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:06

BUSINESS INTELLIGENCE ARCHITECTURE | BI&A

So in order to understand the business intelligence architecture, let us start from the basics. Let us start from the bottom layer. We started with databases or switching center, if you remember earlier, I said telecom companies have switching centers or data centers. So, what is the important ingredient of a data center? It is a database. So in this particular case, we have seen they capture the call detail data, there could be other data pertaining to the business, say an employee database and many other tables may be existing in a database, but they have a database and in the particular problem solving, we found CDD is the data they require, that is part of a database, but there is an enterprise database and let us assume that it does exist for AT&T Long Distance already. So, there is a database. So that is a starting point.

Now, we also saw that for the purpose of the bizocity score, there was an interesting data transformation that they needed to do. That is, they needed to bring certain specific attributes from the data to a different store possibly, for the purpose of analysis for building the bizocity scoring algorithm, they need certain indicators from the database. And we saw that those indicators are typically the number and they wanted indicators like recency, frequency, monetary value, bizocity score, etc etc etc.

Now, in order to do this, and these are the variables that will be used for building the bizocity score or profiling the customer etc. Now, data has to move from database to a sort of analytic database, if I call it or something else you will call it. So, this is what is contained in the data and there could be other data that is stored here for the purpose of analysis. So here this data store, the purpose is analysis or analytics. And the purpose of this data store is transaction.

There is a transactional database. The purpose of a transactional database is to store, it is to capture and store systematically transaction records. Here a transaction record in telecom, is a CDD. A transaction record in a retail business can be, you know, an invoice or the result of a particular transaction, a customer ID, what are the items purchased by the customer, the date, the price, discount and so on. That is a transaction record for a retail business.

For telecom, it is the call detail data. Every business does transactions and ERP systems as we saw previously, automates these transactions and the resulting data or the resulting records gets captured in databases. And that is what we see here. So, databases are for transactions. And when these databases are configured for online transaction, they are also known as online transaction processing or in short OLTP.

You may come across this technical term in your textbook,n OLTP. OLTP is nothing but a database, but designed with certain characteristics that support online multi-user transaction processing. And there is something known as ACID property, which we will see later when we discuss data management. So, OLTP databases are advanced databases configured, designed and configured for the purpose of transaction processing. They are transaction processing database.

And generally, if you want to do data analysis, modeling, etc, you do not directly work with a transactional database. You do not directly query a transactional database for building even MIS reports. You try to create another copy of the database and work with that database. Of course, data has to move from the transaction database to other data stores in a, by batch processing, but you do not keep querying or keep working with a transactional database, which will make the transactional database inefficient. It may delay. So, therefore, it is always recommended to build a data store separately for the purpose of analytics.

And what is that data store? That is the data store that is represented here, this one, we will name it and you will understand the name of it as I explain further. But how, what kind of data goes into this data store, which is meant for storing data for analysis. The data will be retrieved, of course, from databases, the source of data is databases.

So, therefore, there has to be some tool which will extract data from database, but it is not the same raw data that is loaded into this data store, central data store, but the data is transformed. Data is transformed. What is data transformation? Well, data transformation is about changing the data into a form that is useful for analysis. So, data transformation here, in the particular case that we discussed involves aggregation of data, you can see that the data is summed or counted or averaged over a period of time. So, that is one transformation that happens on the data.

And it also depends on the specific application. For example, if it is a database that is distributed in different geographies, the business is spread across the world. And then you may be having databases, which operate in different currency units, different units of measurement, so you need to standardize the data as well. So, you can keep those aspects of data preparation in mind, when you think of transformation. So in other words, data

transformation is the process of preparing data useful for analysis.

The data needs to be stored in a way that they are directly useful for analysis purpose. So, that is the aspect in transformation. And then this data after extracted and transformed should be loaded here. So you can imagine, there is a function here for a tool which has the ability to access, which has the protocols to access different databases, transform them, that is a software in itself, and then load it into or access another data store and load that into that data store. So you can see a short form for this will be ETL or integration service.

