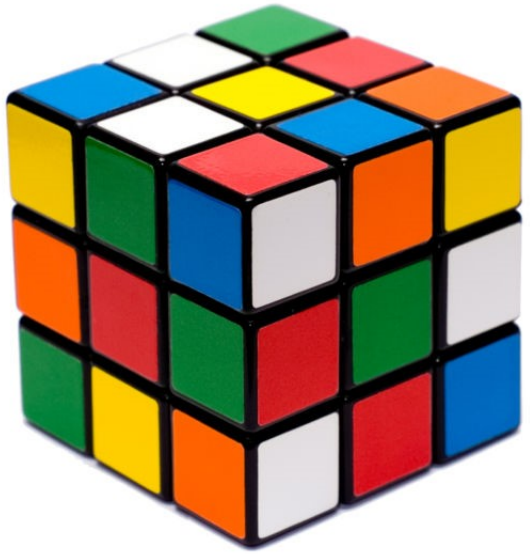
Data Cube

Trishla Shah

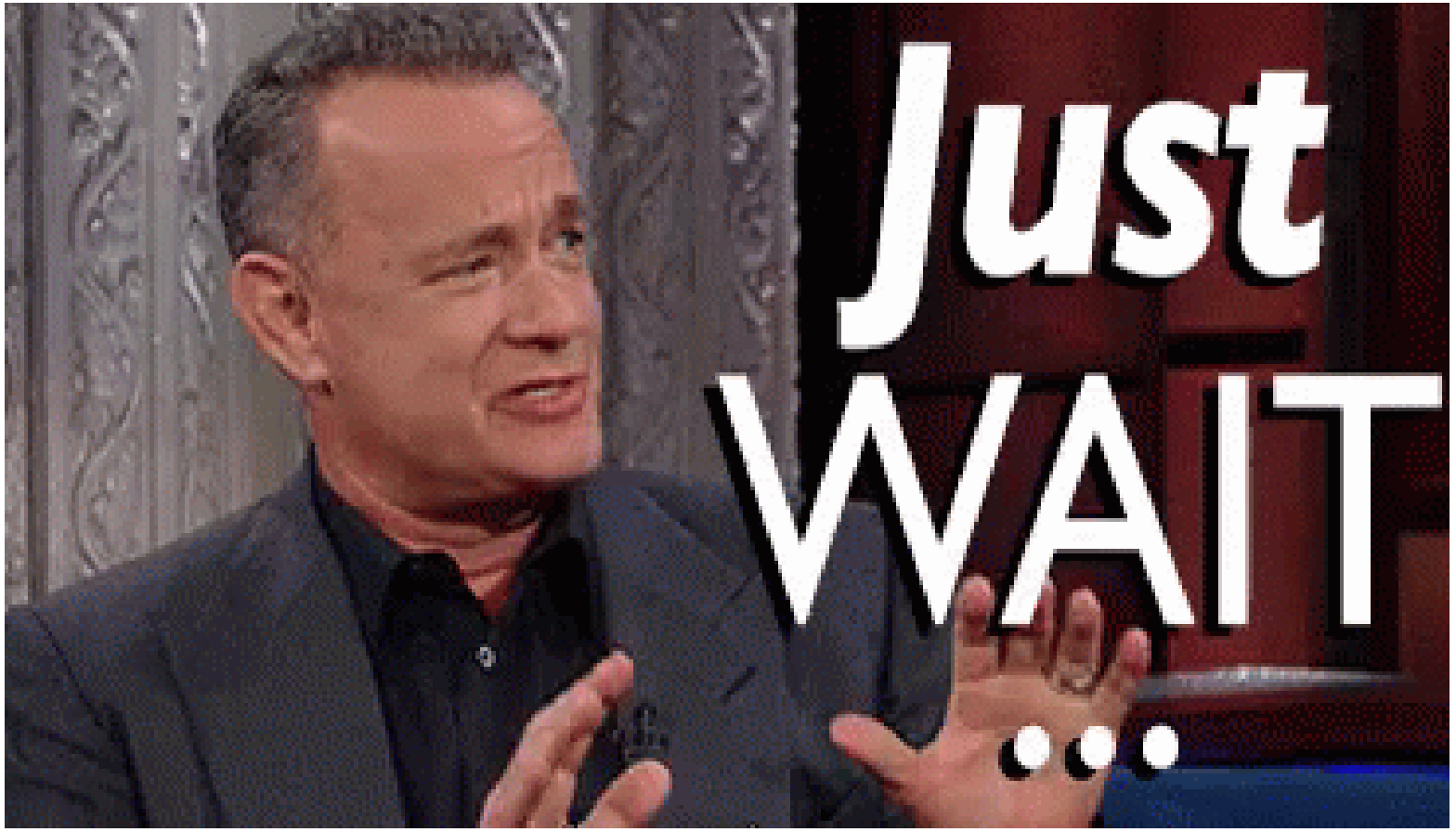
trishla@dal.ca

What is Data Cube?

Data Cube is a multi-dimensional ‘n-D”) array of values.



Why should I learn about Data Cube?



Let’s take an example…

Assume that we have one supermarket…

Where did our data come from ?

- Lots of individual shoppers buying a juice or other items.
- Each transaction stored in database designed to store check out transactions.

