CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

APEX INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Introduction to Data Science (21CST-292)

Faculty: Dr. Jitender Kaushal (E14621)

Associate Professor

Lecture - OLTP and OLAP

Introduction to Data Science: Course Objectives

COURSE OBJECTIVES

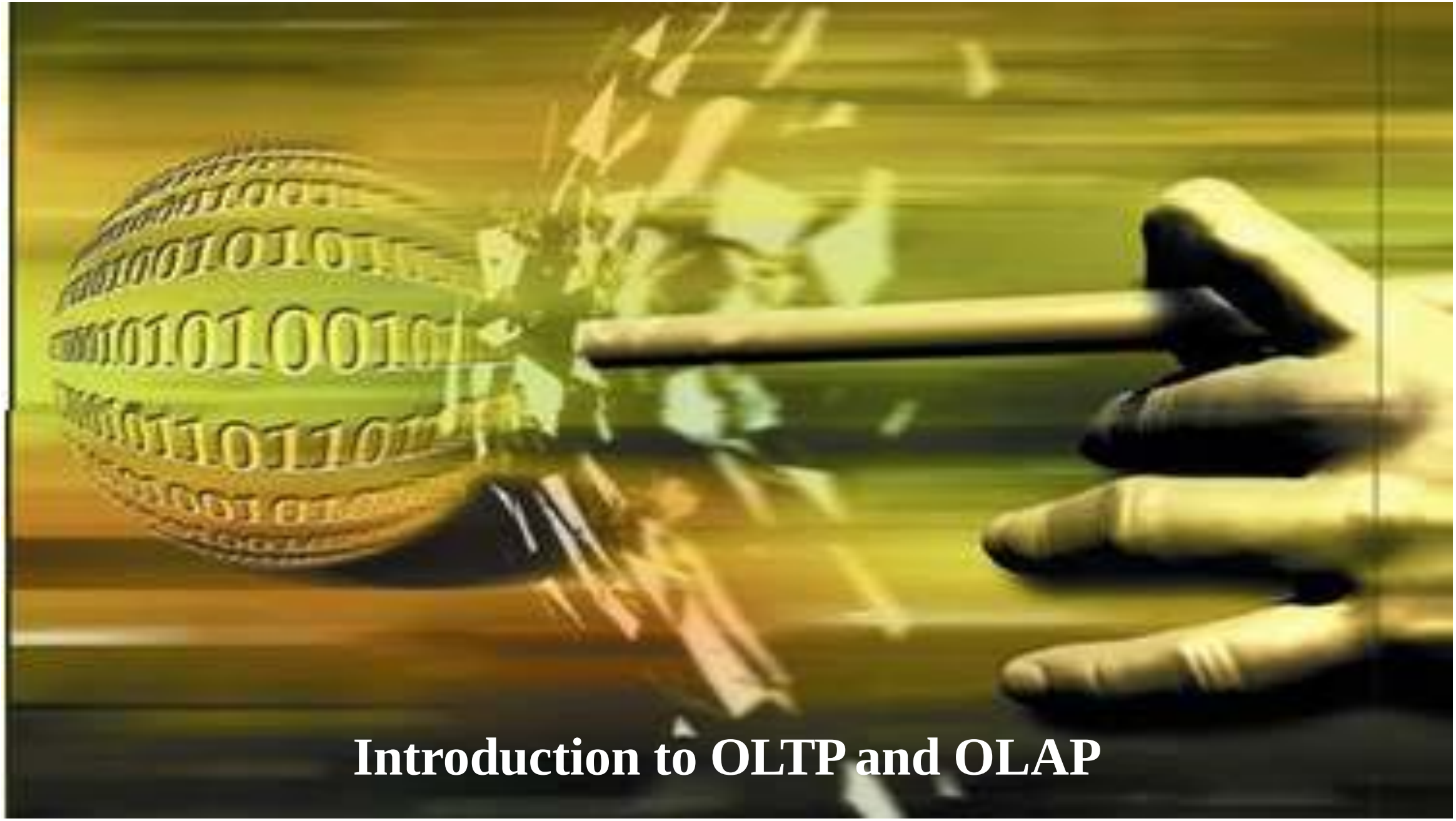
The Course aims to:

- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation and Loading
- To prepare a sample project in Hadoop Environment

COURSE OUTCOMES

On completion of this course, the students shall be able to:-

CO3	To learn and understand Data Life Cycle, Data Preparation.



Introduction to OLTP and OLAP

What is OLTP?

Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.

The primary objective is data processing and not data analysis

Examples: Uses of OLTP are as follows:

•ATM center is an OLTP application.

•OLTP handles the ACID properties during data transactions via the application.

•It’s also used for Online banking, Online airline ticket booking, sending a text message, add a book to the shopping cart.

USAGE & APPLICABILITY

- Used for transaction oriented applications
- Used by lower level employee
- Quick updates and retrievals
- Many users accessing the same data
- Users are not technical persons
- Response rate is very fast
- Single transaction (one application) at a time

- **Online Transactional Processing**, systems handle a large number of transactions happening in real-time. But, what are the transactions?
- **Transactions** are processes that occur in their entirety and in isolation from one another. They either insert, update, or delete data in a database. On successful execution, the changes made by a transaction to a database persist in the database even in the event of a system failure.

