UNIT - 5

5.1 Emerging Trends and Buzz Words

5.1.1 Concepts of ERP (Fundamentals and importance)

Enterprise Resource Planning (ERP) is business management software that allows an organization to use a system of integrated applications to manage the business.

ERP systems are large computer systems that integrate application programs of accounting (i.e. accounts receivable), sales (i.e. order booking), manufacturing (i.e. product shipping) and the other functions in the firm.

A typical ERP system will use multiple components of computer hardware and software to achieve the integration

A key ingredient of most ERP systems is the use of a unified database to store data for various system modules.

Modules of ERP system

1. **Finance**: In this Data is collected from various functional departments and generate financial reports ledger, Trail Balance, Balance Sheets etc.

2. **Human Resource (HR)** : HR Module routinely maintain a complete employee database including contact information, Salary details Attendance, Promotions of all employees.

   - Produce pay check Reports
   - Maintain personnel Record
   - Training
   - Time and Attendance Benefits

3. **Purchasing:** Purchasing module is tightly integrated with the inventory control and production planning modules.

4. **Inventory:** Inventory Module facilitates processes of maintaining the appropriate level of stock in a warehouse.

Importance of ERP System

1. Integrate financial information.

2. Integrate customer order information.

3. Standardize and speed up operations process

4. Reduce inventory.

5. Standardize human resources information.

6. Common definitions.

7. Common database.

8. Update one module, automatically updates others.

Limitations of ERP

1. High Cost.

2. Forced change of processes.

3. Very complex software.

4. Lack of trained people.

5. Flexibility of software system upgrades

6. Implementation timelines.

7. Education and training

8. Resistance to change.

5.1.2 ETL Concepts (Extraction, transformation and loading)

A Company data may be scattered in different locations and in different format.

ETL Allows

- Migrate the data into a data warehouse.
- Convert various format and types to adhere to one consistent system.

ETL is a predefined process for access and manipulate source data and loading into target database.

or

ETL (Extract, Transform and Load) is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse.

The ETL PROCESS:

Extract: - The process of reading data from a database.

Gathering the data

- Raw data that was written directly into the disk.
- Data written to flat files or relational tables from structured source systems.
- Data can be read multiple times, if needed.

Cleansing the data

- Eliminate duplicates or fragmented data.
- Exclude unwanted / unneeded information.

Transform:- The process of converting data from one form to another.

- Preparing the data to be loaded in the data warehouse.
- Converting the extracted data
  - ➢ Using rules and lookup tables.
  - ➢ Combining data.

> ➤ Verification/Validity checks.
> ➤ Standardization.

Load:- Loading the data into data warehouse or data repository

Advantages of ETL Tool

- Simple, faster and cheaper development.
- Most ETL tools provide a metadata repository, synchronizing metadata from various sources.
- Most ETL tools deliver good performance even for very large database.
- Most ETL tools allow reuse of the existing complex programs.
- Most ETL tools have built in connectors for all the major RDBMS systems.

Popular ETL tools are:

- Oracle Warehouse Builder developed by ORACLE.
- Microsoft SQL Server Integration developed by Microsoft.
- SAP BI

5.1.3 Concepts of data warehousing

What is Data warehouse?

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in what they can understand and use in a business context.

Data warehouse is a relational or multidimensional database that is designed for query and analysis.

Database Vs Data warehouse

S.no	Database	Dataware House
1.	Database is a collection of interrelated data. For ex. Information about students, teachers and class stored in a table is an example of Database.	Data warehouse is a database used to store data. It is a central repository of data in which data from various sources is stored.
2.	A database is used for Online Transactional Processing (OLTP).	A data warehouse is used for Online Analytical Processing (OLAP). This reads the historical data for the Users for business decisions.
3.	In a database, the performance is low for analysis queries	In a data warehouse, there is high performance for analytical queries.
4.	Databases are usually normalized, which means that there's a unique copy of any given bit of information, which is important for performance and consistency reasons.	Data warehouses are more irregular, partly by design, as they need to be able to work with so many different sources of information.

Why Data warehouse is required?