Operational Database

- It supports the day-to-day operations of a company

Core operational database functionality:

- Gather Data
- Update Data
- Store Data
- Retrieve Data
- Archive Data

Collectively, operational systems are usually referred to as online transaction processing (OLTP)


EXAMPLE

Buying juice at superstore:

You place juice on conveyor belt

Buying juice at superstore:

Cashier swipes barcode over POS scanner

Buying juice at superstore:

- POS system looks up price of toothpaste
- POS totals cost of transaction + tax
- POS prompts for payment

Buying juice at superstore:

You swipe debit card and enter PIN

Buying juice at superstore:

POS system transfers cost of juice from your bank account Target’s account

Buying juice at superstore:

POS generates receipt and cashier bags

OLTP Example

Buying toothpaste at superstore:

- You place toothpaste on conveyor belt
- POS system looks up price of toothpaste
- POS totals cost of transaction + tax
- POS prompts for payment
- You swipe debit card and enter PIN
- POS system transfers cost of toothpaste from your bank account Target's account
- POS generates receipt and cashier bags
- Purchase

Question

- Total number of OLTP datasets with names.
- Which operations are preformed on each OLTP dataset

Key OLTP Characteristics

- Process a transaction according to rules
- Performs all elements of a transaction in real time
- Continually processes multiple transaction

OLTP Systems

OLTP systems are everywhere

- Order tracking
- Invoicing
- Credit card processing
- Retail POS
- Banking
- Airline reservations

Benefits of OLTP datasets

- OLTP is optimized for managing low level business data
- OLTP systems can be used to answer transactional questions.

For example,

- “Did somebody buy juice today at the Walmart?”

  is a transaction question that can be easily answered by an OLTP system without much fuss.

Then what is the problem?

So What's the problem?

- While this sort of data can be served up quickly, it's not really useful for an analysis of the overall business.
- Raw transactional data not really useful for business intelligence.

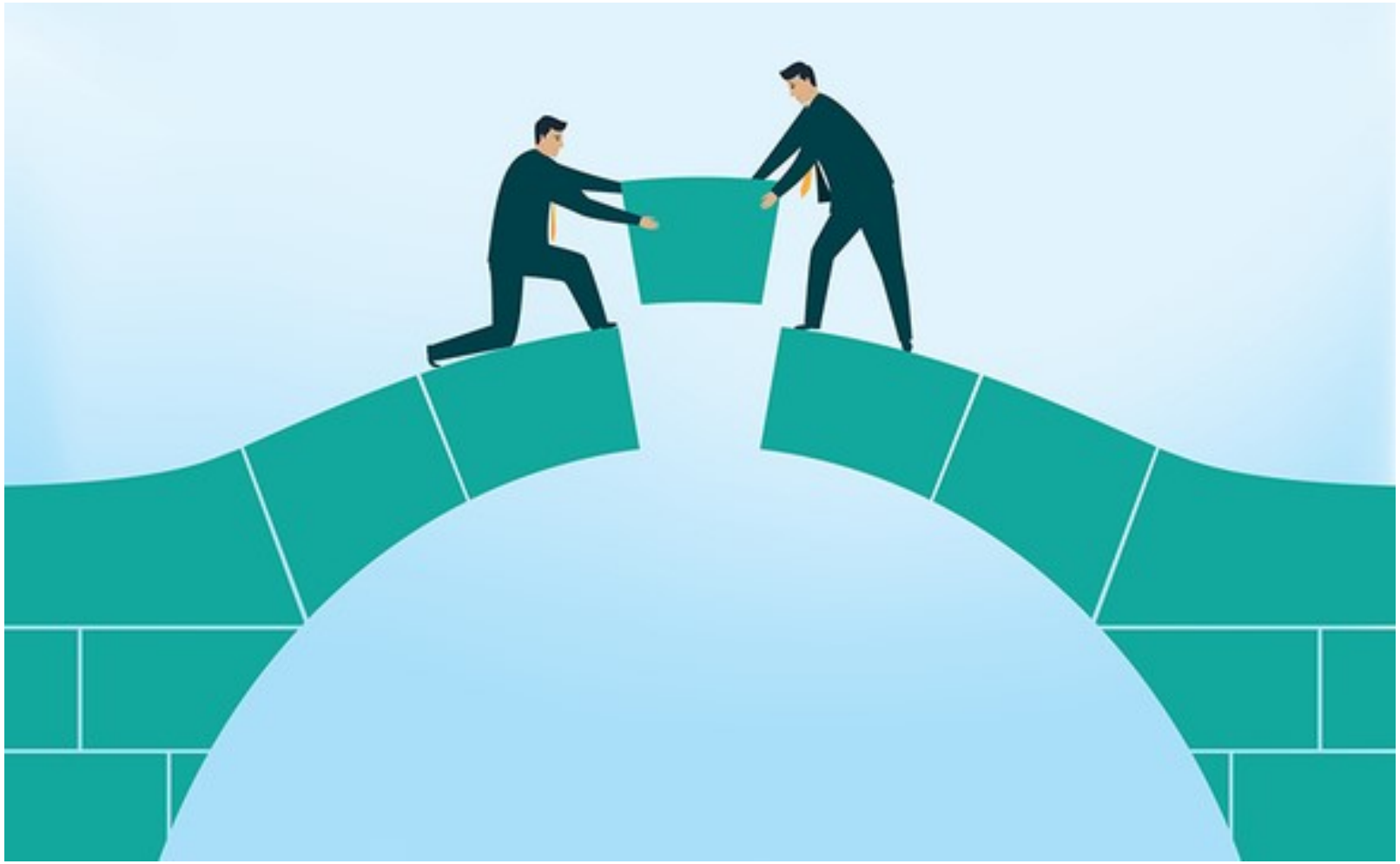
For example if we want to answer following questions:

- "As a trend over the past 6 months, what has been the average dollar sales of juices per target store per week in the Halifax area?"
- Which Targets sell toothpaste as the greatest percentage of their oral care category sales?
- Which targets are seeing the fastest growth in toothpaste sales?