- OLTP governs transactions because they are the critical processes that we encounter in our everyday life.
- Online transactions, e-commerce orderings, online hotel bookings, atm transactions, etc. are all managed by OLTP processes.
- The transactional data is stored in **Relational Databases** that ensure **ACID** properties for transactions. This data is written and queried at a very high pace to prevent any delay in processing.

PRACTICAL EXAMPLE

Imagine you log onto an e-commerce website to book the last pair of your favorite headphones which are currently on sale. Consider the following:

- Multiple people might be trying to book the headphones but none are aware of the processes of the others. **(Isolation)**
- *The order will be considered successful only when the entire steps along with the payment are completed by any user.* **(Atomicity)**
- Once the order is successfully completed by a user, it will be updated in the website database. The headphones will then become unavailable on the website. **(Consistency)**
- *Now, even if the e-commerce website goes down due to a deluge of user traffic, the user still owns the headphone they bought successfully.* **(Durability)**

*OLTP ensures that such transactions are carried out without any inconsistencies in the database with the help of the **ACID (Atomicity, Consistency, Isolation, Durability)** properties*

OLTP
(ONLINE TRANSACTION PROCESSING SYSTEM)

- Stores routine data
- Follows client server model
- Applications
  - Banks
  - Retail stores
  - Airline reservation

OLTP

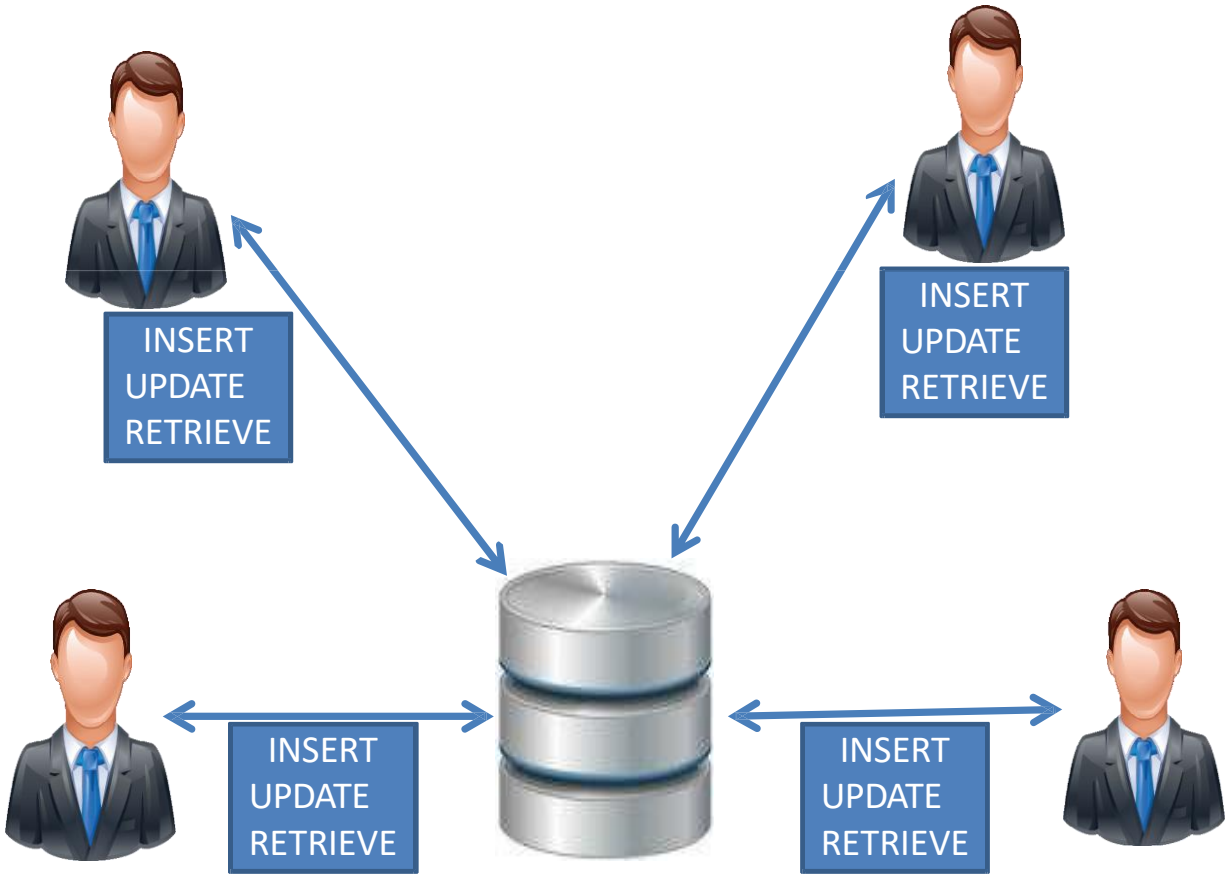
(ONLINE TRANSACTION PROCESSING SYSTEM)