- Consolidation of information from different resources

- Improved query performance

- Separate research and decision support functions from the operational systems

Data Warehouse Architecture
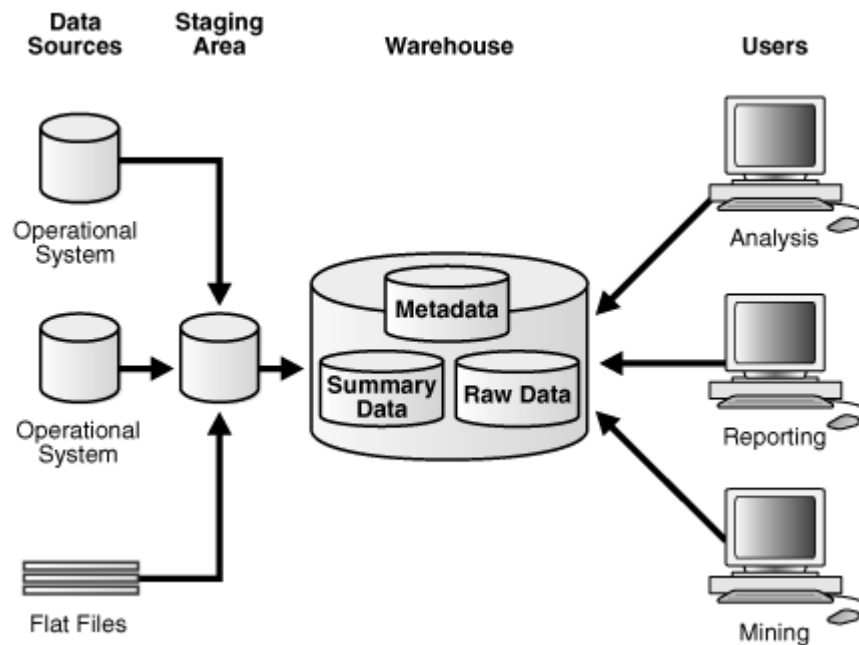
- The data has been selected from various sources and then integrate and store the data in a single and particular format.

- Data warehouses contain current detailed data, historical detailed data, lightly and highly summarized data, and metadata.

- Current and historical data are voluminous because they are stored at the highest level of detail.

- Lightly and highly summarized data are necessary to save processing time when users request them and are readily accessible.

- Metadata are "data about data". It is important for designing, constructing, retrieving, and controlling the warehouse data.

What is Data warehousing?

A process of transforming data into information and making it available to users in a timely enough manner to make a decision.

Data warehousing is a process in which relational and multidimensional database management system designed to support management decision making.

Features of Dataware house

- **Subject Oriented**: Data that gives information about a particular subject instead of about a company's ongoing operations.

- **Integrated:** Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

- **Time-variant:** All data in the data warehouse is identified with a particular time period.

- **Non-volatile:** Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

- Data warehousing is combining data from multiple and usually varied sources into one comprehensive and easily manipulated database.

- Common accessing systems of data warehousing include queries, analysis and reporting.

5.1.4 Concepts of Data science and its application areas

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.

So, Data Science is primarily used to make decisions and predictions, It uses predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

- **Predictive causal analytics** − Predictive casual analytics apply to predict the possibilities of a particular event in the future.

   For example :- if you are providing money on credit, then the probability of customers making future credit payments on time is a matter of concern for you. Here, you can build a model which can perform predictive analytics on the payment history of the customer to predict if the future payments will be on time or not.

- **Prescriptive analytics:** Prescriptive analytics is applied when we want system has the intelligence of taking its own decisions and the ability to modify it with dynamic parameters. In other terms, it not only predicts but suggests a range of prescribed actions and associated outcomes.

   For example: - Google's self-driving car, the data gathered by vehicles can be used to train self-driving cars. You can run algorithms on this data to bring intelligence to it. This will enable your car to take decisions like when to turn, which path to take, when to slow down or speed up.

- **Machine learning for making predictions —** It is applied when transactional data of a finance company and need to build a model to determine the future trend, then machine learning algorithms are the best bet. This falls under the paradigm of supervised learning. It is called supervised because you already have the data based on which you can train your machines.

   For example - a fraud detection model can be trained using a historical record of fraudulent purchases.

- **Machine learning for pattern discovery —** It is applied when don't have the parameters based on which you can make predictions, and then you need to find out the hidden patterns within the dataset to be able to make meaningful predictions. This is

nothing but the unsupervised model as you don't have any predefined labels for grouping. The most common algorithm used for pattern discovery is Clustering.

