Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology, Madras
Week:02
Lecture:07

DATA MANAGEMENT | BI&A

Hello and welcome back to the next session of Business Intelligence and Analytics course. We are moving to a very important topic titled data management. And we are going to intently spend some time in understanding the technology for managing data. Managing here means managing large volumes of data. It is not small amount of data that we deal with in data mining and business intelligence analytics, but it is large volumes of data. We have seen the definition itself carries the term large data volumes.

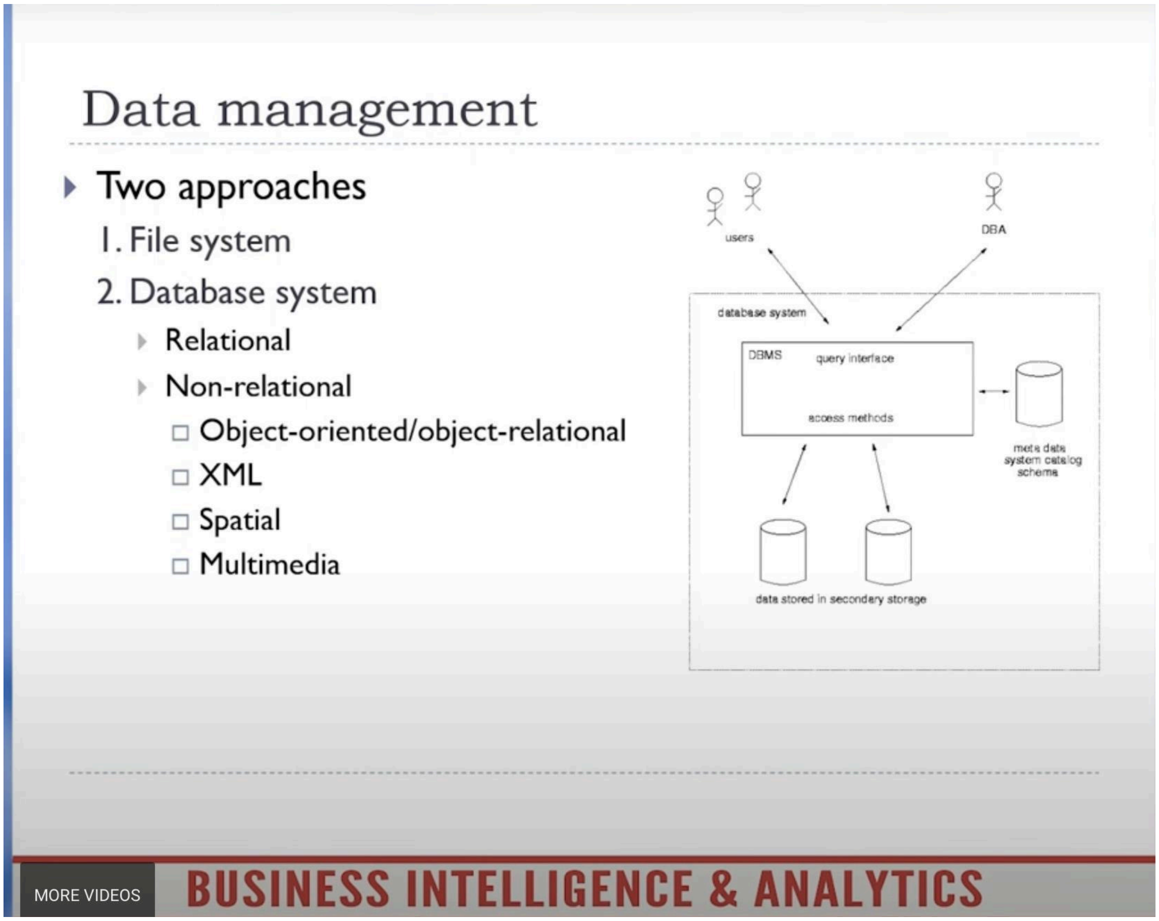
And hence, it is very important for us to have technology to manage it. It is difficult to manually manage large volumes of data. That is where technology comes for helping us. So, data management is an integral part of business intelligence and analytics.

Last class, when we saw the architecture of BI, we found that the bottom layer is nothing but data or raw data. It is through business transactions that raw data enters an enterprise. It is through that raw data that behavior enters a business enterprise. Now, that is an alternate statement. Behavior in the sense, the stakeholders of a business are interacting with the business through information systems like the ERP, like the CRM, like the SCM and so on.

