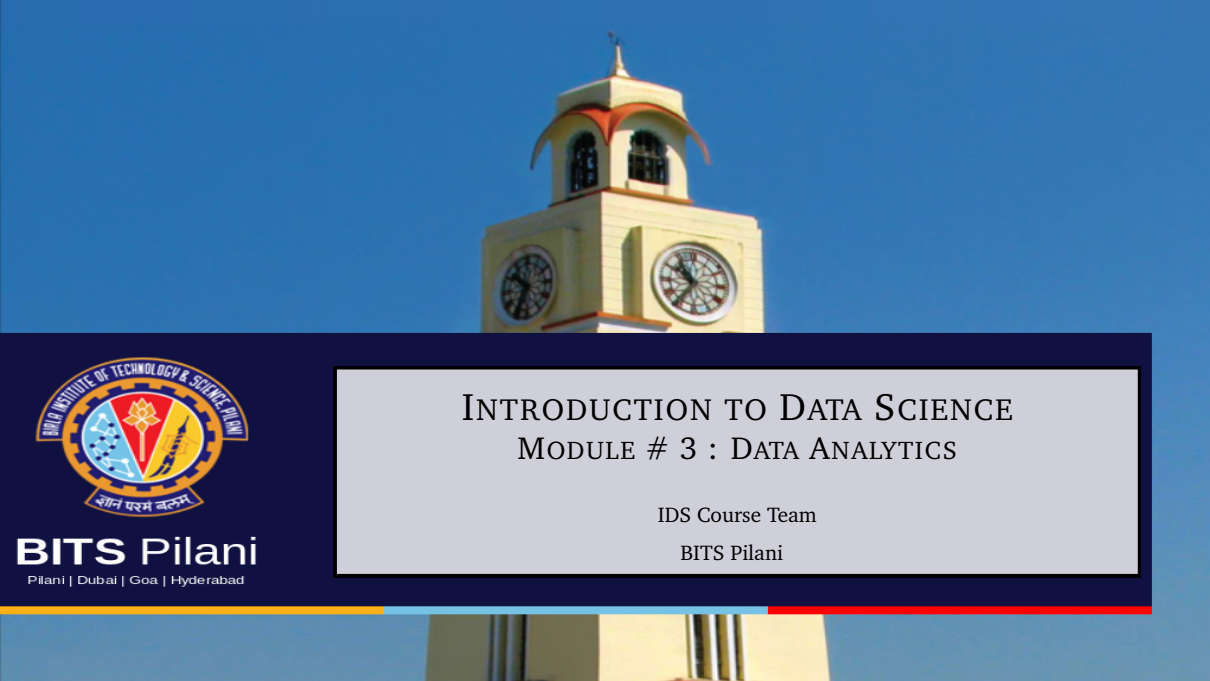


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science
Module # 3 : Data Analytics

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

Table of Contents

Data Analytics

- Data Analytics is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making.

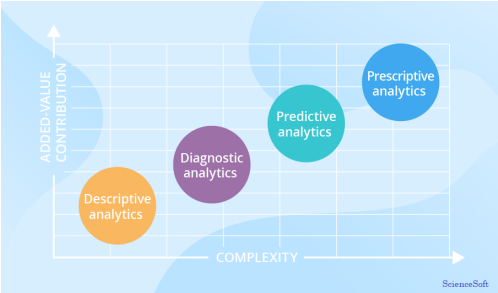


Table of Contents

Data Analytics Methodologies

- Use standard methodology to ensure a good outcome.
  1. CRISP-DM
  2. Big Data Life-cycle
  3. SEMMA
  4. SMAM

Need for a Standard Process

- Framework for recording experience.
  - Allows projects to be replicated
- Aid to project planning and management.
- “Comfort factor” for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on “stars”
- Encourage best practices and help to obtain better results.

Data Science Methodology

10 Questions the process aims to answer

- Problem to Approach
  1. What is the problem that you are trying to solve?
  2. How can you use data to answer the questions?
- Working with Data
  3. What data do you need to answer the question?
  4. Where is the data coming from? Identify all Sources. How will you acquire it?
  5. Is the data that you collected representative of the problem to be solved?
  6. What additional work is required to manipulate and work with the data?
- Delivering the Answer
  7. In what way can the data be visualized to get to the answer that is required?
  8. Does the model used really answer the initial question or does it need to be adjusted?
  9. Can you put the model into practice?
  10. Can you get constructive feedback into answering the question?

