Data Engineering: using Data Analysis Techniques in Producing Data Driven Products

V. I. Nnebedum
Department of Electrical and Electronics
University of Port Harcourt, Nigeria.

ABSTRACT
Data analysis is prominent in data science researches, but by each day data usage is expanding, and in recent times the usage is becoming indispensable and inseparable in all works of life including engineering profession. This is why data engineering as a discipline sprang up - using the data analysis techniques from statistics, machine learning, pattern recognition or neural networks, together with other technologies such as visualization, optimization, database systems, knowledge discovery etc to produce systems needed in diverse business, science and social science domains. This paper is a novel presentation of data analysis and data engineering discipline, focusing on critical issues that are relevant to both, but divulging more the new trend of moving data science beyond data analysis, to data engineering. Data engineering is a multi-disciplinary field with applications in control, decision theory, and in the emerging areas like bioinformatics. Data engineering is needed in critical activities for business, engineering, and scientific organizations, since service oriented architecture and web services has moved into full swing.

Keywords
Data Science, Data Engineering, Data analysis, Data pipelines, Data infrastructure

1. INTRODUCTION
Data analysis is a discipline that existed before data engineering [1]. It is rooted in statistics, which has a pretty long history marked in ancient Egypt, when Egypt was taking a periodic census for building pyramids. Data analysis is a process that begins with retrieving data from various sources, and then analyzing it with the goal of discovering beneficial information. It played important role for governments all across the world, for the creation of censuses and taxation, which were used for various governmental planning activities. The involvement of computers and subsequent advances in computing technology dramatically enhanced what we can do with data analysis. As the collected data size gets larger, new methods of data analysis have been introduced in each stage, out of necessity. Data analysis has played major role in the success of many data-driven or Internet companies like Google, Facebook, LinkedIn, and Amazon. The companies have made their marks by using data creatively – for smart search results, targeted advertisement, list of possible acquaintances etc.

Nonetheless, with this advent of data engineering, it became known that data analysis is just a tool for data engineering. While data analysis implements its process with the focus of building data models, validating and testing data, developing algorithms, exposing knowledge of statistics, machine learning etc, data engineering, with a wider exposition of data science, goes deeper into steps like data pipelines, platform management, productionalized algorithms, scripting languages etc. [2]. A well analyzed data enhances the output of data engineering. For the fact that data analysis processes are applicable in data engineering, some people still interchange the two professions, but the difference is very clear.

In addition to data analysis, other aspects of data studies such as data governance, database management system, data security, data quality, reference and master data management, data warehousing and business intelligence, document, record and content management, meta data management and contact data management, are now useful facets of data engineering. This article x-rays these two disciplines before explaining why the knowledge is growing and tilting more in Data Engineering.

2. DATA ANALYSIS
Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in business, science, and social science domains. There is no single way of conducting data analysis. Different projects need different approaches. Data Analysis greatly uses statistical and logical techniques in describing and evaluating data [3]. One thing engineers are trying to avoid is allowing statistical manipulations overshadow the engineering principle of solving the target problems.

To get a good result, it is crucial to thoroughly investigate the context behind the project, then come out with a proper plan and design methodology. Another essential component is to ensure that data integrity and accurate/appropriate analysis of the research findings are done. Every field of study has developed its acceptable practices for data analysis [4]. The practice is based on the nature of the variables used (e.g. quantitative, qualitative or comparative) and the population from which the data are drawn (e.g. random distribution, independence, sample size, etc.).

A good data analysis generally need to:

- start with a solid plan of the study,
- have a valid set of data,
- conduct a valid data analysis,
- apply appropriate statistical procedures and,
- present a correct interpretation of the results.

