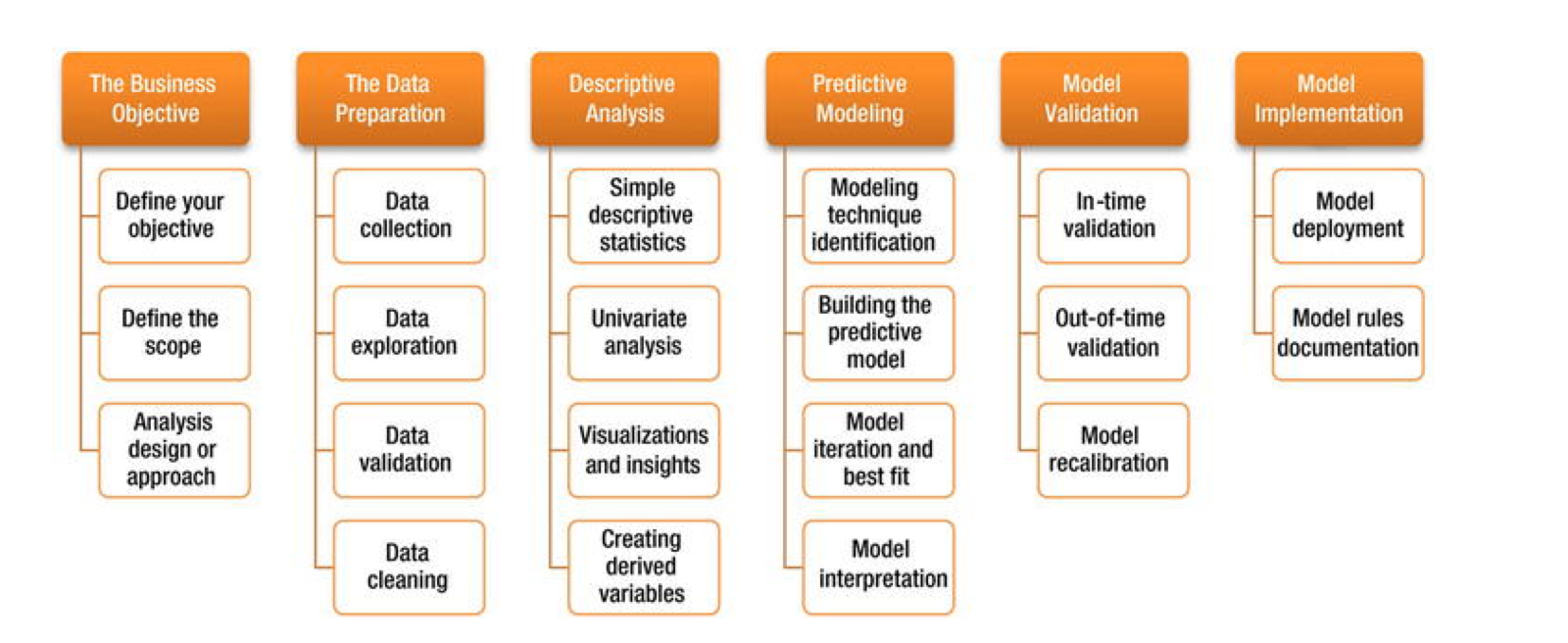
# An approach to Machine Learning and Data Analytics Lifecycle



[Dr. GP Pulipaka](https://medium.com/@gp_pulipaka)

Jul 25, 2016



Data Analytics lifecycle for Statistics, Machine Learning

**Statistical Machine Learning Data analysis life cycle**

During this stage a framework of statistics is explored for data collection, data cleansing, data collection, data validation, and data exploration. Descriptive analytic methods built with simple, univariate analysis, data visualizations, data insights, and derived variables. The predictive modeling is finalized by identifying number of methods and techniques, identifying the best fit model, interpretation of the model, and finally the model is deployed to the production during the operationalization phase of the data analytics project by involving all the stakeholders leveraging several machine learning, deep learning algorithms in R language for the corporations.

