https://doi.org/10.48009/1_iis_2019_157-167

Issues in Information Systems
Volume 20, Issue 1, pp. 157-167, 2019

PROPOSING A NEW GRADUATE DEGREE IN DATA ENGINEERING AND ANALYTICS

Kimberly W. Bartholomew, Utah Valley University, barthoki@uvu.edu
John E. Anderson, Utah Valley University, jandrerson@uvu.edu

ABSTRACT

The proposed program is a 30-credit hour program, which will be delivered as an asynchronously online, part-time program to allow students to complete coursework at their own pace making this program ideal for adult learners who work full time. It is intended for individuals who desire to be responsible for the creation and maintenance of analytics infrastructure. Data Engineers create data architectures and data sets used in data science and analytics. This program draws from constructs found in information systems, data analytics, database modeling and administration, data architecture, business intelligence, management, and security. The proposed curriculum is presented.

Keywords: Data Engineering, Data Engineer, Data Engineer Skill Set, Data Engineer Education

INTRODUCTION AND LITERATURE REVIEW

Data engineers are vital members of any data analytics team who not only need technical skills but communication skills to work across departments and understand what business leaders want to gain from the company's large data sets. A data engineer is mainly concerned with the tools and processes needed to make organizational data ready for data scientists and other members of the team so they can focus on finding new insights (Furbush, 2018). Data scientists used to (and often still) spend 70% of their time doing database and data wrangling work, while the data science algorithmic work comprised a minority of their time. The data engineer job was created to take this workload so that the data scientist could focus on data modeling. After the data scientist has created a new algorithm, the data engineer deploys the code into the production environment and integrates it with business processes.

While job listings may interchangeably use the title data engineering and data science, in most businesses the data engineer and data scientist are different jobs requiring a different skill set: data engineers being involved with constructing and managing the data pipeline architecture, and data scientists applying statistics, machine learning, mathematics, and programming in order to support business decision making. In small to medium sized organizations a data engineer may be asked to act as a data "generalist" responsible for every step of the data process from managing the data to providing analysis and reporting duties (White, 2018). In midsized companies, pipeline-focused data engineers work with data scientists to provide them use of the data collected. In larger organizations, managing the flow of the data from data warehouses and across multiple databases is the main focus of data engineers.

Data Engineers should have the following skills and knowledge (Goldman, n.d):

- Data Ingestion: getting data out of source systems and bringing it into the data lake or data warehouse
- Data Synchronization: detect changes in source data, merge and sync changed data from sources
- Data Transformation: integrate and transform data for specific business uses using SQL and support tools
- Data Governance: maintain the systems needed to ensure data access control needed for good governance
- Performance Optimization: build data pipelines that are efficient and scalable
- Production Organization: monitor the health and performance of data pipelines and ensure fault tolerance of the operational environment.

The area of data engineering has evolved from several mature job profiles:

- Data Engineer focuses on Database Engineering and Big Data Engineering (Data Engineer, Data Warehouse Specialist, Database Architect, Database Administrator, Systems Analyst, ETL Developer, SQL Database Administrator).

- Data engineering has a solid history as a field recognized by the Institute of Electrical and Electronics Engineers (IEEE) and since 1977, the IEEE has published the *Data Engineering Bulletin* as a quarterly journal that focuses on the design, implementation, modeling, and application of database systems and their related technology (Data Engineering Bulletin, n.d.). For 34 years, the IEEE International Conference on Data Engineering has addressed research issues related to designing, building, managing, and evaluating data-intensive systems (ICDE 34th IEEE International Conference, n.d.).

- Data engineers are experts at designing, building, and maintaining the data-based systems in support of an organization's analytical and transactional operations.

  "Engineering is at the heart of a data engineer's job," Jim Deters, CEO Galvanize (Woodie, 2018). Job titles of data engineer or a "big data" engineer are often used interchangeably. "A data engineer is the all-purpose everyman of a big data analytics operation, working between downstream analysts on the one hand, and upstream data scientists on the other…They're called on to ensure that data pipelines are scalable, repeatable, and secure, and can serve multiple constituents in the enterprise." (Woodie, 2014).