- These are interesting questions, with potentially interesting answers, that might affect how both Target and toothpaste manufactures do business but they aren’t questions you can ask an OLTP system.
- OLTP systems can’t be used to answer most analysis questions.
- There are way too many records to search, sort and summarize. Can’t search, sort and summarize large numbers of records
- Can’t handle required calculations (Don’t forget the mathematical calculations that are also required to obtain the answers).
- Imposing these types of queries on the point-of-sale OLTP system on a regular basis would almost definitely interfere with the main business of ringing up shopper’s purchases (Negative impact on OLTP system performance.)

- OLTP systems gather raw data used for multidimensional analysis
- Raw data (millions and billions of transactions) has to be converted into something suitable for analysis.
- Converting raw data to something useful isn’t easy when the data resides within multiple, disparately organized and old legacy systems.

How to bridge the Gap?



- Nowadays company purchase packaged apps including some meaningful reporting capabilities.
- Packaged systems have 2 big limitations:
  1. Can only report on their own data – “Silos of data” or “Stovepipe reporting”
  2. Does not support high-speed multidimensional analysis

What is the alternative solution?

Solution

- Every large company has some sort of BI system to analyze operational data – where data from multiple operational systems, and possibly outside data sources, is pulled together for analysis – is part of every large company's infrastructure.
- Fortunately, all operational systems have export capabilities.
- Most BI systems designed to follow On-line Analytic Processing (OLAP) model.

Key OLAP Criteria

1. Must support multidimensional analysis

- Top manages/analysts have always thought multidimensionally.
- View “by ” qualifiers are usually dimensions.
- Questions like:

  What are our actual sales compared to the forecasted sales by region, by salesperson, by product?

  What is our profitability by product and by customer?

  What is our backlog by product, by customer and by time?

- OLAP systems organize data into multidimensional structures
- OLAP systems also provide tools for users to examine/filter dimensional data

2. Fast retrieval times

- Answer more questions in less times

3. Calculation engine that can handle specialized multidimensional math

- Lets analysts use simple formulas that are auto-performed across dimensions.

Dimensions

- Categorically consistent view of data. For example, all of the members within a dimension, such as products, belong together as a group.

- Two tests for dimensionality:

Can data about members be compared?

  - Sales numbers of one product compared to sales numbers of another product

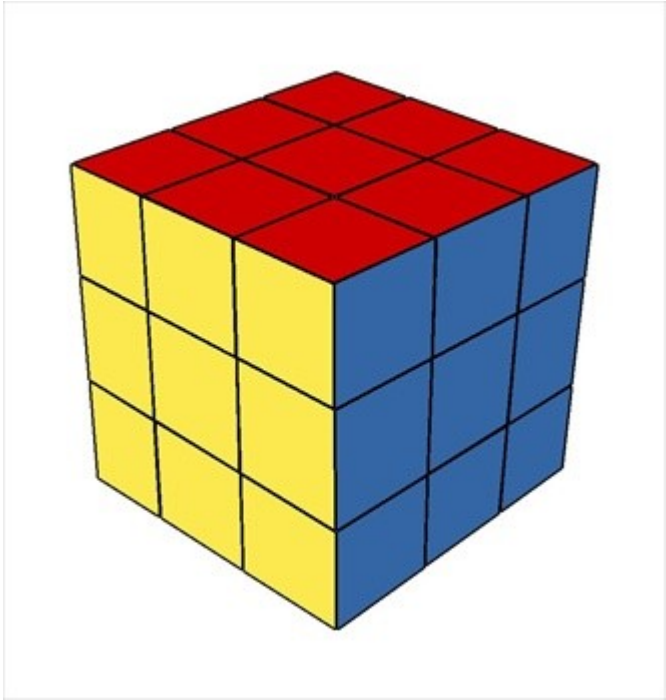
Can data from members be aggregated into summaries?

  - Jan, Feb, Mar aggregated together as Q1

OLAP

- Data cubes can have very large numbers of members

OLAP CUBE: Multi dimensional structure that stores and maintains discrete intersection values