2.1 Data analysis steps
A data analysis has steps for carrying out the analysis of a project. This is illustrated with the model shown in figure 1

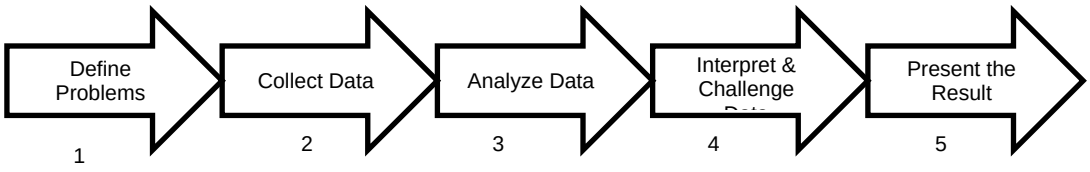

Figure 1: Data Analysis Model

As mentioned earlier, there is no cast rule on how to carry out data analysis. The amount of effort put in each of the steps depends on the type of data analysis intended and the data or problem being investigated. Generally, the steps revolve around:

- Defining or articulating problems and objectives - formulating hypothesis etc.
- Defining workflow and dataset.
- Collecting the data.
- Cleaning the data.
- Analyzing the data.
- Interpreting and challenging the results, and
- Presenting results.

The real analysis is done in step 3 and that is where statistical analysis is done [5]. The actions are to:

- Select the variables and model
- Perform preliminary analyses e.g. using graphs, tables, correlation and stepwise regression analyses etc.
- Build the model e.g. using analysis of variance
- Check the model
- Extract the equation

There are many data analysis tools that perform variety of data analysis tasks mentioned above. They are available open source and in the market. Some require no programming, while others need a combined code, visuals, and text in the same workflow. The technologies will help to apply analytics to data sets and enable users with the task of interpreting results. **SaaS** startups (**DataHero, DataCracker,** and **Statwing**) make it easy to perform simple data wrangling, visual analysis, and statistical analysis. **BigML** and **Datameer's Smart Analytics** make it easy for business users to apply machine-learning algorithms to massive data sets. **MADlib** is an open source library of scalable analytic functions for doing clustering, topic modeling, statistics, and many other tasks. For code, text and graph, **IPython** is popular among data scientists who use the Python programming language for analysis projects. In fact there are a lot of software packages; some are simple and generic like, **SAS, SPSS, Excel, Minitab, Gapminder** etc.

3. DATA ENGINEERING

As stated above, the development and implementation of data analysis products is gradually becoming the responsibility of the **data engineer** than **data analysts**. The data engineering tools have evolved tremendously over the past decade, with incredible amount of collaboration taking place through open source projects. It found its applications in control and decision theory and is needed in activities for business, engineering, and scientific organizations as service oriented architecture and web services.

Hilary Mason, a data scientist and one of the recent researchers in data engineering simply states that data engineering is when the architecture of a system is dependent on the characteristics of the data flowing through that system [8]. Very useful for the 'engineering' design and development of system architecture, data engineering is becoming a necessity in several important and diverse application domains such as geographic information systems, healthcare, fundamental sciences, business and finance.

The goal of the data engineering is to use the available data or even generate more data, to develop tools, and create software system, following structured and well defined protocol of data management - covering conceptual modeling and database design, data models, query languages, query processing and optimization indexing and many more. Good data engineering practice requires both the ability to manipulate data and to understanding the analytic purposes to which the data are going to be used [2].

Data engineering is closely related to Information engineering, Knowledge engineering, Information management and Knowledge management. With same principles as in data analysis, data engineering starts with understanding the problems to be solved, specifying how and where data is to be acquired and managed, formatting the incoming data and specifying how the data will be stored and retrieved.

3.1 Development of Data Engineering

Various individuals performing the role of data engineering have involved themselves in various field of knowledge such as pattern recognition and applied machine learning researchers, decision analysts, statisticians, neural network researchers and so on. Some areas such as data mining and knowledge discovery, exploratory data analysis, intelligent data analysis, seems to have gained much greater knowledge, has performed similar tasks as data engineering [2]. This inter-discipline nature made it is very difficult to point out specifically the role of data engineering. The difference in the disciplines ranges from their origins, the applications they serve, to the algorithms for data analysis they use [4].

