CM2606 Data Engineering

Data Warehousing 02

Week 06 | Piumi Nanayakkara

ROBERT GORDON UNIVERSITY ABERDEEN
TEF Gold
INFORMATICS INSTITUTE OF TECHNOLOGY

Learning Outcomes

- Covers LO1 and LO2 for Module
- On completion of this lecture, students are expected to be able to:
  - Understand and explain the Data Warehouse Architecture
  - Apply OLAP Cube operations where applicable
  - Describe Data Warehouse security aspects in detail.

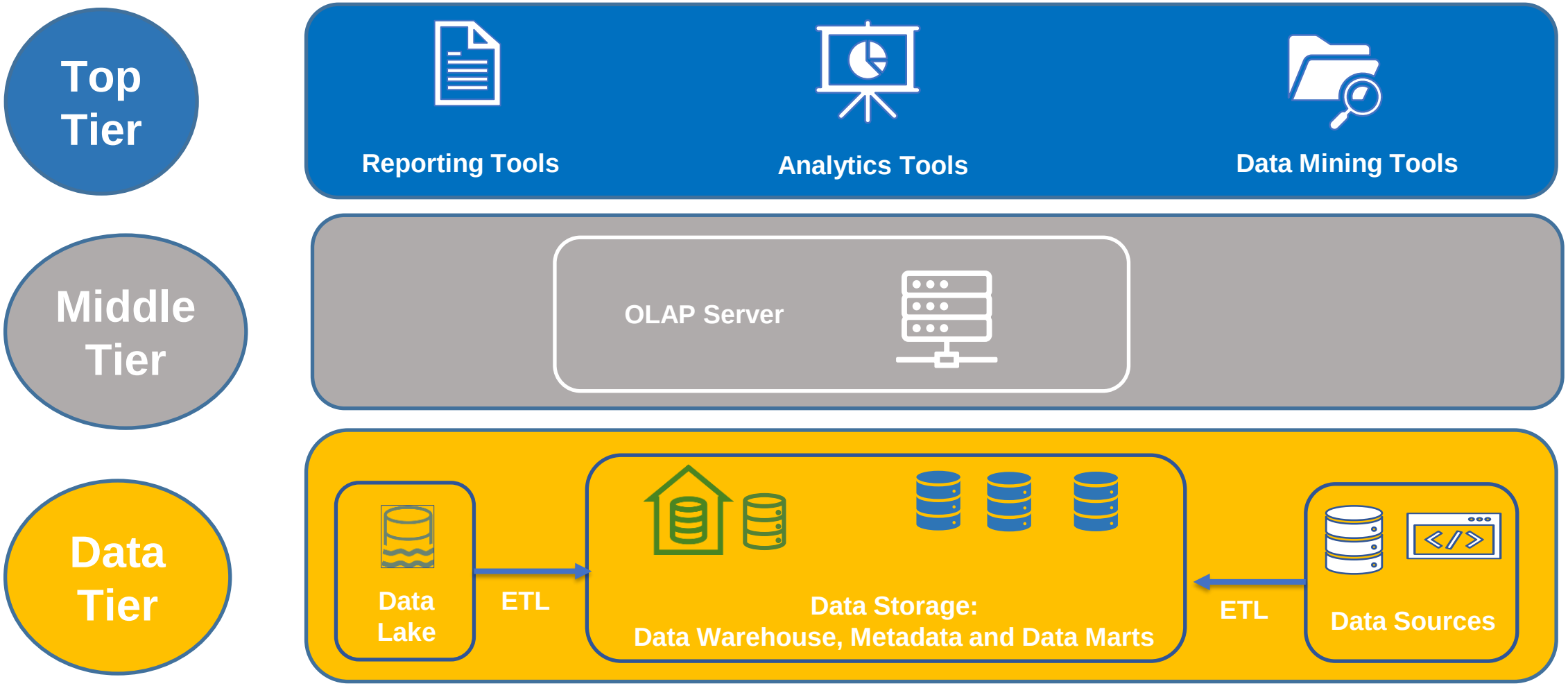
CONTENT

- Data Warehouse Architecture
- OLAP Server
  - Types of OLAP
  - OLTP Vs OLAP
  - OLAP Cubes Operations and Query
  - Vendor Comparison
- Data Mining
- Data Warehouse Security

Data Warehouse Architectures

- Single Tier:
  - Only layer physically available is the source layer Objective is to minimize the amount of data stored by removing data redundancy.
  - Data warehouses are virtual in the sense that multidimensional view of operational data is by specific middleware, or an intermediate processing layer
- Two Tier
  - Separates physically available sources and data warehouse.
  - Not scalable and only supports a nominal number of users.
- Three Tier
  - Most common type of modern DWH design as it
  - Produces an organized data flow from raw data to valuable insights.

