UNIT III

Data Analytics Lifecycle

3.1 Need of data Analytic lifecycle:

- The defined data analytics life cycle should be followed by sequences for effectively achieving the goal using input datasets.
- This data analytic process may include identifying the data analytics problems, designing and collection datasets data analytics and data visualization.

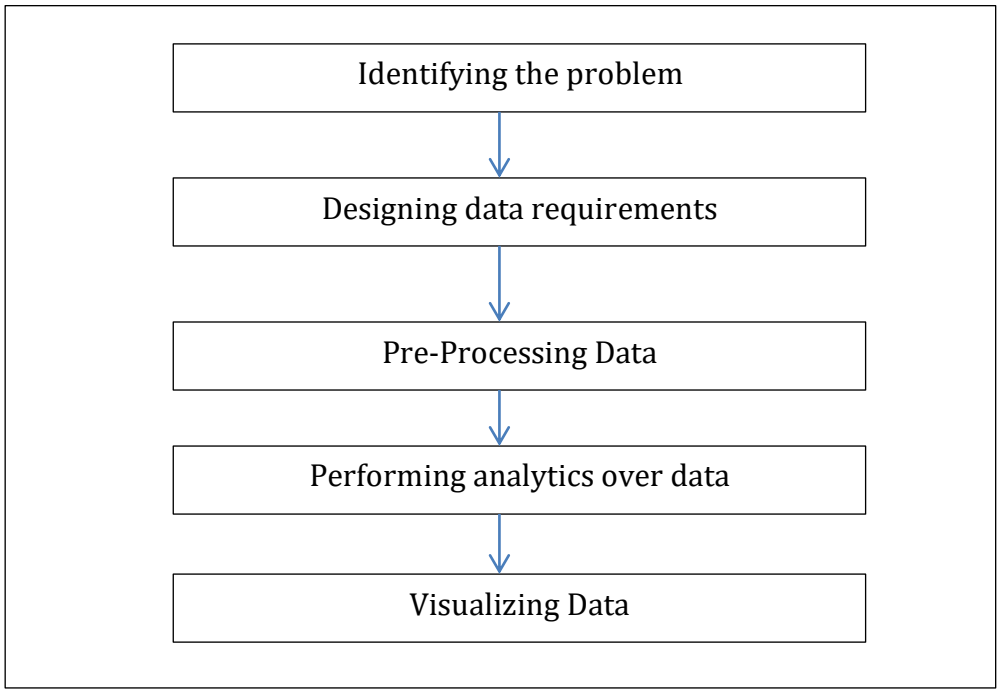

Figure: Data Analytic Lifecycle

1. Identifying the Problem:

- Today's business analytics trends change by performing data analytics over web datasets for growing business. Since their data size increasing gradually day by day, their analytical application need to be scalable for collecting insights from their datasets with the help of analytics we can solve business analytics Problems.
- Assume that we have a large e-commerce website and we want to know how to increase the business. We can identifying the important pages of our websites by categorizing them as per popularity into high, medium and low Based on their popular pages, their types there traffic

sources and there content we will able to decide the roadmap to improve business by improving web traffic as well as content.

2. Designing data requirements:

- To perform data analytics for a specific problem it needs datasets from related domains based on the domain and problem specification.
- The data source can be decided and based on the problem definition the data attributes of these datasets can be defined
- For example, if we are going to perform social media analytics (problem specification) we use the data source as Facebook or Twitter for identifying the user characteristics we need user profile information likes, and posts as data attributes.

3. Preprocessing Data:

- The data analytic, we do not use the sae data sources, data attributes data tools and algorithms all the time as all of them will not use the data in the same format.
- This leads to the performances of data operations such as data cleansing , data aggregation, data augmentation, data sorting and data formatting, to provide a data in a supported format to all the data tools as well as algorithms that will be used in the data analytics.
- In simple terms, Preprocessing is used to perform data operation to translate data into a fixed data format before providing data algorithms or tools.
- The data analytics process will then initiated with this formatted data as the input
- In case of Big Data, the datasets need to be formatted and uploaded to Hadoop distributed file system (HDFS) and used further by various nodes with mapper and reducer in Hadoop Clusters.