Table of Contents

CRISP-DM

- Cross Industry Standard Process for Data Mining
- conceived around 1996
- 6 high-level phases
- Used in IBM SPSS Modeler tool
- Iterative approach to the development of analytical models.

CRISP-DM Phases

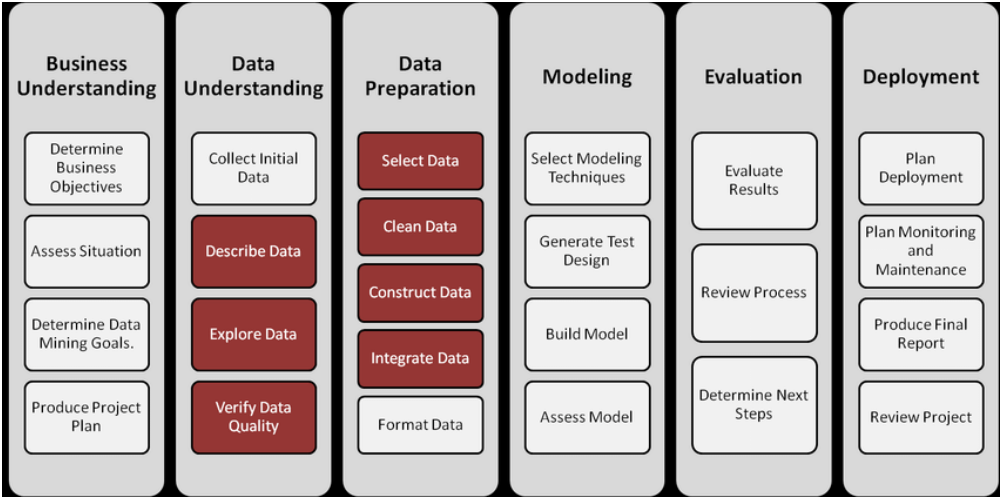
CRISP-DM Phases

- Business Understanding
  - Understand project objectives and requirements.
  - Data mining problem definition.
- Data Understanding
  - Initial data collection and familiarization.
  - Identify data quality issues.
  - Identify initial obvious results.
- Data Preparation
  - Record and attribute selection.
  - Data cleansing.

CRISP-DM Phases

- Modeling
  - Run the data mining tools.
- Evaluation
  - Determine if results meet business objectives.
  - Identify business issues that should have been addressed earlier.
- Deployment
  - Put the resulting models into practice.
  - Set up for continuous mining of the data.

CRISP-DM Phases and Tasks



Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills.
- CRISP-DM provides a uniform framework for
  - guidelines.
  - experience documentation.
- CRISP-DM is flexible to account for differences.
  - Different business/agency problems.
  - Different data

Table of Contents

Big Data Life-cycle

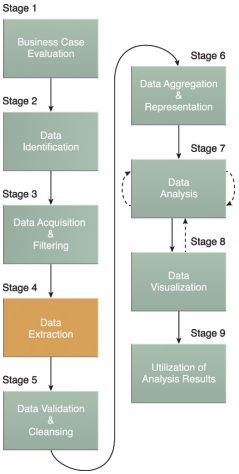
- Data Acquisition
  - Acquiring information from a rich and varied data environment.
- Data Awareness
  - Connecting data from different sources into a coherent whole, including modeling content, establishing context, and insuring search-ability.
- Data Analytics
  - Using contextual data to answer questions about the state of your organization.
- Data Governance
  - Establishing a framework for providing for the provenance, infrastructure and disposition of that data.

Big Data Life-cycle

- Phase 1: Foundations
- Phase 2: Acquisition
- Phase 3: Preparation
- Phase 4: Input and Access
- Phase 5: Processing
- Phase 6: Output and Interpretation
- Phase 7: Storage
- Phase 8: Integration
- Phase 9: Analytics and Visualization
- Phase 10: Consumption
- Phase 11: Retention, Backup, and Archival
- Phase 12: Destruction

PS: Some phases may overlap and can be done in parallel.

Big Data Life-cycle



Big Data Life-cycle

- Phase 1: Foundations
  - Understanding and validating data requirements, solution scope, roles and responsibilities, data infrastructure preparation, technical and non-technical considerations, and understanding data rules in an organization.
- Phase 2: Data Acquisition
  - Data Acquisition refers to collecting data.
  - Data sets can be obtained from various sources, both internal and external to the business organizations.
  - Data sources can be in
    - structured forms such as transferred from a data warehouse, a data mart, various transaction systems.
    - semi-structured sources such as Weblogs, system logs.
    - unstructured sources such as media files consisting of videos, audios, and pictures.

