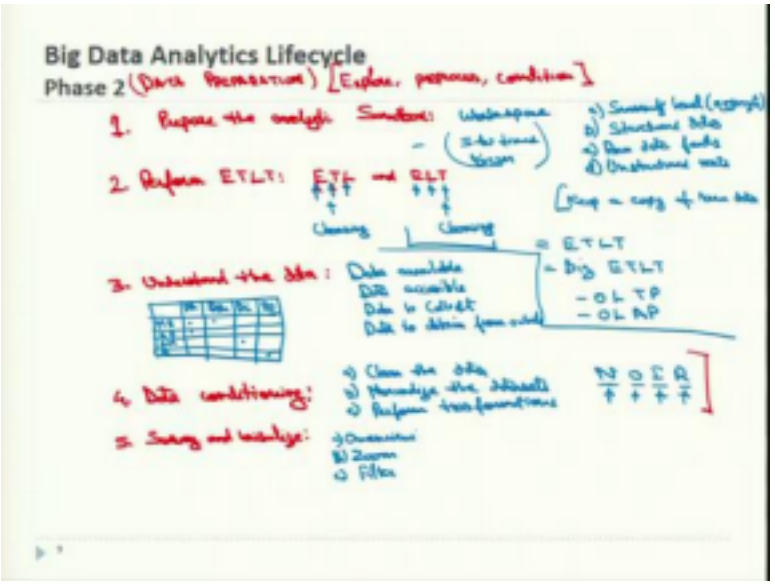
Computer Aided Decision Systems
Industrial Practices Using Big Analytics
Professor Deepu Philip
Department of Industrial and Management Engineering
Professor Amandeep Singh
Imaginary Laboratory
Indian Institute of Technology, Kanpur
Lecture 20
Big Data Analytics Lifecycle – Phase 2

(Refer Slide Time: 0:19)



Welcome, let us continue with the next phase of our Big Data Analytics Lifecycle.

Phase 2: Data Preparation

When I say Data Preparation, it means we try to explore, then we try to pre-process and we try to condition the data as per our Model that we are going to build in the forthcoming phases. So, this is done prior to the modelling or analysis that we are going to do.

So, this is generally done by the preparation of the Analytics sandbox. When I say Analytics Sandbox, it is again the workspace that we are designed to do, then we perform the ETLT system (Extract Transform Load Transform), into the Sandbox, once the data is in the Sandbox, then the team needs to learn about the data and become familiar with it, then learning the data in details is very important to have the project completed within time. The team also must decide how to condition and transform the data into the format to facilitate the subsequent analysis.

So, they may perform some visualizations in the training phase itself that how the data would

visualize, how the results could visualize, because what is going to file comes finally, this overall scenario, overall solution statement is generally there in the mind of the Business Intelligence Analysts itself. So, the person would try to visualize the things, how the things would look, what the graphs would look like or how it should come, if we say 20 per cent improvements in the overall views of a specific program or specific movie that is being broadcasted, so it should the 20 per cent improvements that debt could be put in a graph window, so that could be visualized.

So, there are certain steps in this as well like in phase one, we did in Discovery, so the data preparation phase is generally the most iterative and the one that teams tend to underestimate most often, because they try to jump to the next phase, let us build the model immediately. So, preparation for the data, preparation for the model is very important. It is always important to invest some more time here than it is generally thought of because that will definitely help you to not to come back to this phase once again and that will be saving time in the overall Business cycle or overall lifecycle of the big data analytics systems that we are trying to learn here.

First step, in the Data Preparation phase, is

1) Preparing the Analytics Sandbox- So, when I say Analytics Sandbox, it is our workspace that we are trying to build, when I say workspace, it is the team can explore the data without interfering with the live production databases that is why we put it as a separate sandbox, so that the other parts of the system are not affected like in the Sandbox we have a wooden box as I said, so wooden box the sand does not go out of the box and it does not collector your house. So, this is how the Sandbox word has come here.So, we work with the data analytics here and the regular production, regular retail, regular streaming of the systems are not affected by it. So, we work in this workspace only, so this is a sandbox that is prepared.

So, it is a best practice to collect all kinds of data here and team members need to access the high volumes and varieties of the data for a big data analytics project. So, this can include many parts maybe some reliable aggregated data, I will just say it can have

a) Reliable data or Aggregate data
b) Structured data
c) Raw data feeds,

d) Unstructured text

This text could come from the maybe call logs, maybe blogs which people have given for the specific domain. We are working upon or people might have given some reviews on the online sales portal of the product or so. So, this kind of unstructured text that they give the people if they are identified as the verified or the maybe start user. So, if they are giving a text, that is an important point to be considered.