4. performing analytic over data:

- After data is available in the required format for data analytics algorithms, data analytics operations will be performed.
- Data analytics operations are perform for discovering meaningful information from data to take better decisions towards business with data mining concepts.
- It may either use descriptive or predictive analytics for business intelligence.
- Analytics can be performed with various machine learning as well as custom algorithms concepts, such as regression, classification, clustering, and model based recommendation for Big Data, the same algorithms can be translated to MapReduce algorithms for running them on Hadoop clusters

by translating there data analytics logic to the MapReduce job which is to be run over Hadoop Clusters.

- These model need to be further evaluated as well as improved by various evaluation stages of machine learning concepts Improved or optimized algorithms can provide better insights.

5. Visualizing Data:

- Data visualization is used for displaying the output of data analytics. Visualization is an interactive way to represent the data insights
- This can be done with various data visualization software as well as R packages

3.2 Key Role for successful analytic projects:

There are different key role for successful analytics projects. Following are the key role for successful analytics projects. These are descriptive of the various roles and main stockholders of an analytics projects.

1. Business User:

- Someone who benefits from the end results and can consult and advised project team on values of end results and how these will be operationalized.
- Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfils this role.

2. Project Sponsor:

- Person responsible for the genesis of the projects providing input for the projects and case business problems, generally providing the funding and will gauge the degree of values from final outputs of the working team.
- This person sets the priorities for the project and clarifies the desired outputs.

3. Project Manager:

- Ensure key milestones and objectives are met on time and at expected quality.

4. Business Intelligent analysts:

- Business domain expertise with deep understanding of the data KPI,s key metrics and business intelligence form a reporting perspective
- Business Intelligence Analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

5. Data Engineer:

- Deep Technical skill to assist with tuning SQL queries for data management, extraction and support data ingest to analytics and box.
- Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox.
- DBA sets up and configures the databases to be used; the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

6. Database Administrator (DBA):

- Provisions and configures the database environment to support the analytics needs of the working team.
- These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.
- Provisions and configures the database environment to support the analytics needs of the working team.
- These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

7. Data Scientist:

- Provide subject matter expertise for analytical techniques, data modelling, applying valid analytical techniques to given business problem and ensuring overall analytical objectives are met.

3.3 Various Phases of data analytic lifecycle:

- There are different phases are data analytic lifecycle, following are the phases