For example : -In a telephone company and you need to establish a network by putting towers in a region. Then, you can use the clustering technique to find those tower locations which will ensure that all the users receive optimum signal strength.

Why Data science is required?

- Traditionally, the data that we had was mostly structured and small in size, which could be analyzed by using the simple BI tools.

- But today most of the data is unstructured or semi-structured.
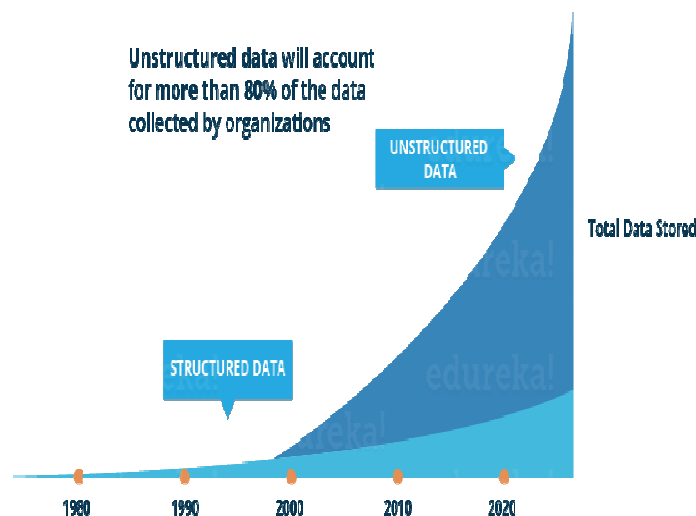
This data is generated from different sources like financial logs, text files, multimedia forms, sensors, and instruments. Simple BI tools are not capable of processing this huge volume and variety of data. This is why we need more complex and advanced analytical tools and algorithms for processing, analyzing and drawing meaningful insights out of it.

How Data Science is being used in various domains.

- To understand the precise requirements of your customers from the existing data like the customer's past browsing history, purchase history, age and income. No doubt you had all this data earlier too, but now with the vast amount and variety of data, you can train models more effectively and recommend the product to your customers with more precision.

- To understand the role of Data Science in decision making, Like the self-driving cars collect live data from sensors, including radars, cameras and lasers to create a map of its surroundings. Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn – making use of advanced machine learning algorithms.

- To understand the role of data science in predictive analytics, like weather forecasting as an example. Data from ships, aircrafts, radars, satellites can be collected and analyzed to build models. These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities. It will help you to take appropriate measures beforehand and save many precious lives.

Application areas of Data Science

Internet Search

There are many search engines like Yahoo, Bing, Ask, Google etc. All these search engines make use of data science algorithms to deliver the best result for our searched query in fraction of seconds. Considering the fact that, Google processes more than 20 petabytes of data every day. Had there been no data science, Google wouldn't have been the 'Google' we know today.

Digital Advertisements (Targeted Advertising and re-targeting)

The entire digital marketing spectrum. Starting from the display banners on various websites to the digital bill boards at the airports – almost all of them are decided by using data science algorithms.

This is the reason why digital ads have been able to get a lot higher CTR (Click –Through Rate) than traditional advertisements. They can be targeted based on user's past behavior time.

Image Recognition

You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. Similarly, while using whatsapp web, you scan a barcode in your web browser using your mobile phone. In addition, Google provides you the option to search for images by uploading them. It uses image recognition and provides related search results.

Speech Recognition

Some of the best example of speech recognition products is Google Voice, Siri, etc. Using speech recognition feature, even if you aren't in a position to type a message, your life wouldn't stop. Simply speak out the message and it will be converted to text.

Price Comparison Websites

At a basic level, these websites are being driven by lots and lots of data which is fetched using APIs and RSS Feeds. Price Grabber, Price Runner, Jungle, Shopzilla, Deal Time are some examples of price comparison websites. Now days, price comparison website can be found in almost every domain such as technology, hospitality, automobiles, durables, apparels etc.

Fraud and Risk Detection

One of the first applications of data science originated from Finance discipline. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paper work while sanctioning loans. They decided to bring in data science practices in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

5.1.5 Concepts of data analytics and related tools

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories.

