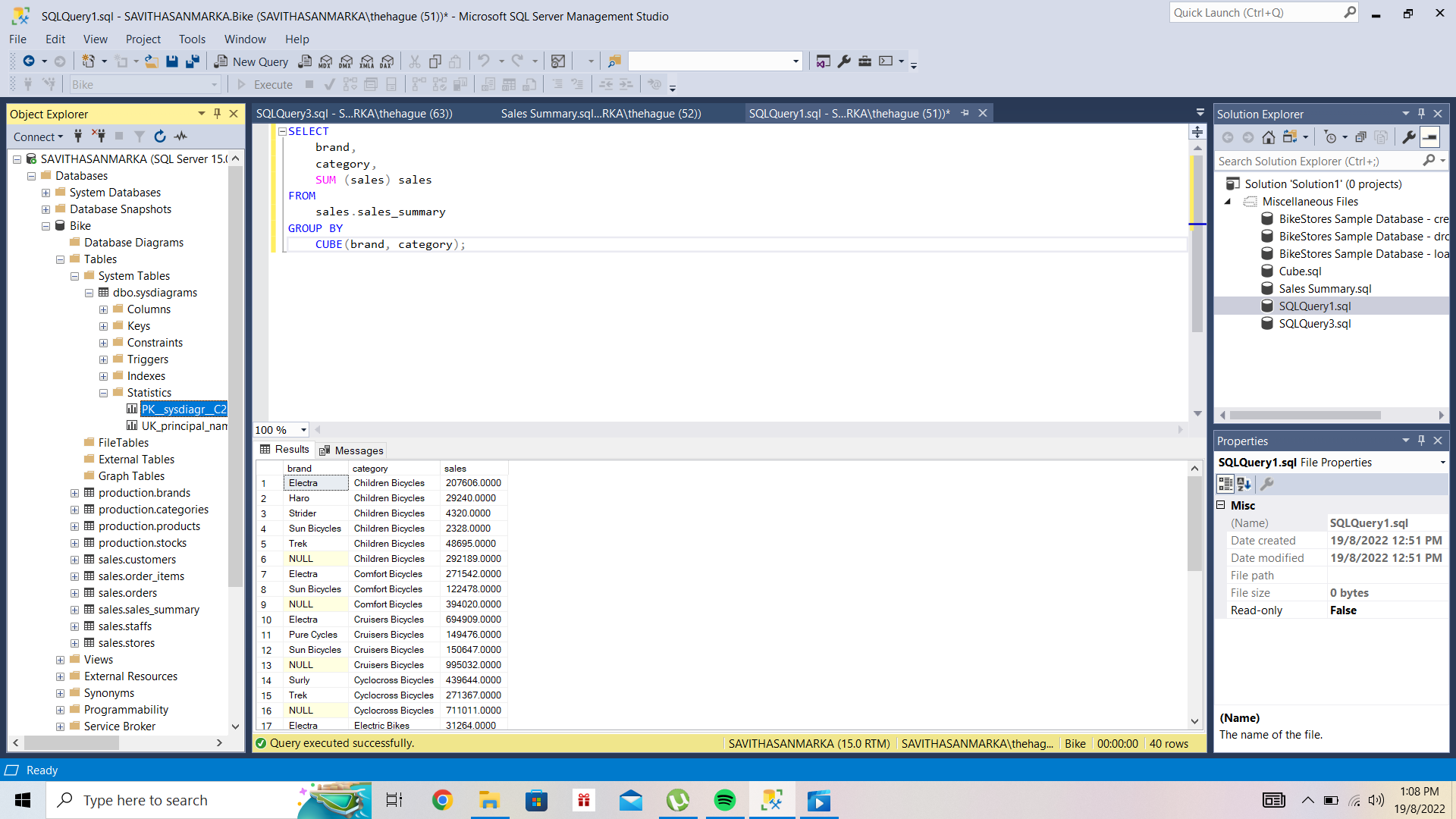


Star Schema Database

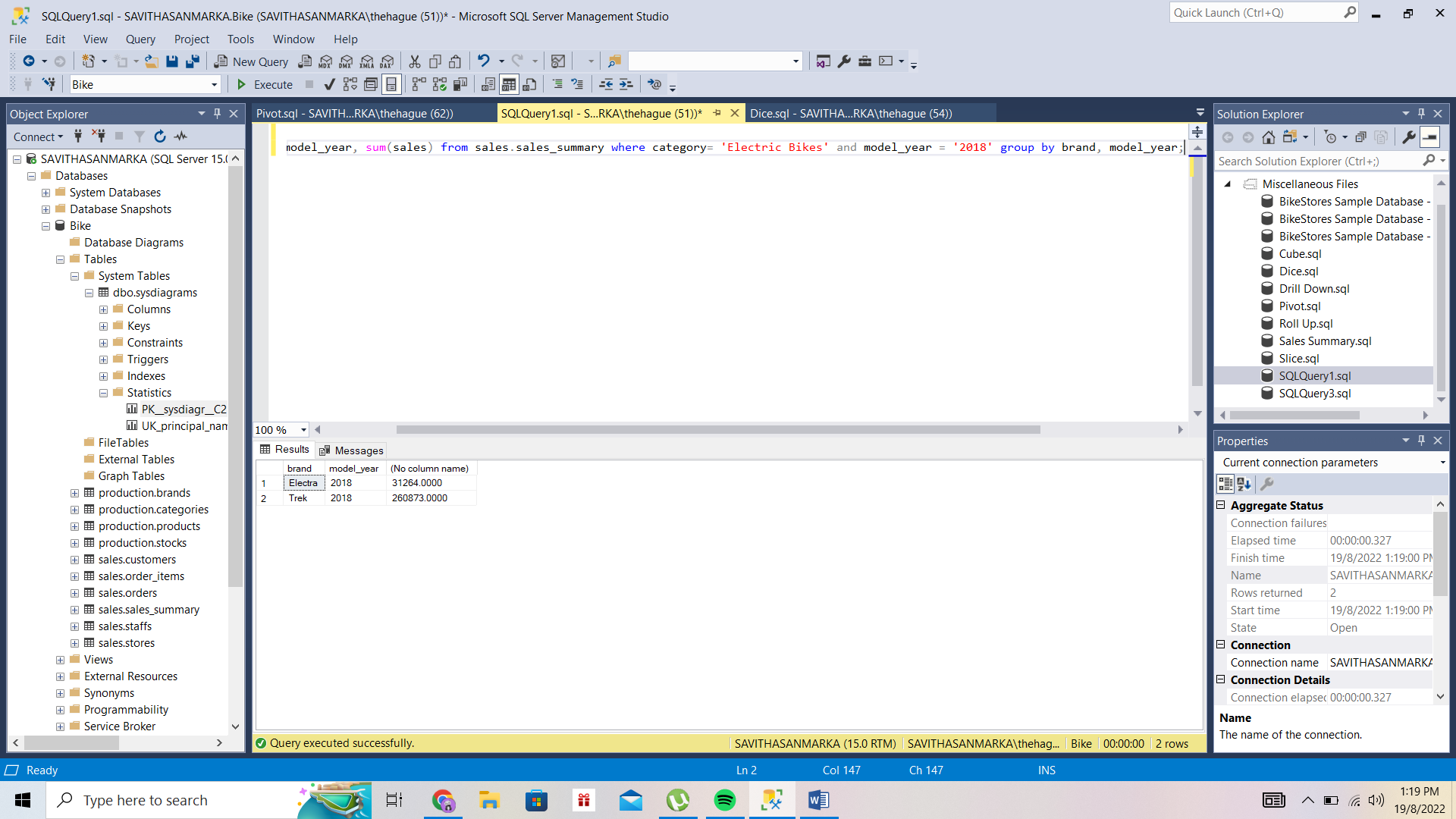
Data cube representations (cuboids) of the data according to the selected attributes (fields) for OLAP operations. (information package diagram)’