Data Warehouse Architecture



Data Warehouse Architecture

- Data Tier:
  - Storage and the tools and utilities to feed data into the bottom tier
- Middle Tier
  - Implemented either based on Relational or Multi Dimensional Model
  - Can have multiple OLAP servers of different models
  - Depends on the volume and type of the data to be processed
- Top Tier
  - Serve as the face of the Data Warehouse system
  - Display results provided by OLAP, as well as additional tools for data mining.

Data Warehouse Metadata

- Specifies the source, usage, values, and features of data warehouse data.
  - It also defines how data can be changed and processed.

- Generally, contains data on:
  - Datawarehouse Content: tables, attributes, and keys
  - Data Sources: Row data source details
  - Data Refresh: Frequency, schedules and data pipeline run history
  - Data Transformations within DW: Lineage of transformations

- There are two types:
  - Technical Meta Data: Contains information about warehouse to be used by Data warehouse designers and administrators.
  - Business Meta Data: Contains detail that gives end-users a way easy to understand information stored in the data warehouse.

OLAP Server

- Act as the bridge between end users who are performing OLAP and the data storage which receives data from ETL pipelines.

- OLAP server performs multiple functions:
  - Data transformation: Uses operations such as filtering, summarizing, restructuring and data combining
  - Storage: the data should be stored in such a way as to maximize system flexibility, manageability and overall availability
  - Data presentation and access: server presents convenient data view and simple individual access for users

OLTP vs OLAP

OLTP	OLAP
Deals with current data	Deals with historical data
Operational purposes	Analytical purposes
Normal Form modelling	Dimensional modelling
Primitive and highly detailed data	Summarized and Consolidated data
Reads and Writes	Mostly reads
Usually in GB Scale	Go up to TB or PB scale
Fast, high performant	Flexible
Application oriented	Subject oriented

OLAP Server - Types

- **MOLAP (Multidimensional OLAP)**
  - Pre calculated data is stored in a **multidimensional cube**.
  - Fast data retrieval for intended use cases
  - Limited Information with pre calculated data

- **ROLAP (Relational OLAP)**
  - Data stored in the relational database
  - Can handle large volumes of data
  - Querying using SQL: slower performance and limited features

- **HOLAP (Hybrid OLAP)**
  - Combine the advantages of MOLAP and ROLAP.
  - For summary-type information, HOLAP leverages cube technology for faster performance.
  - When detail information is needed, HOLAP can "drill through" from the cube into the underlying relational data

OLAP Cube

- Implementation of MOLAP

- A multi-dimensional array of data.
  - E.g., Sales of a company per product, per store period during a certain time period where product, time period and stores are the dimensions
  - Some dimensions could be hierarchical e.g., Months >> Quarters >> Years