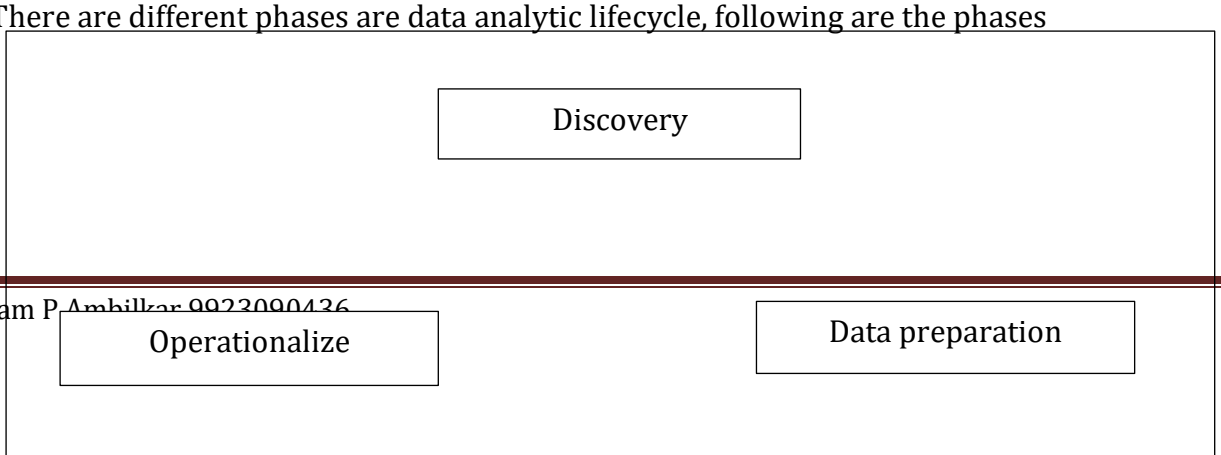

Figure: Phases of Data analytic Lifecycle

1. Discovery
2. Data preparation
3. Model planning
4. Model Building
5. Communication Results
6. Operationalize

1. Data Discovery:

- In this phase, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.
- The team assesses the resources available to support the project in terms of people, technology, time, and data.
- Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

2. Data Preparation:

- In this phase requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.
- The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT.
- Data should be transformed in the ETLT process so the team can work with it and analyze it.
- In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data.

3. Model Planning:

- In this phase, model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

4. Model building:

- In this phase the team develops data sets for testing, training, and production purposes.
- In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
- The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and work flows (for example, fast hardware and parallel processing, if applicable).

5. Communicate Results:

- In this phase, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1.
- The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

6. Operationalize:

- In this phase, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.
- Once team members have run models and produced findings, it is critical to frame these results in a way that is tailored to the audience that engaged the team. Moreover, it is critical to frame the results of the work in a manner that demonstrates clear value.
- If the team performs a technically accurate analysis but fails to translate the results into a language that resonates with the audience, people will not see the value, and much of the time and effort on the project will have been wasted.

3.3.1. Discovery:

The first phase of the Data Analytics Lifecycle involves discovery .In this phase, the data science team must learn and investigate the problem, develop context and understanding, and learn about the data

sources needed and available for the project. In addition, the team formulates initial hypotheses that can later be tested with data.

1. Learning the Business domain:

- Understanding the domain area of the problem is essential.
- In many cases, data scientists will have deep computational and quantitative knowledge that can be broadly applied across many disciplines.
- An example of this role would be someone with an advanced degree in applied mathematics or statistics. These data scientists have deep knowledge of the methods, techniques, and ways for applying heuristics to a variety of business and conceptual problems. Others in this area may have deep knowledge of a domain area, coupled with quantitative expertise.
- An example of this would be someone with a Ph.D. in life sciences. This person would have deep knowledge of a field of study, such as oceanography, biology, or genetics, with some depth of quantitative knowledge. At this early stage in the process, the team needs to determine how much business or domain knowledge the data scientist needs to develop models in Phases 3 and 4.
- The earlier the team can make this assessment the better, because the decision helps dictate the resources needed for the project team and ensures the team has the right balance of domain knowledge and technical expertise

2. Resources:

- As part of the discovery phase, the team needs to assess the resources available to support the project. In this context, resources include technology, tools, systems, data, and people.
- During this scoping, consider the available tools and technology the team will be using and the types of systems needed for later phases to operationalize the models.
- The skills and computing resources, it is advisable to take inventory of the types of data available to the team for the project. Consider if the data available is sufficient to support the project's goals.
- The team will need to determine whether it must collect additional data, purchase it from outside sources, or transform existing data. Often, projects are started looking only at the data available. When the data is less than hoped for, the size and scope of the project is reduced to work within the constraints of the existing data.
- Ensure the project team has the right mix of domain experts, customers, analytic talent, and project management to be effective. In addition, evaluate how much time is needed and if the team has the right breadth and depth of skills.

3. Framing the Problem:

- Framing the problem well is critical to the success of the project. Framing is the process of stating the analytics problem to be solved. At this point, it is a best practice to write down the problem statement and share it with the key stakeholders.
- Each team member may hear slightly different things related to the needs and the problem and have somewhat different ideas of possible solutions.
- For these reasons, it is crucial to state the analytics problem, as well as why and to whom it is important. Essentially, the team needs to clearly articulate the current situation and its main challenges.