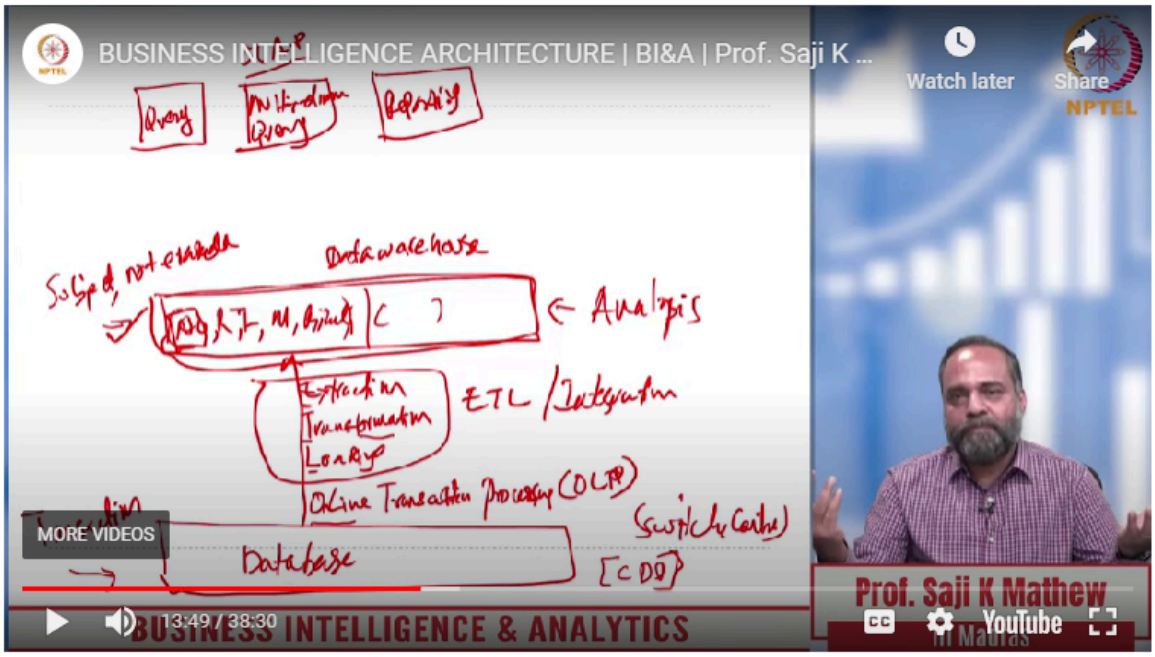
These are data integration tools, or ETL tools as they are sometimes called. And the purpose of ETL tools is to create data from raw data that is useful for modeling, or that is useful for analytics, in our words. And what is an appropriate name for this data store, which is a derived data store from databases and other sources, which contains useful data for analysis. So, this data is again with respect to, say certain key or certain identifier. So, you can say certain subject, this is subject oriented.

And this data is stored for the purpose of analysis and they will be stored for a long time, long time, so they are not erased. They are not erased, because this is history of the organization. And the, since the data is very selected and filtered for, only for the purpose of analysis, you do not transfer all the garbage in raw data to this data center or this data base, but it is only selected relevant data useful for analysis. And this data store is called data, data warehouse. A data warehouse is a warehouse of useful data for analysis.

The purpose of a data warehouse is to store data for analytics or data analysis. It is not to retain data for transaction records or legal purposes or anything of that kind. It is basically for the purpose of analytics. Now, we also need to think of different tools that you can use to work with the data warehouse. For example, you can have query tools, you can simply query to profile, you can run a simple query to profile customers in terms of their RFM and bizocity score, all this can be done by query.

And you can also do multi-dimensional query. We will discuss this separately in one session, multi-dimensional query or OLAP, online analytical processing, which is one technology that can work with the data warehouse to create useful insights, useful multi-dimensional insights for business decision makers. Multi-dimensional here meaning not simple queries, but query about a certain business performance indicator, like the sales performance or the sales volume, or whatever KPI is interesting to the business, that KPI can be analyzed with respect to time, with respect to a particular product, with respect to a particular market, with respect to a particular promotion, you

can bring in multiple dimensions to look at your measures, business performance measures. And that is actually that system which is developed for it is known as OLAP, online analytical processing. OLTP is the database, OLAP is the analytical interface for conducting, you know, more complicated or more advanced queries.



