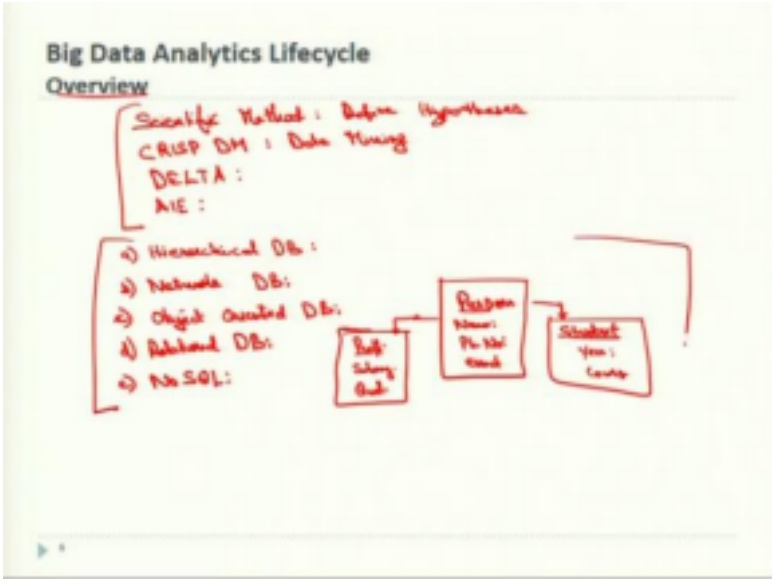
Computer Aided Decision Systems
Industrial Practices Using Big Analytics
Professor Deepu Philip
Department of Industrial and Management Engineering
Professor Amandeep Singh
Imagineering Laboratory
Indian Institute of Technology, Kanpur
Lecture 19
Big Data Analytics Lifecycle - Phase 1

Welcome back to the second lecture of week 5, in which I am discussing Big Data Analytics Lifecycle. When we talk about the Life Cycle, the overview of the Lifecycle will be given first in this lecture then I will give you six phases of the Life Cycle, which was devised from a study.

- The Life Cycle, when we say Lifecycle is a starting to the end of any product, we call it cradle to grave.

Now, here it is a Big Data Analytics process. When I say Lifecycle, the first part would be Discovery; we need to discover what is the dataset? What are we going to do? Slowly, we select the model, we start preparing the model, we put the data in the model, finally we try to validate that and this keeps on going in an iterative way. So, we will put that as a cyclic process, the Life Cycle.

(Refer Slide Time: 01:05)



Before that, the overview of what kind of approach we have to Data Analytics. There are certain methods such as:

i) Scientific methods- Scientific method is the general Data science, where we have the Descriptive, Predictive, Prescriptive Analytics. We use tools such as those commonly available. The tools like MATLAB which are available or we use R programming to just have mean, median or so, we try to have forecasting tools. We try to have a moving average or we try to have an exponential forecast etc. So, these are the general scientific methods which are there for many decades being used or maybe for the centuries it is being used, we first define the Hypothesis here.

a) Define Hypothesis- Hypothesis is a statement that we say if it is not true the data would be presenting some other result, so we define a null Hypothesis, then we say that if the statement is not true then the data is giving some result, so, we will come to that point.

ii) Other than the Scientific methods, we have nowadays, methods such as 'CRISP DM', which is useful in providing input on the ways we frame the analytics problem so this is a popular approach for Data mining.

iii) Then we have DELTA. So, this framework gives us an approach for the Big Data Analytics projects, in which the organization skills, datasets, leadership agreement, everything become part of it.

❖ AIE is the approach that is majorly used by the softwares, which is quite common. For example, the Tableau uses these types of the datasets how it connects to it. So, in this the framework measures intangibles and provides guidance to developing chain models. These are only a few which I have mentioned here, there are multiple other approaches towards Big Data Analytics, it depends upon the kind of the database that we have. Databases could be available in the various forms.

❖ Hierarchical Database- Hierarchical Database means it has a hierarchy, for instance in IIT Kanpur we have a director, with the director where there are connected faculties, students in academics, then we have a registrar. Registrar come takes care of the administration, so this comes the hierarchy. In the faculties we have first professor HHG grade, then professor, then we have associate and assistant professors. We have PhD students, PG students, master students and undergraduate students. For undergraduate students also we have the branch wise students there, different departments are there. So, this is how the hierarchy goes.

❖ Network Database- In Network Databases, we have the director, the department, the

administration, and the students have also made clubs in which the inter department communication takes place, students have a robotics club. They have a programming club, in which the people or the students from all the professors as well from the electrical, computer science, and mechanical departments collectively work together, so this makes a network now.