Big Data Life-cycle

- Phase 3: Data Preparation
  - Collected data (Raw Data) is rigorously checked for inconsistencies, errors, and duplicates.
  - Redundant, duplicated, incomplete, and incorrect data are removed.
  - The objective is to have clean and useable data sets.
- Phase 4: Data Input and Access
  - Data input refers to sending data to planned target data repositories, systems, or applications.
  - Data can be stored in CRM (Customer Relationship Management) application, a data lake or a data warehouse.
  - Data access refers to accessing data using various methods.
  - NoSQL is widely used to access big data.

Big Data Life-cycle

- Phase 5: Data Processing
  - Processing the raw form of data.
  - Convert data into a readable format giving it the form and the context.
  - Interpret the data using the selected data analytics tools such as Hadoop MapReduce, Impala, Hive, Pig, and Spark SQL.
  - Data processing also includes activities
    - ★ Data annotation – refers to labeling the data.
    - ★ Data integration – aims to combine data existing in different sources, and provide a unified view of data to the data consumers.
    - ★ Data representation – refers to the way data is processed, transmitted, and stored.
    - ★ Data aggregation – aims to compile data from databases to combined data-sets to be used for data processing.

Big Data Life-cycle

- Phase 6: Data Output and Interpretation
  - In the data output phase, the data is in a format which is ready for consumption by the business users.
  - Transform data into usable formats such as plain text, graphs, processed images, or video files.
  - This phase is also called the **data ingestion**.
  - Common Big Data ingestion tools are Sqoop, Flume, and Spark streaming.
  - Interpreting the ingested data requires analyzing ingested data and extract information or meaning out of it to answer the questions related to the Big Data business solutions.

Big Data Life-cycle

- Phase 7: Data Storage
  - Store data in designed and designated storage units.
  - Storage infrastructure can consist of storage area networks (SAN), network-attached storage (NAS), or direct access storage (DAS) formats.
- Phase 8: Data Integration
  - Integration of stored data to different systems for various purposes.
  - Integration of data lakes with a data warehouse or data marts.
- Phase 9: Data Analytics and Visualization
  - Integrated data can be useful and productive for data analytics and visualization.
  - Business value is gained in this phase.

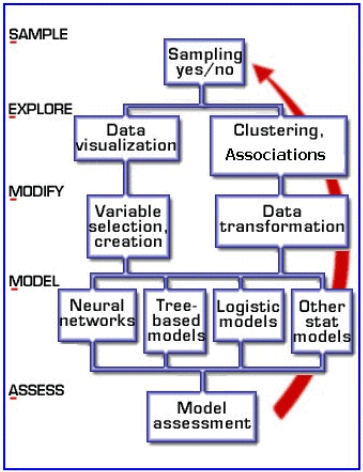
Big Data Life-cycle

- Phase 10: Data Consumption
  - Data is turned into information ready for consumption by the internal or external users, including customers of the business organization.
  - Data consumption require architectural input for policies, rules, regulations, principles, and guidelines.
- Phase 11: Retention, Backup, and Archival
  - Use established data backup strategies, techniques, methods, and tools.
  - Identify, document, and obtain approval for the retention, backup, and archival decisions.
- Phase 12: Data Destruction
  - There may be regulatory requirements to destruct a particular type of data after a certain amount of times.
  - Confirm the destruction requirements with the data governance team in business organizations.

Table of Contents

SEMMA

- SAS Institute
- Sample, Explore, Modify, Model, Assess
- 5 stages



SEMMA Stages

1. Sample
   - Sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.
   - Optional stage
2. Explore
   - Exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
3. Modify
   - Modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

SEMMA Stages

1. Model
   - Modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
2. Assess
   - Assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