Then since if you have query, then you can have reporting also. Reporting uses queries, but structure the results in a format, in a way that is can be easily visualized. There could be visualization tools, graphical tools. There are very advanced visualization techniques today, all of them would work with a data warehouse, a central data store. And then we saw that a part of the analysis that is carried out in the AT&T Long Distance is the bizocity score.

Bizocity score is a result from a particular model, be it a Logitech model or a decision tree model or anything else. There is a particular model that is built and tested. And when you input, input data to the model, it will output or it will predict and return a bizocity score. But the model is developed, who will develop, what is a tool that can develop this kind of models, not a query tool, not an OLAP tool, not a reporting tool, they are all very descriptive. They provide descriptive analytics, but here you need more advanced analytics.

And that suite of tools would be known as data mining tools. The mining tools or in some places, you will see advanced analytics in some literature. But data mining is

nothing but using large volumes of data from the relevant data or variables from a data warehouse, model them using advanced techniques, statistics or algorithms and build those models, test those models and deploy those models for business use. So, that is the other analytical interface. You can call them as the intelligence part. All this is for business intelligence.

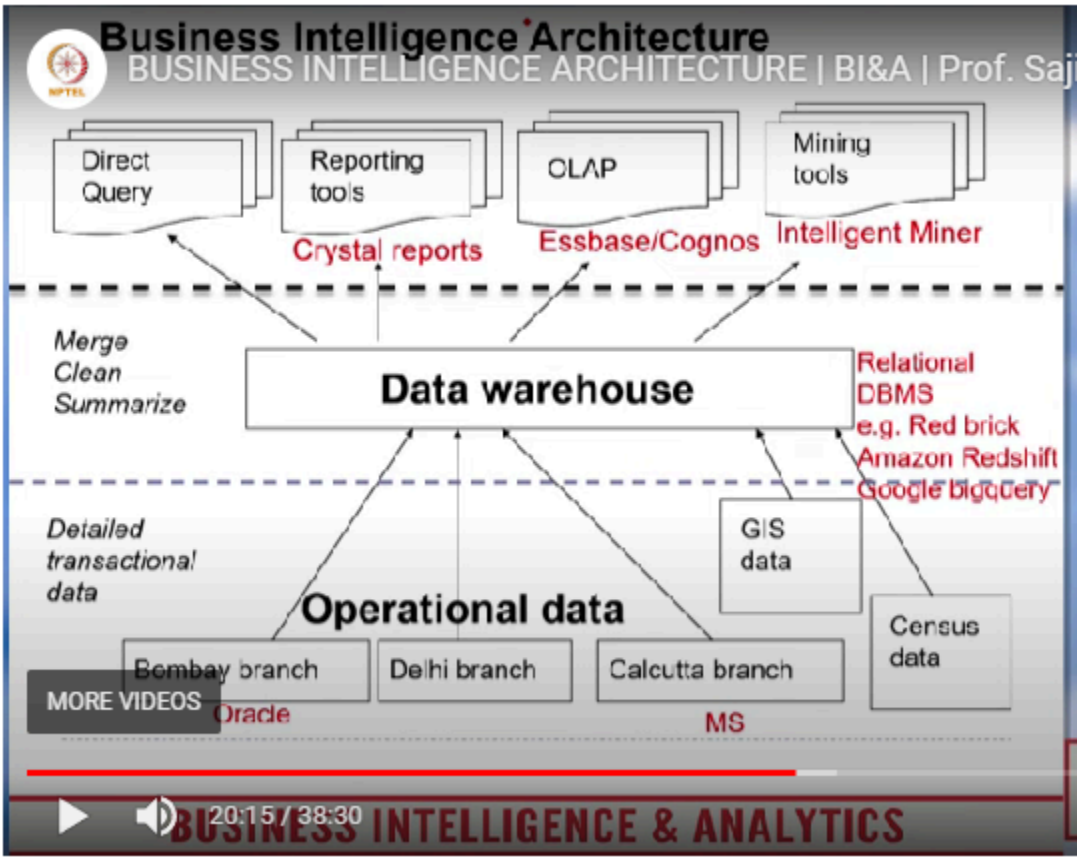
So, if you look at this diagram, which we created from a business problem with that we discussed, they you can see multiple layers of computing, multiple layers of computing, starting from database and data integration software, data warehouse to a layer of tools that are used for analyzing data. So up till here, I would say it is about data. And this is about intelligence. In as we saw already, intelligence is what business requires for decision making.