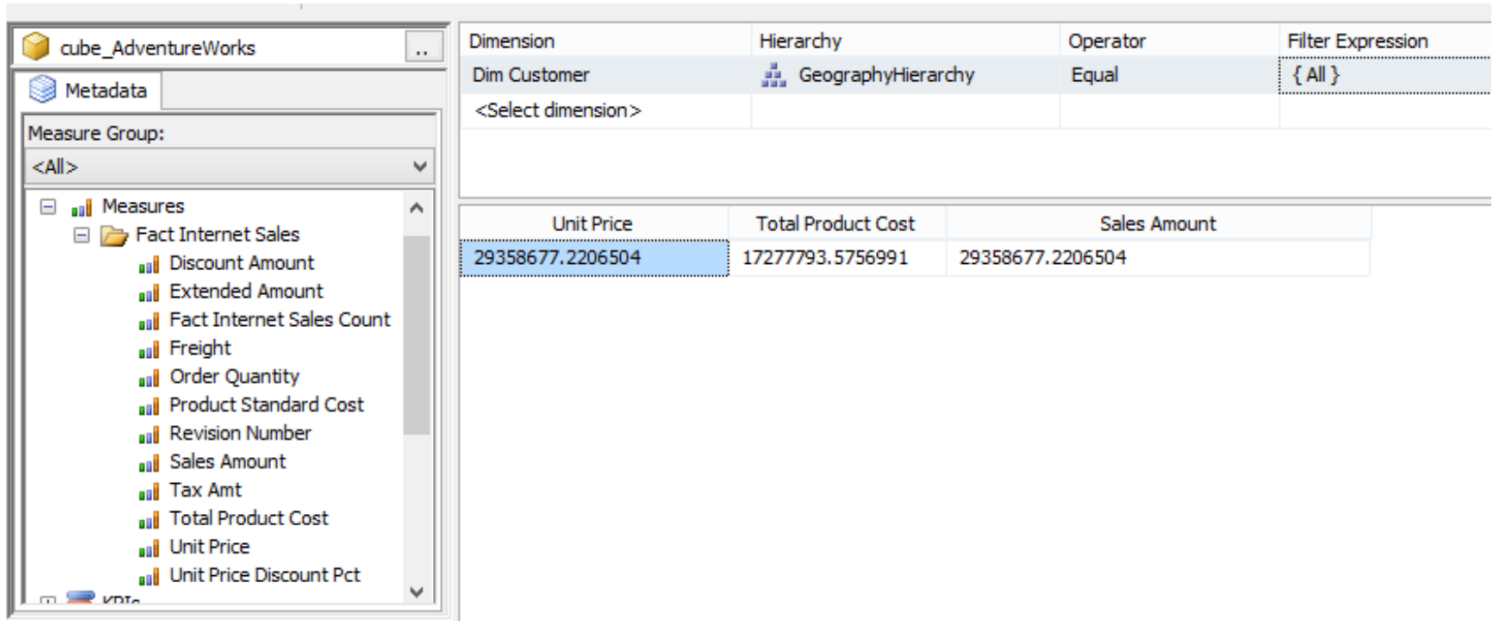
OLAP Operations

- **Roll-up**
  - Summarize the data by climbing up the hierarchy or by performing dimension reduction. Climbing along a dimension's concept hierarchy is also called roll-up, e.g. for time dimension, change from month-level to year-level, etc.
- **Drill-down**
  - Reverse of roll up.
  - We move from higher level summary to lower level summary or introducing new dimensions.
- **Slice**
  - Data selection based on predicate of one dimension.
- **Dice**
  - Slice on two or more dimensions.

- Roll-up

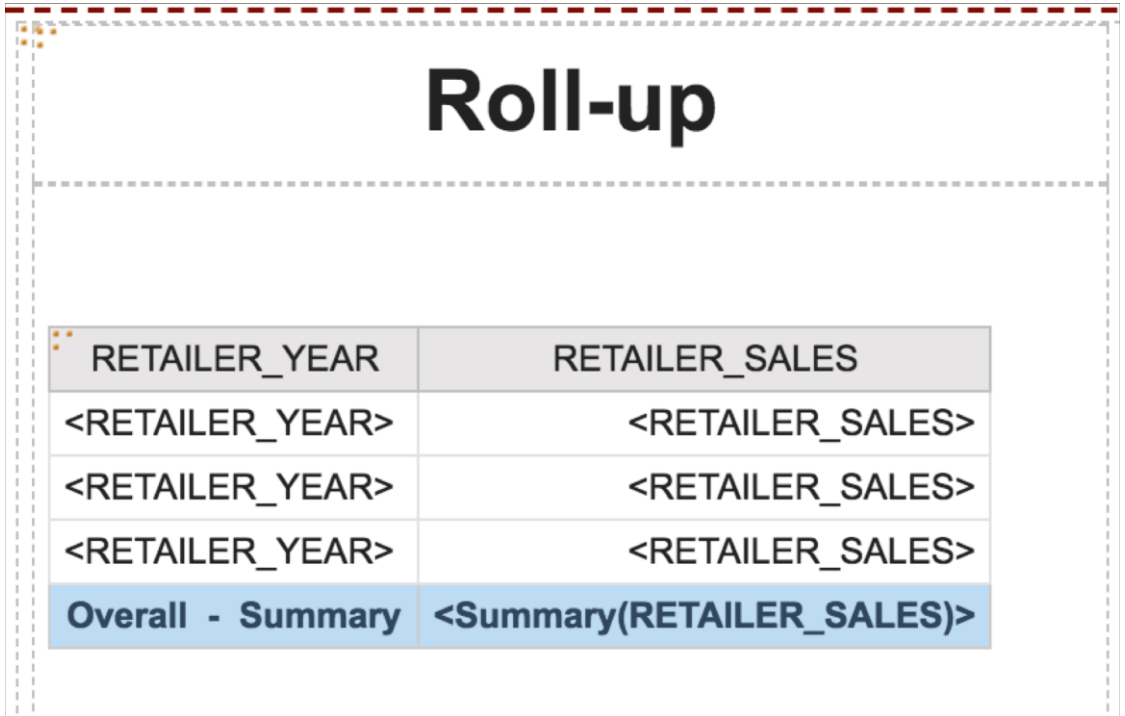
  Find out the total sales and production cost for all geographies.

Solution : Add ‘Sales Amount’ and ‘Total Production Cost’ measures to the data area. In the dimension area, select customer dimension → then select geography hierarchy and set filter to all geographies.