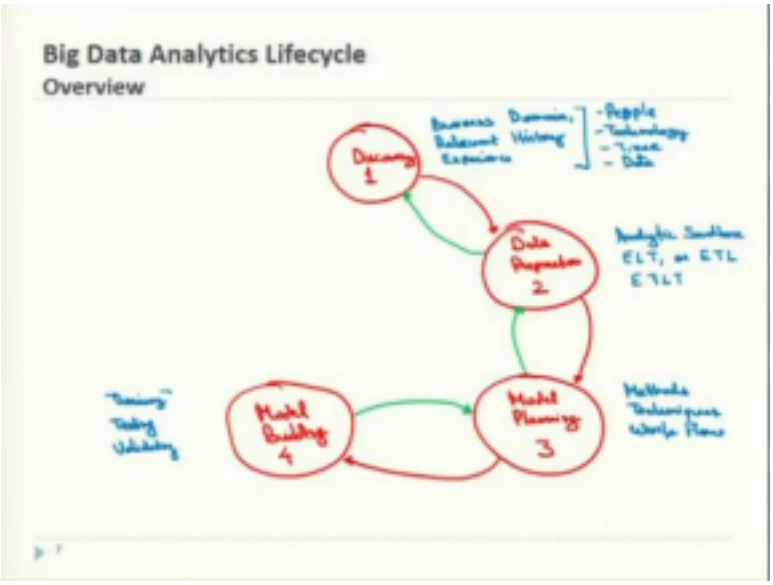
Now another kind of the database, I am quickly going to get through is the general people who are from computer background or who have learned Object-Oriented Programming. They would more appreciate it as an:

- Object-Oriented Database- In this kind of Database, we have information about a person. For example, there is a person, the person could be a student. When I say person it would be the information, here would be the name of the person, the phone number, the email, this is one object. Now, this object for example, if it is a student; the student's year, its course, it becomes another object. Then we have a person who could also be a professor. Here the professor's data might be its salary or its qualifications or so, this is an Object-Oriented Database.
- Relational Database- Relational Database is the most detailed database in a way, so, this database leads in the production line along with the management systems. So, in this kind of the database, each other piece of information has a relationship. For instance, if I say roll number of the student, marks of the students, average marks of the students, these marks of the students related to the other class are marks, so, what is the average marks of other classes? The word the key becomes average. Average marks in class one and class two, Average marks in branch one and branch two. So, there are certain relationships. What is the key of the relationship? What is the intersection point? That is only to be defined, so, this helps us in Data Normalization as well.
- NoSQL Databases- NoSQL when I say it comes to the non-SQL or non-relational databases. So, NoSQL Database provides us a mechanism for the storage or retrieval of the data. This data is modeled in means other than tabular relations used in relational databases. So, NoSQL databases includes simplicity in design, it is simpler horizontal scaling of the cluster of the machines or the systems or the people, so, this kind of data structure has certain advantages, for example, it can work in a NoSQL

Databases such as we can work in Cassandra, we can work in the NoSQL systems which require this kind of data, but disadvantages would be it is an open source database, the GUI is not directly available, and the backup is a weak, because some NoSQL databases cannot be stored for a longer time or cannot be stored very securely unless we have our own system built for it, so, the document size is large. So, depending

upon the different kinds of databases, I will keep on putting different examples from here.

(Refer Slide Time: 08:24)



Now, I will be talking about these in my Data Analytics Lifecycle. So, in the Life Cycle, if I define the phases, I would put them into a cyclic view, which can interact with each other, the first phase will interact with the second phase could go to the third phase, but yes, reverse communication could also happen whenever some data point is required, for example, in the model execution stage itself something is required, Discovery stage required this thing or this data point was also required, so, it might come back there.

Phase 1- Discovery

The Discovery phase is the starting phase. Discovery is nothing but we are trying to explore. We are trying to only say what is the Business Domain? What is the Relevant History? Have you attempted similar projects in the past? For instance, Walmart first took the attempt to have the historical data and took the instances from there to forecast what sales will be in the

next Halloween or the phase? What will sales forecast be in the next Christmas or so? They had thousands of stores, so we are at which point locally? What is happening? Which event is happening and at what shelf? What amount of things could be stored in history? They pick it from there. So, Discovery means what is the history that we have and what is the experience of the people or experience of the team already here? One has experienced that in history, have we taken similar attempts in the past or not?