And who would create that intelligence, the set of tools that are described here. In different categories, we are only building categories of tools here. So, all together, I would look at this diagram and I will call this as the business intelligence architecture, business intelligence architecture. If we looked at the objectives of this course, one of the objectives is to understand what is an architecture for creating a BI& A or business intelligence and analytics practice in an organization. So, looking at it from a business point of view, essentially, we are saying, if AT&T Long Distance or any organization is seriously pursuing business intelligence and analytics as a strategy, not just one of problems, they just want to hire a consultant, solve a problem and then move on, they do not want to make huge investment, then that is a different case.

But if, it is analytics is a part of business strategy, and then in that case, data has to be preserved, relevant data has to be extracted, they have to be stored for long term basis. And therefore, you need tools like ETL tools, you know, data stores like the data warehouse, which are major investments as far as organizations are concerned. It is not only this physical systems that you need, but you also need people, you can see that it is not just, you know, people are there everywhere, you know, in database to ETL, to data warehouse design, implementation, maintenance. And then here you need the so called analytics or data scientist. These are the people who understand both technology as well as business.

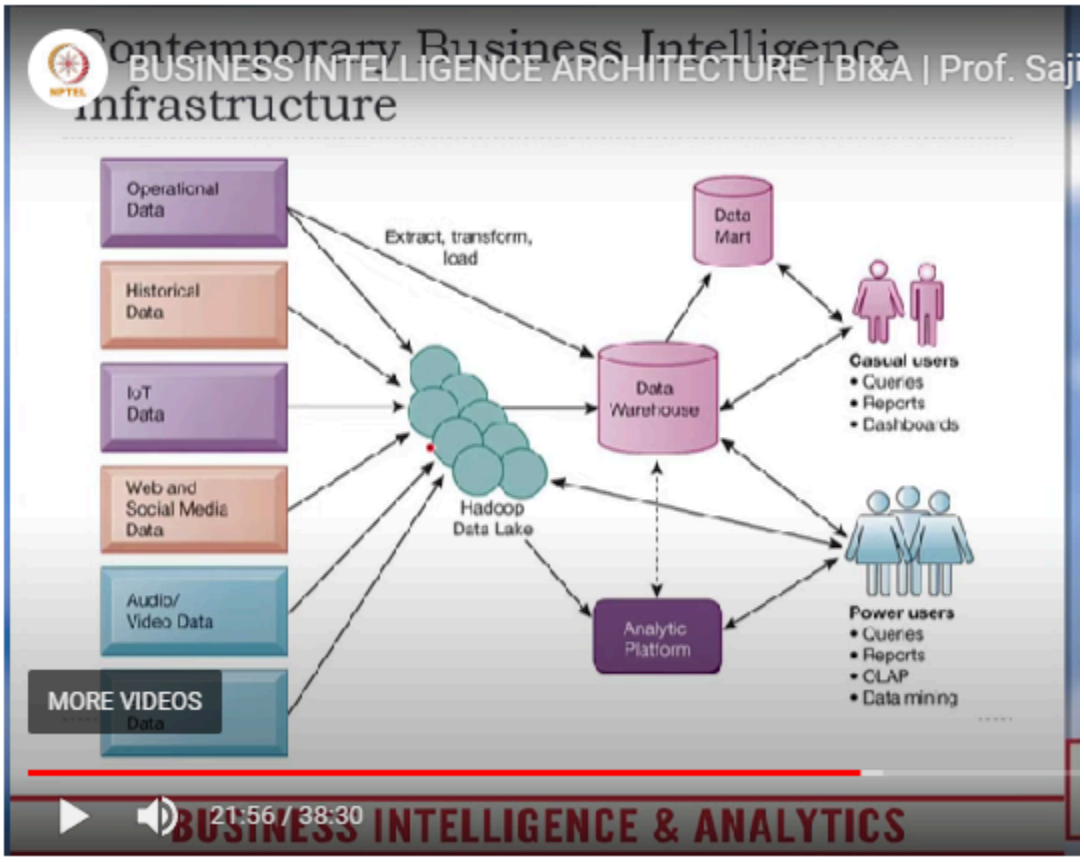
This goes to business use, the data driven decisions is actually the next layer. You can see their business is waiting for insights from data. So therefore, you need to hire data scientists who understand. So of course, in this particular case, engineers from Bell Labs came in and they provided the solution. But in order to build it as a practice, you need to hire people and build what you call in administration, we call a analytics capability, you have to build a capability, capability meaning resources, technical as well as people resources, so that this can deliver solutions to business.