LABOR MARKET DEMAND

According to the Bureau of Labor and Statistics, employment in computer and information technology occupations are projected to grow 13% faster than the average for all occupations through 2026 (U.S. Bureau of Labor Statistics, n.d.). In the state of Utah, over 6,996 technology companies that employ over 143,000 people paying average salaries 88% higher than the average Metropolitan Statistical Areas (MSA) wage (CompTIA, 2019). This growth is expected to continue as Utah projects a 33% increase in Science, Technology, Engineering and Math (STEM) jobs by 2018 (Carnevale, Smith & Melton, 2018). Nationally, research suggests that higher-paying IT job openings at all levels, from entry to professional will continue to increase (U.S. Bureau of Labor Statistics, 2019).

According to Glassdoor's 2019 50 Best Jobs in America study, career paths that would strongly align with the proposed Professional Master's Degree of Data Engineering and Analytics: DevOps Engineer (#6), Data Engineer (#8), Business Analyst (#26), Business Development Manager (#29), Data Analyst (#15), Systems Engineer (#39), and Systems Administrator (#50) (Glassdoor, 2019). Nationwide, more than 90,000 positions currently available on Glassdoor's jobs site for a search of "Data Engineer" jobs. The fact that many of these positions include the terms 'engineer', 'manager' or 'administrator' indicates that this proposed graduate degree is responsive to demand in the marketplace for people who will be prepared to accept and fulfill roles requiring leadership and responsibility.

CURRICULUM

Method of Delivery

The proposed Master of Data Engineering and Analytics at Utah Valley University will be offered completely in the online format. Because many alumni of the Information Systems and Technology department, as well as Computer Science are currently working full-time in the field, an online graduate degree will allow them to commit to this program because of its flexible nature over the more traditional face-to-face programs. This proposed program will also be a part-time program to be completed over a two-year time frame that includes one course taken during the summer semester.

Courses
The program is similar to the Master in Information Systems with a Database Administration Concentration (DePaul University, 2019) at DePaul University, Chicago, IL, and also a Master in Big Data and Business Intelligence at the University of Greenwich, London, UK (University of Greenwich, 2019).

The proposed program includes courses in Relational Database Management Systems and Data Warehousing, Data Shaping & Integration, NoSQL Data Engineering Essentials, Data Engineering Cloud, Data Engineering Real-Time Streaming, Applied Data Analytics, Advanced Data Visualization Applications, Applied Big Data Analytics, and a Capstone Project. Integrated into course curriculum, will be high-demand tools and technologies for data engineering professionals (Glassdoor, n.d.) which may include: R, Python, Hadoop, Map reduce, Hive, Apache Spark, Kafka, Cassandra and other NoSQL databases, Cloud Computing, Docker, D3, Apache Pig, Tableau, GitHub, and SAS Enterprise Miner.

Since this is a new academic area, and we desired a more technical-applied focus, we developed the coursework primarily by looking at industry training and certification programs in data engineering. We specifically used the sources as shown in Table 1:

Table 1. Industry Training and Certification Programs in Data Engineering

Curricula Source	Curricula Findings
Data Engineering on Google Cloud Platform Specialization on Coursera https://www.coursera.org/specializations/gcp-data-machine-learning?utm_source=googlecloud&utm_medium=institutions&utm_campaign=GoogleCloud_Training_Data_ML_DE https://cloud.google.com/certification/data-engineer	
4 Courses:	build end-to-end data pipelines Tools: Tensorflow, Bigquery, Bigtable, Dataflow
• Google Cloud Platform Big Data and Machine Learning Fundamental	Extracting, Loading, Transforming, cleaning, and validating data Creating and maintaining machine learning and statistical models Querying datasets, visualizing query results and creating reports **Pre-requisites:** should have roughly one (1) year of experience with one or more of the following: • A common query language such as SQL • Extract, transform, load activities • Data modeling • Machine learning and/or statistics • Programming in Python
• Leveraging Unstructured Data with Cloud Dataproc on Google Cloud Platform	unstructured data using Spark and ML APIs **Pre-requisites:** Google Cloud Platform Big Data & Machine Learning Fundamentals (or equivalent experience) • Some knowledge of Python
• Serverless Data Analysis with Google BigQuery and Cloud Dataflow	extremely large datasets using Google BigQuery **Pre-requisites:** Google Cloud Platform Big Data and Machine Learning Fundamentals • Experience using a SQL-like query language to analyze data • Knowledge of either Python or Java
• Serverless Machine Learning with Tensorflow on Google Cloud Platform	machine learning models using Tensorflow and Cloud ML **Pre-requisites:** Completed Google Cloud Fundamentals- Big Data and Machine Learning course OR have equivalent experience ● Basic proficiency with common query language such as SQL ● Experience with data modeling, extract, transform, load activities ● Developing applications using a common programming language such Python ● Familiarity with Machine Learning and/or statistics