Lattice of cuboid for the selected attributes (fields) for OLAP operations

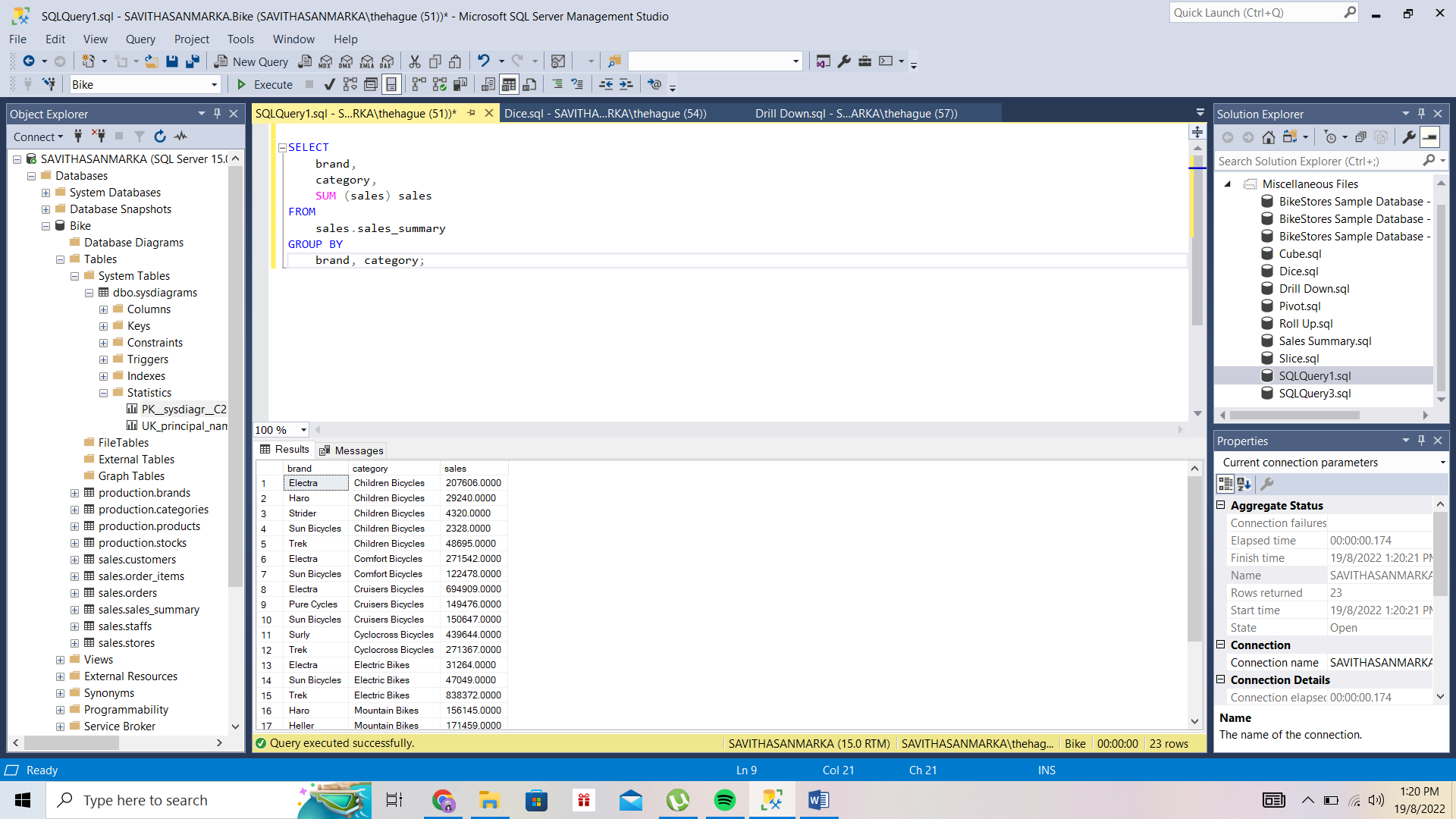