So, that is the business side of business intelligence where a rational to invest in solutions of this kind needs to be developed, so that business is convinced about the investment. So, let us move on and look at these things, these lessons we learned more comprehensively. So, this slide is a more neat representation of the architecture, which I drew and explained to you. And you can see that a BI architecture is a layered architecture and you can identify at least three layers here, two layers of data. And what is not represented here is the ETL. And ETL is also part of this architecture.



And in this particular diagram, I have also shown some of the specific products, technology or the commercial products that are available for building this solutions. And so, in terms of implementation, when the architecture becomes your infrastructure, you know, infrastructure is physical, architecture is a plan, it is not physical. And therefore, when it becomes an infrastructure, then you have to buy and integrate and, you know, sort of create the infrastructure that is required to drive your BI and analytics practice. Now, I would say this particular representation of the BI architecture is very conventional, very traditional and it does not take into account the current developments in analytics and data science.

So, let me actually give you an overview of what is current, without really going into the details of the current developments, but to, my aim is to give you an overview. So a more contemporary BI infrastructure would look like this, where the left side is, is the data sources. This is actually presented horizontally in my representation, I represented it vertically. So, you are starting with the data on the left side, you can see that operational data, which is again OLTP, it is not changing here. And this may be in the same database, historical data and other data, IoT data.



I am sure many of you are familiar with Internet of Things, which are data sources today; data sources, which are continuously capturing large volumes of data, often sensor based. So if you are looking at healthcare analytics, if you are wearing a Fitbit or an Apple Watch, it is continuously capturing your health related data. And it is actually with your permission, of course transferring that data into some database or data warehouse. And of course, at the back end, they are analyzing your health data. Of course, there are privacy issues there, let me go, not go there.

So IoT represents those devices, which are internet connected and which are capturing continuous data. It could be weather data, it could be smart meters used in energy today. And there are, you know, in manufacturing, you know, there are several sensors which

are smart and IoT based today. So, that is another source of data.

And social media data. So, generally, the social media is outside of an organization. So, if you are a conventional manufacturing company, you are not into social media, you are a manufacturing company. So, the diagram which I represented in the previous slide is more relevant for that organization. But today, data about you, about your transactions, you can find in your databases, but data about what the public, what society, what your consumers are talking about your products and services outside, is not stored inside. So therefore, you also have to go sometimes for external sources of data.

And this kind of data would fall under the big data, large volumes of data and different types of data or different data types like the video, audio, pictures and so on, which you upload to social media and other forums. They are also useful for analysis. They are also useful to understand how consumers are responding to a latest phone or a latest car that your company has launched. What is customer sentiment about those products? So sentiment analysis, and there are different techniques for using social media data to support business decisions. So, that is another source of data in my architecture, this was not included.

Then in this diagram, you know, this is I would not say this is a well done diagram, it is having a lot of overlaps like the social media data itself can consist of audio, video data and you know, in itself it is external data and so on. So, there are different categories of data out there. I would call them, this is internal, this could be external, depending on where an IoT device is, if it is external like, then it is external, if it is inside your organization, it is internal. Now, in order to handle or in order to store, transmit and also analyze this kind of big data, there is a separate technology today known as the Data Lake and Hadoop. Hadoop some of you must have heard about it, which is an open source project to manage big data, which are, we call this also unstructured data. This is not typically relational database, which is used to store this type of social media data, they are unstructured in format.

So, you need different technology for that. Hadoop is a technology that is used to manage data that is unstructured and that is big array or the big data. And in this particular representation, you can see the data lake data can also inform or that can also become part of the data warehouse. So you see, the conventional database or the data warehouse also involved in a current architecture of business intelligence in addition to the big data. And then, of course, as we saw already, the data from the data warehouse is used for analytics and the insights are passed on to different types of users. Data mart is also shown here, I did not explain what is a data mart earlier.