This all is taken in terms of do we have people who have taken these things in the Business Domain and relevant history? Or do we have experienced people already? Or do we have technology? That is to collect the data to collect what is there on the shelf. Do we have assistance? Like do we have QR code already instituted? Do we have RFID batches or so? So, What kind of technology do we have? Do we have the time or what time frame are we trying to fix? Or what kind of data do we need to have? What kind of databases do we have? So, this is the first phase.

I will talk about this phase in detail, I will talk about the sub phases or steps in this phase itself. For instance, in the Discovery phase we first need to understand what is the Business Domain. Then we need to understand what are the tools or resources that we have. Then we frame the problem. The problem is also framed in this very first phase itself, then we identify the key people there, key stakeholders. Do we need to have really the Business Data Analysts or Business Intelligence Analysts in the discovery phase itself? Or do we only need the Data Scientist only? That is also decided in that Discovery phase itself.

So, then we have a meeting with the project sponsor. Project sponsor meeting means what is the overall viewpoint of the sponsor? Why is the person funding us? What are the results that he or she expects?

This is also taken in the Discovery phase itself, the Discovery phase in the Big Data Analytics is one of the most detail phases, where we try to get as much information as we can, because we do not do like the other Data Analytics or Traditional Data Analytics system, we do not work upon the specifics, we work upon the detail analysis, that is why it is called Big Data and this big data using the Normalization or using the small data retrieval systems and the small languages that we have. We will try to organize that structure and go to the next phase.

Phase 2- Data Preparation

So, Data Preparation means do I really have a good amount of the data? Do I really have a good amount of the or good quality of the data to start building the model? So here, the Discovery phase interacts with the data preparation phase, in which we prepare the Analytics Sandbox. Sandbox, if you understand, it is generally the kids have a wooden box which might be two meters by two-meter or might one meter, we have small tools there, kids come and stay there, so that the sand does not spread in your house or in your garden, you put the sand within that box and kids start playing in it.

When the children are playing in the sandbox itself, they are given different tools, they can build houses, they are given different shoals or saws to use the sand, so, this is the kind of Analytics Sandbox. Analytic Sandbox means my system boundary, my system is this boundary, we will collect this amount of data and we will work to define an Algorithm. If the Algorithm will use tools ABC, we will use only SQL systems to retrieve the data or delete or add. We will have PHP, we will connect it to some markup language like HTML or so, we will have a web or we will have a website developed out of it. So, this is my framework, or I use some systems like for a large dataset Hadoop or so. So that I can develop a bigger sandbox.

This sandbox is once decided, we can expand or contract it in the later stages as well but when once we define it, we will try to create an algorithm which is well tested, trained and validated, so that when this is projected upon the future, or the more datasets it gives us the transformational results, so, Analytic Sandbox is prepared here.

So, here is what we do: generally it is Extract Load and Transform (ELT) or we also do Extract Transform and Load (ETL). Combining these two together, we do Extract Transform Load Transform (ETLT). ELT and ETL are sometimes abbreviated to this, the data is transformed into the ETLT process to make sure that the team can work with and analyze it. In this phase, the team also needs to be familiar with the data thoroughly and take steps into the condition of the data.

Phase 3- Model Planning

Once the sandbox is prepared, we plan the model. This is phase three, phase two interacts with phase three, in which we determine the methods, techniques and workflow. So, this is made to follow the subsequent model building phase, now the team gets aware of the data to

learn about the relationship between the variables of the data, so they subsequently select the key variables or the most suitable model out of it. So, what is this? We will take an example and we will do more explaining.

Phase 4- Model Building

As I said, because we are building the model, we need to have different datasets. For instance, if I have 20 GB of data, the 80 percent of the data is generally taken for the regular testing for the experimentation. The design of experiments is designed in such a way, 80 percent of data is used. 20 percent remaining could be used for the testing and validation. Training is, first we try to do small tests, initial tests or the pilot tests so that we see that other algorithm is running to a small extent, when we actually try to test eighty percent of the big data is for testing, then we do the validation for training, then testing and validating. So, all these steps are designed or thought of or planned in the Model Building stage.

So, in addition to this, the team builds and executes models based upon the work done in the Model Planning phase. So, that was planned in phase three that is executed here. So, they also consider whether the existing tools are working well or not. So, phase three interacts with phase four. So, if the team says the models which we have planned in phase three are not working great, in the training phase itself it is identified, then reverse introduction also takes place (we will put that with the green color).