The growing complexity of the big data and the emerging technical landscape of connected data platform bring some complex challenges to the organization to support the executive decision-support systems. Several consulting firms adopt the best practices with their methodology and accelerators for implementing the data science projects leveraging data analytics lifecycle best practices. This brief paper will review some of the practices and approaches.

According to the definition of McKinsey, the corporations require modern big data architectures equipped with data analytics sandboxes to perform rapid prototyping solutions on sandboxes by processing the large volumes of data flowing through disparate data sources such as genomics, smart grids, geospatial location based data, radiological medical imaging, data from the sensors, telematics, and mobiles, and data from surveillance sources. Genomics play a vital role to predict the outbreaks of chronic diseases that can impact individuals through precision medicine and personalized medicine of big data by building a proactive healthcare system. Harnessing the power of big data goes through the data collection, data consolidation, data mining, machine learning, data visualization, business intelligence dashboards, and dissemination of the KPIs in the organization to derive the value from the volume, veracity, velocity, and variety of the big data. A number of analytic tools leveraging data marts, spreadsheets, analytic sandboxes for prototyping, and data warehouses aid the data analysis on big data in an enterprise.

**BI and data science practices**

BI analytics life cycle and data science life cycle differ in the implementation approach. The business intelligence analytics lifecycle provides dashboards for measuring the key performance indicators of the organization to meet the yearly targets in measuring the business performance of the enterprise. However, some of the data sets for business intelligence could be of large-scale. Extraction of the data from the corporate data warehouse from various data sources occurs in building the business intelligence reports. Business intelligence can narrow down the problems in the enterprise with a solution-based approach. The data science life cycle involves developing predictive forecasting models leveraging methods such as time series analysis for advanced planning and optimization of the organization with the aid of statistical framework with what if analysis.

The data analytics encompasses six phases that are data discovery, data aggregation, planning of the data models, data model execution, communication of the results, and operationalization.These six phases of data analytics lifecycle are iterative with backward and forward and sometimes overlapping movement. The key stakeholders of the data science project perform various roles such as a business analyst, business intelligence analyst, data engineer, administrator of the database, project manager, an executive project sponsor, and a data scientist. The business user or business analyst can define the metrics and results from the data science project implementation. The business analyst can be involved from the stage of defining the value of the data initiative. The project sponsor identifies the business conundrum and gets involved from the requirements gathering stage. The project manager ensures there is quality in the deliverables of the final data product and ensures to deliver the project on-time and on-budget leveraging all the resources on the project. The business intelligence analyst is the expert stakeholder in building and defining the dashboards and key performance indicators of the organization. The database administrator configures the database and provisions the services to the data analytics team including granting the authorizations. The data engineer is the subject matter expert with the SQL and NoSQL queries for data ingestion and processing.

The lifecycle of the data analytics lays out the framework for best practices from the genesis of the project to the completion of the project. The framework was established after surveying a large number of data scientists that make up the data science team. Some of the approaches considered for building the data analytics lifecycle framework best practices are, a) scientific method; b) CRISP-DM; c) DELTA framework; d) Applied information economics approach; e) MAD skills.

**Discovery**

During the phase of data discovery, the stakeholders constantly analyze the business trends, similar data analytics case studies, and domain of the business industry. An assessment is done on the in-house resources, the in-house infrastructure, and technology. Once the evaluation is complete, the stakeholders begin to build the hypothesis for resolving the critical business conundrums keeping in view of the customer experience and market.

**Data preparation**

Once the discovery is complete with a walkthrough of the business models, metrics, and results, the data aggregation occurs during the preparation phase by transforming the data from legacy system to the target data analytics platform into the sandbox for prototyping. Most of the stakeholders involve during this stage to pave the way for processing and conditioning the data for preliminary results from the sandbox.