When we discuss data warehouse in some detail, I will give some more insights about what is a data mart, but in short, data mart could be a subset of a data warehouse, which is created for specific business domain or business function. For example, marketing department requires a specific subset of the data warehouse where data is related to customers. They may also add more data, enrich this data mart with other external sources of data, relevant to them. For example, census data is a data that would be useful for customer segmentation.

So, they may bring in that external data. They will also use customer behavioral data from data warehouse and create a customized data store for analytics purpose known as Data Mart.

Now, let me give you some contemporary view of analytics. One is the data science view. Today, analytics and data science are two terms that is used to describe almost the same thing. Using data to support business or using data for data driven decision making, as it is called.

So ultimately, the purpose of data science and analytics appears to be the same. As far as my reading suggests, the purpose is data driven decision making by organizations. So, and data science view would consist of three layers as you can see, and if you include this also then four layers. One is the data science. So, data scientists are the interface between business domain and technology.

They understand data algorithms. They also understand business problems. Those are the data scientists. So, and the data science terminology has evolved from computer science. So, essentially a data science team in a typical organization would consist of masters of computer scientists, who are having good knowledge, sound knowledge of big data, databases and algorithms, statistics, etc. What they may have less knowledge is about the business domain, which they learn, which they make up through continuous interaction with the industry.

That is what data science, data scientists are. Then you see that there is a data engineering layer, which is the step up till the data warehouse, a data warehouse and Hadoop, etc., or the data lake, etc. All these have to be created for the purpose of data analysis. So all that is related to data capture, data transmission, data preparation of structured and unstructured data falls under data engineering. So there could be other side effects or after effects of data engineering, which is depicted here.

So, data engineering is essentially about data or big data is specifically mentioned here because that falls under the computer science category or the computer science group of knowledge and expertise. And that is why the data science has grown as a separate

discipline today, sub-discipline today, because advanced technology for data engineering is used, which involves big data. So, analytics is more referring to the statistical analysis of data where probably big data is not directly implied. But you do not have clear definitions yet and therefore, I am giving my own understanding.

So, I close with certain definitions. Data mining is about extracting or mining knowledge from large amounts of data. So, you can see that data mining is about knowledge. It is not about data and it may sound like a misnomer. Data mining is not about pulling data from database.

But data mining is about advanced modeling. It is knowledge discovery, as it said in computer science. And there are other related terms and I give you different definitions. But what you do not miss in any definition is the use of large volumes of data to extract knowledge. That is what data mining is and that is the engine.

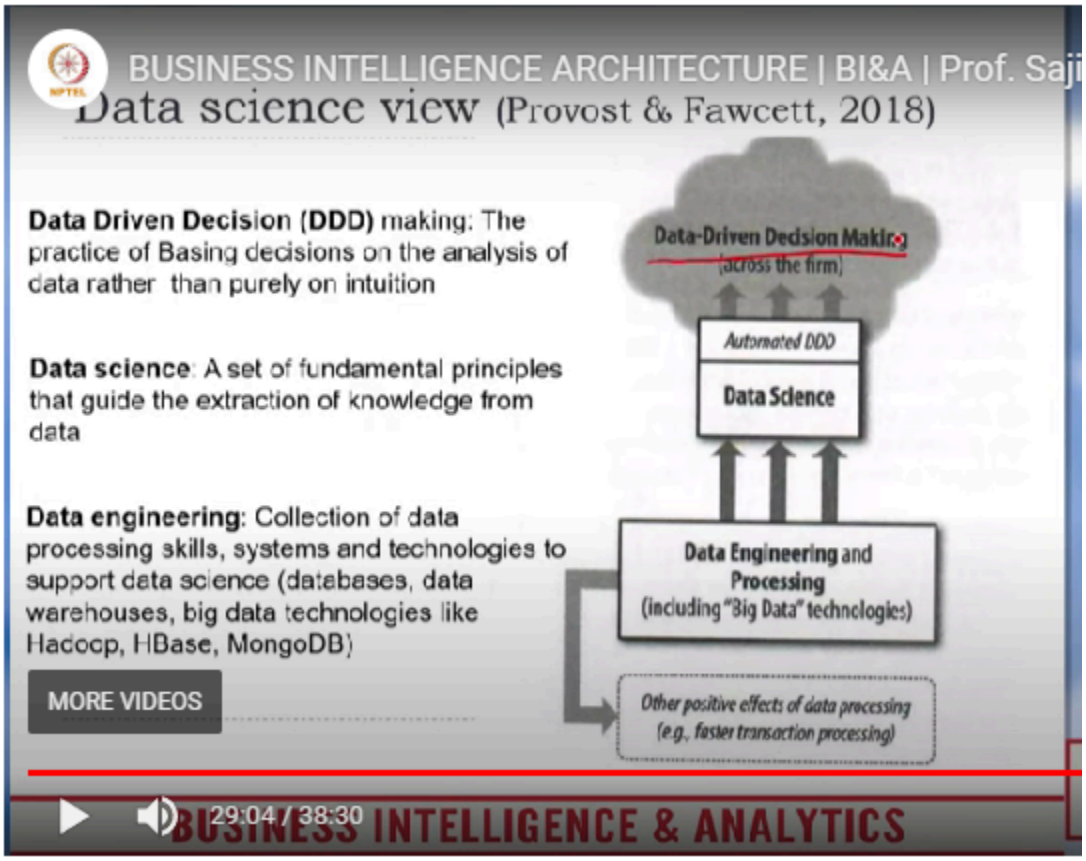
That is the engine for decision making. So, that is what drives decision making, like a bizocity score is the result of data mining. That is what drives decision making. Online transaction processing or OLTP, I showed you what it is, when I explained to you the architecture of business intelligence. It is the database and it has properties like atomicity, consistency, isolation, durability, etc.We will study this in a subsequent session.