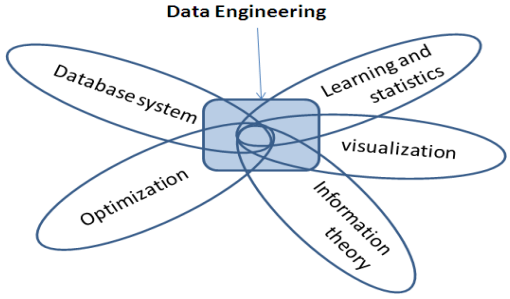

Figure 2: The component disciplines of data engineering

Figure 2 shows the multi-disciplinary nature of data engineering comprising of learning and statistics, database system, visualization, optimization and information theory.

This multi-disciplinary nature slowed the development of the discipline because the individual components lacked cross disciplinary training. In business management for instance, engineers and scientists saw no need to introduce expert knowledge from their field to address key problems. This made individuals fields traditionally develop their own data analysis tools specially suited for their own needs, and do not often call the computer scientists or statisticians for help.

3.2 Aspects of data engineering

The component facets, in data engineering are as follows:

1. Representation and manipulation of data: Conceptual data models, Knowledge representation techniques and Data manipulation languages and techniques.
2. Architectures of data, expert systems: New architectures for data expert systems, design and implementation techniques, languages and user interfaces.
3. Construction of data: Data design methodologies and tools, data acquisition methods, integrity/security/ maintenance issues.
4. Applications and management issues: Data administration issues, data engineering practice, office and engineering applications.
5. Tools for specifying and developing data using tools based on Linguistics or Human Machine Interface principles.
6. Communication aspects involved in implementing, designing and using a specified method.

2.2 Data Engineering Model

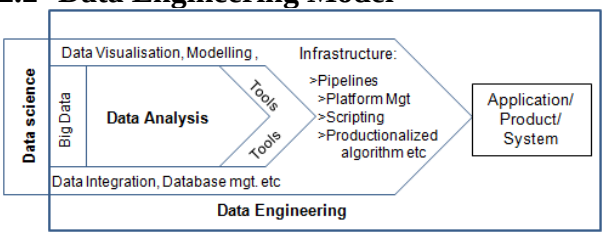

Figure 3: The data engineering model

The data engineering model above pictorially presents the moving or expanding of data science from data analysis to data engineering. This shows that **data scientist** builds data models, data visualization etc for validating and testing data, developing algorithms, with knowledge of statistics, machine learning etc, **data analyst** with deep domain knowledge in data generation and reporting, data exploration, pattern discovery, correlation and hypotheses testing etc, and **data engineer** will have all the two experiences but concentrate more on data pipelines, platform management, productionalized algorithms and scripting languages and infrastructure design [1] – generally called data engineering technology.

4. DATA ENGINEERING TECHNOLOGY

4.1 Data pipelines

Pipeline, in computing, is known as a set of data processing elements connected in series, so that the output of one element is the input of the next one. The elements of a pipeline are often executed in parallel or in time-sliced fashion and some amount of buffer storage is often inserted between elements.

Data pipeline therefore means the connection of steps from data acquisition and transformation to production deployment of the algorithm, all in series. Data pipeline is very necessary in big data management.

To bring this home, let's assume we have a unified university system holding students data. The student's multivariable information is collated through the point elements (e.g. course advisers, the department, the faculty/school, the University) before they are sent to the unified university system in a serial manner [5]. From data acquisition to data transformation the work is done in parallel at each of the point elements. A good data pipeline will reduces data redundancy, captures closer to real-time data, streamlines the data collection process and allows individual university to exchange information about students for instance

The following steps are carried out in a typical data pipeline building:

- Data acquisition – identifying the optimal data sources.
- Exploratory analysis and feature engineering -- using statistical techniques and visualizations to gain deep familiarity with the data.
- Data munging – cleansing and transforming the data to a form more appropriate for machine learning.
- Choose algorithm – selecting and using the correct model selection, choosing the appropriate algorithm for the problem and able to tune the algorithm parameters.
- Creating the training, cross validation, and test data sets and train the algorithm.
- Use cross validation to tune the algorithm further.
- Run the algorithm on the test set - seeing how the algorithm performs on new data.
- Ensembling - getting the best machine learning results from multiple algorithms or ensembles.
- Validation - seeing how the algorithm runs in a production environment.