Data Cube representation



Dice Operation

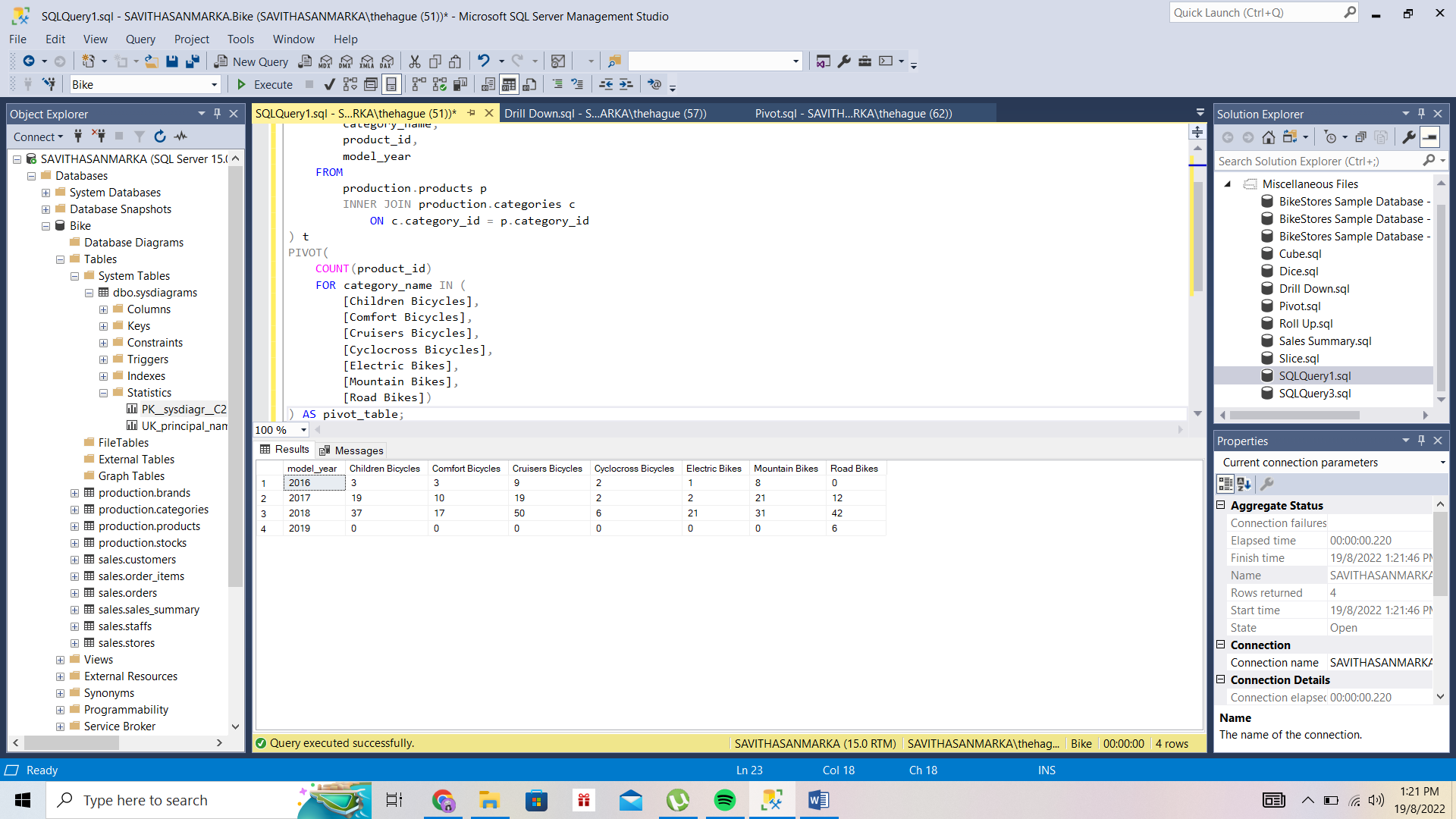