TRANSACTIONS

- Single event that changes something
- Different types of transactions
  - Customer orders
  - Receipts
  - Invoices
  - Payments

- Processing of transactions include storage and editing of data
  - When transaction is completed then the records of an organization are changed

TRANSACTIONS



TRANSACTIONS

Cash at register gone up

Inventory of video game gone down

Ordering of new video game for the store

OLTP Segmentation

- They can be segmented into:
  - Real-time Transaction Processing
  - Batch Processing

Real-time Transaction processing

- Multiple users can fetch the information
- Very fast response rate
- Transactions processed immediately
- Everything is processed in real time

Batch Processing

- Where information is required in batch
- Offline access to information
- Presorting (sequence) is applied
- Takes time to process information

Characteristics of OLTP

- It handles real-time transactions.
- These systems modify data in the database.
- They handle transactions that are governed by the ACID properties.
- These systems store data in Relational Databases.
- The implementation of OLTP transactions is usually very fast, in the order of milliseconds.

Advantages and Challenges of an OLTP System

Advantages of an OLTP System

- Simplicity – It is designed typically for use by clerks, cashiers, clients, etc.
- Efficiency – It allows its users to read, write and delete data quickly.
- Fast query processing – It responds to user actions immediately and also supports transaction processing on demand.

Challenges of an OLTP System

- Security – An OLTP system requires concurrency control (locking) and recovery mechanisms (logging).
- OLTP system data content not suitable for decision making – A typical OLTP system manages the current data within an enterprise/organization. This current data is far too detailed to be easily used for decision making.

The Queries that OLTP Cannot Answer

- The supermarket store is deciding on introducing a new product. The key questions they are debating are: **“Which product should they introduce?” and “Should it be specific to a few customer segments?”**

- The supermarket store is looking at **offering some discount on their year-end sale**. The questions here are**: “How much discount should they offer?”** and **“Should it be different discounts for different customer segments?”**
- The supermarket is looking at **rewarding its most consistent salesperson**. The question here is: “How to zero in on its most consistent salesperson (consistent on several parameters)? All the queries stated above have more to do with analysis than simple reporting”
- Ideally these queries are not meant to be solved by an OLTP system.

What is OLAP?

Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time.

The primary objective is data analysis and not data processing.

Any type of Data warehouse system is an OLAP system. The uses of OLAP are as follows:

•Spotify analyzed songs by users to come up with a personalized homepage of their songs and playlist.

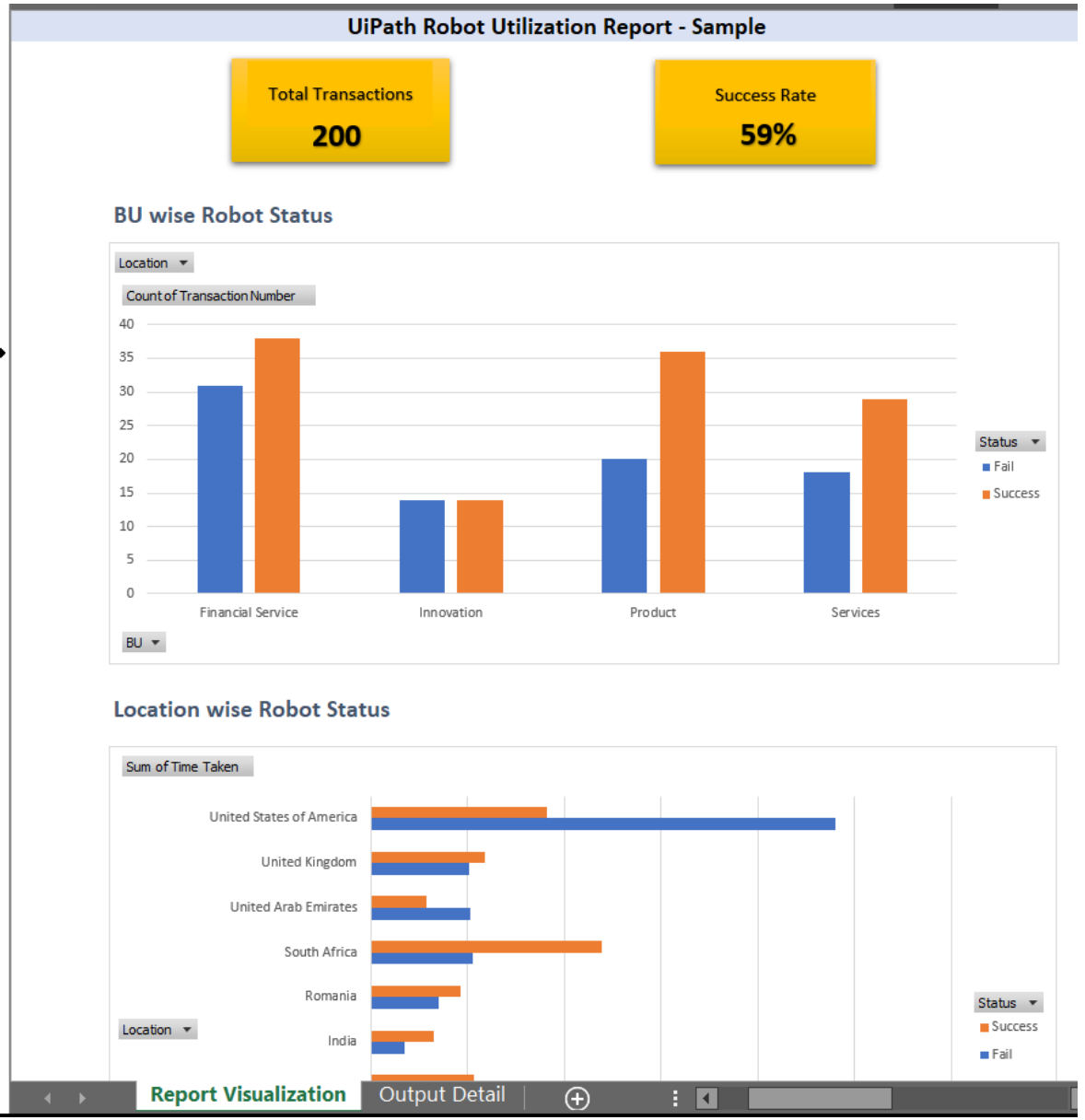
•Netflix movie recommendation system.