So, there are different enterprise software systems today. So, when people or stakeholders like customers use enterprise systems for transactions, business transactions, the transaction data enters a database. And therefore, the nature of interaction of customer with the business is captured in the form of a transaction record. We found that in the case of telecom, it is a call detail data or CDD or sometimes also known as CDR, call detail record CDR or CDD. So they actually, so the databases capture the customer interactions or stakeholder interactions including employee, it could be other stakeholders, that data gets captured in databases.

So, that is oftentimes the integral part of an organization today because enterprise systems like ERP comes with the database or it requires a database. And that is a major advantage, of course, for going for enterprise system software. But from that, how do you move forward and keep your data or store your data in a way that is available and useful or relevant for analytics, is another question. It is not about simply storing transaction records, but it is about filtering and bringing relevant data, oftentimes in an aggregate level, not at an individual level to a different data store for the purpose of analytics. So, we saw that architecture previously.

Today, and one more session, we will be spending on databases; how databases manage data. And specifically, we will look at how relational databases manage data or what is this whole thing called Relational Database Management System or RDBMS, which is widely and very extensively used in industry, thanks to enterprise software like ERP, because they all come with a standard database of relational form of databases. And therefore, they are very prevalent. And they will be prevalent at least for some more time as we see it because enterprise software continue to be used by organizations. There are other types of data as we saw yesterday, or in the previous session that also are used for analytics, like the big data, IoT data, social media data and so on.



Data management

- Two approaches
  1. File system
  2. Database system
     - Relational
     - Non-relational
       - □ Object-oriented/object-relational
       - □ XML
       - □ Spatial
       - □ Multimedia

So, today, there is a Data Lake concept, which actually brings multiple formats of data together into a lake. And then that data is stored, transmitted, transformed, and analyzed for the purpose of insights, or that is the big data analytics or data science. So, all this is about data. So, let us in today's session, look closely at the database format of data structure or structured data, and then go on to understand how data stored in a database can be accessed or can be selected using certain commands and then be structured or reported etc. So for that, there is a specific language called structured query language.

So, we will be having a more detailed demonstration to you about database, particularly relational database, and how to use SQL to query the databases. And what are the limitations of simple queries when it comes to business questions, when business ask more complex questions, simple queries will not be enough, there will be multi-dimensional queries. So that will take us to the next topic of data management and data presentation using online analytical processing or OLAP. So, we will see these topics one by one. So, predominantly, we are dwelling at the data level.

And also, when we talk about an SQL or OLAP, essentially from an analytic lens, we are looking at descriptive data analysis, descriptive data analytics. So, let us move on to understand these topics.

So, here, let me be very clear, I am talking about data management at a enterprise level, at an organizational level, where an organization may have a centralized data center, or it may be following a cloud service. So, wherein the data is managed by a cloud service provider, but there is a specific data center and database that is separately available for an organization. So there are two approaches, one is classical and other is modern and the postmodern or current approach to data management is, you know, big data, which does not follow a structured database system like the relational database, that I would that is not mentioned here, but I would call that as a more postmodern or current form of data management.

But that is not dealt with in this course, except for one session, which is on text mining. So traditionally, in the classical form of data management, data capture and data storage, data transmission all involved file systems. We discussed this while discussing the evolution of information systems in organizations, starting with legacy systems or centralized information systems. So, in which the legacy systems did not have a database, legacy systems for business transactions will produce files. So each transaction record is a file or a flat file as it is often called in COBOL.

So, the flat files were stored as records of historic data, in file format. And that was not as easy and as structured as a relational database. And in the mid 80s came the relational database and Oracle as a company was founded, we saw that and so that there started the era of relational form of databases, which as I told you already is the most prevalent form of data storage; not data storage, data management. So, here, we need to differentiate between a physical data storage for which there are storage devices. So, the cylinder symbol that is shown here stands for disk storage.

That is a physical storage medium, where data is actually stored in the form of zeros and ones. But when we talk about database, we are not talking about physical storage, we are talking about database management. Database management means how data is stored, retrieved and updated, deleted from a data storage medium. Data is finally stored here. But how do we actually manage this input and output, inputting and outputting from the storage medium.

So, managing that data transactions is what a database management system or DBMS do. And when the DBMS is relational, we are going to see what is relational database soon in another session, but that is known as RDBMS. So essentially, RDBMS is a complex software; essentially, it is a software, when you install a MySQL or any DBMS system in your server or in your laptop, you are basically installing a very complex software. And that can be configured to manage data. That can be configured to manage data.