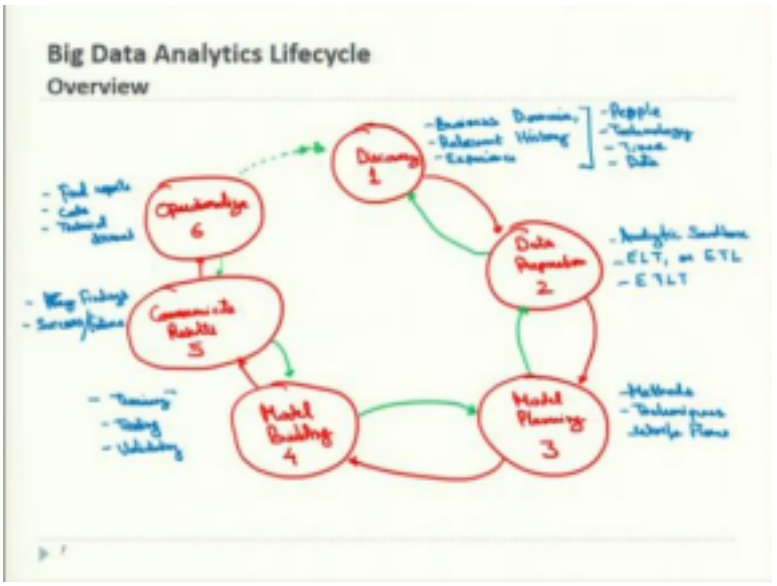
So reverse interaction, if the Model Planning itself says the data preparation in which we have created the data and developed the sandbox, is to be expanded more than a reverse interaction again also plays splits. If the data preparation itself says in the Discovery phase that, we found that the people who are there need more trained people who would have the specific kind of the skill sets, maybe the person who should not be more aware about the Apache spark program. So, then we need more people here, then the timeframe which is set for this project needs to have ten percent more advancement in it, then we need to have time here the datasets are more required, then also we have the reverse interaction, but as far as possible, the reverse interaction should be avoided, so, this is what I wanted to mention in the previous model itself.

(Refer Slide Time: 19:19)

I will have to connect it to the previous slide here, so, these two systems I told you about in the previous lecture. Data Analytics Systems, this system might have taken I would say 20 days and if I say the amount spent on this is let me say Rs 10 Lakhs.

So, this Big Data Analytics system because it is using the ETL and more systematic way and we are also using the most sophisticated programs here, it would take time in hours only, let me say it has taken over 30 hours. And if the time is less, though the investment in the time of the people who are working on it invest in times of if we have purchased small softwares that will be higher, but overall expense for this would be quite less, let me say 2 lakhs. This is the speed, that is why, though the velocity of the Data is high, we still get systems in time.

(Refer Slide Time: 20:23)



So, that is why the interaction or the reverse interaction between the phases is avoided or should not be there as far as possible, but yes, it is made open, it is let open, so that if it is required, we should be able to get information from the previous phases as well.

So, that is why any data or any dataset, that is not used, that is not taken to the further phases is still stored, so, that is now the advantage that we have gotten that we have Doctor Deepu Philip also discussed. Now space is no problem at all, we can have data on the cloud, we can have data on the small size of the pen drive, the small side of the SSD drives in which big terabytes of the data could be stored and kept for future use.

Phase 5- Communicate the Results

We communicate the results, the major stakeholders, the different roles are different that team the people that we have discussed in the previous lecture, the Data Engineer reports to the Database Administrator, which further reports to the Business Intelligence Analysts, the project manager can have an overall overview of the system, the sponsor is given the executive summary all the time Business user is giving the feedback. So, this is how we try to get the results so we identify the key findings from the system.

The key findings, where we saw how the project is successful and with success, I will also put failure because it is a general rule failure is no and no is a new opportunity. Why the system has failed, if this is also learned, we can try to not produce a similar mistake in the further models that we develop. So, why has the system failed? Or why the part of the system has failed? That also is noted here. So, the key findings to quantify the business value or to develop a narrative to summarize or to convey findings to the stakeholders is important here.