OLAP - Online Analytical Processing

- OLAP differs from traditional databases in the way data is conceptualized and stored.
- In OLAP data is held in the **dimensional form** rather than the relational form.
- OLAP's lifeblood is **multi-dimensional data**.
- OLAP tools are based on the multi-dimensional data model. The multi-dimensional data model views data in the form of a data **cube**.
- **Online Analytical Processing (OLAP) is a technology that is used to organize large business databases and support business intelligence.**
- OLAP databases are **divided into one or more cubes**. The cubes are designed in such a way **that creating and viewing reports become easy**.
- OLAP databases are divided into one or more cubes, and each cube is **organized and designed by a cube administrator** to fit the way that you **retrieve and analyze data** so that it is easier to create and use the **PivotTable reports and PivotChart reports** that you need.

Transaction Number	Location	BU	Status	Time Taken
JIP7542789	India	Product	Success	1:01:00
JIP7542790	Canada	Product	Success	0:59:00
JIP7542791	South Africa	Product	Fail	0:57:00
JIP7542792	United States of America	Services	Success	0:55:00
JIP7542793	South Africa	Services	Fail	0:53:00
JIP7542794	United States of America	Innovation	Fail	0:51:00
JIP7542795	South Africa	Product	Success	0:49:00
JIP7542796	United States of America	Product	Success	0:47:00
JIP7542797	South Africa	Services	Fail	0:45:00
JIP7542798	India	Services	Success	0:57:00
JIP7542799	India	Product	Success	0:55:00
JIP7542800	United States of America	Product	Fail	1:01:00
JIP7542801	South Africa	Services	Success	2:01:00
JIP7542802	India	Services	Fail	1:01:00
JIP7542803	India	Financial Service	Fail	1:19:00
JIP7542804	India	Product	Success	1:37:00
JIP7542805	India	Services	Success	0:45:00
JIP7542806	India	Services	Fail	0:57:00
JIP7542807	India	Services	Success	0:55:00
JIP7542808	India	Product	Success	0:45:00
JIP7542809	United States of America	Product	Fail	0:57:00
JIP7542810	India	Services	Success	0:55:00
JIP7542811	United States of America	Innovation	Fail	1:01:00
JIP7542812	United States of America	Product	Fail	1:19:00
JIP7542813	United States of America	Financial Service	Success	1:37:00
JIP7542814	United States of America	Financial Service	Success	0:57:00
JIP7542815	India	Financial Service	Fail	0:55:00
JIP7542816	Germany	Innovation	Fail	0:57:00
JIP7542817	Australia	Innovation	Success	0:55:00
JIP7542818	Australia	Innovation	Success	1:01:00
JIP7542819	Australia	Services	Fail	2:01:00
JIP7542820	Australia	Financial Service	Success	2:01:00
JIP7542821	Germany	Services	Fail	1:19:00
JIP7542822	Germany	Product	Fail	1:37:00
JIP7542823	Germany	Financial Service	Success	2:01:00
JIP7542824	Germany	Product	Fail	2:01:00
JIP7542825	Australia	Product	Fail	2:01:00
JIP7542826	Australia	Financial Service	Success	2:01:00
JIP7542827	Australia	Financial Service	Fail	2:01:00
JIP7542828	France	Financial Service	Fail	0:55:00

