From Data Management to Data Engineering

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

Kamalu, A.U
Department of Electrical and Electronics University of Port Harcourt, Nigeria

Dike, J.N
Department of Electrical and Electronics University of Port Harcourt, Nigeria

ABSTRACT

No doubt, data management is very important in the development and execution of plans and programs that control and enhance the value of data, but most recently, engineers and allied professionals are beginning to play bigger role stabilizing and utilizing data management skills in developing a new field of study in data science called data engineering. This paper is a tutorial presentation on data management and data engineering, focusing more on the critical issues that are relevant to both studies but justifying the recent paradigm move towards data engineering.

Key words

Data Management, Data Engineering, Data analysis, Data pipelines, Productionalized algorithms, Data platform.

1. INTRODUCTION

The distinctions between data management and data engineering are still not clear because major aspects of data management studies, including data analysis, data architecture and data modeling are embedded in data engineering and IT practices. Big-data technology is becoming indispensible even in all aspect of life. It is found that the total value of data is not delivered if certain aspects of engineering aspect like infrastructure or platform designs are not well handled. Hence data engineering begins to proffer methods of using data management tools (already existing) to build complete data driven process for actualizing viral business and engineering applications. The data engineer, with wider exposure of data science, thus goes deeper into steps like data pipelines, infrastructure and platform management, productionalized algorithms, scripting languages etc.

In recent researches and businesses, data management had been in focus, resulting to the development of many software tools used to enhance the value of data. Groups, research centers and various departments of higher institutions have developed and built curriculum on data management. But much recently, attentions are drifting from *managing* the data to now *engineering* the data. The differences are here discussed.

2. DATA MANAGEMENT

Data Management International, DAMA - Data Management Body of Knowledge (DAMA-DMBOK) in April 2009 published The DAMA Guide, under the guidance of DAMA-DMBOK Editorial Board by identifying data management field of study and the subtitled areas [9] to include:

- Data governance (data asset, data governance and data steward).
- Database management system (data maintenance, database administration and management).
- Data security (data access, data erasure, data privacy, data security).
- Data quality management (data cleansing, data integrity, data enrichment, data quality and data quality assurance).
- Reference and master data management (data integration, master data management and reference data).
- Data warehousing and Business Intelligence (business intelligence, data mart, data mining, data movement and data warehousing).
- Document, record and content management (document and record management).
- Meta data management (meta-data, discovery, publishing and registry).
- Contact data management (business continuity planning, marketing operations, customer data integration, Identity management, ERP and CRM software etc).

The Body (DAMA-DMBOK) then defined data management as "…. the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.” The study of data management is broad. It is well established in some Universities. Various professionals, including engineers are now into data management [5]. The identified data management titles and subtitles listed above are explained in [9] and [10].

Today, some bodies like IEEE - Computer Society [7] are beginning to popularize data engineering as a great research area. They have come together to provide a forum to standardize, professionalize, and otherwise. advance the discipline of enterprise architecture.

ISO 15926 Standard is developed for engineering data. It concerns with design and engineering project data and industrial catalogues data transfer between various information warehouses. The standard is following "all-to-all" integration principle and is designed to be used during the whole life cycle of any engineering project [8].

3. DATA ENGINEERING

Data technologies have evolved tremendously over the past decade, with an incredible amount of collaboration taking place through open source projects. Data engineering is a multi-disciplinary field with applications in control, decision theory, and even in the emerging areas like bioinformatics and medicine. Data engineering is needed in critical activities - for business, engineering, and scientific organizations, as the move to service oriented architecture and web services moves into full swing.

Hilary Mason, a data scientist and one of the recent researchers, in her work in data engineering [6], simply states that data engineering is when the architecture of a system is dependent on the characteristics of the data flowing through that system. This shows that data, or rather big-data, is becoming indispensible in engineering and IT world today [1]. That is why data engineering, very useful in the software engineering design and development of system architecture, is now a necessity in several important and diverse application domains such as Geographic Information Systems (GIS), healthcare, fundamental sciences, business and finance.