Phase 6- Operationalize

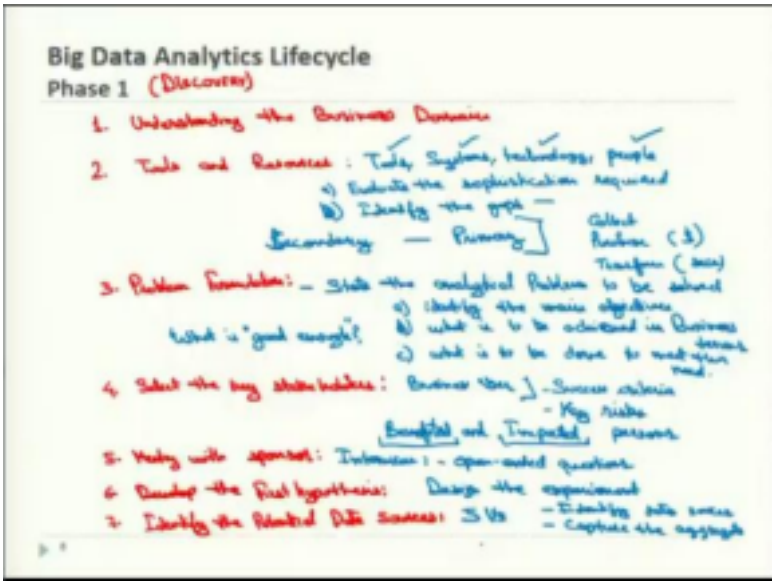
The word Operationalize means giving the final reports. It is phase six. So, phase four helps us to communicate the results, within the system, within the stakeholders. The stakeholders now try to Operationalize the system, so that they can give the results to the final executives. So, we deliver final reports here, that is we try to draft the final reports or we try to get the final code or algorithm ready and we also prepare the technical documents. So, this is how it works. Here also reverse interaction might also take place when we communicate the results, the reverse Model Building testing could also happen.

Once the results are communicated and they are Operationalize, then there are very less chances, I will only put a very dotted line here, so it is very rare chance, that one it is Operationalize because we have gone back and forth between phase one and phase two between phase two and three, three and four and so we are going back and forth to make sure that is the phase five, we are quite sure and system is almost perfect to communicate the results.

Once the results are communicated, then we Operationalize, while we Operationalize and presented the executive summary to the people to execute that finally in the market, then coming back is completely forbidden, I say completely forbidden, but there could be 0.1 percent of the chances that results if it fails further as well still after so much of the testing, something is missed or something is overdone. The systems might have to retake the system overall call.

So, I have only put a dotted line here, also the Operationalize system could only give feedback to the Discovery. This is broadly the lifecycle phases of Big Data Analytics. Now, I will try to discuss each phase in detail and then we will take a case study to understand how one of the companies use this kind of the phases for actual big data analytics report making for themselves.

(Refer Slide Time: 25:45)



Phase 1- Discovery

In the Discovery phase, the big data analytics teams now have to learn or investigate the

problem and develop what is the context? What are we trying to target? Are we going to improve the sales of the system? Or are we just going to improve the productivity throughput in my factory or in my system? Or are we going to reduce the queue length in a retail shop? Or are we going to do some in relative manufacturing? The quality has to be high.

So, it depends upon what is the overall context? What are we trying to target? Those hypotheses will be designed here, those hypotheses related treatments could be made. And then we go to the next phase and try to prepare the data for that. In a Discovery phase, the first and the foremost requirement here is

1) Understanding the Business Domain. So, Data Scientists generally have deep computational or quantitative knowledge, or they have advanced degrees in applied mathematics or statistics. And the Business Data, Business Intelligence Analysts, or Data Engineers sometimes could be a person with a PhD or so. So, the team with these kinds of people need to understand how much business or how much domain knowledge a Data Scientist needs to have to model a system further.

So, to understand the system to have the business domain understanding, what kind of skill set should we have, what kind of assessment a team should have, they should have the right knowledge between the expertise that the team has and the technical knowledge.
Once they have understood the system, they have to now collect the,

2) Tools or resources. When they say tools and resources, now the team needs to assess the resources available to support the project. The resources might include the tools, I would just dot it down here, it should have what are the tools? So, what are the systems in my company? Or what technology do we have? And what people do we have who will be stakeholders or the key stakeholders in the system? So, using this system, they try to:

a) Evaluate the sophistication of the system required: Because when we are talking about the tools and resources, are we going to use this algorithm or this program that we are developing for the present problem statement only? Or we are going to use this further for the overall for the next 10 years for the next five years? So, to evaluate after evaluation or when we are evaluating it, we also

b) Identify the gaps in the datasets, in the tech tools, in the skills: So, whatever those gaps

are to be filled. So, what kinds of skills and rules are needed? Does the requisite level expertise exist in normalization today or not? So, when we answer these questions, it will overall affect or influence the techniques that the team selects and the kind of implementation the team chooses to pursue in the subsequent phases of the total data analytics lifecycle.