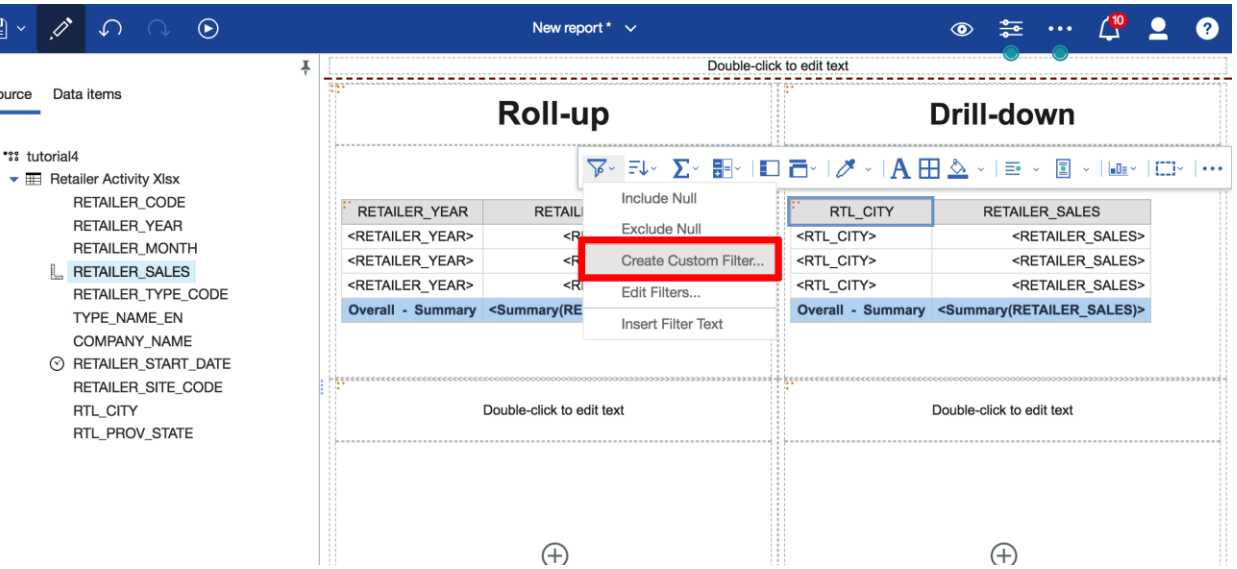
Roll-up

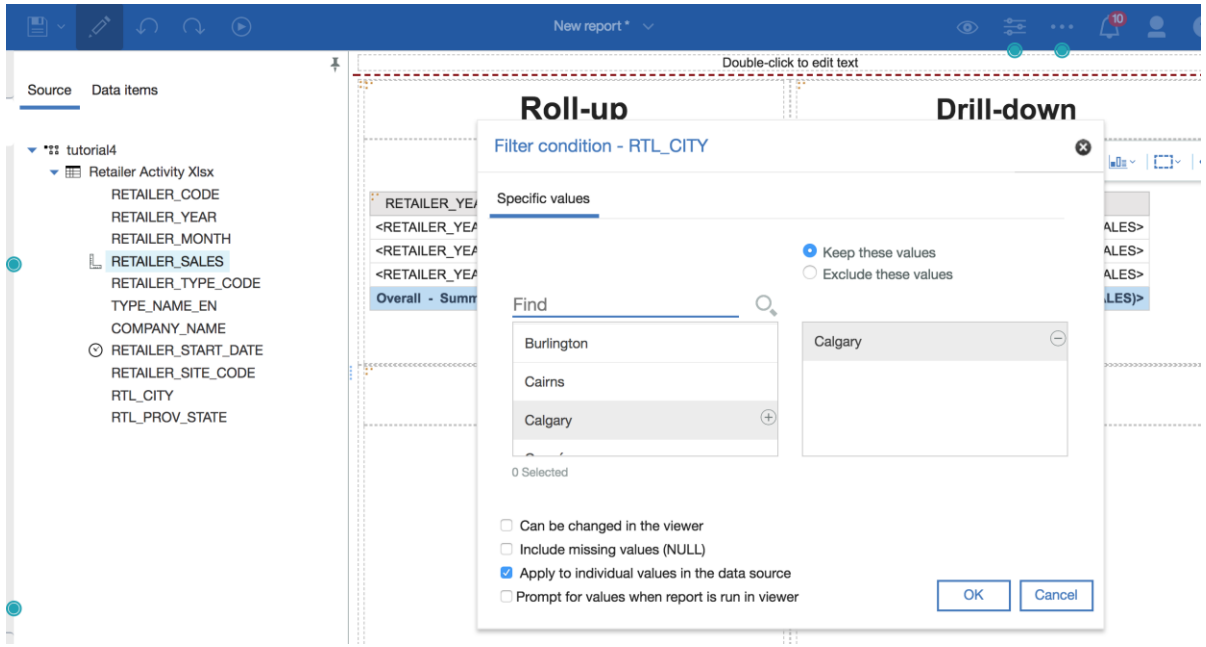
- Find out the total sales for all years
- Drag and Drop RETAILER_SALES and RETAILER_YEAR column to the list



Drill-down

- Find total sales in Calgary. Create custom filter for RTL_CITY column and select Calgary.