Many IT groups provide access to only a particular segment of the data for a specific purpose. So, this is the general data analytics try to work with the specific groups of the data, conversely, to this, the data science team in the big data analytics program wants to access everything because from their viewpoint more data is better. So, this is how it differs from the traditional IT programs. So, it is using Historical data, so the data science team needs to have to give IT a justification to develop an Analytics Sandbox which is separate from the traditional Sandbox, so that the overall regular working of the system or of the company is not affected by this. So, it undertakes more ambitious data science projects and moves beyond Traditional data analysis or so.

So, for Business Intelligence to perform more Robust and Advanced Predictive Analytics, the data size has to be as big as possible. So, we will clean the data and pick the data whatever is required. So, generally it is 5 to 10 times bigger than the general datasets, which are used in the regular or Traditional IT programs or the regular data science programs.

2) Performing ETLT- Now ETL and ELT. How does it come to ETLT? This we will discuss here. To make sure that the Analytics Sandbox has a large amount of the bandwidth and reliable network connections to the underlying data sources to enable the uninterrupted or unheralded read and write. Since ETL uses the (Extract Transform Load) program, when we Extract the data, Transform it as per a requirement, then we Load into our system or the analytic system that we are using we cannot have the raw data, raw data is cleansed here. When we transform the data, it is cleaned here. So, this T is cleansing. The big data analytics teams would like to cleanse the data by themselves that is why they would like to extract the data first, then load and then only they would like to cleanse here. So, they will clean the data here and also keep a copy of the raw data.

So, the Analytics Sandbox approach distorts nicely, it advocates Extract Load and Transform,

this is our sandbox approach, the second approach. The reason for this approach is that there is significant value in preserving the raw data and including it in the sandbox before any transformations take place. So, that is why ELT is taken care of.

So, let us take an example of the credit card frauds or let us take an example of the outliers which are very highly impact in the market like credit card frauds or maybe the terrorism attacks that have happened, there might be slight outliers. When we transform that data when the data is clean because of considering these points as outliers, this fraud has happened maybe once in 6 months.

So, if we clean the data based upon the time frame in which this fraud has happened but size of this fraud might be very large or maybe it is in the specific period of the year. Each 6 months, at the end of June and at the end of December, it happens, this is the period related to that. This raw data would be cleaned already if we try to put an ETL process.

So, ETL (Extract Transform and then Load) is transformed and the outliers are cleaned already, but the big data analytics team would have to have access to the raw data itself, so that they can also see the outliers there itself. So, they would like to first Extract, then Load the data, then they would like to Transform by themselves. The outliers also they will see are the outliers coming time and again each year, if those are there, they would like to have a small plot on that as well. So, that is how it does the ELT processor as well.

So, following the ELT approach gives the team access to clean the data to analyze after the data has been loaded into the database and give access to the data in its original form for finding hidden nuances in the data. So, this approach is the part of the reason that Analytics Sandbox can quickly grow large, so that is why electric sandbox resizing keeps on happening as and when we keep on working with the system in the data preparation itself, it is more important to spare more time to define or to fix the system boundary or the size of our sandbox.

So, the team may want to clean and aggregate the data or may keep the copy of the original data as it is mentioned here. So, this process, ETL+ELT, is equal to the ETLT process. So, where we Extract, Transform, Load and then we Transform. So, in Big data we call it Big ETLT, so the programs or the technologies such as Hadoop of mad produce, helped us to do these kinds of operations on them. So, these technologies can be used to perform parallel data

ingestion and introduce a huge number of files or datasets in parallel for a very short period of time. Hadoop is very useful for data loading as well as for data analysis in the subsequent phases.

So, ETLT is a step that is advisable to make inventory of the data, we perform the sort of the Gap analysis, so then the certain transaction, certain processing of the data like such as we have Online Transaction Processing, Online Overland Transaction Processing (OLTP) or we have Online Analytical Processing (OLAP). So, these cubes or these databases are or maybe data fields are taken, so then application program interfaces are developed also here. So, this is how it goes.

3) Understand the data- We need to learn or we need to be familiar with the data itself that means the team has access to the data and identifies the additional data, the team can leverage, but perhaps does not have access to today the data which will be in the future that is to be put. So, the team clarifies the data that the data science team has access to now. So, we will say, data available, then data accessible, then data to collect and data to obtain from third parties from third parties. So, they can build maybe a small table here based upon this, so we can have different kinds of the sections here. I can say this is the movies streamed, the serials streamed, then the short movies streamed, then the reviews given by the people or so. So, I would say data available, then data accessible or data to collect. So, data to take from the outside.