MCSA: Data Engineering with Azure	
Exam 70-775/Course 20775 Perform Data Engineering on Microsoft HD Insight	implement big data engineering workflows on HDInsight implement batch data processing, real-time processing, and interactive processing
Exam 70-776/Course 20776 Perform Big Data Engineering on Microsoft Cloud Services	design analytics solutions and build operationalized solutions on Azure experience in data engineering issues with Azure SQL Data Warehouse, Azure Data Lake, Azure Data Factory, and Azure Stream Analytics Design and Implement Complex Event Processing By Using Azure Stream Analytics Design and Implement Analytics by Using Azure Data Lake Design and Implement Azure SQL Data Warehouse Solutions Design and Implement Cloud-Based Integration by using Azure Data Factory Manage and Maintain Azure SQL Data Warehouse, Azure Data Lake, Azure Data Factory, and Azure Stream Analytics
Amazon AWS Certified Big Data – Specialty https://d1.awsstatic.com/training-and-certification/eligibilityupdates/AWS%20Certified%20Big%20Data%20-%20Specialty_Exam%20Guide_v1.3_FINAL.pdf https://aws.amazon.com/training/course-descriptions/bigdata-fundamentals/	
	Implement core AWS Big Data services according to basic architectural best practices Leverage tools to automate Data Analysis Pre-requisites: Recommended AWS Knowledge: 2 years' experience using AWS technology, AWS Security best practices, define AWS architecture and services and understand how they integrate with each other, define and architect AWS big data services and explain how they fit in the data lifecycle of collection, ingestion, storage, processing, and visualization. Recommended General IT Knowledge: 5 years' experience in a data analytics field, understand how to control access to secure data, understand the frameworks that underpin large scale distributed systems like Hadoop/Spark and MPP data warehouses, understand the tools and design platforms that allow processing of data from multiple heterogeneous sources with difference frequencies (batch/real-time), capable of designing a scalable and cost-effective architecture to process data
Domain 1: Collection	1.1 Determine the operational characteristics of the collection system 1.2 Select a collection system that handles the frequency of data change and type of data being ingested 1.3 Identify the properties that need to be enforced by the collection system: order, data structure, metadata, etc. 1.4 Explain the durability and availability characteristics for the collection approach
Domain 2: Storage	2.1 Determine and optimize the operational characteristics of the storage solution 2.2 Determine data access and retrieval patterns 2.3 Evaluate mechanisms for capture, update, and retrieval of catalog entries 2.4 Determine appropriate data structure and storage format
Domain 3: Processing	3.1 Identify the appropriate data processing technology for a given scenario 3.2 Determine how to design and architect the data processing solution 3.3 Determine the operational characteristics of the solution implemented
Domain 4: Analysis	4.1 Determine the tools and techniques required for analysis 4.2 Determine how to design and architect the analytical solution 4.3 Determine and optimize the operational characteristics of the Analysis

Domain 5: Visualization	5.1 Determine the appropriate techniques for delivering the results/output 5.2 Determine how to design and create the Visualization platform 5.3 Determine and optimize the operational characteristics of the Visualization system
Domain 6: Data Security	6.1 Determine encryption requirements and/or implementation technologies 6.2 Choose the appropriate technology to enforce data governance 6.3 Identify how to ensure data integrity 6.4 Evaluate regulatory requirements
Cloudera CCP Data Engineer	
CCP Data Engineer Exam (DE575)	Candidate given five to ten customer problems each with a unique, large data set, a CDH cluster, and four hours. For each problem, you must implement a technical solution with a high degree of precision that meets all the requirements. You may use any tool or combination of tools on the cluster (see list below) -- you get to pick the tool(s) that are right for the job. You must possess enough industry knowledge to analyze the problem and arrive at an optimal approach given the time allowed. You need to know what you should do and then do it on a live cluster under rigorous conditions, including a time limit and while being watched by a proctor.
Data Ingest	The skills to transfer data between external systems and your cluster. This includes the following: Import and export data between an external RDBMS and your cluster, including the ability to import specific subsets, change the delimiter and file format of imported data during ingest, and alter the data access pattern or privileges. Ingest real-time and near-real time (NRT) streaming data into HDFS, including the ability to distribute to multiple data sources and convert data on ingest from one format to another. Load data into and out of HDFS using the Hadoop File System (FS) commands.
Transform, Stage, Store	Convert a set of data values in a given format stored in HDFS into new data values and/or a new data format and write them into HDFS or Hive/HCatalog. This includes the following skills: Convert data from one file format to another Write your data with compression Convert data from one set of values to another (e.g., Lat/Long to Postal Address using an external library) Change the data format of values in a data set Purge bad records from a data set, e.g., null values Deduplication and merge data Denormalize data from multiple disparate data sets Evolve an Avro or Parquet schema Partition an existing data set according to one or more partition keys Tune data for optimal query performance
Data Analysis	Filter, sort, join, aggregate, and/or transform one or more data sets in a given format stored in HDFS to produce a specified result. All of these tasks may include reading from Parquet, Avro, JSON, delimited text, and natural language text. The queries will include complex data types (e.g., array, map, struct), the implementation of external libraries, partitioned data, compressed data, and require the use of metadata from Hive/HCatalog. Write a query to aggregate multiple rows of data Write a query to calculate aggregate statistics (e.g., average or sum) Write a query to filter data Write a query that produces ranked or sorted data Write a query that joins multiple data sets Read and/or create a Hive or an HCatalog table from existing data in HDFS