Data pipelines that couple nicely with popular machine learning libraries and tools are still very few. Some of the tools available today are, workflow tools like **Oozie** and **Azkaban**, for the **Hadoop** environment. Many big data tools/technologies such as **Cassandra** and **Pig**, are open source, enabling organizations to experiment before making a large investment in the infrastructure technologies.

4.2 Platform management

A data engineering platform serves as a unifying platform for collecting, organizing, and activating first, second and third party data from any source, including online, offline, or mobile. A good data platform should have the ability to collect unstructured data set from mobile web and app, web analytic tools, CRM, point of sale, social, online video, and other available offline data sources. A data engineering product is built with a well-managed platform. Example of good unified technology platforms are **Hive, Solr, Hbase** and **Mahout.**

4.3 Scripting

Data engineering software applications are packaged using good scripting language. This is because data engineering programs are written for runtime environment that can

interpret and automate the tasks which could alternatively be executed one-by-one by a human operator. The big data environments associated with data engineering makes it mandatory for web application. Web design is automated through scripting web pages within a web browser. The spectrum of scripting languages ranges from very small and highly domain-specific languages to general-purpose programming languages are available open source. Example of dynamic high-level general-purpose scripting language, are **Perl** and **Python.**

5. CONCLUSION

Researches in data studies have grown extensively since the last decade mainly because of tremendous data explosion, diverse sources of release, security, diverse configurations, complexity in the application domain, and for many other reasons. Started with the popular data analysis, now moving to data engineering, all are geared towards completing the life cycle of data. Data engineering as a discipline has played a prominent role in engineering, business and other aspects of the society.

Engineers generally use their knowledge of science and other appropriate experience or tacit knowledge to find suitable and potential solutions to problems. Data engineers, in that line, with the knowledge of data analysis and plus other aspects of data skills like data pipelines, platform management, productionalized algorithms, scripting languages etc, provide suitable and potential solutions to problems around diverse application domains such as GIS, healthcare, fundamental sciences, business and finance.

Data engineering as a discipline is still evolving. Many institutions of higher learning are yet to classify it in their programme. 'Distilling' it from other known courses is still ongoing. Hence with little contributions in this paper, the definition of the data engineering, in relation to data analysis, will be clearer to many and many research bodies.

6. REFERENCES

[1] Hector Cuesta (2013) Practical Data Analysis; Packt Publishing; ISBN: 978-1-78328-099-5

[2] Calvin Andrus, Jon Cook, Suresh Sood; 2016; "Data Science: An Introduction"; WikiBook (last modified on 1 November 2016): https://en.wikibooks.org/wiki/Data_Science:_An_Introduction

[3] Judd, Charles and, McCleland, Gary (1989). "Data Analysis". Harcourt Brace Jovanovich Publication ISBN 0-15-516765-0*;* http://en.wikipedia.org/wiki/Data_analysis

[4] Resnik, D. (2000). Statistics, ethics, and research: an agenda for educations and reform. Accountability in Research. 8: 163-88

[5] Chris Olsen, Roxy Peck, Jey L Devore; Introduction to Statistics and Data Analysis; Chegg Books EISBN-13: 9781305445963

[6] J. Scott Long, 2009, "The Workflow of Data Analysis Using Stata", Stata Press, ISBN-13:978-1-59718-047-4

[7] DJ Patil, Hilary Mason (2015) Data Driven Publisher: O'Reilly Media

[8] Brian Shive (2013), "Data Engineering", Technics Publications, LLC; ISBN-13: 978-1935504603

[9] Yupo Chan, John Talburt Terry M. Talley, 2010, Data Engineering: Mining, Information and Intelligence (International Series in Operations Research & Management Science) ISBN-13: 978-1441901750 ; Publisher: Springer

[10] Viktor Mayer-Schönberger, Kenneth Cukier (2013) "Big Data: A Revolution That Will Transform How We Live, Work, and Think", publisher- Eamon Dolan/Houghton Mifflin Harcourt; ISBN-13: 978-0544002692