4. Identifying key stack holders:

- Another important step is to identify the key stakeholders and their interests in the project.
- The team can identify the success criteria, key risks, and stakeholders, which should include anyone who will benefit from the project or will be significantly impacted by the project.

5. Developing initial Hypothesis:

- Developing a set of IHs is a key facet of the discovery phase. This step involves forming ideas that the team can test with data.
- Generally, it is best to come up with a few primary hypotheses to test and then be creative about developing several more.
- These IHs form the basis of the analytical tests the team will use in later phases and serve as the foundation for the findings
- The team should perform five main activities during this step of the discovery phase:

i .Identify data sources:

- Make a list of candidate data sources the team may need to test the initial hypotheses outlined in this phase.
- Make an inventory of the datasets currently available and those that can be purchased or otherwise acquired for the tests the team wants to perform.

ii. Capture aggregate data sources:

- This is for previewing the data and providing high-level understanding. It enables the team to gain a quick overview of the data and perform further exploration on specific areas.
- It also points the team to possible areas of interest within the data.

iii. Review the raw data:

- Obtain preliminary data from initial data feeds. Begin understanding the interdependencies among the data attributes, and become familiar with the content of the data, its quality, and its limitations.

iv. Evaluate the data structures and tools needed:

- The data type and structure dictate which tools the team can use to analyze the data. This evaluation gets the team thinking about which technologies may be good candidates for the project and how to start getting access to these tools.

v. Scope the sort of data infrastructure needed for this type of problem:

- In addition to the tools needed, the data influences the kind of infrastructure that's required, such as disk storage and network capacity.

3.3.2. Data Preparation:

- The second phase of the Data Analytics Lifecycle involves data preparation, which includes the steps to explore, pre-process, and condition data prior to modelling and analysis.
- In this phase, the team needs to create a robust environment in which it can explore the data that is separate from a production environment. Usually, this is done by preparing an analytics sandbox.
- To get the data into the sandbox, the team needs to perform ETLT, by a combination of extracting, transforming, and loading data into the sandbox. Once the data is in the sandbox, the team needs to learn about the data and become familiar with it. Understanding the data in detail is critical to the success of the project.
- The team also must decide how to condition and transform data to get it into a format to facilitate subsequent analysis.
- The team may perform data visualizations to help team members understand the data, including its trends, outliers, and relationships among data variables.
- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often. This is because most teams and leaders are anxious to begin analysing the data, testing hypotheses.

1. The preparing Analytical sandbox:

- The first sub phase of data preparation requires the team to obtain an analytic sandbox (also commonly referred to as a workspace), in which the team can explore the data without interfering with live production databases.

- Consider an example in which the team needs to work with a company's financial data. The team should access a copy of the financial data from the analytic sandbox rather than interacting with the production version of the organization's main database, because that will be tightly controlled and needed for financial reporting.
- When developing the analytic sandbox, it is a best practice to collect all kinds of data there, as team members need access to high volumes and varieties of data for a Big Data analytics project. This can include everything from summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs or web logs, depending on the kind of analysis the team plans to undertake.

2. Performing ETLT:

- As the team looks to begin data transformations, make sure the analytics sandbox has ample bandwidth and reliable network connections to the underlying data sources to enable uninterrupted read and write.
- In ETL, users perform extract, transform, load processes to extract data from a datastore, perform data transformations, and load the data back into the datastore.
- However, the analytic sandbox approach differs slightly; it advocates extract, load, and then transform. In this case, the data is extracted in its raw form and loaded into the datastore, where analysts can choose to transform the data into a new state or leave it in its original, raw condition.
- 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.
- For instance, consider an analysis for fraud detection on credit card usage. Many times, outliers in this data population can represent higher-risk transactions that may be indicative of fraudulent credit card activity.
- Using ETL, these outliers may be inadvertently filtered out or transformed and cleaned before being loaded into the data store.
- ln this case, the very data that would be needed to evaluate instances of fraudulent activity would be inadvertently cleansed, preventing the kind of analysis that a team would want to do.
- Following the ELT approach gives the team access to clean data to analyze after the data has been loaded into the database and gives access to the data in its original form for finding hidden nuances in the data.