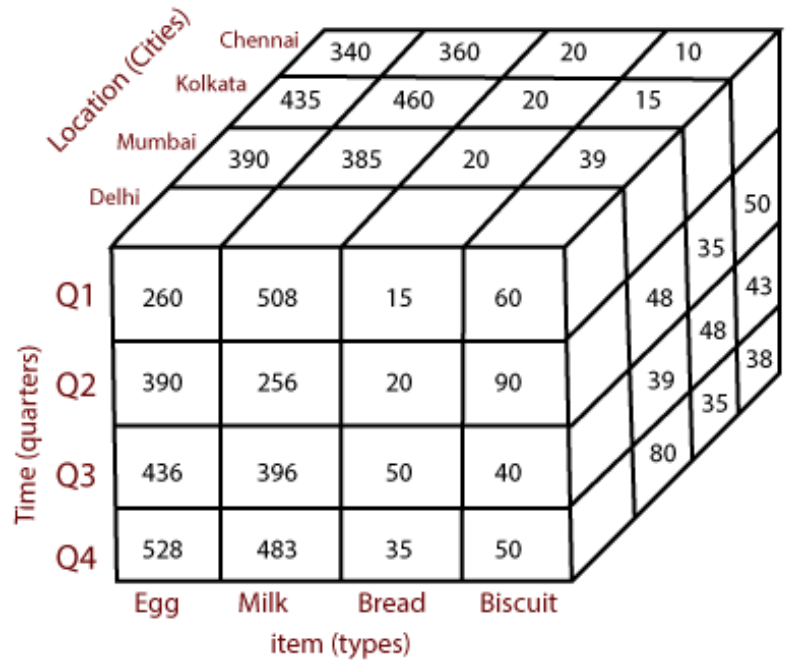
ReportData

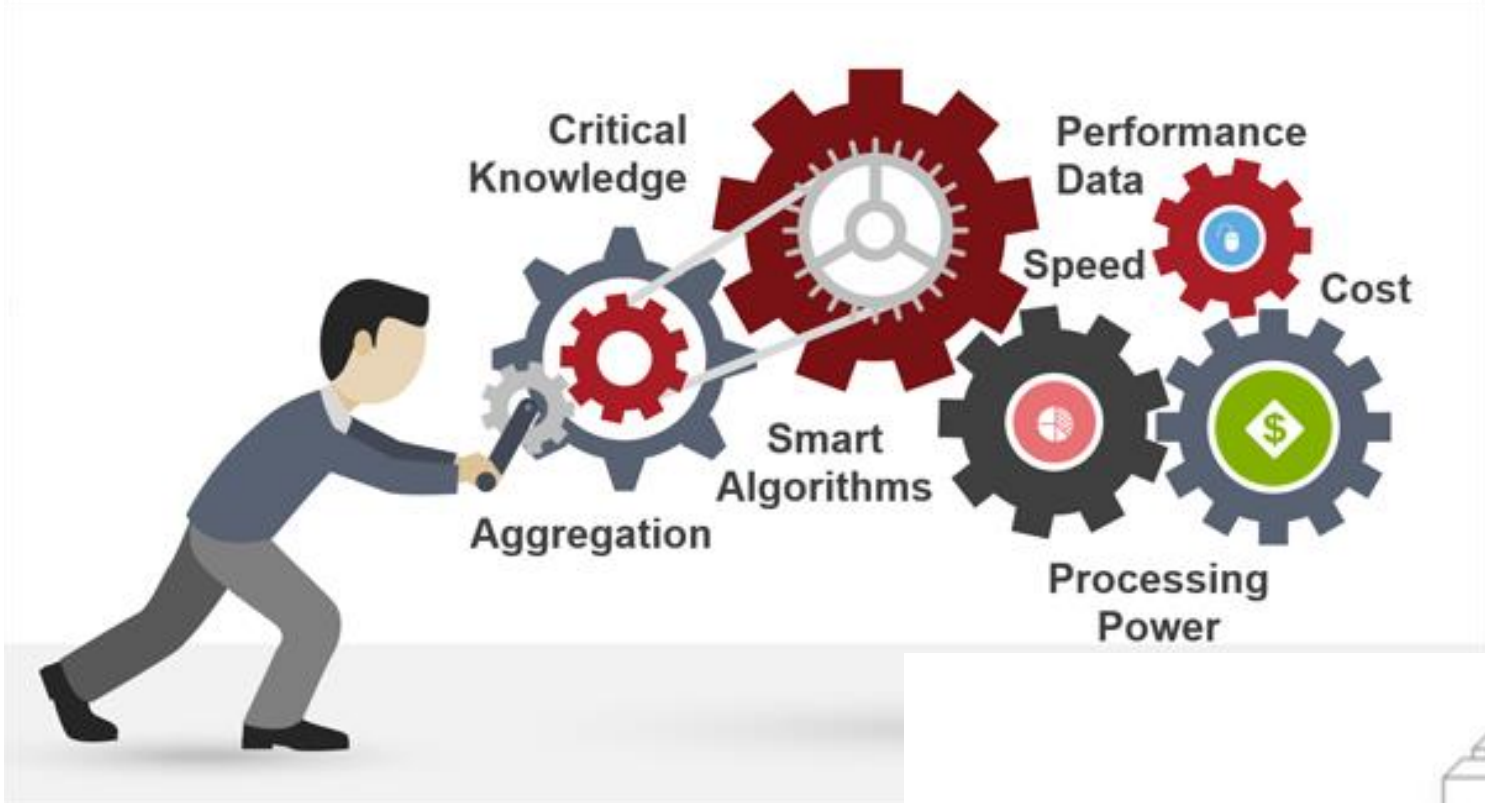


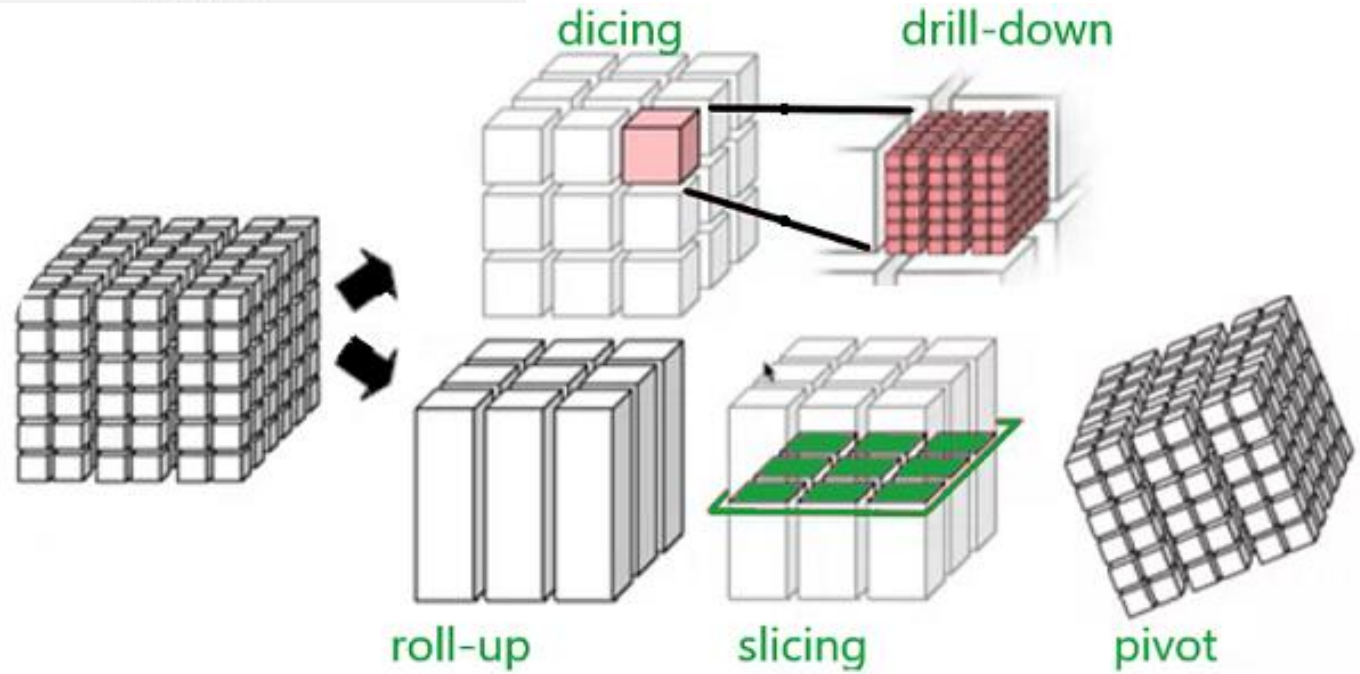
PivotTable reports and PivotChart reports

OLAP (Online Analytical Processing)

- OLAP is a category of software that allows users **to analyze information from multiple database systems at the same time**. It is a technology that **enables analysts to extract and view business data from different points of view.**
- Analysts frequently need to **group, aggregate and join data**. These operations in relational databases are resource intensive. With OLAP, **data can be pre-calculated and pre-aggregated, making analysis faster**.
- Provides a multidimensional **view** of data.
- Used for analysis of data
- Data can be viewed from different perspectives
- Determine why data appears the way it does
- **Drill approach** is used to further dig down deep into the data.




Critical
Knowledge
Performance
Data
Speed
Cost
Smart
Algorithms
Aggregation
Processing
Power


dicing
drill-down
roll-up
slicing
pivot

Characteristics of OLAP

- It deals with historical data.
- These systems do not make changes to the data.
- It stores data in data warehouses in multidimensional form.
- It is used for data analysis purposes.
- The data is never modified.

OLAP - Example

- Let us consider the data of a supermarket store, “AllGoods” store, for the year “2022”.
- This data as captured by the OLTP system is under the following column headings: Section, Product-CategoryName, YearQuarter, and SalesAmount. We have a total of 32 records/rows.
- The Section column can have one value from amongst “Men”, “Women”, “Kid”, and “Infant”.
- The ProductCategory Name column can have either the value “Accessories” or the value “Clothing”.
- The YearQuarter column can have one value from amongst “Q1”, “Q2”, “Q3”, and “Q4”.
- The SalesAmount column records the sales figures for each Section, ProductCategory Name, and Year Quarter.