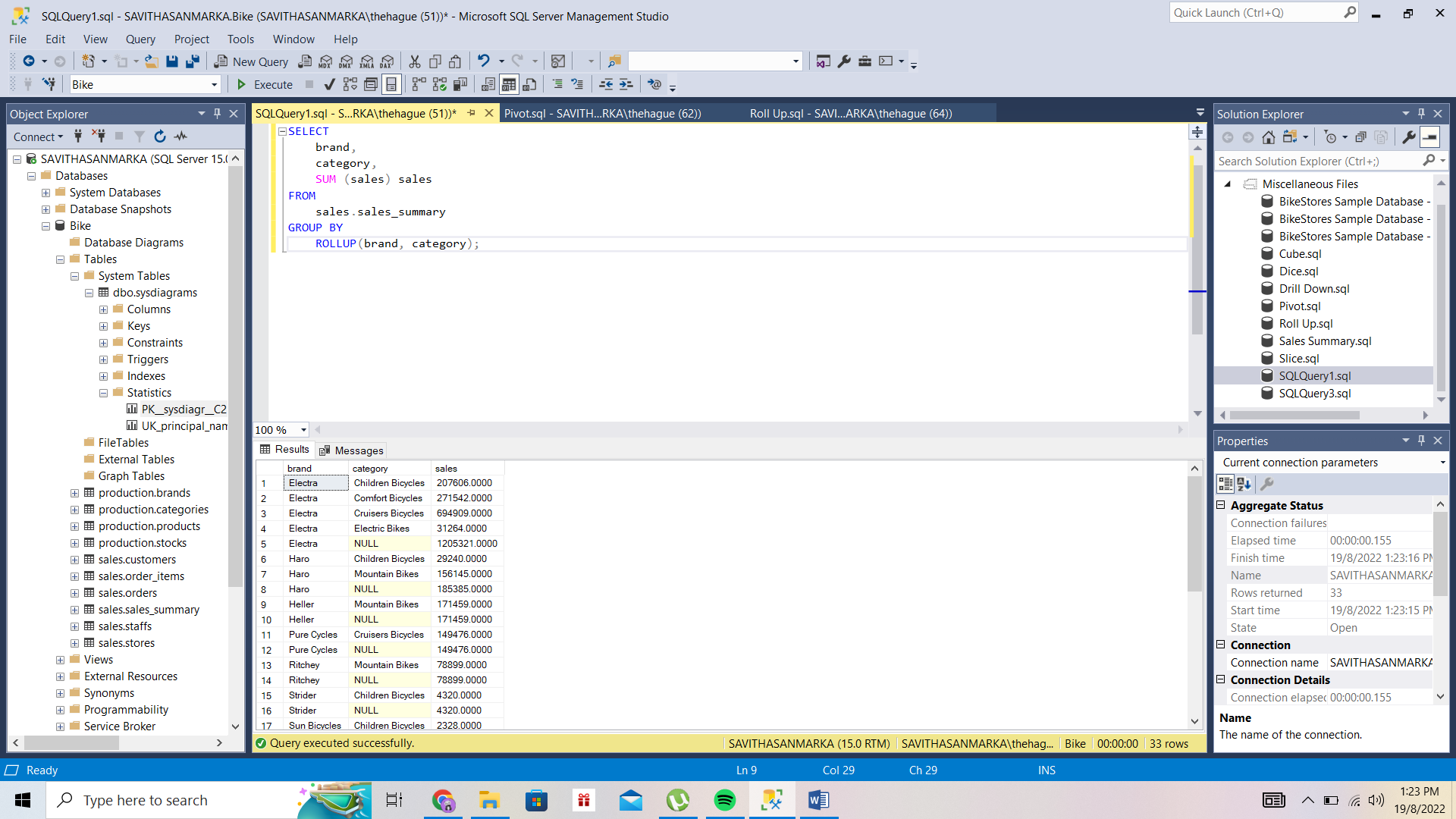
Lattice of cuboid for the selected attributes (fields) for OLAP operations.