Data may be stored, still stored in your own device. But what you actually do is this database management. So, and you can see database management involves a software for managing, creation, updation, deletion, and access of data. And it also involves people. You need a database administrator, there are database users and so on.

So it is a system, we call it an information system. It is a database, it is an information system, which captures, stores and make relevant data available when users ask for it using a language. So users, these interactions with the database is managed by a specific language known as structured query language. This is structured query language. That is a language users, be it administrator, be it business users, they would be using in the absence of a interface, you could use an SQL, structured query language to interact with the database system.

So, that is what is known as DBMS. DBMS is what is depicted here. Now as I said, here I am talking about enterprise data management. In books related to data mining analytics, you will also hear about data management in a different context as to how algorithms manage data. Data mining algorithms, oftentimes deal with very large volumes of data.

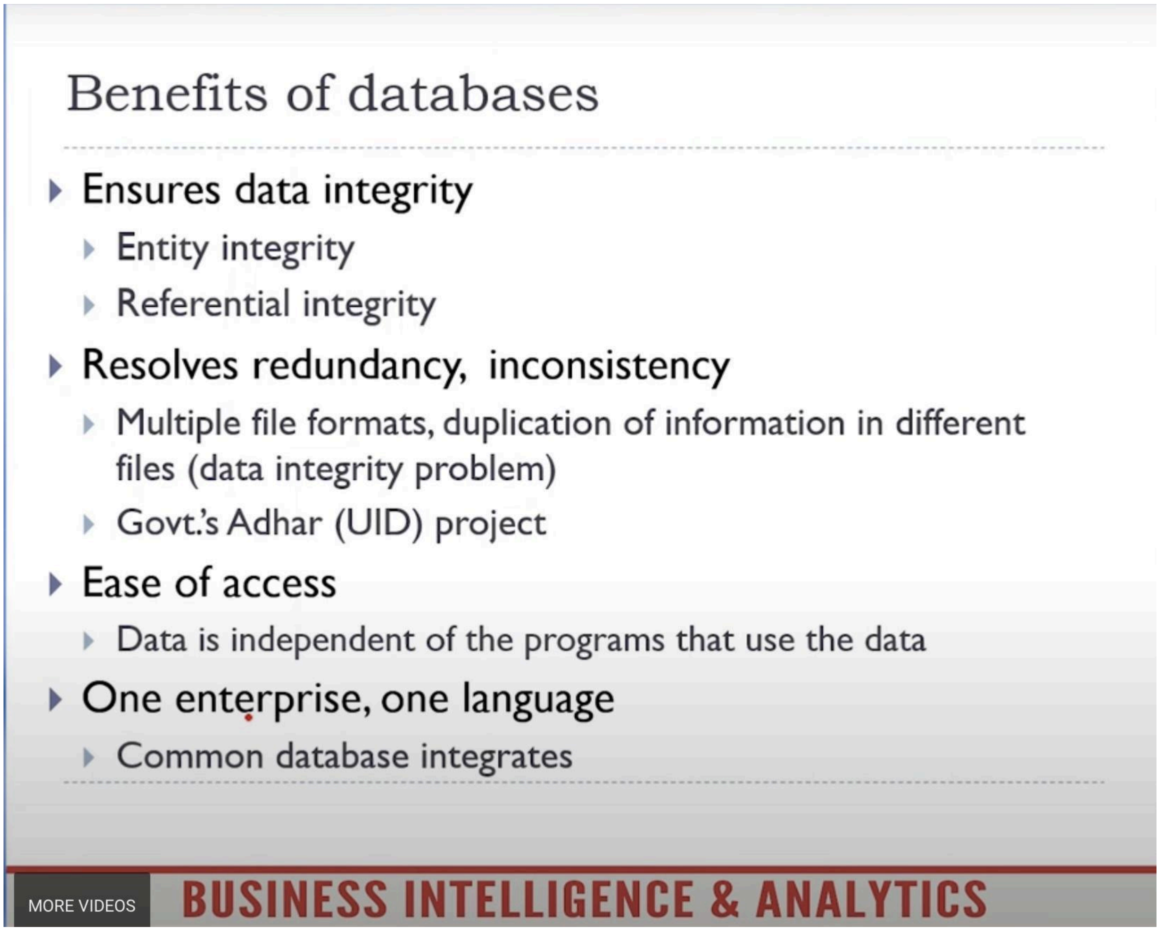
So, in that case, managing this large volume of data for the purpose of processing or running the algorithm in itself is a challenge. So, there are different techniques developed for data management within a algorithm or within a tool. And that is a separate topic in itself. And that is not what I am talking about here. So, you will hear about scalable algorithms and so on.