OLAP - Example

Section	ProductCategoryName	YearQuarter	SalesAmount
Men	Accessories	Q1	3000.50
Men	Accessories	Q2	1000.50
Men	Accessories	Q3	3500.50
Men	Accessories	Q4	2556.50
Women	Accessories	Q1	1250.50
Women	Accessories	Q2	1000.50
Women	Accessories	Q3	1500.50
Women	Accessories	Q4	1556.50
Kid	Accessories	Q1	1234.50
Kid	Accessories	Q2	5678.50
Kid	Accessories	Q3	1233.50
Kid	Accessories	Q4	1567.50
Infant	Accessories	Q1	1555.50
Infant	Accessories	Q2	2000.50

One Dimensional

Consider the table shown in the earlier slide - It displays "AllGoods" store's sales data by Section, which is one-dimensional .

Figure 3.4 shows data in two dimensions (horizontal and vertical), in

OLAP it is considered to be one dimension as we are looking at the

SalesAmount from one particular perspective, i.e. by Section.

Table 3.4 One-dimensional data by Section

Section	*SalesAmount*
Infant	22124.00
Kid	34070.00
Men	18313.00
Women	16941.00

Table 3.5 One-dimensional data by ProductCategoryName

ProductCategoryName	*SalesAmount*
Accessories	33837.00
Clothing	57611.00

Table 3.6 One-dimensional data by YearQuarter

ProductCategoryName	*SalesAmount*
Q1	16924.00
Q2	22046.00
Q3	26663.00
Q4	25815.00

Table 3.5 presents the sales data of the "AllGoods" stores by ProductCategoryName. This data is again in one dimension as we are looking at the SalesAmount from one particular perspective, ie.ProductCategoryName.
Table 3.6 presents the"AllGoods" sales data by yet another dimension, i.e. YearQuarter. However, this data is yet another example of one-dimensional data as we are looking at the SalesAmount from one particular perspective, i.e. by YearQuarter.

Two Dimensional

One-dimensional data was easy. What if, the requirement was to view Company's data by calendar quarters and product categories? Here, two-dimensional data comes into play. The two-dimensional depiction of data allows one the liberty to think about dimensions as a kind of coordinate system.

Table 3.7 gives you a clear idea of the two-dimensional data. In this table, two dimensions (YearQuarters and ProductCategoryName) have been combined.

Table 3.7 Two-dimensional data by YearQuarter and ProductCategoryName

YearQuarter	Accessories	Clothing	SalesAmount
Q1	7041	9883	16924
Q2	9680	12366	22046
Q3	9660	17003	26663
Q4	7456	18359	25815
Total	33837	57611	91448

In Table 3.7, data has been plotted along two dimensions as we can now look at the SalesAmount from two perspectives, i.e. by YearQuarter and ProductCategoryName. The calendar quarters have been listed along the vertical axis and the product categories have been listed across the horizontal axis. Each unique pair of values of these two dimensions corresponds to a single point of SalesAmount data. For example, the Accessories sales for Q2 add up to $9680.00 whereas the Clothing sales for the same quarter total up to $12366.00. Their sales figures correspond to a single point of SalesAmount data, i.e. $22046.

Three Dimensional

What if the company's analyst wishes to view the data — all of it — along all the three dimensions (Year-Quarter, ProductCategoryName, and Section) and all on the same table at the same time? For this theanalyst needs a three-dimensional view of data as arranged in Table 3.8. In this table, one can now look atthe data by all the three dimensions/ perspectives, i.e. Section, ProductCategoryName, YearQuarter. If theanalyst wants to look for the section which recorded maximum Accessories sales in Q2, then by giving aquick glance to Table 3.8, he can conclude that it is the Kid section.

Table 3.8 Three-dimensional data by Section, ProductCategoryName, and YearQuarter

ProductCategoryName	*YearQuarter*	*Men*	*Women*	*Kid*	*Infant*	*Total*
Accessories	Q1	3000.5	1250.5	1234.5	1555.5	**7041**
	Q2	1000.5	1000.5	5678.5	2000.5	**9680**
	Q3	3500.5	1500.5	1233.5	3425.5	**9660**
	Q4	2556.5	1556.5	1567.5	1775.5	**7456**
Clothing	Q1	2000.5	4536.5	1000.5	2345.5	**9883**
	Q2	1230.5	2345.5	6789.5	2000.5	**12366**
	Q3	1456.5	3200.5	8889.5	3456.5	**17003**
	Q4	3567.5	1550.5	7676.5	5564.5	**18359**
Total		**18313**	**16941**	**34070**	**22124**	**91448**

Can we go beyond Three Dimensional?

✓Well, if the question is "Can you go beyond the third dimension?" the answer is YES!

✓If at all there is any constraint, it is because of the limits of your software. But if the question is "Should you go beyond the third dimension?" we will say it is entirely on what data has been captured by your operational transactional systems and what kind of queries you wish your OLAP system to respond to.