The goal of the data engineering is to use the available data or even generate more data, with the tools provided by data management and other discipline, to develop and create system products, following structured and well defined engineering protocol - 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

Data engineering is multi-discipline. Various individuals - engineers and non-engineers are performing the role of data engineering as shown in figure 2. They have considered themselves in various field of knowledge such as pattern recognition and applied machine learning researchers, decision analysts, statisticians, neural network researchers and so on. There are some other areas such as data mining and knowledge discovery (DMKD), exploratory data analysis, intelligent data analysis etc that seems to have gained greater knowledge of data and have performed similar tasks as data engineering [2]. This inter-discipline nature of data engineering 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].

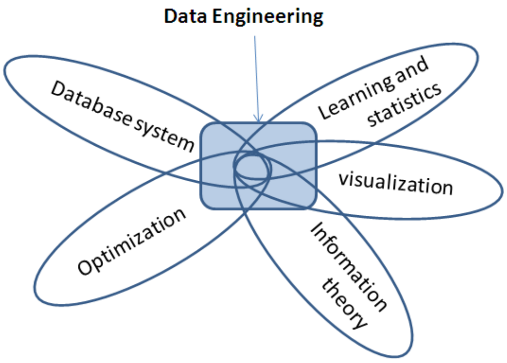

Figure 1: Some component disciplines of data engineering

This multi-disciplinary nature of data engineering slowed the development of the discipline (data engineering) 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 the individuals fields to traditionally develop their own data analysis tools specially suited for their own needs, and haven not often called in the computer scientists or statisticians for help. Nowadays that idea is being dropped. The introduction and better understanding of software engineering as a discipline played major role in this. The explosion of data in this millennium ‘forced’ many professionals to come together to think of how to manage and utilize the size characteristics. Data engineering is then the cohesion of various fields of study preferring solutions on how to “engineer” data.

3.2 Aspects of data engineering

To touch all the component facets in data engineering, the work will involves the following:

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.

3.3 Need for Data Engineering

The goal of the data engineer is to use the available data (or even generate more data), to develop tools, and create software system and applications, 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 so on. Thus data engineering principle is needful just to achieve a well coordinated and standard result toward solving business, industrial or societal problems.

The process involves, thorough working upfront to understand the nature of the data, transform the data as they are processed, then effectively focusing on the designing of the infrastructural platform.

3.4 Data engineering process

Before focusing on the infrastructural platform design plus other aspects, data engineering is involved in acquiring, ingesting, transforming, storing, and retrieving data. Data engineering starts with an understanding of the general nature of the problems to be solved, formulating data acquisition and management plan that specifies where the data are coming from, the format of the incoming data (text, numbers, images, video), and how the data will be stored and retrieved (file system, database management system).

While Data Scientist for instance is more involved in building data models, validating and testing data, developing algorithms, having knowledge of statistics, machine learning etc, Data engineer will have a broad experiences of data science, but concentrating more on such areas like data pipelines, platform management, productionalized algorithms, scripting languages etc.

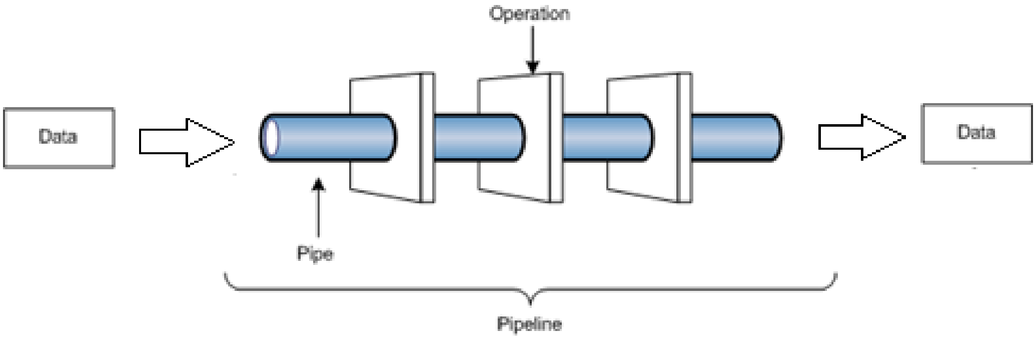

Figure 2: Data pipeline architecture

3.5 Data pipelines

A well planned and designed data pipeline is very crucial in achieving productionalized algorithms. Data pipelines are built to enable data **buffering** and enhance **parallelized** data execution, among other reasons.

Pipeline, in computing, contains a set of data processing elements (or operations) put 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. The operations are chunks of logic and pipes, as shown in figure 2, connect the operations with each other to allow smooth and fast data flow.