3. Learning about the data:

- A critical aspect of a data science project is to become familiar with the data itself. Spending time to learn the nuances of the datasets provides context to understand what constitutes a reasonable value and expected output versus what is a surprising finding.
- Clarifies the data that the data science team has access to at the start of the project
- Highlights gaps by identifying datasets within an organization that the team may find useful but may not be accessible to the team today. As a consequence, this activity can trigger a project to begin building relationships with the data owners and finding ways to share data in appropriate ways. In addition, this activity may provide an impetus to begin collecting new data that benefits the organization or a specific long-term project.
- Identifies datasets outside the organization that may be useful to obtain, through open APIs, data sharing, or purchasing data to supplement already existing datasets

4. Data conditioning:

- Data conditioning refers to the process of cleaning data, normalizing datasets, and performing transformations on the data.
- A critical step within the Data Analytics Lifecycle, data conditioning can involve many complex steps to join or merge data sets or otherwise get datasets into a state that enables analysis in further phases.
- Data conditioning is often viewed as a Preprocessing step for the data analysis because it involves many operations on the dataset before developing models to process or analyze the data.

5. Survey and Visualize:

- After the team has collected and obtained at least some of the datasets needed for the subsequent analysis, a useful step is to leverage data visualization tools to gain an overview of the data. Seeing high-level patterns in the data enables one to understand characteristics about the data very quickly.
- One example is using data visualization to examine data quality, such as whether the data contains many unexpected values or other indicators of dirty data.
- Another example is skewness, such as if the majority of the data is heavily shifted toward one value or end of a continuum.

6. Common Tools for the data preparation phase:

- Several tools are commonly used for this phase:

- **Hadoop**: can perform massively parallel ingest and custom analysis for web traffic parsing, GPS location analytics, genomic analysis, and combining of massive unstructured data feeds from multiple sources.
- **Alpine Miner:** provides a graphical user interface (GUI) for creating analytic work flows, including data manipulations and a series of analytic events such as staged data-mining techniques (for example, first select the top 100 customers, and then run descriptive statistics and clustering) on Postgres SQL and other Big Data sources.
- **Open Refine (formerly called Google Refine)**: Open Refine (formerly called Google Refine) is "a free, open source, powerful tool for working with messy data."
- It is a popular GUI-based tool for performing data transformations, and it's one of the most robust free tools currently available.
- Similar to Open Refine, Data Wrangler is an interactive tool for data cleaning and transformation. Wrangler was developed at Stanford University and can be used to perform many transformations on a given dataset.
- In addition, data transformation outputs can be put into Java or Python.
- The advantage of this feature is that a subset of the data can be manipulated in Wrangler via its GUI, and then the same operations can be written out as Java or Python code to be executed against the full, larger dataset offline in a local analytic sandbox.

3.3.3 Model Planning:

- In this Phase the data science team identifies candidate models to apply to the data for clustering, classifying, or finding relationships in the data depending on the goal of the project.
- It is during this phase that the team refers to the hypotheses developed in Phase 1, when they first became acquainted with the data and understanding the business problems or domain area.
- These hypotheses help the team frame the analytics to execute in Phase 4 and select the right methods to achieve its objectives.
- Some of the activities to consider in this phase include the following:
    - Assess the structure of the datasets. The structure of the data sets is one factor that dictates the tools and analytical techniques for the next phase.

- Depending on whether the team plans to analyze textual data or transactional data, for example, different tools and approaches are required.
- Ensure that the analytical techniques enable the team to meet the business objectives and accept or reject the working hypotheses.
- Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow. A few example models include association rules and logistic regression.
- Other tools, such as Alpine Miner, enable users to set up a series of steps and analyses and can serve as a front-end user interface (UI) for manipulating Big Data sources in PostgreSQL.