So, it is advisable to take inventory of all types of the data, when I say Inventory of the data, where we saw resources, we had the Secondary data first. So, first it is only approached upon the catalogs, the Historical data or whatever you have in inventory, if it is not available, we can also approach the Primary data. The team must determine whether it must collect the Additional data, purchase it from outside or transform the existing data, so it could just collect (secondary or primary) then purchase the data or transform the existing data.

Transformation would need time but it would have only your own skills, for instance if we need to have the average sales which has happened in the last maybe in the pre COVID period itself what was that scenario? And post COVID what is the scenario of the sales of a specific item let me say the face mask itself? So, if the Data is available in the countries in your own area that is fine. Otherwise, we can get Data from the places where the COVID most affected.

For instance, we can get data from China, we can Data from India, the US can forecast data from India itself, so, you can purchase the datasets from India. Or it can also transform the data that they have gotten from their country itself. It would take some time. Purchasing needs more money, I will put dollars and seconds.

Data collection if it is available, it has in the secondary form that it could be collected directly from there. So, they should be the right mix of domain experts, customers, analytical talent, project management, so this is what team is there, in the people that are there, in the tools we have discussed already. In the systems, which means what is the sandbox that you have built? What total size, total time that you have planned for? And what technology are you using?

Tools and technology when we talk about then we need to negotiate for the resources at the outside of the project. So, it is generally more useful to scope the goals or objectives and feasibility in the beginning itself so the technology is selected in that way itself.

3) Problem formulation: Problem formulation means we try to state the analytical problem

that is to be solved. So what problem is big data type to solve? For instance in Walmart there were millions of people every day who were purchasing, so having the right product, at the right place, in right time, this was a problem statement. So, there are millions of products as well. So, what products are more focused? So, there was one instance that is reported in one of the studies I was reading, that the sales of the one product which was thought to be set to peak was not even there, like the product did not even sell in that period of one or two days. So, then they identified that there was something that happened that the rack whether it was to be shelved was covered with the display of that product and the back in the product was not even shelfed, so, if the product is not even shelved the customer will not find it there and customer will not take the product home anyway, so these kinds of these small problems, small errors could come.

So, problem formulation means we frame the problem, so, in case of retailing like this, what target do we have? Or for instance, as it took the example of maybe Amazon Web Services or Netflix, so the problem statement that is to be solved could be so what is going to come next to the box office that could be the problem formulation.

Do we need to have a serial with more emotional domains? Or do we need to have a beautiful actress in that? For example in comedy series they identified using the big data analytics only that with comedy serials or comedy movies where the animals or the animated characters talk that is more sold, for example, the movies like Toy Story, where the toy itself talk to each other to those kinds of comedy systems were more sold.

This was also taken from the Data that what people like more, what kinds of things or what kinds of small when we have small short movies to 3 minutes or 5 minutes those advertisements are first set when the live show that advertisement of views of that rise to maybe millions of so it is identified that, this movie is more expected but what is actually going to hit and what is going to fail there is no prediction of that but we are still coming closer using right problem formulation.

So, problem formulation depends upon the kinds of systems that we have, again, if I take an example of maybe a manufacturing concern, so, like the companies like Shell, so, which is an oil extraction company, so, they had a problem in the face of growing population and the large

requirement of the non-renewable sources still, so, what attempts should be made to generate more energy from the renewable or alternative sources of energy?

So, this is a vast waste of energy from the non-renewable sources that are oil , gas and coal that they are extracting only. So, how is this affected by international politics? What are the international policies? Or what is the difficulty in the exploration? Or what are the hydrocarbons which are generated out of it? So, hydrocarbon generated versus the renewable sources of energy, the amount of hydrocarbons generated in the non-renewable sources of energy, or using the renewable sources of energy and setting up a system for that, it would include or invest require the investment of the lot of the money or if suppose, we are using the hydrocarbon, how is it effected? So, this trade off, that is sustainability, green ecology balance. These kinds of problem formulations could be found and Shell has actually used this kind of system, the big data set system that I will use as a case study to formulate the problem.

