Life Cycle Phases of Data Analytics

The Data analytics lifecycle was designed to address Big Data problems and data science projects. The process is repeated to show the real projects. To address the specific demands for conducting analysis on Big Data, the step-by-step methodology is required to plan the various tasks associated with the acquisition, processing, analysis, and recycling of data.

Phase 1: Discovery

- The data science team is trained and researches the issue.
- Create context and gain understanding.
- Learn about the data sources that are needed and accessible to the project.
- The team comes up with an initial hypothesis, which can be later confirmed with evidence.

Phase 2: Data Preparation

- Methods to investigate the possibilities of pre-processing, analysing, and preparing data before analysis and modelling.
- It is required to have an analytic sandbox. The team performs, loads, and transforms to bring information to the data sandbox.
- Data preparation tasks can be repeated and not in a predetermined sequence.
- Some of the tools used commonly for this process include - Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning

- The team studies data to discover the connections between variables. Later, it selects the most significant variables as well as the most effective models.
- In this phase, the data science teams create data sets that can be used for training for testing, production, and training goals.
- The team builds and implements models based on the work completed in the modelling planning phase.
- Some of the tools used commonly for this stage are MATLAB and STASTICA.

Phase 4: Model Building

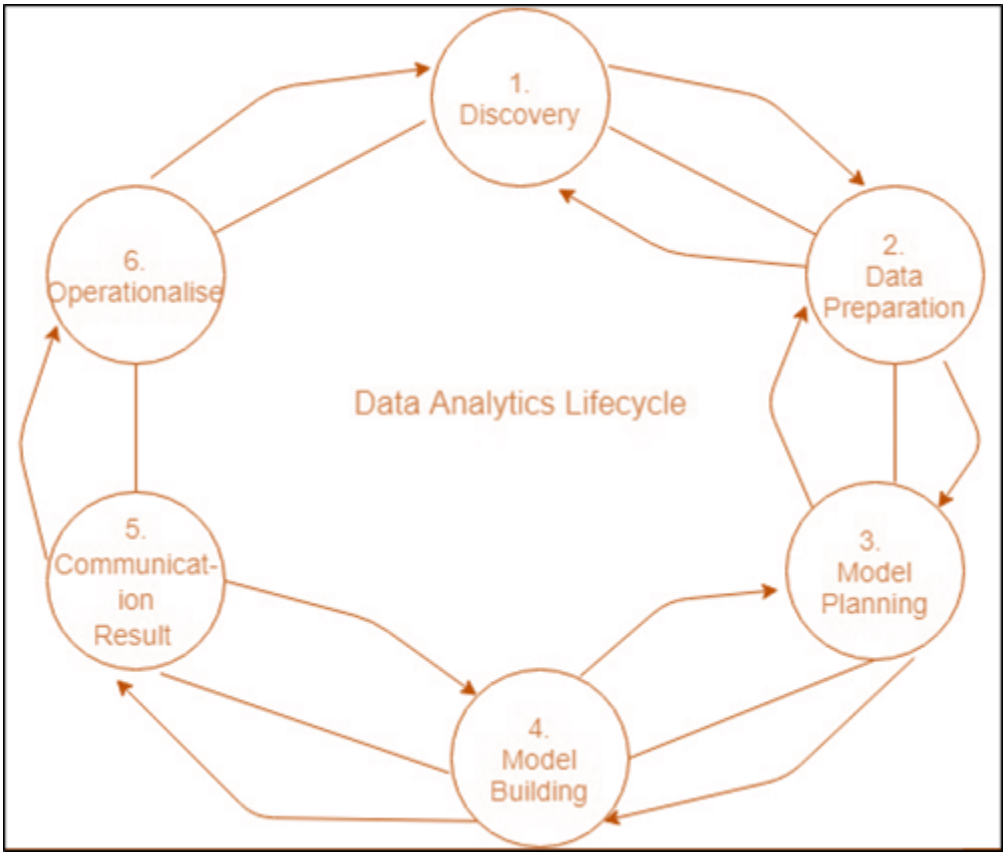
- The team creates datasets for training, testing as well as production use.
- The team is also evaluating whether its current tools are sufficient to run the models or if they require an even more robust environment to run models.
- Tools that are free or open-source or free tools Rand PL/R, Octave, WEKA.
- Commercial tools - MATLAB, STASTICA.

Phase 5: Communication Results

- Following the execution of the model, team members will need to evaluate the outcomes of the model to establish criteria for the success or failure of the model.
- The team is considering how best to present findings and outcomes to the various members of the team and other stakeholders while taking into consideration cautionary tales and assumptions.
- The team should determine the most important findings, quantify their value to the business and create a narrative to present findings and summarize them to all stakeholders.

Phase 6: Operationalize

- The team distributes the benefits of the project to a wider audience. It sets up a pilot project that will deploy the work in a controlled manner prior to expanding the project to the entire enterprise of users.
- This technique allows the team to gain insight into the performance and constraints related to the model within a production setting at a small scale and then make necessary adjustments before full deployment.
- The team produces the last reports, presentations, and codes.
- Open source or free tools such as WEKA, SQL, MADlib, and Octave.



Tools Used

OpenRefine

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Main features

- Faceting

Drill through large datasets using facets and apply operations on filtered views of your dataset.

- Clustering

Fix inconsistencies by merging similar values thanks to powerful heuristics.

- Reconciliation
  Match your dataset to external databases via reconciliation services.
- Infinite undo/redo

Rewind to any previous state of your dataset and replay your operation history on a new version of it.

- Privacy

Your data is cleaned on your machine, not in some dubious data laundering cloud.

- Wikibase

Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.

Apache Hadoop

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Alpine Miner

Alpine Data Labs Gives Oracle Businesses the Power of Predictive Analytics

Support for all data analytics operations, including exploring, transforming, predictive modeling, data mining, scoring, automated model fitting and automated model exporting operations.

An intuitive drag-and-drop interface that makes the predictive analytic process straightforward and accessible. Business users and business experts can work side-by-side with analytics experts or use the interface themselves.

The fastest end-to-end Big Data Predictive Analytics (BDPA) process. Alpine Miner embeds statistical algorithms in the database to leverage the innate capabilities of peta-scale parallel processing databases.

MATLAB and STASTICA

Statistics and Machine Learning Toolbox provides functions and apps to describe, analyze, and model data. You can use descriptive statistics, visualizations, and clustering for exploratory data analysis; fit probability distributions to data; generate random numbers for Monte Carlo simulations, and perform hypothesis tests. Regression and classification algorithms let you draw inferences from data and build predictive models either interactively, using the Classification and Regression Learner apps, or programmatically, using AutoML.For multidimensional data analysis and feature extraction, the toolbox provides principal component analysis (PCA), regularization, dimensionality reduction, and feature selection methods that let you identify variables with the best predictive power.

WEKA

An open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.

GNU Octave

- Powerful mathematics-oriented syntax with built-in 2D/3D plotting and visualization tools
- Free software, runs on GNU/Linux, macOS, BSD, and Microsoft Windows
- Drop-in compatible with many Matlab scripts

Apache MADlib

- Open source, commercially friendly Apache license
- For PostgreSQL and Greenplum Database
- Powerful machine learning, graph, statistics and analytics for data scientists

SQL

SQL analytics tools are software that enable advanced data analysis using SQL (Structured Query Language) in relational databases. These tools provide functionalities for complex querying, data manipulation, and analysis, often incorporating features for data visualization, reporting, and real-time analytics.