OLAP is for multi-dimensional queries. We have seen that multi-dimensional queries are required by business decision makers to analyze KPIs from multiple dimensions. And look at the definition given by Thomas Davenport for business analytics.

It is a long, long statement. It is a kilometer long. Definition should be crisp and to the point. But he is not able to do it because there is no common understanding at about what is business analytics, what is data science etc. There are, these are overlapping concepts. So therefore, there are certain keywords you can find in business analytics like use of data, data driven decisions, use of algorithms and advanced algorithms, essentially to empower decision makers.

That is what business analytics is. So, keep in mind business analytics is for business. It is not for learning, just for learning data analysis techniques. That will be statistics or that will be algorithms. But analytics is related to business also. So before we close, what are the drivers of business value and adoption? As I emphasized throughout the previous sessions, business analytics should create business value. It should augment, it should create either new value or it should influence the top line or bottom line of business. It should bring in money or it should reduce cost. So therefore, so when you increase efficiency, actually it is reducing cost.

So, it must have a value proposition to the business. And that is very important. Only then business will be interested to invest in BI infrastructure and people. And what are some of the supporting research that can rationalize investment in business analytics? I have cited two research references here, wherein scholars have shown that data driven decisions, for example, from a research. So, it is more generalizable. It is not about one case, but it is based on sample data, which is, which shows that standard deviation.



Let me clearly read that. Based on quantitative analysis, as I said, these scholars found that more data driven decision a firm is, the more productive it is. One standard deviation higher on DDD is associated with 46 percent increase in productivity. So, there is a significant correlation between data driven decision practice and business productivity.

And therefore, there is business value. This is one research finding. Other which is more recent is, personalized recommender systems increases consumers propensity to buy 12.4 percent and basket value, basket meaning the market basket or your card. Its value increases by 1.7 percent. This is again based on quantitative research meaning a large sample data. And therefore, this kind of research again support data driven decisions or investment in business analytics. And in other terms, the business value proposition of business analytics is supported by research, recent research. And I wanted to conclude with the business value of analytics. So, we discussed a lot of concepts today related to

business intelligence and analytics. Particularly, we looked at the thought process involved in translating business problems to analytics problems.

We also looked at the corresponding investment in infrastructure that is required to implement this kind of solutions. And then we also looked at definitions of concepts. And then also made a case for investment in business analytics systems, because these are huge investments.

The commercial products are costly. The people are also expensive. Business analytics professionals like you or the future you, are very costly for business. Therefore, put together, they should produce more value for business, only then it makes sense for business to invest in them. Thank you and see you in the next session.