So, we can mark from the movie theme the data is available, for the serial streams the data is available. So, the available is and is also accessible, for them. For the serial stream, sometimes the serial each week it is happening how many people are watching each week, the data is not available directly on each Monday what has happened on Sunday sometimes suppose if it is there. So, we say it is only available but not yet accessible. So, data is to be collected from the outside party, that is how many people have marked their short movies or so.

So, how many people have given reviews is to be taken. A third party is given the contract to get the reviews from the people or the feedback from the people physically by having interviews or from questionnaires, this is to be collected outside. So, this kind of table could also be prepared.

4) Condition the data- So Conditioning means majorly,

a) Clean the data- Cleansing means we are transforming the data as per our requirement.

b) Normalized datasets- Normalizing means we are reducing the number of tables to the minimum possible, so that the whole information will be available but there is no repetition, no redundancy, so we try to normalize that.

c) Perform major other Transformations- These Transformations are as per our programs which are available. Maybe if we have Hadoop, we have to transform the data as per its requirement.

So, a critical step in the Big Data Analytics Lifecycle, the data conditioning can involve many complex steps to join or merge datasets or otherwise get datasets into a state that enables the analysis in further phases. So, it is a step that is also performed by the Traditional IT analytics system as well. So, data conditioning is always required, so that we have a database administrator who has the data available to provide to the data engineer of scientists as an event required.

So, a data based administrator has to do this step or perform this step and it has to have the skills to actually understand what to delete and extract normalization. So, data conditioning is there. So, what are data sources, how the data is cleansed and what is the consistency of the data, so then what is the level of the scales that we are having, do we have consistency or missing points there and also, we use the scale known as NOIR. So do we have data available in the qualitative forms only yes or no, good or bad, red or green or maybe black or white, true or false or is it available in the quantitative form?

So, can we have the data in the quantitative numbers, which could be used to multiply it? Do we have intervals or do we have the ratio data or is it only the order of the data available or are the names of the people normalized or not? For example, names in India would be Vinay Kumar, Ramchandra, it could be Suresh Sharma or names in Punjab would be Manpreet Singh or Harpreet Kaur, Prabhleen Kaur or so. So, if the names are normalized, it is person A or B, that is all. So, nominal is the name that is why I said a few names here.

So, we review the content columns and input the checks that what is the content that is required, would these contents be useful to have proper visualization, could be plotted in a pie chart or do we need to have a histogram or line diagram out of it. So, this is how we try to condition the data. So, as in the training itself we try to have a small visualization of how the data would look like, what the datasets would be finally providing in the future.

5) Survey and visualize-. So, once the team has collected or gotten at least some of the data nets which are needed for subsequent analysis, we try to leverage the data visualization tools. So that is why we put visualization here, survey and visualize the data. So, that is more seen here data Conditioning, we have put the data in the proper scale that is there and Nominal Ordinal Interval or Ratio (NOIR), then we try to visualize we overview first zoom and filter, I would say in visualization I will put it here,

a) Overview

b) Zoom

c) Filter

So, then we try to find the areas of interest which are there, zoom and filter to find more detailed information about a particular data and then we try to find the data behind the particular area.

So, reviewing data to ensure that the calculations remain consistent, so that the data distribution, if those are there what kind of distributions are we following, so can we normalize not can we standardize that because many distributions can be projected or can have normal approximation of those distributions. So, that those are majorly used in the software such as Minitab or design expert three, if they perform the analysis of variance, they try to assume that the variances are normally distributed.

So, these normal distributions etc, so do we have that we need to see in the visualization itself. So, access to the granularity of the data that is the range of the values, level of aggregation. Granularity means what are the data points which are available, how granular, how the different data elements are available, what is the spread of them, that also is important to understand.

So, then we represent the data in the population of interest, so we also need to understand does the data really represent the population of interest? then time related variables may be the data variable for daily, weekly or monthly maybe in the peak hours we might need data for each second what is the sale in America or each second what is the sale worldwide or each millisecond what is the sale in the evening, so something like that.

So, what time frame is there, so accordingly data is surveyed and visualized, but these are the major five steps in the second phase that is Data Preparation, there are certain tools which are there for the data preparation, for example we have a Hadoop I have talked multiple times,

then we have Alpine Miner, we have Spark, we have Teradata. I will talk about the tools as well in one of the lectures and we will first discuss all the phases. Thank You.