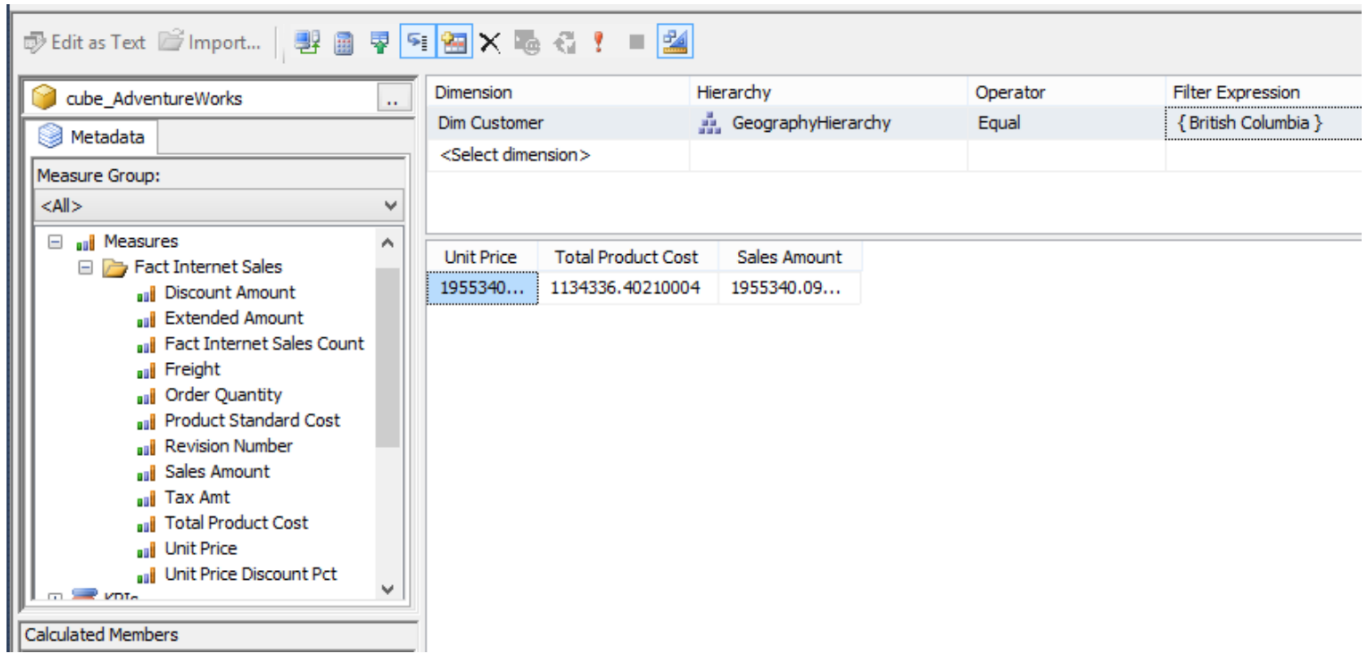




- Drill-down

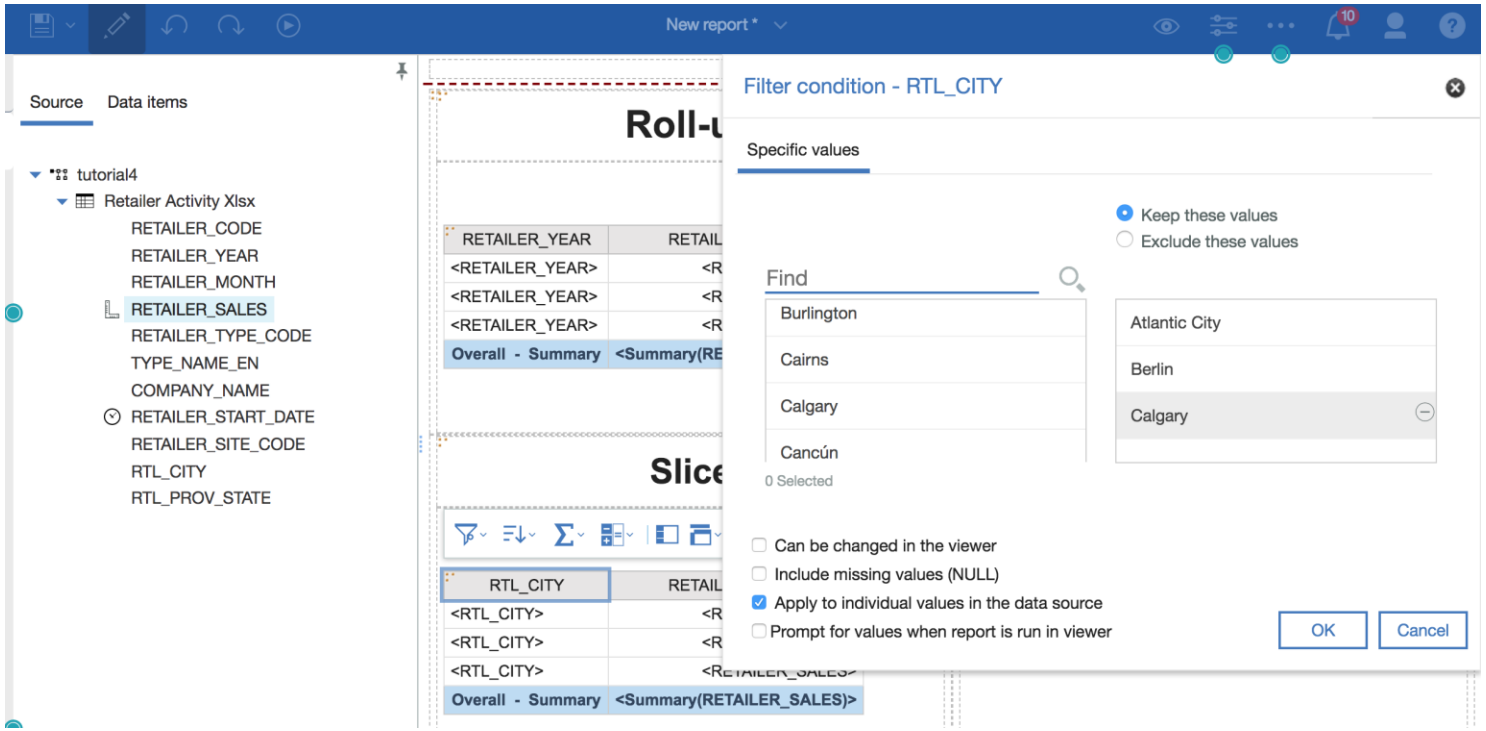
Find out the total sales and production cost in ‘British Columbia’.