1. Data Exploration and Variable Selection:

- Although some data exploration takes place in the data preparation phase, those activities focus mainly on data hygiene and on assessing the quality of the data itself.
- The objective of the data exploration is to understand the relationships among the variables to inform selection of the variables and methods and to understand the problem domain.
- As with earlier phases of the Data Analytics Lifecycle, it is important to spend time and focus attention on this preparatory work to make the subsequent phases of model selection and execution easier and more efficient.
- A common way to conduct this step involves using tools to perform data visualizations.

2. Model Selection:

- In the model selection subphase, the team's main goal is to choose an analytical technique, or a short list of candidate techniques, based on the end goal of the project.
- In the case of machine learning and data mining, these rules and conditions are grouped into several general sets of techniques, such as classification, association rules, and clustering.
- When reviewing this list of types of potential models, the team can winnow down the list to several viable models to try to address a given problem.

3. Common Tools for the Model Planning Phase:

- Many tools are available to assist in this phase. Here are several of the more common ones:
- R has a complete set of modelling capabilities and provides a good environment for building interpretive models with high-quality code. In addition, it has the ability to interface with databases via an ODBC connection and execute statistical tests and analyses against Big Data via an open source connection.

- These two factors makeR well suited to performing statistical tests and analytics on Big Data. As of this writing, R contains nearly 5,000 packages for data analysis and graphical representation.
- New packages are posted frequently, and many companies are providing value-add services for R (such as training, instruction, and best practices), as well as packaging it in ways to make it easier to use and more robust.
- This phenomenon is similar to what happened with Linux in the late 1980s and early 1990s, when companies appeared to package and make Linux easier for companies to consume and deploy.
- UseR with fi le extracts for offline analysis and optimal performance, and use RODBC connections for dynamic queries and faster development.
- SQL Analysis services can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.
- SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JOB (, and OLE DB. SAS itself is generally used on file extract s, but with SAS/ACCESS, users can connect to relational databases (such as Oracle or Teradata) and data warehouse appliances (such as Green plum or Aster), files, and enterprise applications.

3.3.4. Model Building:

- In this Phase, the data science team needs to develop data sets for training, testing, and production purposes. These data sets enable the data scientist to develop the analytical model and train it ("training data"), while holding aside some of the data ("hold-out data" or "test data") for testing the model.
- During this process, it is critical to ensure that the training and test datasets are sufficiently robust for the model and analytical techniques.
- A simple way to think of these datasets is to view the training dataset for conducting the initial experiments and the test sets for validating an approach once the initial experiments and models have been run.
- In the model building phase, an analytical model developed and fit on the training data and evaluated (scored) against the test data. The phases of model planning and model building can overlap quite a bit, and in practice one can iterate back and forth between the two phases for a while before settling on a final model.
- Although the modeling techniques and logic required to develop models can be highly complex, the actual duration of this phase can be short compared to the time spent preparing the data and defining the approaches. In general, plan to spend more time preparing and learning the data and

crafting a presentation of the findings. Phases 3 and 4 tend to move more quickly, although they are more complex from a conceptual standpoint.

- The data science team needs to execute the models defined in Phase 3. During this phase, users run models from analytical software packages, such as R or SAS, on file extracts and small data sets for testing purposes. On a small scale, assess the validity of the model and its results. For instance, determine if the model accounts for most of the data and has robust predictive power.
- At this point, refine the models to optimize the results, such as by modifying variable inputs or reducing correlated variables where appropriate. In Phase 3, the team may have had some knowledge of correlated variables or problematic data attributes which will be confirmed or denied once the models are actually executed.
- When immersed in the details of constructing models and transforming data, many small decisions are often made about the data and the approach for the modeling. These details can be easily forgotten once the project is completed.
- Therefore, it is vital to record the results and logic of the model during this phase. In addition, one must take care to record any operating assumptions that were made in the modeling process regarding the data or the context.
- Creating robust models that are suitable to a specific situation requires thoughtful consideration to ensure the models being developed ultimately meet the objectives outlined in Phase 1.