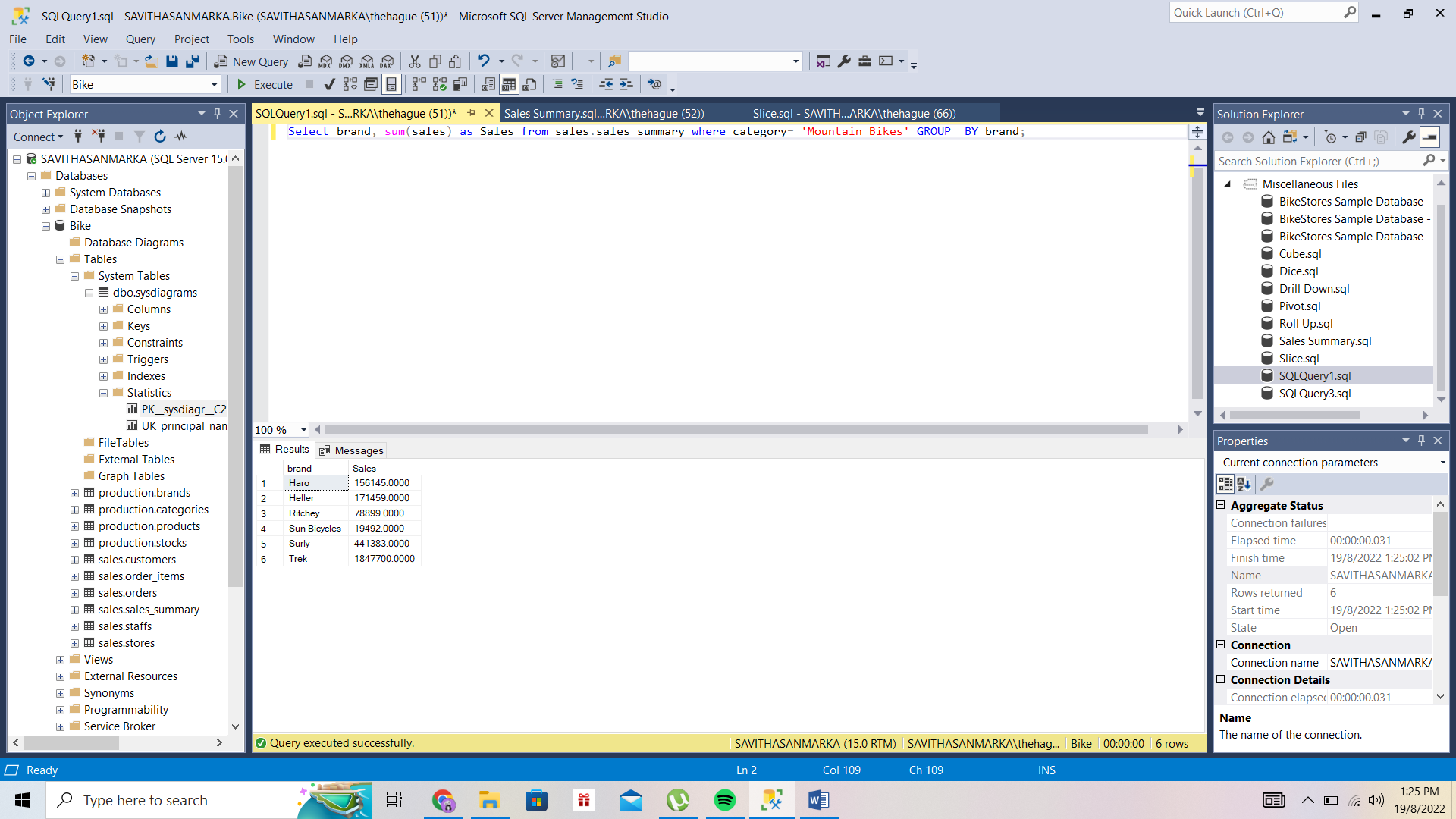
Drill Down

The hierarchical structures for the selected attributes (fields) for OLAP operations.

Pivot Operation



Roll Up operation

Slice Operation

Chapter 4: Discussion and Analysis

The detailed description and explanation for each of the OLAP operations applied to the data and the results generated from the operations.

OLAP

We have Implemented our OLAP in SQL Server by using the Sales dataset of bikes.

First of all, We will use the sales.sales\_summary table created in the GROUPING SETS tutorial for the demonstration. This object or table is used for demonstration of the different operation is OLAP in SQL Server. We can use the following statement to create it.

SELECT

b.brand\_name AS brand,

c.category\_name AS category,

p.model\_year,

round(

SUM (

quantity \* i.list\_price \* (1 - discount)

),

0

) sales INTO sales.sales\_summary

FROM

sales.order\_items i

INNER JOIN production.products p ON p.product\_id = i.product\_id

INNER JOIN production.brands b ON b.brand\_id = p.brand\_id

INNER JOIN production.categories c ON c.category\_id = p.category\_id

GROUP BY

b.brand\_name,

c.category\_name,

p.model\_year

ORDER BY

b.brand\_name,

c.category\_name,

p.model\_year;

CUBE Operation

SQL CUBE is a data structure, is like a matrix or a two-dimensional array that is used to perform aggregate values and summary reports on multiple axes or dimensions and provides a more detailed analysis and report by performing grouping of data along with more than one column and creating multiple grouping sets while using just a single query.

SELECT

brand,

category,

SUM (sales) sales

FROM sales.sales\_summary

GROUP BY CUBE(brand, category);

In the above query the cube Operation is being performed on the 2 fields brand and category, and the Sum of sales is the fact column that shows the sum of sales. It together forms a 2-dimensional array or a cube.