✓Now that we understand multi-dimensional data, it is time to look at the functionalities and characteristics of an OLAP system. OLAP systems are characterized by a low volume of transactions that involve very complex queries. Some typical applications of OLAP are: budgeting, sales forecasting, sales reporting, business process manage

✓Example: Assume a financial analyst reports that the sales by the company have gone up. The next question is "Which Section is most responsible for this increase?" The answer to this question is usually followed by a barrage of questions such as "Which store in this Section is most responsible for the increase?" or "Which particular product category or categories registered the maximum incréase?" The answers to these are provided by multidimensional analysis or OLAP;

Can we go beyond Three Dimensional?

✓Let us go back to our example of a company's ("AllGoods") sales data viewed along three dimensions: Section, ProductCategoryName, and YearQuarter.

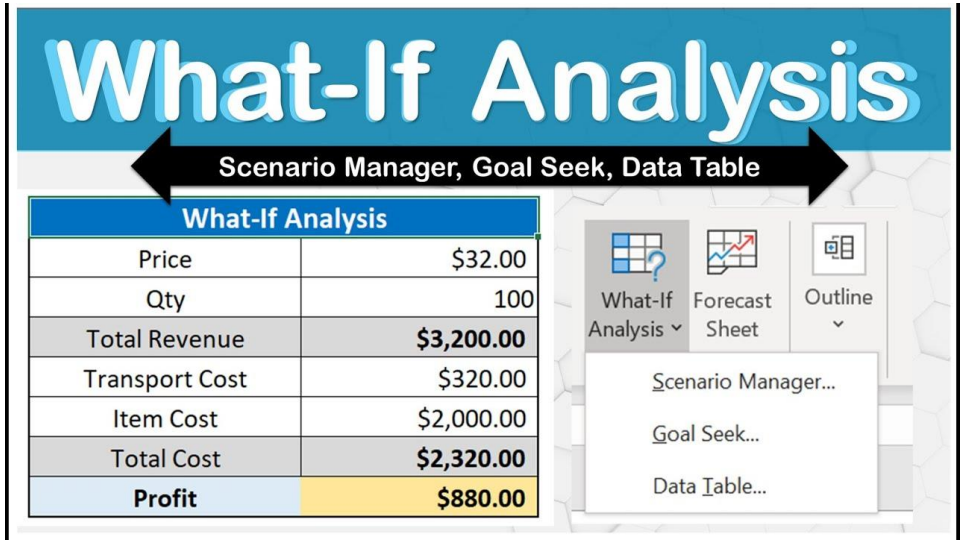
✓Given below are a set of queries, related to example, that a typical OLAP system is capable of responding to:

- What will be the future sales trend for "Accessories" in the "Kid's" Section?
- Given the customers buying pattern, will it be profitable to launch product "XYZ" in the "Kid's" Section?
- What impact will a 5% increase in the price of produces have on the customers?

Characteristics of OLAP

- Easy-to-use End-user interface
  - Easy to use graphical interfaces
  - Familiar interfaces with previous data analysis tools
- Client-Server Architecture
  - Provides flexibility
  - Can be used on different computers
  - More machines can be added

Advantages of an OLAP System



- Multi-dimensional data representation.
- Consistency of information.
- “What if ” analysis.
- Provides a single platform for all information and business needs – planning, budgeting, forecasting, reporting and analysis.
- Fast and interactive ad hoc exploration.

OLTP vs. OLAP

	OLTP	OLAP
	Online Transaction Processing	**Online Analytical Processing**
Focus	Data in	Data out
Source of data	Operational/Transactional Data	Data extracted from various operational data sources, transformed and loaded into the data warehouse
Purpose of data	Manage (control and execute) basic business tasks	Assists in planning, budgeting, forecasting and decision making
Data contents	Current data. Far too detailed – not suitable for decision making	Historical data. Has support for summarization and aggregation. Stores and manages data at various levels of granularity, thereby suitable for decision making
Inserts and updates	Very frequent updates and inserts	Periodic updates to refresh the data warehouse
Queries	Simple queries, often returning fewer records	Often complex queries involving aggregations
Processing speed	Usually returns fast	Queries usually take a long time (several hours) to execute and return
Access	Field level access	Typically aggregated access to data of business interest

OLTP vs. OLAP

	OLTP	OLAP
	Online Transaction Processing	**Online Analytical Processing**
Database Design	Typically normalized tables. OLTP system adopts ER (Entity Relationship) model	Typically de-normalized tables; uses star or snowflake schema
Backup and Recovery	Regular backups of operational data are mandatory. Requires concurrency control (locking) and recovery mechanisms (logging)	Instead of regular backups, data warehouse is refreshed periodically using data from operational data sources
Joins	Many	Few
Derived data and aggregates	Rare	Common
Data Structures	Complex	Multi-dimensional
Few Sample Queries	• Search & locate student(s) • Print student scores • Filter students above 90% marks	• Which courses have productivity impact on-the-job? • How much training is needed on future technologies for non-linear growth in BI? • Why consider investing in DSS experience lab?

THANK YOU