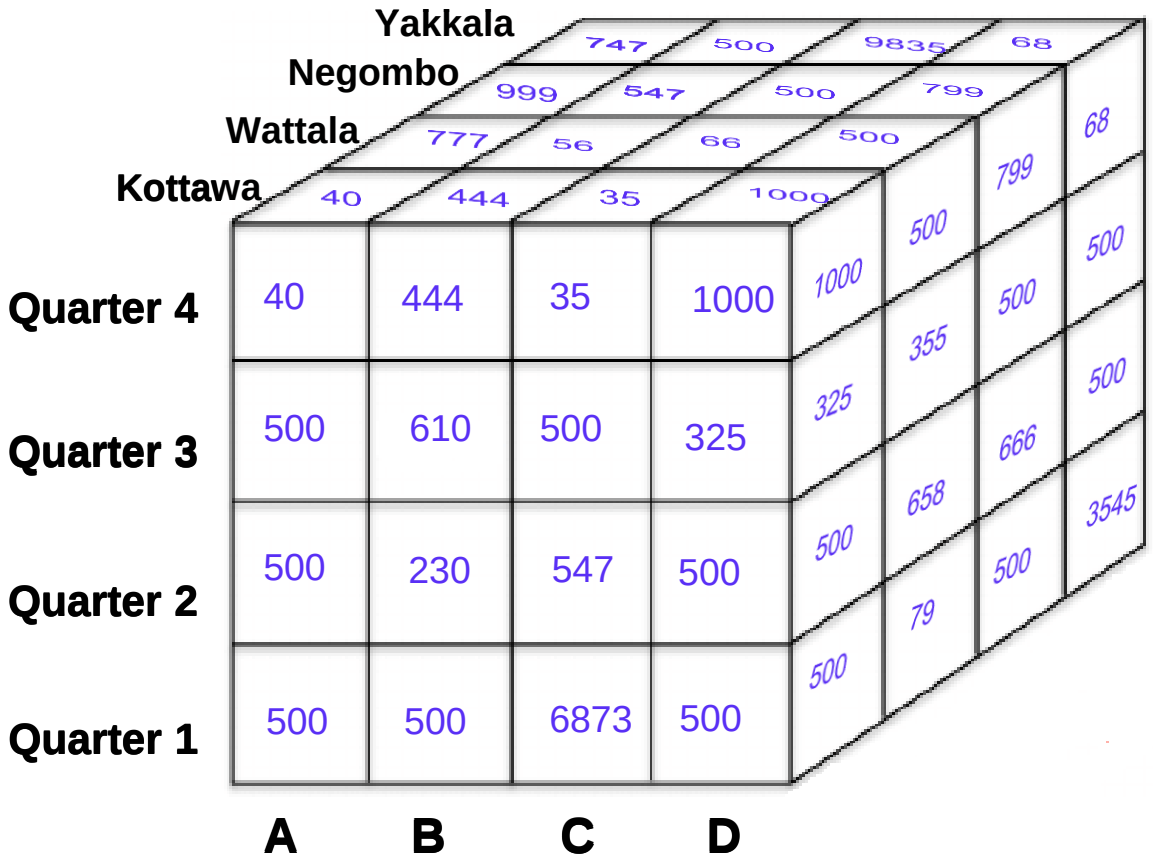
- Each cell of the cube holds a number that represents some **measure / fact** of the business. E.g., Sales

- Data can typically be derived from a dimensional data model stored in a relational data warehouse.
  - Considering a star / snowflake schema measures derived from fact table and dimensions derived from dimension tables.

OLAP Cube Performance

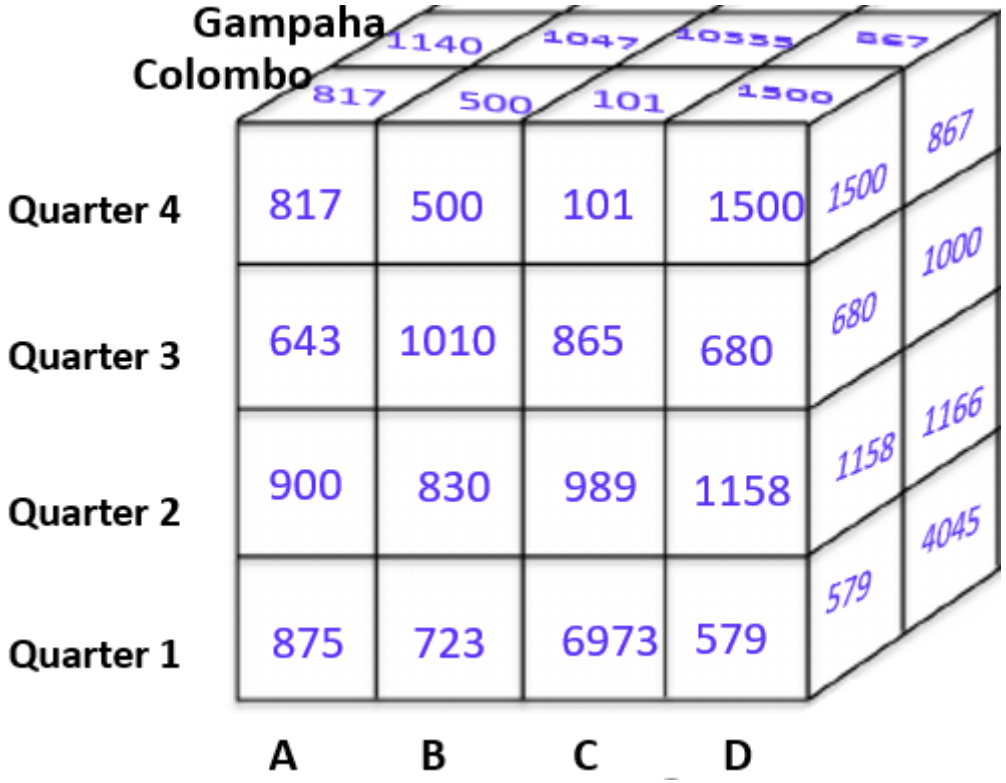
- Cubes are implemented with many optimizations such as:
  - Advanced data compression techniques
  - Various embedded data structures to help with indexing and filtering and scanning
  - Heuristics to figure out what to load into memory, what to pre-aggregate, and what to persist to disk.
  - Advanced array-processing techniques and algorithms for managing data and calculations.
    - process calculations in a fraction of the time required of relational-based products