Roll Up Operation

By descending idea hierarchies, or dimension reduction, the roll-up operation—also known as drill-up or aggregation operation—performs aggregation on a data cube. By moving up the location hierarchy from the level of the city to the level of the country, the roll-up process aggregates the data.

When a roll-up is accomplished through a reduction in dimensions, one or more dimensions are taken out of the cube. Consider a sales data cube, for instance, which has two dimensions: location and time. The temporal dimensions that show in an aggregate of the total sales by place, rather than by location and by time, can be removed to execute a roll-up.

SELECT

brand,

category,

SUM (sales) sales

FROM sales.sales\_summary

GROUP BY ROLLUP(category);

The Above query perform the Roll Up Operation on the category column and gives the results.

Drill Down

This is a reverse of the ROLL UP operation. The data is aggregated from a higher level summary to a lower level summary data.

SELECT

brand,

category,

SUM (sales) sales

FROM sales.sales\_summary

GROUP BY category;

No such implementation of Drill Down exists in the SQL Server, so we just can remove the RollUp clause from the query and the resultant will be Drill Down from the above query.

Slice Operation

An individual value for one or more members of the dimension is represented by a slice, which is a subset of the cubes. When a consumer requests a selection on one side of a three-dimensional cube, for instance, a slice operation is carried out, creating a two-dimensional site. In order to create a subcube, the Slice procedures make a pick on one dimension of the provided cube.

Select brand, sum(sales) as Sales

from sales.sales\_summary

where category= 'Mountain Bikes'

GROUP BY brand;

Slicing can be done by applying Where and Group By clauses in SQL Server. In the above query, slice is being applied to the brand where brand is “Mountain Bike” and the output is being grouped by brand.

Dice Operation

A new sub-cube is provided by the dice after it chooses two or more dimensions from a given cube. Take a look at the following dice operation diagram. Based on the following selection criteria, the dice operation on the cube involves three dimensions.

Select

brand, model\_year, sum(sales)

from sales.sales\_summary

where category= 'Electric Bikes' and model\_year = '2018'

group by brand, model\_year;

In the above query dice is being applied to the category and the model\_year column in the sales\_summary table.

Pivot

Rotation is another name for the pivot action. A pivot is a visualisation procedure that turns the data axes visible to offer a different way to present the data. The rows and columns might be switched about, or one of the row-dimensions might be shifted into the column dimensions. These steps are needed to make a query a pivot table:

• Choose a base dataset for pivoting first.

• Second, use a derived table to produce a temporary result.

• Lastly, apply the PIVOT operator.

SELECT \* FROM

(

SELECT

category\_name,

product\_id,

model\_year

FROM

production.products p

INNER JOIN production.categories c

ON c.category\_id = p.category\_id

) t

PIVOT(

COUNT(product\_id)

FOR category\_name IN (

[Children Bicycles],

[Comfort Bicycles],

[Cruisers Bicycles],

[Cyclocross Bicycles],

[Electric Bikes],

[Mountain Bikes],

[Road Bikes])

) AS pivot\_table;

Conclusion

Large databases can be accessed using the OLAP family of analysis and report-generating tools. It provides the multidimensional storage of entire reports or partially aggregated data for quick, easy access and analysis. The foundation of OLAP techniques are databases that enable multidimensional views of business data. OLAP is useful for visualization of relationships between predesignated variables.

In order to obtain a wider analysis of the data, OLAP applications are used. Typically, numerous sources are updated in batch manner into the OLAP database. OLAP is designed with reporting and analysis in mind. Users of online transaction processing (OLTP) applications, in contrast, can add to, edit, or retrieve specific customer records. OLTP databases are intended for updating transactions.

Data mining and OLAP may be used together, although OLAP is not a replacement for data mining. While data mining tools look for patterns in the data, OLAP tools are strong and quick tools for producing reports on data. Users of OLAP are limited in what queries they can ask because OLAP can only respond to those that the data formats were intended to answer. They are unable to revisit the initial information and look for fresh answers. Data mining is therefore more effective than OLAP.

Strengths and weaknesses of the developed data warehouse and OLAP framework/queries

Strength

* The database is centralized and the data is being fetched from the single database.
* Speed of operation is high
* Database or the results are adaptable and can be made accordingly

Weakness

* Short of Interactive analysis to the database
* Pre modelling is must for this OLAP

Turnitin Report