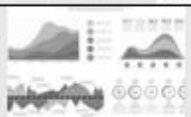
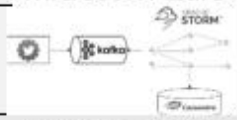
Workflow	The ability to create and execute various jobs and actions that move data towards greater value and use in a system. This includes the following skills: Create and execute a linear workflow with actions that include Hadoop jobs, Hive jobs, Pig jobs, custom actions, etc. Create and execute a branching workflow with actions that include Hadoop jobs, Hive jobs, Pig jobs, custom action, etc. Orchestrate a workflow to execute regularly at predefined times, including workflows that have data dependencies
Data Science Council of America – Associate Big Data Engineer https://www.dasca.org/data-science-certifications/associate-big-data-engineer	
	The ABDE™ by the Data Science Council of America (DASCA) 3rd–party, vendor–neutral certification Audience: Undergraduate Degree in Information Technology/ Computer Science OR a Diploma in Computer Programming/ Software Engineering from an accredited institution. Program Pre–requisites: strong formal exposure and knowledge of the basic concepts of programming, and ideally, are hands–on with the tools and techniques of object oriented programming – knowing Core JAVA and being able to understand SQL statements. They should also be ideally, aware of the basics of scripting languages like PERL or RUBY, and have a good understanding of the Linux and Unix environments. also comfortable with handling databases and spreadsheets. basic understanding and exposure to Big Data uses in business and general technology trends. Introduction to Data Science & Big Data 15% Storing and processing data in Hadoop 20% Decoding Sqoop and Flume 8% Yarn, Hive, and Pig 12% Decoding Machine Learning 8% Big Data Analytics and R 10% Integrating R and Hadoop 10% Social Media, Mobile and Big Data Solution Engineering 5% Big Data tools for Engineers 5% Essential Python 7%

Big Data for Data Engineers Specialization on Coursera https://www.coursera.org/specializations/big-data-engineering	
	four concise courses you will learn the basics of Hadoop, MapReduce, Spark, methods of offline data processing for warehousing, real-time data processing and large-scale machine learning. creating batch and real-time data processing pipelines, doing machine learning at scale, -deploying machine learning models into a production environment Join some of best Skills gained: Apache Hadoop, Recommender Systems, MapReduce, Apache Spark
Big Data Essentials: HDFS, MapReduce and Spark RDD	HDFS, MapReduce and Spark; - be guided both through systems internals and their applications; - learn about distributed file systems, why they exist and what function they serve; - grasp the MapReduce framework; - apply the framework to process texts and solve sample business cases; - learn about Spark, the next-generation computational framework; - build a strong understanding of Spark basic concepts; - develop skills to apply these tools to creating solutions in finance, social networks, telecommunications and many other fields
Big Data Analysis: Hive, Spark SQL, DataFrames and GraphFrames	Warehouse your data efficiently using Hive, Spark SQL and Spark DataFframes. - Work with large graphs, such as social graphs or networks. - Optimize your Spark applications for maximum performance. - Writing and executing Hive & Spark SQL queries; - Reasoning how the queries are translated into actual execution primitives (MapReduce or Spark transformations); - Organizing your data in Hive to optimize disk space usage and execution times; - Constructing Spark DataFrames and using them to write ad-hoc analytical jobs easily; - Processing large graphs with Spark GraphFrames; - Debugging, profiling and optimizing Spark application performance
Big Data Applications: Machine Learning at Scale	- Identify practical problems which can be solved with machine learning - Build, tune and apply linear models with Spark MLLib - Understand methods of text processing - Fit decision trees and boost them with ensemble learning - Construct your own recommender system
Big Data Applications: Real-Time Streaming	Streaming processing systems and NoSQL databases: Kafka, Cassandra and Redis. Pre-requisites: know Hadoop, SQL, bash, Python and Spark