Scalable algorithms actually represents algorithms that can actually manage large volumes of data using certain sampling techniques. That is a separate topic. So I, since you will be reading textbooks or reading a lot of resources and materials, you will hear about the term data management in different contexts. So, you should be aware in what context this particular term is used. Here, I am talking about data management in the context of organizations.

So, relational form of data management is what we are going to focus on in this course as it is very prevalent and it is very useful and very structured. That is the structured approach to data management. There is unstructured data, which is about big data, which will be managed separately, like Hadoop and related technologies are used for managing unstructured data, which I am not dealing with in this course. Then there are a host of other techniques, other than relational databases known as non-relational. And if you have to add a three-third today, that will be the big data management.

So, Hadoop and other technologies are examples for that. So, let us move on. Now, since we are talking about databases, it is important to know why you need database. Why cannot data be managed manually? Of course, you have obvious answers when the volumes are large.Manual processing is not easy. And it is not efficient. It takes, consumes long time and it could lead to errors and so on. But why cannot data be managed or stored using spreadsheets, like the Microsoft Excel or G sheet or you have spreadsheet technologies available today. And that is also very structured. So, a spreadsheet is structured into worksheets.

So, you can have a worksheet for, one worksheet for one category of data, next say for customers, second worksheet for another category of data, say for employees, third worksheet for another category of data, say items and so on. So it is, actually a spreadsheet is a structured way of storing and managing data. And we also know, as users of spreadsheets, that you can embed formulas in spreadsheets, you can filter data, you know, filtering and sorting is a very powerful technique in spreadsheets to almost similar to the SQL queries where you can actually filter the data that you want from a large set of raw data that is stored. So, when a spreadsheet has all these features, which a database typically has, and you do not have to write any code, you see, filtering is very intuitive, you can actually apply filters on each columns, the way you want and you can specify those conditions, it is very much menu driven, you do not have to learn an SQL language and this complex table structure etc. in, not very complex, but it requires some technical knowledge.



Benefits of databases

- Ensures data integrity
  - Entity integrity
  - Referential integrity
- Resolves redundancy, inconsistency
  - Multiple file formats, duplication of information in different files (data integrity problem)
  - Govt.'s Adhar (UID) project
- Ease of access
  - Data is independent of the programs that use the data
- One enterprise, one language
  - Common database integrates

So, if you use spreadsheets, and why spreadsheets, why not spreadsheets for large volumes of data, why database? And that is a question that is answered in this particular slide. So, the first topic is actually about integrity, Integrity of data. So, what is data integrity? Data integrity is about the credibility of the data, it is related to the credibility of the data, it is related to the accuracy of the data and so on. So, and so data integrity to be ensured, there are two aspects to data integrity, one is entity integrity, other is referential integrity.

So in this context, I refer to the Aadhaar project of the Government of India. Why did government initiate a large project, arguably the largest database in the world, arguably the largest platform in the world, which perhaps has almost 95% of the citizen data, citizen's identity data has been captured and stored in the Aadhaar database, or what is called as the Aadhaar registry. So, that particular database carries wealth of fundamental data about citizens of India. So, why was it required? If you ask that question, it is about a specific aspect of data, which is the need to uniquely identify each record. What is this problem of unique identification? The Aadhaar is also known as UID, unique identification. And an organization was created for that, Unique Identification Authority of India.

And so this was created because there was no unique identification of citizens prior to that. There were, in other words, multiple IDs for the same person. And then what happens? If government has to deliver its services to citizens; for the government, there is only one citizen and that citizen should be receiving certain, certain services. It could be ration, it could be your cooking gas. And there is certain quota for each individual and the individual should receive only that quota and the quota should go to only that individual, not somebody else. So, a uniquely identified individual can be provided with this services, which is relevant and which is a right of that citizen or individual.

But suppose there are many IDs for the same person, then what happens? Then what happens is, so the same service, two people can claim. That becomes a conflict. This is a data integrity problem, because a person is not uniquely identified. And think of a voter ID, for example. So, a person generates a voter ID in Karnataka and also in Kerala and also in Madhya Pradesh. You can go and vote in three places, although you are not supposed to do that, you have three IDs, you are not uniquely identified.