OLAP Cube – Sales Quantities



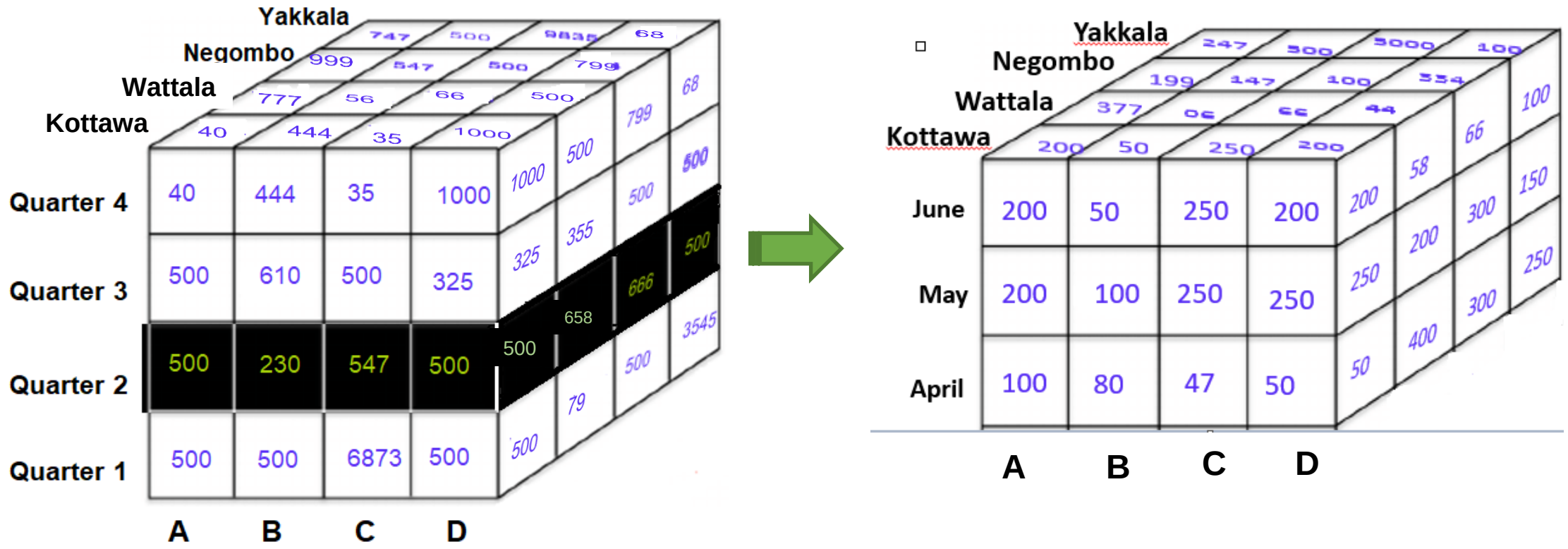
OLAP Operations: Roll Up

- Also known as drill-up or consolidation, use to summarize operation data along with the dimension.



