Title: **Introduction to Analytics** Date:

Topic: **Section 2: Data Analytics Lifecycle** Continued from:

Discovery Phase:

Notes

Identify & describe the discovery phase

The Discovery phase is the initial phase where the team **learns** the business domain and relevant history (e.g., is organization first project?,) **assesses** the resources available to support the project, and **conduct** project activities.

Explain the purpose of the discovery phase

The purpose of the Discovery Phase is to work on the project activities such as, **framing** *(the process of stating the analytics problem to be solved)* the business problem or opportunity, **identifying** key stakeholders (e.g., project sponsors,) **setting** project objectives and scope, **formulating** the initial hypotheses (IHs), and **identifying** potential data (e.g., volume for amount, variety for selection, velocity for speed).

Define the questions of interest

Questions of interest in the Discovery phase involves understanding the business problem and formulating hypotheses. Specific **questions** may include is enough information here to draft an analytic plan? and address well-defined business problems?

Assess the resources or resource constraints

The Discovery phase **Resources** such as, technology (e.g., programming tools, types of systems), time (e.g., analyze 1yr or 10-yrs of data), data (e.g., collect additional data, purchase outside data, used to test hypotheses,) and people (e.g., data scientist, does team have the skills now?) are assessed. This ensures that necessary resources are available to support the project.

Define outcomes

During the Discovery phase is where the team define outcomes. The outcomes are defined based on the analytic plan *(a peer review that shows understanding of the analytic problem and ways to address it)* and the **success or failure criteria developed** *(a way to avoid unproductive effort and remain aligned with the project sponsors).* Moving forward to the next phase can't happen until these outcomes have been completed.

Summary

Title: Introduction to Analytics Date:

Topic: Section 2: Data Analytics Lifecycle Continued from: Section 2.1

Data Preparation Phase:

Notes

Identify & describe the phase

The Data Preparation phase is the second phase for the team to execute extract, transform, load, transform (ETLT) processes, familiarize themselves with the data thoroughly, and take steps to condition the data.

Explain the purpose of the phase

The purpose of the Data Preparation phase is to **obtain** an analytic sandbox *(a.k.a workspace)* without interfering with live production databases. The team will **prepare** data (e.g., ELT, ETL processes) for analysis, **consider factors** like data availability, algorithms, data complexity, and outcome frequency. The extract, load, transform (ELT) process is the preferred data transformation method because certain data wouldn't be inadvertently cleansed, and still allow the data to be in its original form.

Identify data sources

Data Preparation data sources is influenced by data availability. The analytic sandbox will house high volumes and a variety of data *(e.g., raw, unstructured, external sources, departmental databases, etc.)*

Identify common tools

Tools used in the Data Preparation phase includes ***Hadoop*** *(which handles large unstructured data)*, **Alpine Miner** *(which creates analytic workflows)*, **OpenRefine** *(which is good for working with messy data)*, **and Data Wrangler** *(which interactively cleans and transforms data.)*

Identify steps

Data Preparation steps are to set up sandboxes, make an inventory of the data and compare it, conduct (ETLT) processes, condition data, investigate distributions, clarify inconsistencies, and survey and explore data visually.

Summary

Title: **Introduction to Analytics** Date:

Topic: **Section 2: Data Analytics Lifecycle** Continued from: **Section 2.2**

Model Planning Phase:

Notes

Identify & describe the phase

The Model Planning phase is the third phase that **identifies** appropriate **models** for clustering, **classification**, or to **uncover relationships**, and **workflows**.

Explain the purpose of the phase

The purpose of the Model Planning phase is to plan and determine suitable models for subsequent phases.

Identify the activities of the phase

The activities in the Model Planning phase are **ensuring accessibility to dataset structures, variable selections, exploring other approaches, partition datasets** (for training, validating, and testing), **selecting analytical techniques** *(which is a short list of candidates),* **and aligning with business goals**. These activities focus mainly on data hygiene and on assessing the quality of the data itself.