So practically, when you implement databases, what you solve is a problem called identification or a need for unique identification. What does the Aadhar database do? It creates a unique ID. And that is the only ID that government will use in delivering its important services. Of course, there is a conflict of this identity database versus data privacy and some of you may be aware of this. So, let me not enter there, but government is capturing only basic level data, what do you call foundational data, you can call it.

And beyond that, it is not captured. So, it serves certain purpose. And that purpose is to identify individual for the purpose of government's government services. And the same identity database can be used for other services like telecom by telecom service providers. But, you know, today we know that it cannot be mandated, but this is a very, very reliable or credible form of ensuring identity. So, entity integrity is about giving a unique ID and that unique ID in database terms is known as the primary key.

So, what is Aadhar ID? It is a primary key for an individual in database terms. And once there is a primary key, all other fields related to you, get related to the primary key. So, all in database, we say all fields or all attributes should be fully related to the primary key. And that is how a individual's record is created. And there cannot be multiple records for the same individual because the key is the same.

A primary key is unique and it is not, it is not distinct or there are no different primaries key for the same person. So, therefore, the uniqueness of records are maintained. That is known as entity integrity. There is also a concept of referential integrity, which is related to data integrity that is, that is, in order to illustrate it with an example, suppose a person retires from an organization. So, employee X retires from an organization, but employee X's records would be removed or deleted from the employee database. So, the HR department does it very promptly, the record is deleted.

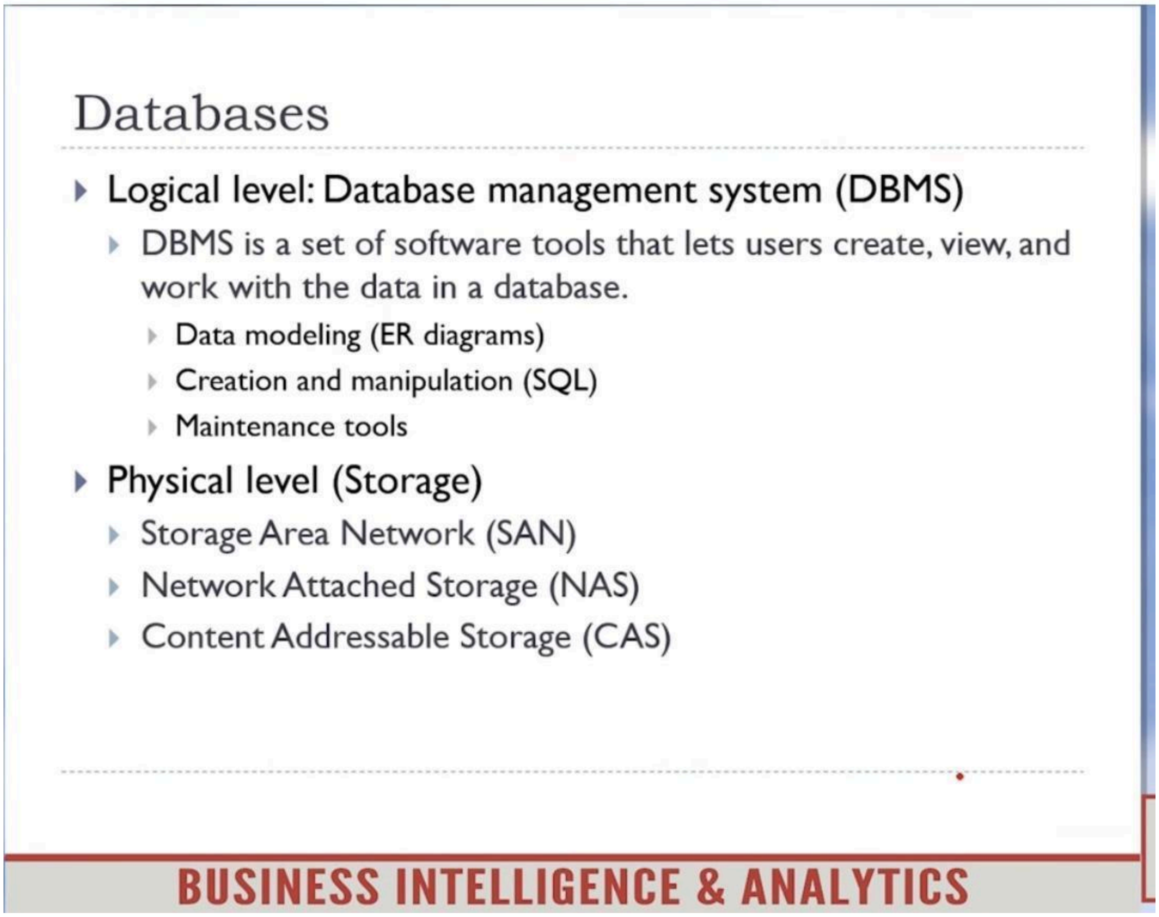
And suppose each department maintains a separate data store, maybe it is not database, it could be file system or spreadsheets or whatever. So it is, assume it is a small or medium organization, they do not have the concept of central database, but it is only about individual systems. So HR deleted, but operations department did not delete or marketing department did not delete. So for them, the employee is still an employee.