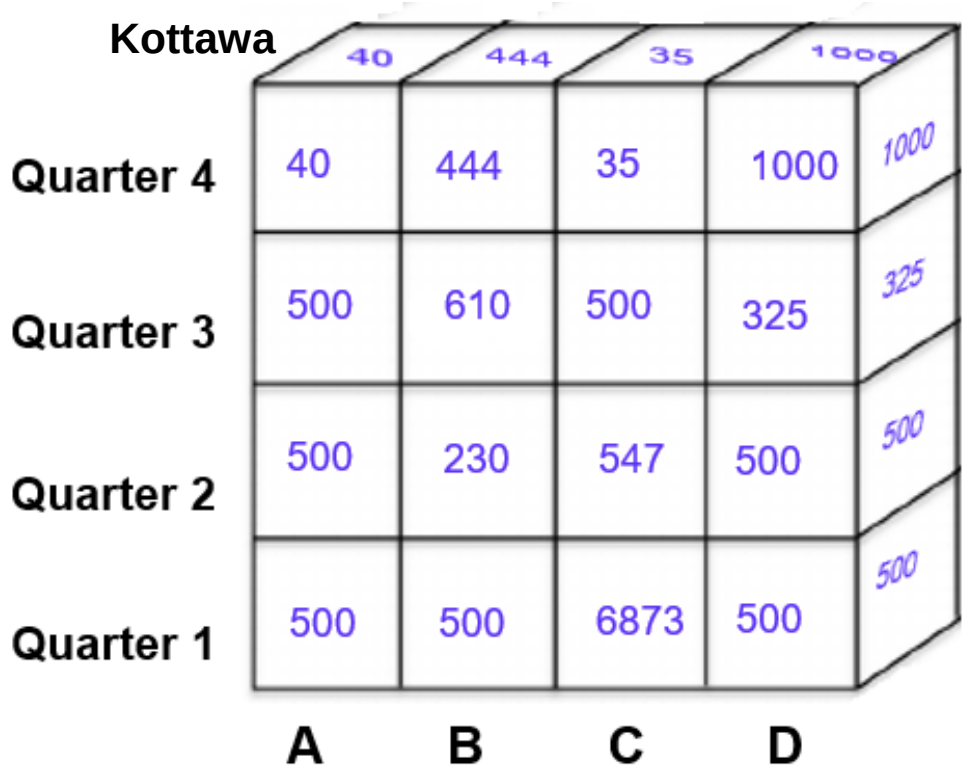
OLAP Operations: Drill Down

- Perform the analysis in deeper among the dimensions of data.



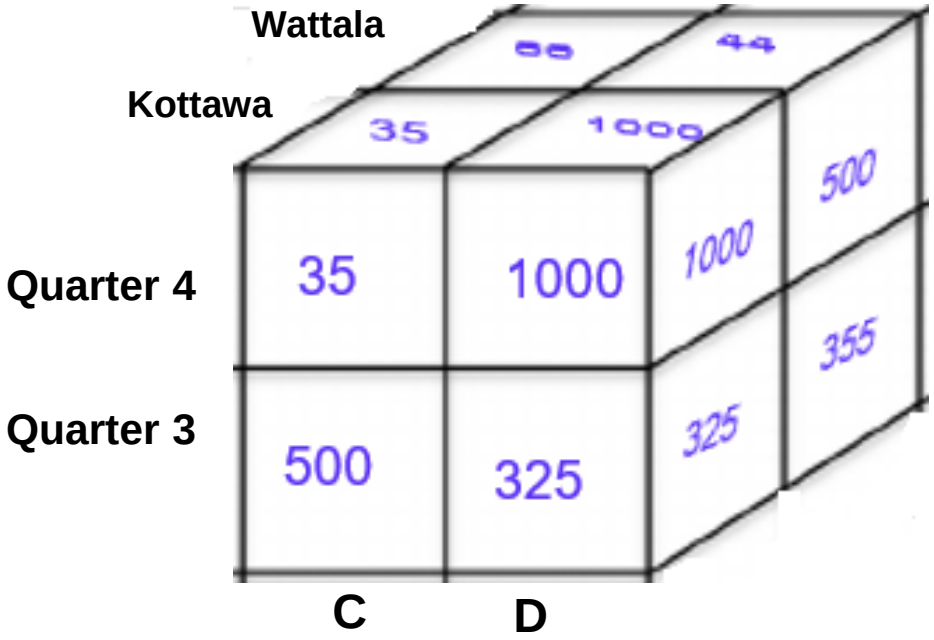
OLAP Operations: Slice

- Perform the analysis to take one level of information for display, such as “Sales in Kottawa outlet”.