Identify common tools

Common tools in the Model Planning phase are *SASS/ACCESS* *(which connects users to relational databases and data warehouses),* **R** *(which builds interpretive models with high-quality code),* **SQL Analysis Services** *(which performs in-database data mining functions.)*

Identify common models

Common models in the Model Planning phase are **Clustering** - k-means, hierarchical, etc.; **Classifications** - logistic regression, decision trees, automatic relevance determination (ARD), etc.; **Uncovering Relationships** - linear regression, correlation analysis, etc.; and **Specific Models** - association models, neural networks, etc...

Summary

Title: Introduction to Analytics Date:

Topic: Section 2: Data Analytics Lifecycle Continued from: Section 2.3

Model Build/Execution Phase:

Notes

Identify & describe the phase

The Model Execution phase is the fourth phase that **involves developing datasets for training, testing, and production**, builds and executes models based on the planning phase, and evaluating the need for more robust tools or environments.

Explain the purpose of the phase

The purpose of the Model Execution phase is to develop datasets, refine models, and assess validity, construct and evaluate models, address questions about accuracy, inputs, transformations, and run-time requirements.

Identify the activities of the phase

Activities in the Model Execution phase include users running models from analytical software packages on file extracts and small datasets for testing purposes, refine the models to optimize the results, record any operating assumptions that were made in the modeling process regarding the data or the context, and record the results and logic of the model.

Identify common tools

Common tools used in the Model Execution phase can be either commercial *(e.g.,* ***SAS Enterprise Miner, SPSS Modeler****)* or open-source tools (e.g., **Octave**, **MADlib.**)

Summary

Title: *Introduction to Analytics* Date:

Topic: *Section 2: Data Analytics Lifecycle* Continued from: *Section 2.4*

	Notes
Communicate Results Phase: *Identify & describe the phase*	The Communicate Results phase is the fifth phase where the team **determines** the outcomes of the **success and failure criteria** *(a benchmark used to determine whether the analysis has met its objectives)* established in phase 1, **identifies key findings**, **quantifies the business value**, and develops a narrative to **summarize** and **share** the **findings** to stakeholders.
Explain the purpose of the phase	The purpose of the Communicate Results phase is to implement and maintain the analytics solution in a production environment, communicate the findings and outcomes to the various team members and stakeholders, and determine whether the data will prove or disprove the hypotheses.
Identify the activities of the phase	Activities in the Communicate Results phase include analyzing the data, determining if the results are statistically significant and valid, determining which model(s) address the analytical challenge, record all the findings and share the top three with stakeholders, reflect on the implications and measure the business value, and make recommendations or improvements.
Identify common tools	The Communicate Results phase tools are for presenting clear results to stakeholders such as, **Tableau**, **Power BI**, **Microsoft PowerPoint**, and **D3.js** (which creates Web-based Visualizations).

Summary

Title: *Introduction to Analytics* Date:

Topic: *Section 2: Data Analytics Lifecycle* Continued from: *Section 2.5*

Operationalize Phase:	Notes
Identify & describe the phase	The Operationalize phase is the final phase that **communicates project benefits**, **sets** up the **pilot project**, and **deploys** in **production**.
Explain the purpose of the phase	The purpose of the Operationalize phase is to **test** the **model** in a **controlled environment**, **make** necessary **adjustments** and **integrating** it into practical uses within the organization.
Identify key outputs for stakeholders	Key outputs of the Operationalize phase include the **Business Users**, typically tries to determine the benefits and implications of the findings to the business; **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); **Project Manager**, needs to determine if the project was completed on time and within budget and how well the goals were met; **Business Intelligence (BI) Analyst**, needs to know if the reports and dashboards he manages will be impacted and need to change; **Data Engineer** / **Database Administrator (DBA)**, typically need to share their code from the analytics project and create a technical document on how to implement it; **Data Scientist**, needs to share the code and explain the model to her peers, managers, and other stakeholders.

Summary