1. Common Tools for the model Building phase:

There are many tools available to assist in this phase, focused primarily on statistical analysis or data mining software. Common tools in this space include, but are not limited to, the following:

- **Commercial Tools:**
    - **SAS Enterprise Miner:** allows users to run predictive and descriptive models based on large volumes of data from across the enterprise. It interoperates with other large data stores, has many partnerships, and is built for enterprise-level computing and analytics.
    - **SPSS Modeler:** (provided by IBM and now called IBM SPSS Modeler) offers methods to explore and analyze data through a GUI.
    - **Matlab:** provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.
    - **Alpine Miner:** provides a GUI front end for users to develop analytic workflows and interact with Big Data tools and platforms on the back end.

- **STATISTICA** and **Mathematica** are also popular and well-regarded data mining and analytics tools.

- **Free or Open Source tools:**
  - **Rand PL/R:** R was described earlier in the model planning phase, and PL!R is a procedural language for PostgreSQL with R. Using this approach means that R commands can be executed in database.
  - This technique provides higher performance and is more scalable than running R in memory.
  - **Octave:** a free software programming language for computational modeling, has some of the functionality of Matlab. Because it is freely available, Octave is used in major universities when teaching machine learning.
  - **WEKA:** is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
  - **Python**: python is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib.
  - **SQL:** SQL in-database implementations, such as MADlib, provide an alternative to in-memory desktop analytical tools. MADiib provides an open-source machine learning library of algorithms that can be executed in-database, for PostgreSQL or Greenplum.

3.3.5 Communication Result:

- After executing the model, the team needs to compare the outcomes of the modeling to the criteria established for success and failure.
- In this phase, the team considers how best to articulate the findings and outcomes to the various team members and stakeholders, taking into caveats, assumptions, and any limitations of the results account.
- Because the presentation is often circulated within an organization, it is critical to articulate the results properly and position the findings in a way that is appropriate for the audience.
- The team needs to determine if it succeeded or failed in its objectives. Many times people do not want to admit to failing, but in this instance failure should not be considered as a true failure, but rather as a failure of the data to accept or reject a given hypothesis adequately.

- This concept can be counterintuitive for those who have been told their whole careers not to fail. However, the key is to remember that the team must be rigorous enough with the data to determine whether it will prove or disprove the hypotheses outlined in Phase 1 (discovery).
- Sometimes teams have only done a superficial analysis, which is not robust enough to accept or reject a hypothesis. Other times, teams perform very robust analysis and are searching for ways to show results, even when results may not be there.
- It is important to strike a balance between these two extremes when it comes to analysing data and being pragmatic in terms of showing real-world results. When conducting this assessment, when conducting this assessment, determine if the results are statistically significant and valid.
- If they are, identify the aspects of the results that stand out and may provide salient findings when it comes time to communicate them. If the results are not valid, think about adjustments that can be made to refine and iterate on the model to make it valid.
- During this step, assess the results and identify which data points may have been surprising and which were in line with the hypotheses that were developed in Phase 1.
- Comparing the actual results to the ideas formulated early on produces additional ideas and insights that would have been missed if the team had not taken time to formulate initial hypotheses early in the process.
- By this time, the team should have determined which model or models address the analytical challenge in the most appropriate way.
- In addition, the team should have ideas of some of the findings as a result of the project. The best practice in this phase is to record all the findings and then select the three most significant ones that can be shared with the stakeholders.
- In addition, the team needs to reflect on the implications of these findings and measure the business value.
- Depending on what emerged as a result of the model, the team may need to spend time quantifying the business impact of the results to help prepare for the presentation and demonstrate the value of the findings.
- Assess intangibles in business and quantify the value of seemingly unmeasurable things. Now that the team has run the model, completed a thorough discovery phase, and learned a great deal about the datasets, reflect on the project and consider what obstacles were in the project and what can be improved in the future.