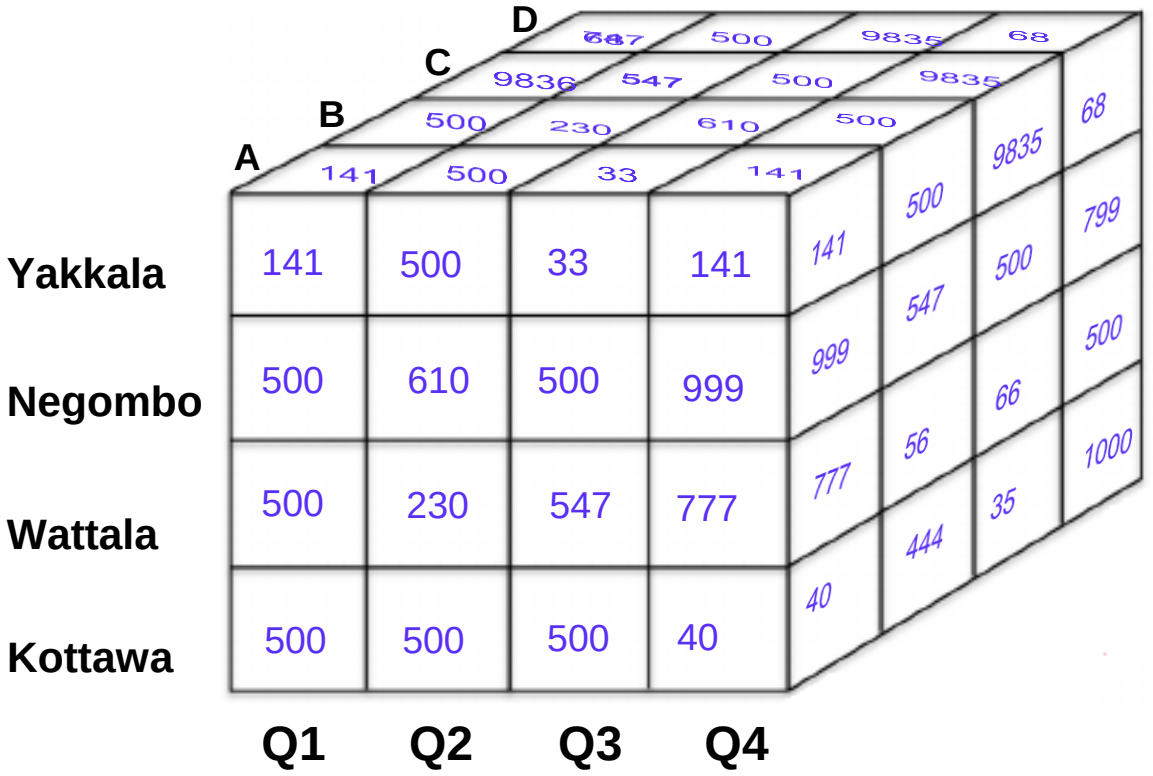
OLAP Operations: Dice

- Select data from multiple dimensions, such as “sales of Products C & D in Colombo District outlets during second half of the year.”



OLAP Operations: Pivot

- Perform the analysis that can gain a new view of data by rotating the data axes of the cube.



Querying OLAP Cubes

- Multidimensional Expressions (MDX)
  - Industry standard query and calculation language used to retrieve data from multidimensional data structures in OLAP databases
  - Introduced by Microsoft
  - Based on the XML for Analysis (XMLA) specification.
  - Possible to query data from SQL server and get a dataset with axis and cell data
  - Also supports data definition and data manipulation

- SQL
  - Oracle OLAP cube data is directly accessible to SQL by a set of relational views.
  - These views represent an OLAP cube as a star schema

MDX Query

```
SELECT { [Measures].[Unit Sales], [Measures].[Store Sales] } ON COLUMNS, {
[TIME].[2019], [TIME].[2020] } ON ROWS FROM Sales WHERE (
[Store].[KOTTAWA].[GAMPAHA] )
```

Result:

	Unit Sales	Store Sales
2019	900	17000
2020	555	12876

Disadvantages of MOLAP

- Relevant data must be transferred from relational systems
  - potentially “redundant” re-creation of data in another (multidimensional) database.
- Once data has been transferred, there may be no simple means for updating the MOLAP “engine”
- Also, MOLAP products are typically proprietary systems.
  - For some IT departments, introducing a new database system is an issue.