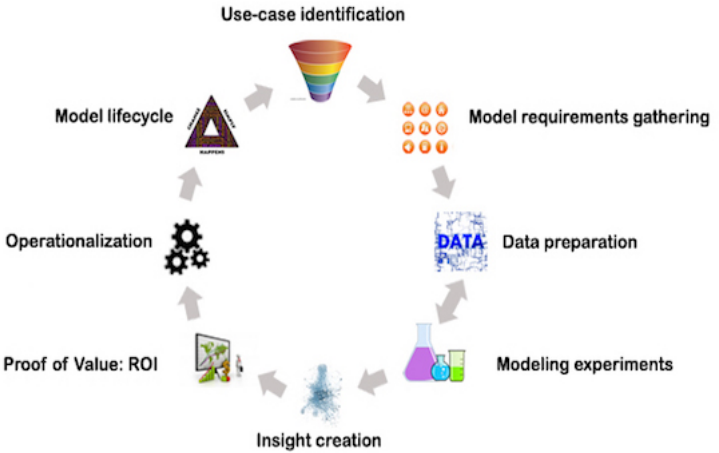
SEMMA

- “SEMMA is not a data mining methodology but rather a logical organization of the functional tool set of SAS Enterprise Miner for carrying out the core tasks of data mining.
- Enterprise Miner can be used as part of any iterative data mining methodology adopted by the client. Naturally steps such as formulating a well defined business or research problem and assembling quality representative data sources are critical to the overall success of any data mining project.
- SEMMA is focused on the model development aspects of data mining.”

Table of Contents

SMAM

- Standard Methodology for Analytics Models



SMAM Phases

Phase	Description
Use-case identification	Selection of the ideal approach from a list of candidates
Model requirements gathering	Understanding the conditions required for the model to function
Data preparation	Getting the data ready for the modeling
Modeling experiments	Scientific experimentation to solve the business question
Insight creation	Visualization and dash-boarding to provide insight
Proof of Value: ROI	Running the model in a small scale setting to prove the value
Operationalization	Embedding the analytical model in operational systems
Model life-cycle	Governance around model lifetime and refresh

Table of Contents

Challenges in Data Driven Decision-Making

- 1. Discrimination
  - Algorithmic discrimination can come from various sources.
  - Data used to train algorithms may have biases that lead to discriminatory decisions.
  - Discrimination may arise from the use of a particular algorithm.
  - Algorithms can result in discrimination as a result of misuse of certain models in different contexts.
  - Biased data can be used both as evidence for the training of algorithms and as evidence of their effectiveness.

Challenges in Data Driven Decision-Making

- 2. Lack of transparency
  - Transparency refers to the capacity to understand a computational model and therefore contribute to the attribution of responsibility for consequences derived from its use.
  - A model is transparent if a person can easily observe it and understand it.
  - Three types of opacity (i.e. lack of transparency) in algorithmic decisions
    - ★ Intentional opacity – The objective of this type of opacity is to protect the algorithm inventors' intellectual property.
    - ★ Knowledge opacity – This type of opacity is due to the fact that the most people lack the technical skills to understand how algorithms and computational models are constructed.
    - ★ Intrinsic opacity – This type of opacity arises from the nature of certain computer learning methods (e.g. deep learning models).

Challenges in Data Driven Decision-Making

- 3. Violation of privacy
  - Misuse of users' personal data and on data aggregation by entities such as data brokers, may have direct implications for people's privacy.
- 4. Digital literacy
  - Devote resources to digital and computer literacy programs from children to the elderly.
  - This enables the society to make decisions about technologies that we do not understand.
- 5. Fuzzy responsibility
  - As more and more decisions that affect millions of people are made automatically by algorithms, we must be clear about who is responsible for the consequences of these decisions. Transparency is often considered a fundamental factor in the clarity of attribution of responsibility.

Challenges in Data Driven Decision-Making

- 6. Lack of ethical frameworks
  - Algorithmic data-based decision-making processes generate important ethical dilemmas regarding what actions are appropriate in light of the inferences made by algorithms.
  - It is therefore essential that decisions be made in accordance with a clearly defined and accepted ethical framework.
  - There is no single method for introducing ethical principles into algorithmic decision processes.
- 7. Lack of diversity
  - Data-based algorithms and artificial intelligence techniques for decision-making have been developed by homogeneous groups of IT professionals.
  - Ensure that teams are diverse in terms of areas of knowledge as well as demographic factors

References

- https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projecthtml
- https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome
- https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en
- http://jesshampton.com/2011/02/16/semma-and-crisp-dm-data-mining-methodologies/
- https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html
- https://medium.com/illumination-curated/big-data-lifecycle-management-629dfe16b78d
- https://www.esadeknowledge.com/view/7-challenges-and-opportunities-in-data-based-decision-making-193560

Thank You