- Make recommendations for future work or improvements to existing processes, and consider what each of the team members and stakeholders needs to fulfill her responsibilities. For instance, sponsors must champion the project.
- Stakeholders must understand how the model affects their processes. (For example, if the team has created a model to predict customer churn, the Marketing team must understand how to use the churn model predictions in planning their interventions.) Production engineers need to operationalize the work that has been done.
- In addition, this is the phase to underscore the business benefits of the work and begin making the case to implement the logic into a live production environment.
- As a result of this phase, the team will have documented the key findings and major insights derived from the analysis.
- The deliverable of this phase will be the most visible portion of the process to the outside stakeholders and sponsors, so take care to clearly articulate the results, methodology, and business value of the findings.

3.3.6 Operationalize:

- In the final phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem of users.
- The team scored the model in the analytics sandbox, represents the first time that most analytics teams approach deploying the new analytical methods or models in a production environment. Rather than deploying these models immediately on a wide-scale basis, the risk can be managed more effectively and the team can learn by undertaking a small scope, pilot deployment before a wide-scale rollout.
- This approach enables the team to learn about the performance and related constraints of the model in a production environment on a small scale and make adjustments before a full deployment.
- During the pilot project, the team may need to consider executing the algorithm in the database rather than with in-memory tools such as R because the run time is significantly faster and more efficient than running in-memory, especially on larger datasets.
- While scoping the effort involved in conducting a pilot project, consider running the model in a production environment for a discrete set of products or a single line of business, which tests the model in a live setting.

- This allows the team to learn from the deployment and make any needed adjustments before launching the model across the enterprise.
- Be aware that this phase can bring in a new set of team members- usually the engineers responsible for the production environment who have a new set of issues and concerns beyond those of the core project team.
- This technical group needs to ensure that running the model fits smoothly into the production environment and that the model can be integrated into related business processes. Part of the operationalizing phase includes creating a mechanism for performing on-going monitoring of model accuracy and, if accuracy degrades, finding ways to retrain the model.
- If feasible, design alerts for when the model is operating "out-of-bounds." This includes situations when the inputs are beyond the range that the model was trained on, which may cause the outputs of the model to be inaccurate or invalid.
- If this begins to happen regularly, the model needs to be retrained on new data.
- Often, analytical projects yield new insights about a business, a problem, or an idea that people may have taken at face value or thought was impossible to explore.
- Four main deliverables can be created to meet the needs of most stakeholders.

The key outputs for each of the main stakeholders of an analytics project and what they usually expect at the conclusion of a project.

1. Business User:

- Typically tries to determine the benefits and implications of the findings to the business.

2. Project Sponsor:

- Typically asks questions related to the business impact of the project, the risks and return on investment (ROI), and the way the project can be evangelized within the organization (and beyond).

3. Project Manager:

- Project Manager needs to determine if the project was completed on time and within budget and how well the goals were met.

4. Business Intelligence Analyst:

- Needs to know if the reports and dashboards he manages will be impacted and need to change.

5. Data Engineer and Database Administrator (DBA):

- Typically need to share their code from the analytics project and create a technical document on how to implement it.

6. Data Scientist:

- Needs to share the code and explain the model to her peers, managers, and other stakeholders.

Although these seven roles represent many interests within a project, these interests usually overlap, and most of them can be met with four main deliverables.

- Presentation for project sponsors: This contains high-level takeaways for executive level stakeholders, with a few key messages to aid their decision-making process. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
- Presentation for analysts, which describes business process changes and reporting changes. Data scientists will want the details and are comfortable with technical graphs (such as Receiver Operating Characteristic [ROC) curves, density plots, and histograms.
- Code for technical people.
- Technical specifications of implementing the code.