Future of OLAP Cubes

- Too complex and costly to implement for small companies
  - Giants have already moved away from this
- Unlike early days computing power / memory is cheaper
  - Resources to process a complex SQL query is affordable
- Might be replaced by Columnar databases
  - Use Materialized views instead of Cubes
  - Read efficient
  - Compress better → more data may be loaded into memory when you're running an aggregation query

OLAP Server Vendors

- Microsoft Analysis Services
  - The Microsoft SQL Server 7.0 (1998) OLAP Services supports a hybrid OLAP server
  - Microsoft then released Analysis services (2000, 2005) which offered Data Mining capabilities on top of OLAP features
- Oracle Database OLAP Option
  - This option is marketed as an extra-cost option to supplement the "Enterprise Edition" of its database
  - Provides native multidimensional storage and speed-of-thought response times. rich support for analytics such as time series calculations, forecasting, advanced aggregation.
- Apache Druid
  - Commonly used in applications to analyze high volumes of real-time and historical data
  - Used in production by technology companies such as Alibaba, Airbnb, Cisco, eBay, Netflix, Paypal, Pinterest, Twitter

OLAP Server Vendors

- IBM Planning Analytics
  - Full analytics platform containing database server, an ETL tool, server management and monitoring tools and several front tier application.
  - Formerly IBM Cognos TM1. Now database server referred as TM1.
  - Data is stored in in-memory multidimensional OLAP cubes, generally at the leaf level, and consolidated on demand.
  - Computations on the data are performed in near real-time, without the need to precalculate, due to a highly performant database design and calculation engine.

OLAP Server Vendors

	License / Pricing	Type	Querying
Apache Druid	Apache 2.0 / Free	Hybrid	Druid SQL
IBM Planning Analytics	Proprietary	MOLAP	XMLA/MDX
Microsoft Analysis Services	Proprietary	Hybrid	XMLA/MDX/SQL
Oracle Database OLAP Option	Proprietary	ROLAP	MDX/SQL

Data Mining

- Refers to mining knowledge from a huge amount of data
  - This includes finding hidden patterns in data and making predictions into the future

- Vs. OLAP
  - Deals with detailed level data as opposed to summary data in OLAP
  - Focus to make predictions for future where as in OLAP focus is to analyze current status and improvise the current process
  - Complements each other as well:
    - OLAP suggests sales reduction in a certain branch, this branch's detailed transactions can be analyzed using data mining
    - Data Mining predicts 5% increase in sales which can be used by OLAP to arrive and total net sales values

Materialized Views

- A normal database view acts as a virtual table on top of the base table(s) containing original data containing resultant data set from a query.
- In contrast a materialized view caches the qury result as a concrete ("materialized") table which would be updated from the original base tables from time to time.
- Provides efficient access, at the cost of extra storage and of some data being potentially out-of-date.
- Heavily used in data warehousing since frequent queries of the actual base tables can be expensive.

Data Warehouse Security: Data Classification

- Data Warehouses could contain:
  - Data that are subject to disclosure where there's minimal use in applying security as well as
  - Sensitive data such as PII data, Financial data, health records which needs to be handled carefully.

- Upon classification splitting it into separate tables and storing such information on different servers could be considered

Data Warehouse Security: Standards

- Based on the business there could be certain security compliance requirements that needs to be adhered.
  - HIPAA: The Health Insurance Portability and Accountability Act – 1996
  - GDPR: The General Data Protection Regulation 2016

- FIPS 140-2 certified software should be for data encryption.
  - FIPS 140-2 is a U.S. government computer security standard for cryptographic modules that ensures the highest levels of security.

Data Warehouse Security: Data at Rest

- Set up data warehouse to be read-only by default
  - Prevents any dangerous SQL write statements and accidental deletes
- Authorization Rules / Access Control
  - Can be implemented in a hierarchical approach or role-based approach Access can be restricted at schema level, table/view level or row level
  - “Need to Know” or “Least Privilege” principles offer the highest security
- Data Encryption
  - Only aggregated/summarized data can be exposed to users though this will limit what type of analysis can be performed
  - Might degrade performance

Data Warehouse Security: Data in Transit

- Networks / hardware should be secured
  - Network security solutions like firewalls and network access control
- Communication Protocols
  - use encrypted connections (HTTPS, SSL, TLS, FTPS, etc.)
- Cloud Migration
  - VPNs provide required security when moving data by isolating the communication between on-premises databases and the cloud-based data warehouse

READING

- Surajit Chaudhuri and Umeshwar Dayal. 1997. An overview of data warehousing and OLAP technology. SIGMOD Rec. 26, 1 (March 1997), 65–74