**Model planning**

The data science team including the stakeholders builds the framework for model building by determining the techniques and methods particularly addressing the business problem. At this stage, the business problem is clearly emphasized and distributed across the departments breaking the data silos. Introducing scientific methods and corresponding key parameter variables are chosen at this stage to solve the business problem. Without the discovery phase and data preparation, a premeditated selection of scientific method will not apply to the business problem.

**Communicate results**

Communication is another vital methodology involving all the stakeholders to build a data-driven organization by infusing the data analytics culture into all the departments. The stakeholders of data science project summarize the lessons learned from the project, measuring the metrics of the project matching the definition of success introduced during discovery phase.

**Operationalization**

The stakeholders operationalize the results from the sandbox and plan to deploy the data science project into quality and subsequently operationalize into the production environment. The stakeholders of the project team document the code built, technical specifications, and functional specifications, data flow diagrams, data architecture models from the sandbox environment and document the results into a common document repository database such as SharePoint or eRoom or any other relevant document platforms.

**Dell Data models for data analytics lifecycle phases**

Big data analytics provides key answers to the business conundrums by extracting the value from velocity, veracity, velocity, and a variety of the data. The analytics solutions find a way to identify the correlation between the patterns of the data leveraging in-memory computing platforms, HDFS storage systems with distributed parallel computing. The stakeholders of the organization get involved throughout all the stages of the data analytics lifecycle in the organization that combines business and information technology partners to extract the value from the big data. Both business and IT stakeholders should work collaboratively to define the success of the data analytics project in the initial stage of the data discovery to ensure the technical methods employed address the business problems. Defining a scientific method for big data technology is like putting the cart before the horse. From each phase of the data lifecycle and successful metrics defined for the project, the stakeholders should wisely select the tools and techniques required for big data analytics platform deployment. Building data maturity models in the organization require building the trust in the organization, which is a foundation for the success of the data analytics project. The element of trust allows the stakeholders to perform the *what, where, and how*descriptive, prescriptive, and predictive analytics throughout the lifecycle of the data analytics.

**Data aware**

Most of the organizations under this data model create excel spreadsheets from various legacy systems and compile it for meeting the reporting. The data analytics are scattered across the organization with several BI and. The objective of the organization in this data model is to deliver a standardized reporting.

**Data proficient**

The stakeholders in the organization begin to track the KPIs of the organization and question the integration of applications and the data warehousing solutions. The business and the information technology units align with each other to build a prototype initiative for the data analytics platform. It is highly recommended for the executives of the organization needs to be involved at this stage to fund the data initiative and set up a center of excellence for data analytics. The data analytics competency center will expand the capabilities to leverage both structured and unstructured data to create innovatory solutions for the organization through the prototyping solutions.

**Data savvy**

In this phase, the organizations have already started leveraging the value from the data and the value becomes the inputs to support executive decision-support systems of the organization. The executive sponsor is the stakeholder for the organization and expands the capabilities of data analytics center to all the units of the business. The firm leverages the value from the data as a key differentiator from the rest of the organizations. The stakeholders start the focus on several other initiatives to enrich the business capabilities by solving the problems through data warehouses, data lakes, statistical framework, data mining, predictive analytics, and text mining on a day-to-day basis.

**Data driven**

The only way the stakeholders in the organization will make decisions is through the data. None of the decisions in the organizations are taken without the data. The data strategy expands globally as a scalable solution to meet all the regions. Machine learning is heavily leveraged in all business processes and all the business and technological units of the organization leverage the forecasting models and perform the analysis from the social media sentiment to position strategically within the ecosystem of businesses operating at scale.

**References**

EMC Education (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data* (1 ed.). Hoboken, New Jersey: Wiley.

Kadre, S. (2015). *Practical Business Analytics Using SAS: A Hands-on Guide*. New York City, New York: Apress.

Onis, T. D. (2016). The Four Stages of the Data Maturity Model. Retrieved July 5, 2016, from <http://www.cio.com/article/3077871/big-data/the-four-stages-of-the-data-maturity-model.html>