Solution : Add ‘Sales Amount’ and ‘Total Production Cost’ measures to the data area. In the dimension area, select customer dimension → then select geography hierarchy and open the hierarchical filter and select ‘British Columbia’



Slice

- Find out total sales in Atlantic City, Berlin and Calgary. Create custom filter for RTL_CITY column and select 3 cities(Atlantic City, Berlin and Calgary).



- Slice

Find out the total sales and production cost in Canada, France & United States

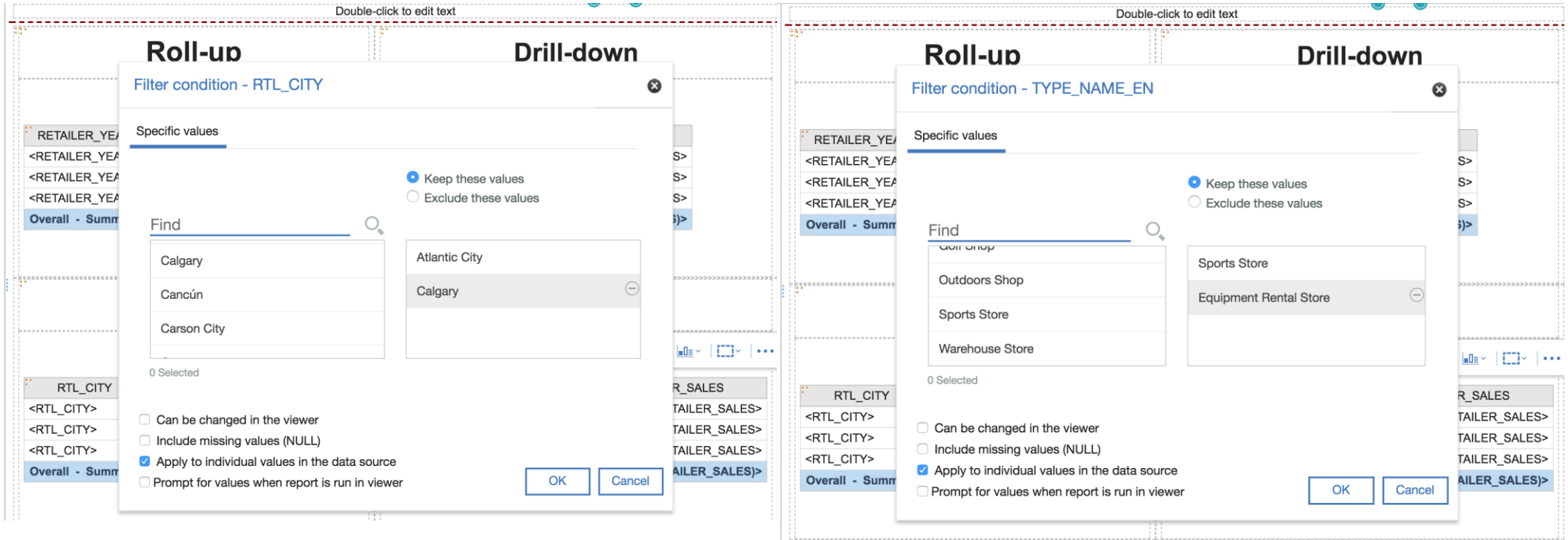
Solution : Add ‘Sales Amount’ and ‘Total Production Cost’ measures to the data area. In the dimension area, select customer dimension → then select geography hierarchy and open the hierarchical filter and select ‘Canada’, ‘France’ and ‘United States’.

Dimension	Hierarchy	Operator	Filter Expression
Dim Customer	GeographyHierarchy	Equal	{ Canada, France, United States }
<Select dimension>			

Unit Price	Total Product Cost	Sales Amount
14011652.0872046	8194485.06149869	14011652.0872046

Dice

- Find out the sales in Atlantic City and Calgary for Sports Store and Equipment Rental Store. Select custom filter for RTL_CITY and choose Atlantic City and Calgary. Also, select custom filter for TYPE_NAME_EN and choose Sports Store and Equipment Rental Store.



- Dice

Find out the total sales and production cost in Canada & United States for Bolts, Nuts and Vests Products.

Solution : Add ‘Sales Amount’ and ‘Total Production Cost’ measures to the data area. In the dimension area, select customer dimension → then select geography hierarchy and open the hierarchical filter and select ‘Canada’, and ‘United States’. Now select Product dimension → then select Product hierarchy, open the hierarchical filter and select all bolts, nuts & vests from the product hierarchy.