So, problem formulation broadly is 'framing of the problem that is stating' the problem statement, that is 'stating the analytical problem to be solved'. Generally, this is a problem statement. So, with all the key stakeholders each member may have a different opinion or a different viewpoint or what is a problem statement? Data Scientists must only be thinking of what kind of programs? How would you write this program? Business Analysts are talking about how the business would run? Sponsor is thinking about the market, but still taking an opinion or taking an input from all of them, the problem statement is made, who makes the statement, generally Business Intelligence Analysts along with the project manager states the problem. Business Intelligence Analysts understand what a database is? How is the System being there? Project manager knows what is the timeframe, what are the milestones? And what is the funding that is coming? So, that is stated here. So, in this statement, there are majorly three major key points;

a) identify the main objective of the project. If I say reducing hydrocarbons, it is a broad hypothesis, like I call it first hypothesis or initial hypothesis, it might have been divided into multiple hypothesis hydrocarbons emission by controlling the pollution reducing hydrocarbon pollution while having the green energy generated at the source or reducing hydrocarbons while using the green energy in the factories where we try to build the machines. So, it depends upon the small sub hypotheses that are being built here. So, the main objective is to reduce hydrocarbons. That is the first

hypothesis that we develop. We will state it based upon this, but this is the first point it is here then we identify,

b) what is to be achieved in the Business terms? Then we say,

c) what is to be done to meet this need? Right now, I am only talking about the Business viewpoint only, what is to be done to meet this need. Here comes the lower level of the Business Analytics or lower level of the Big Data theme, that is Data Scientist, Data Engineers, the people who are actually working upon these, who will apply writing the program, so they all work upon that, a database administrator also has to give what datasets are required. So, they work upon this third step, so, we try to make the boundary of the system, problem formulation itself we have identified or we have made a sandbox. Now to fix the size of the sandbox we will say what is good enough? because we are going to work upon the hydrocarbon saving in the shell company that was targeted to.

So, are they going to work for the overall sales in the world or first they are going to talk about only the sales in the U.S or they are going to talk about their sales in the U.S, maybe in Minnesota itself? So, the problem statement, what is "good enough"? I will put "good enough" in quotes, that is always defined here, what is my boundary? Because the system is big, the system of systems also becomes bigger, so the sandbox in itself is a system of the systems, so, this system of the systems also has to have the outer boundaries, what are these? These are to be defined here.

So, that is where we formulate the problem, so, it is almost taking the best-case scenario approach assuming that everything will proceed as planned, so, that is how the problem is formulated. So, it is almost impossible to plan everything that will emerge in the project, but definitely we can state something that is quite based upon the experience. So, understanding when it is best to stop trying this is also important because if the failure happens, if it happens once, it can be changed, if it happens two it can be again a change, if it happens three times, the failure happens three times it becomes a pattern that means it has to be stopped.

So, that is where do we stop, where do we fail, that is also here in the problem formulation, so, that is to be thought of if we fail, what is a failure criterion? If you fail three times, we will stop it there. So, then we also avoid the unproductive effort and remain aligned to the project sponsors. If it fails, we try to get back to the new problem statement or the modified

problem statement.

4) We select the key stakeholders. The key stakeholders and what are their interests in the project? Here, Business users are definitely part of it. Other than that, the key code stakeholders, so what is the success criteria? Because it might be different from different people, overall, for the Business sponsor it is the money that pays back or money that comes back in the specific time limit.

So, this is the payback criteria or the net worth value that it has. So that is a business criteria or success criteria for the Business sponsor, but for the Data Scientists what is the success criteria? So, that is why we need to select the key stakeholders to do the coding or the program that the data scientist is making? Does he or she have to be part of the major or key stakeholders or not? Sometimes they are told to make the program and they just follow what the Database administrator or the project program manager tells them.

So, success criteria are made based upon what we need exactly, so, that is why the people who are interested in that so those who are taken. Yes, but there might be some risks. We will also list key risks here. For instance, like in the COVID era IIT Kanpur Med. Tech. facility developed oxygen concentrators of their own which were distributed to the country using a startup company, it was developed and distributed. To manufacture this oxygen concentrator within the facility with a minimum available number of resources, the key risks which were there; the risk analysis, the risk priority number was set. So, how do we set those risks? In this the person who was actually assembling the oxygen concentrator machine, his or her inputs were also taken to see what is the wall quality as per our user experience. So, key risks associated with the person who is actually working as a data scientist or the person who is working those are also to be noted. So that is why I need to select the right key stakeholders.
So, when we say key risks, the key stakeholders so this should include anyone who will benefit from the product. So, I would say the person who benefited and the person who is impacted. Benefit means one who gets the advantage or measures out of it once who is impacted, who gets some demerits out of it, maybe if it fails or so. For instance, if the person, who is already busy taking care of the queue at the system where the cash counter or so, is asked to provide the data, he has to be given time, an hour a day or maybe fifteen minutes a day, to provide us this data in this form.