Goals of performing Data Analysis

- To analyze data.
- Extract actionable and commercially relevant information to boost performance.
- There are a number of extraordinary analytical tools that are free and open source so that you can leverage it to enhance your business and develop skills.

Data Analytics Tools and Techniques

Here are some of the useful data analytics tools and techniques that can be used for performing better:

- Visual Analytics

There are different ways to analyze the data. One of the simplest ways to do is to create a graph or visual and look at it to spot patterns. This is an integrated method that combines data analysis with human interaction and data visualization.

- Regression Analysis

It is a statistical tool for investigating the relationship between variables. For instance, the cause and effect relationship between product demand and price.

- Correlation Analysis

A statistical technique that allows you to determine whether there is a relationship between two separate variables and how strong that relationship may be. It is best to use when you know or suspect that there is a relationship between two variables and wish to test the assumption.

- Time Series Analysis

It is the data that is collected at uniformly spaced time intervals. You can use it when you want to assess changes over time or predict future events on the basis of what happened in the past.

Types of Data Analytics

1. Prescriptive analytics :- Uses optimization techniques and suggests what should be done to optimize results

2. Descriptive analytics :- Uses data to understand past and present i.e. describes a set of data.

3. Predictive analytics :- Analyze current and historical facts to make predictions about future events. In essence, to use the data on some objects to predict value for another objects.

Data Analytics Tools

Data analysis is used in different domains like science, business, and social science. With the increasing need of data analysis some tools that directly analyze the data and derive conclusions are in demand in the market.

The R Project

R is a free software environment for statistical computing and graphics. R is a general statistical analysis platform that runs on the command line.[4] R can find means, medians, standard deviations, correlations and much more, including linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing. R also graphs, charts and plots results. There is excellent functionality in R, including quite a number of visualization options as well as numerical and spatial analysis.

Limitations: The fact that R runs on the command line means that users will have to take the time to learn which commands do what, and not all users will be comfortable with a text-only interface.

Runs on: Linux, Mac OS X, Unix, Windows XP or later.

Tableau Plateau :

Data visualization tools allow anyone to organize and present information intuitively. It is exceptionally powerful in business because it communicates insights through data visualization. This tool can turn data into any number of visualizations, from simple to complex. You can drag and drop fields onto the work area and ask the software to suggest a visualization type, then customize everything from labels and tool tips to size, interactive filters and legend display.

Limitations of Tableau Plateau

- All your data is public.
- There is a limitation on data size
- The only way to read is via O Data sources, Excel or txt.

NodeXL: NodeXL is visualization and analysis software of networks and relationships. It uses a technology referring to the discipline of finding connections between people based on various data sets.

This Excel plug-in displays network graphs from a given list of connections, helping you analyze and see patterns and relationships in the data.

NodeXL merges the older and current definitions of Software Network Analysis. Its "optimized for analyzing online social media -- it includes built-in connections to query

the APIs of Twitter, Flickr and YouTube, allowing you to draw networks of users and their activity. It also handles e-mail and conventional network analysis files.

- This is one of the data analysis tools in excel that helps in following areas:

1. Data Import
2. Graph visualization
3. Graph Analysis
4. Data Representation

Runs on: Excel 2007 and 2010 on Windows.

5.1.6 Concepts of Business analytics

Analytics is the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions.

Business analytics techniques break down into two main areas. The first is basic business intelligence. This involves examining historical data to get a sense of how a business department, team or staff member performed over a particular time. This is a mature practice that most enterprises are fairly accomplished at using.

The second area of business analytics involves deeper statistical analysis. This may mean doing predictive analytics by applying statistical algorithms to historical data to make a prediction about future performance of a product, service or website design change. Or, it could mean using other advanced analytics techniques, like cluster analysis, to group customers based on similarities across several data points.

Specific types of business analytics include:

Descriptive analytics, which tracks key performance indicators to understand the present state of a business;

Predictive analytics, which analyzes trend data to assess the likelihood of future outcomes; and

Prescriptive analytics, which uses past performance to generate recommendations about how to handle similar situations in the future.

Advantages of Business Analytics

- Eliminate Guesswork

- Get faster answer to your question

- Get Insight into customer behavior.

- Identify cross selling and up selling opportunities.

- Get key business metrics reports when and where you need them.