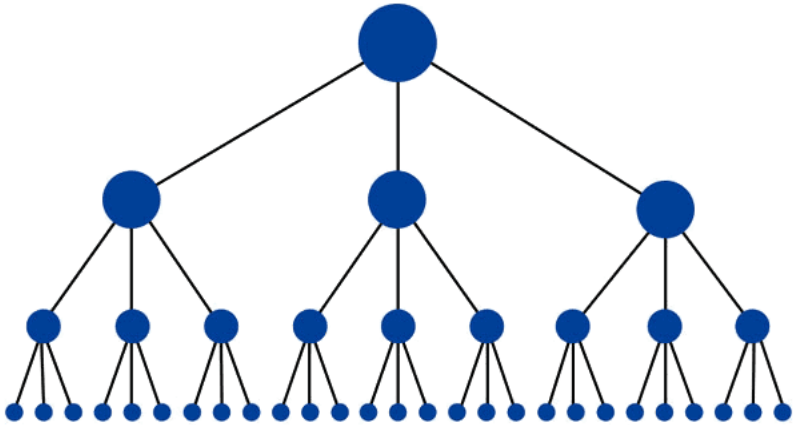
Dimension	Hierarchy	Operator	Filter Expression	Param
Dim Customer	GeographyHierarchy	Equal	{United States, Canada}	☐
Dim Product	ProductHierarchy	Equal	{Chainring Bolts, Classic Vest, L, Chainring Nut, Classic Vest, ...	☐
<Select dimension>				

Unit Price	Total Product Cost	Sales Amount
20447	7647.17799999995	20447

Question

- What is the difference between OLAP and OLTP?
- When should we use OLAP cube?

Hierarchies



- Typical analysis tasks:
  - Unit Sold, Average Price, Dollar Sales
  - 100 products
  - 24 months
  - 200 major cities

Total Data points: 1,440,000

Not all products sold in all cities during all months

- Hierarchy – Organizes data by levels
- Each level in the hierarchy is the aggregate of the levels beneath it
- Examples,

    Monthly data rolls up to quarters and years

    Cities roll up to regions and states

    Products roll up to product lines and groups

- Calculation, like Average Price, can be back-calculated at each hierarchy level.
- Hierarchies le you drill-down into data to explore interesting patterns and anomalies.

Other reasons to use Hierarchy?

- Start by exploring broad trends
- Become more focused as analysis progresses
- Top-down thinking is natural way for humans to organize complex info

Ad hoc Analysis

- Point-and-click drill-down is made usable by OLAP's rapid response model
- Lets managers and analysts perform ad hoc analysis
- Paper-based reporting gives fixed answers to fixed questions
- OLAP-based ad hoc analysis lets virtually any question be answered quickly.

Attributes

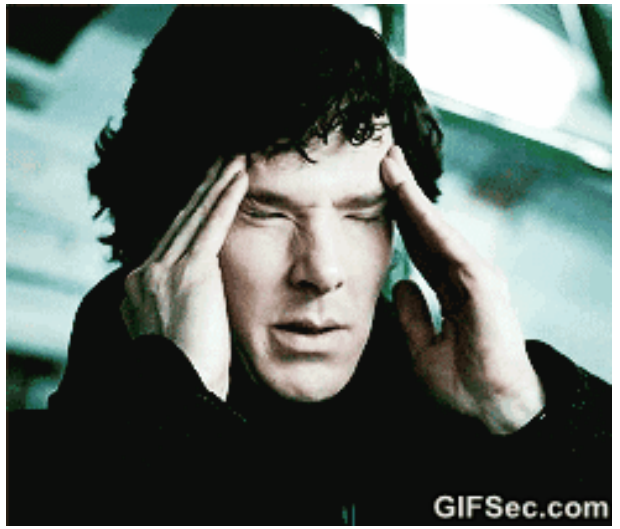
- Descriptive non-hierarchical information

- Examples:
  - Model number
  - Size
  - List Price
  - Color
  - Flavor
  - Street addresses

Measures

- Any quantitative expression contained in an OLAP system.
- A measure is the data that’s being analyzed across multiple dimensions
- Example: Sales of soda by month, by product, and by city

Measures



- Four important properties of a measure:
  - Always a quantity or expression that yields a quantity
  - Can take any quantitative format
  - Can be derived from any original data source or calculation
  - At least one measure required to perform OLAP Analysis

- Measures known by different names depending on application:

  - Metric/Key performance Indicator (KPI)
    - Represents important measures you should pay attention to
  - Benchmark
    - Refers to a measure used for making comparisons, for example, the average cost of goods sold
  - Ratio
    - A ratio is a measure where the result is calculated specifically from dividing one measure by another, such as sales per salesperson

Summary

- Analysis gap between raw data and BI can be bridged by combining OLTP systems with BI systems
- OLAP systems provide ad hoc analysis, slicing and dicing, pivoting dimensions, and drilling down through hierarchies.
- OLAP provides significant capabilities over standard single-dimensional analysis.

Pivot Table

- You actually creates an OLAP Cube without realizing it.
- When you create an OLAP Pivot Table, from a data model, an OLAP Cube is automatically created in the computer’s memory and is used to power an OLAP Pivot Table.
- Any time you can refresh the Cube using the current values in the source tables.
- With very large data sets, it could take an appreciable amount of time for Excel to reconstruct the cube.

- The vison for the Excel data model is that ordinary Excel users can create a ready-to-go OLAP Cube almost instantly. So, it is called Self-Service BI.
- Unlike traditional Business Intelligence solutions, it can be implemented by ordinary Excel users, and provides instant results because the OLAP Cube is generated automatically

Thank you!!!