Individual operations are designed to do data processing, data transforming and data filtering, as algorithms. They are easily reusable, modifiable and free of boilerplate codes. Data engineering is involved in the production and deployment of the algorithms as well as coupling them with most data science popular machine learning tools. The time spent in building a pipeline in the first time can be considerable, but it is worth the effort since reusing the pipeline for future iterations will be much more streamlined. The steps stated below are typical for building data pipeline.

- 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.

3.6 Platform/Infrastructure Management

Proper platform design and development is essential in big data utilization. The key target of a data engineering platform is 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.

There are lots of sophisticated tools available for this. The choice and implementation of the tools is the responsibility of the data engineering team. The technologies have evolved tremendously over time, with an incredible amount of collaboration taking place through open source projects.

Some of the unified technology platforms [1] good for data engineering are:

- **Hadoop** - currently the most widely used open source framework for processing data. Hadoop is an implementation of the **MapReduce programming model** that is popularized by Google. Hadoop is inherently batch-oriented, and aimed at processing streaming data,

- **Kafka, Flume**, and **Scribe** - programs used for collecting streaming data from many sources; aggregate the data; and feed them to a database, a system like Hadoop.
- **Azkaban** and **Oozie** - good job schedulers. They manage and coordinate complex data flow.
- **Hive** and **Pig** - languages for querying large non-relational data stores. Hive is similar to SQL while Pig is a data-oriented scripting language.
- **Voldemort, Cassandra,** and **HBase** - designed for good performance on very large datasets.

3.7 Scripting

As data driven operations and virtual data stores are on the increase in the industries and business today, data engineering programs are thus written for runtime environment that can interpret and automate tasks. Data environments are automated through scripting. Scripting language supports the writing of scripts. Scripting include software applications, web pages and web browser. Scripting languages ranges from very small and highly domain-specific languages to general-purpose programming languages. Examples of such dynamic high-level general-purpose scripting language commonly used in data engineering are **Perl** and **Python.**

4. CONCLUSION

Data engineering is evolving but is becoming popular. While data management takes care of the development, execution and supervision of plans and programs that control, protect, deliver and enhance the value of data and information assets, Data engineering, with broad experiences in data science knowledge deeply put in more efforts in data pipelines, platform and infrastructural management, productionalized algorithms, scripting languages etc to produce systems that helps data-driven
Industries and business bit the test of time.

Research and development in data studies, has grown extensively since last decade mainly because of tremendous data explosion and diverse sources of data generation. Data engineering is now playing greater role in engineering, and is a necessity in several important and diverse application domains such as geographic information systems, healthcare, fundamental sciences, business and finance.

REFERENCES

[1] D.J Patil, 2011, Building Data Science Teams, O'Reilly Media, Inc., ISBN: 978-1-449-31623-5

[2] Data Science: An Introduction / Thinking Like a Data Engineer, WikiBooks: http://en.wikibooks.org/ wiki/Data_Science:_An_Introduction/Thinking_Like_a _Data_Engineer.

[3] W.L. Buntine *et al*, 1996, **What is Data Engineering,** Nikos Drakos Computer Based Learning Unit, University of Leeds.

[4] Ben Kneen, 2011, Data Management Part I: What Are Data Management Platforms? *Ad Ops Insider Blog:* http://www.adopsinsider.com/online-ad-measurement-tracking/data-management-platforms/what-are-data-management-platforms/

[5] W. Hoschek, *et al*, 2000, Data Management in an International Data-Grid Project. In *First IEEE/ACM Int'l Workshop on Grid Computing*, Bangalore, India, Dec 2000.

[6] Hilary Mason, 2013, Data Engineering; HilaryMason Blog: http://www.hilarymason.com /blog/data-engineering/

[7] IEEE Computer Society: http://www.ieee.org/membership_services/membership /societies/index.

[8] ISO 15926 Reference Data Engineering Methodology: http://techinvestlab.ru/files/ RefDataEngenEnglish/ RefDataEngen_ver_3_ English.doc

[9] Maureen Johnson, 2009, DAMA Guide to the Data Management Body of Knowledge (Cd-rom vers), Technics Publications LLC (US), ISBN 9780977140084.

[10] Keith Gordon, 2007, Principles of Data Management, BCS Learning & Development Limited UK, ISBN13: 9781902505848