USERS OF BUSSINESS ANALYTICS

1. Students

2. Business man.

3. Accountants.

4. Organization.

5. Companies.

6. Group of Industries.

7. Small Firm

8. Auditors.

Software Used for Business Analytics

1. SPSS

SPSS Stands for Statical package for the Social Science. SPSS is a Windows based program that can be used to perform data entry and analysis and to create tables and graphs. SPSS is capable of handling large amounts of data and can perform all of the analyses covered in the text and much more. SPSS is commonly used in the Social Sciences and in the business world. The data in the Data Editor is saved in a file with the extension .sav.

2. MS-Excel

Microsoft Excel is a spreadsheet application developed by Microsoft for Microsoft Windows and MAC OS. It features calculations, graphic tools, pivot tables etc.

5.1.7 Concept of cloud

The term Cloud refers to a Network or Internet. In other words, we can say that Cloud is something, which is present at remote location. Cloud can provide services over network, i.e., on public networks or on private networks, i.e., WAN, LAN or VPN. Applications such as e-mail, web conferencing, customer relationship management (CRM), all run in cloud.

Cloud Computing refers to manipulating, configuring, and accessing the applications online. It offers online data storage, infrastructure and application. Cloud Computing is both a combination of software and hardware based computing resources delivered as a network service.

Working models for cloud computing:

- Deployment Models
- Service Models

Deployment Models

Deployment models define the type of access to the cloud, i.e., how the cloud is located? Cloud can have any of the four types of access:

- Public
- Private
- Hybrid
- Community

PUBLIC CLOUD: The Public Cloud allows systems and services to be easily accessible to the general public. Public cloud may be less secure because of its openness, e.g., e-mail.

PRIVATE CLOUD: The Private Cloud allows systems and services to be accessible within an organization. It offers increased security because of its private nature.

COMMUNITY CLOUD: The Community Cloud allows systems and services to be accessible by group of organizations.

HYBRID CLOUD: The Hybrid Cloud is mixture of public and private cloud. However, the critical activities are performed using private cloud while the non-critical activities are performed using public cloud.

Service Models

Service Models are the reference models on which the Cloud Computing is based. These can be categorized into three basic service models as listed below:

1. Infrastructure as a Service (IaaS)

2. Platform as a Service (PaaS)

3. Software as a Service (SaaS)

Infrastructure as a service (IaaS) is a cloud computing offering in which a vendor provides users access to computing resources such as servers, storage and networking.

Organizations use their own platforms and applications within a service provider's infrastructure.

Key features

- Instead of purchasing hardware outright, users pay for IaaS on demand.
- Infrastructure is scalable depending on processing and storage needs.
- Saves enterprises the costs of buying and maintaining their own hardware.
- Because data is on the cloud, there can be no single point of failure.
- Enables the virtualization of administrative tasks, freeing up time for other work.

PaaS

Platform as a service (PaaS) is a cloud computing offering that provides users with a cloud environment in which they can develop, manage and deliver applications. In addition to storage and other computing resources, users are able to use a suite of prebuilt tools to develop, customize and test their own applications.

Key features

- PaaS provides a platform with tools to test, develop and host applications in the same environment.
- Enables organizations to focus on development without having to worry about underlying infrastructure.
- Providers manage security, operating systems, server software and backups.
- Facilitates collaborative work even if teams work remotely.

SaaS

Software as a service (SaaS) is a cloud computing offering that provides users with access to a vendor's cloud-based software. Users do not install applications on their local devices. Instead, the applications reside on a remote cloud network accessed through the web or an API. Through the application, users can store and analyze data and collaborate on projects.

Key features

- SaaS vendors provide users with software and applications via a subscription model.
- Users do not have to manage, install or upgrade software; SaaS providers manage this.
- Data is secure in the cloud; equipment failure does not result in loss of data.
- Use of resources can be scaled depending on service needs.
- Applications are accessible from almost any internet-connected device, from virtually anywhere in the world.

Advantages

- Lower computer costs
- Improved performance:
- Reduced software costs
- Instant software updates
- Improved document format compatibility
- Unlimited storage capacity
- Increased data reliability
- Universal document access
- Latest version availability
- Easier group collaboration
- Device independence

Disadvantages

- Requires a constant Internet connection

- Does not work well with low-speed connections

- Features might be limited

- Can be slow

- Stored data can be lost

- Stored data might not be secure