So that person might think that he is impacted, his overall regular routine monitoring work is impacted. So, if he is thinking it is impacted, that is to be taken care of. So, benefited impacted both the keys stakeholders are listed here, so team may feel that it needs to wait for the approval of someone who thinks that he or she is only the advisor, this person who is taking care of the cash could consider that, this is only my advice that is required, but the team is considering this person has the approval. If this person provides the right data, what is the major item that is sold between the peak hours in the evening, for instance between 7 to 8 pm, what is the sale of the maybe child diaper like or small products like those? So, if you are trying to study that, the person has to provide the data, though the data is faded there, let me say there is no actual computer feed or so. So, the person has to provide the data. So, the person's approval is required, yes, the person will happily provide this data each day in the evening or some. So, the advisor and the approver are to be definitely, clearly identified and definitely marked upon.

5) Meeting with the sponsor: Generally, the stakeholders who are selected now, the stakeholders, the person who are approver, they are definitely called for a meeting with a sponsor. The advisors could also be called meeting the sponsor is important because what are the expectations of the person who is funding this project, who is providing the finance and what are his viewpoints? What are his thoughts? What does he expect? And, in what, which timeline? The questions are asked by the person who is versed with the project already. The interview happens in front of the whole team, this meeting is very important. So, the questions, it is a kind of an interview only. So, the questions are open ended. When I say open ended questions, it should not be, do you wish to get the project completed within 3 months or not? Answer is yes or no. What is the time that you expect the project to be completed? The person who would answer 3 months or maybe within 2.5 months or between 3.5 months, you should have completed the open-ended questions. There should not be answered yes or no, true or false or so, open ended questions.

Then the small probing of the details or follow up questions could also be asked by the interviewer. The interviewer also has to have good experience in drafting the questions properly and try to repeat, sometimes repeating the questions means confirming what the person is saying, this is repeating when you might have seen small interviews maybe in the

TV programs or so. The interviewer repeats the answer or reframes the answer in a different way, this is to make sure that the audience clearly understands what the main person who is being interviewed has said, that answer is repeated time and again. Sometimes one time, sometimes two times, it is repeated so that it is very clear what is wanted or what is told.

So, these small skills, interview skills, what Business problem is the team is trying to solve? So, what is the desired outcome of the product? What sources are available? So, what are the industries you are trying to connect? What are the timelines? So, who could provide the insights to the project? So, who has the final decision-making power here? So then regarding questions regarding that time, people, risk, resources, size and attributes of Data, all these questions could be asked so that the people are clear about what they are going to do.

6) Develop the first hypothesis: The hypothesis is a statement as I said we will make one statement that our team can compare its answer with the outcome of the experiment. So, it is best to come up with a few primary hypotheses here. So, these few primary hypotheses are answered, so, we gather and assess the hypothesis from the stakeholders that what are your hypotheses, this statement is made within this timeline, with this performance, would the things come or not, then we design the experiments, so the experiments are designed based upon this.

The electron design of experiments will be taken in the forthcoming weeks, we will learn that taking or doing the small number of experiments out of the big number of the sets of experiments that could have been done. So, this small set of experiments gives a similar output maybe with the 95 percent or 99 percent confidence level. So, this is also important, this saves time and resources to a large extent.

7) Identified potential data sources: Resources here means the data sources. Identify potential data sources. When I say data sources, consider the volume, the velocity, the varsity, the time of the data, that the time span of the data in a way, the type of the data, so, this again the five V's that we discussed.

So, those are all taken care of here. So, we identify the data sources, we capture the aggregate dData sources, then review the raw data. So, I will jot down the points here with 5 Vs itself. We try to identify data sources because this is the last step of the Discovery phase. Now, understanding who are the stakeholders? Understanding the Business Domains, understanding

the tools that we have in formulating the problem and meeting with the sponsor. Now, we will see what data sources are required and how do we capture the aggregate Data sources or so?

Then, we review the raw data, evaluate the data structures and tools needed, and we set the scope of the data infrastructure. So, unlike many traditional stage gate processes, where the team can only advance when the next criteria is only met, so, Data Analytics process can have a parallel retrieval or extraction of the data.

So, it is not once we only get this data, then only the process will start, that is why we prioritize the data based upon the secondary data itself we start our training sets itself. So once the training is started, as in when we receive the other data within the next available time, we keep on input or keep on ingesting that data there as well. So, when we try to have our algorithm run.

These are the steps where we have just discovered the data. Now we will prepare, prepare the model, then build the model, try to get the results, communicate them and operationalize them. We will discuss that in the next lectures of this week. Thank you.