So, you see the kind of conflict and confusion comes. So, because the, when a record was deleted in one place, that deletion did not reflect in the record of another system, where the same person or the same, the same object was present. So, this problem is addressed through referential, to ensure referential integrity, database systems enforce a rule that when a record is deleted in a master table, it will reflect in all the records where the primary key, using a primary key or the primary key occurs. Let me reword it.If a particular record is deleted in a master table and that particular record has a primary key. So, then any table that refers to that master table, the corresponding records of the reference table will also be deleted or will also be deleted, once the master record is deleted.

So, there is referential integrity. So, the other departments do not have to individually do it, it is, if it is reflected in a master table. That is the idea of referential integrity. So, two aspects of data integrity are enforced by databases- entity integrity and referential integrity and these two together is known as data integrity. This is an important feature of databases. And secondly, of course, the other important purpose of databases is to resolve redundancy and inconsistency. So, in the absence of say, unique identification or a primary key there, as we said, there could be multiple records about the same person and also the same data may have to repeat again and again, when you do not structure the data using keys.

This is something that we, I will show you in the next slide as to how redundancy can be reduced when you use a database system like the relational database. And, of course, the third and fourth advantages you can see the ease of access and where applications, different applications can be written to access the same database. So, you can see that in enterprise systems, there can be an SCM, there can be an ERP, there can be a CRM etc, which connects to the same database, an enterprise database is central, common. So, you have a common data definition, for example, customer is defined as customer in the same way an ERP defines it, an SCM defines it and CRM defines it and a customer is uniquely identified, say by a customer ID. So, how customer will be identified is commonly defined for all departments, all systems.

So, that brings some sort of sanity to the entire enterprise in terms of the data language they speak and database is common as you can see. So, you can write different applications, but the data can be stored in the same database. So, database becomes a common asset, database becomes a sort of integrating entity for the entire enterprise. So, these are philosophical or conceptual schemas that is used when implementing enterprise systems. So, the main advantage there is that database becomes a common asset.



Databases

- Logical level: Database management system (DBMS)
  - DBMS is a set of software tools that lets users create, view, and work with the data in a database.
    - Data modeling (ER diagrams)
    - Creation and manipulation (SQL)
    - Maintenance tools
- Physical level (Storage)
  - Storage Area Network (SAN)
  - Network Attached Storage (NAS)
  - Content Addressable Storage (CAS)

BUSINESS INTELLIGENCE & ANALYTICS

So, that is other advantages. So, instead of using very, very different data storage techniques in different departments or units. So, but let us try to understand how databases are actually developed and implemented. The implementation part, I am not covering in this session, but I would give you some idea about how databases are developed, so that you are able to think how data gets structured in databases, because when we use the term structured approach to data storage, you must understand what is the sort of design, thinking that or what is the design approach that goes into developing the database. So, in this slide, you first see database has two levels, one is a logical level, other is a physical level. And I think I talked about it previously, the logical level is where you develop a management system for managing data.

So, it involves data modeling. I will show you what is data modeling in the next slide, as to how you develop a database structure. And then you use a language to sort of update or delete or retrieve data from the databases, that is for the sort of transactions with the database. And then of course, you need to define who can access and what is the access right of each user and security levels etc. All that is part of the maintenance of the database. So, you can see that there is a database profession called database administrators profession.

And this is the physical level, which is the physical storage, the device at the device level. So, this is not something that we are dealing with, we are looking at here, in terms of database management.