Technical Skills for Data Engineers https://www.mastersindatascience.org/careers/data-engineer/	
	Statistical analysis and modeling Database architectures Hadoop-based technologies (e.g. MapReduce, Hive and Pig) SQL-based technologies (e.g. PostgreSQL and MySQL) NoSQL technologies (e.g. Cassandra and MongoDB) Data modeling tools (e.g. ERWin, Enterprise Architect and Visio) Python, C/C++ Java, Perl MatLab, SAS, R Data warehousing solutions Predictive modeling, NLP and text analysis Machine learning Data mining UNIX, Linux, Solaris and MS Windows
	Business Skills for Data Engineers **Creative Problem-Solving:** Approaching data organization challenges with a clear eye on what is important; employing the right approach/methods to make the maximum use of time and human resources. **Effective Collaboration:** Carefully listening to management, data scientists and data architects to establish their needs. **Intellectual Curiosity:** Exploring new territories and finding creative and unusual ways to solve data management problems. **Industry Knowledge:** Understanding the way your chosen industry functions and how data can be collected, analyzed and utilized; maintaining flexibility in the face of big data developments.

The program consists of five database/data engineering courses, three analytics courses, a course which brings together the data analytics in a big data environment, and capstone course consisting of a team project for an actual client. The proposed courses for the Data Engineering program are shown in Table 2.

Table 2. Master of Data Engineering courses and topics and tools

Curriculum Courses	Probable Topics and Tools
Classical Data Engineering: Relational Data and Warehousing 	RDBMS, Data Warehousing, ETL
Data Analytics in the Reporting Function 	Reporting / Research design / Descriptive analysis
Applied Analytics: Predictive & Prescriptive  	Predictive analysis / Prescriptive analysis / Network analysis / Collaborative filtering in R and Python
Advanced Data Visualization Applications 	Tableau / Javascript / D3 / Illustrator / R/ggplot2 / Highcharts / Visit
NoSQL Data Engineering Essentials 	Hadoop / Map reduce / Spark
Applied Big Data Analysis 	Hive, Spark SQL, Spark DataFrames and GraphFrames, Spark MLLib
Data Engineering: Cloud 	Cloud Storage / Ethereum Blockchain / Docker / CouchDB / OpenStack Swift / Apache Solr / BVLC Caffe / Nvidia Digits / Keras / IBM Watson / GATK / AWS / Azure / Google cloud
Data Engineering: Real-Time Streaming 	Kafka, Cassandra and Redis
Applications of Natural Language Processing with Deep Learning 	information extraction, machine translation, sentiment analysis, and summarization
Capstone 	Project scoping, planning and management / Data acquisition and analysis / Communication / Teamwork / Influence in organizations / Design thinking for data analytics

A wordcloud of frequent terms suggests which subjects are important to the data engineering role are shown in Figure 1.

Figure 1. Wordcloud of frequent Data Engineering terms

Conclusion

In this paper, we presented a proposal for a new graduate program in data engineering and analytics. We provided a definition of the job and role of a data engineer based on a literature review. We also looked at the demand for data engineers. Lastly, we presented a suggested curriculum for the proposed graduate program.

REFERENCES

Carnevale, A.P., Smith, N., & Melton, M. (2018). STEM state-level analysis projections of STEM jobs and education requirements through 2018. Georgetown University Center on Education and the Workforce. Retrieved from https://cew.georgetown.edu/wp-content/uploads/2014/11/stem-states-complete-update2.pdf

CompTIA. (2019.) Cyberstates 2019: The definitive guide to the U.S. tech industry and tech workforce, CompTIA Properties, LLC. Retrieved from https://www.cyberstates.org/pdf/CompTIA_Cyberstates_2019.pdf

Data Engineering Bulletin. (n.d.). IEEE Computer Society. Retrieved from https://tc.computer.org/tcde/data-engineering-bulletin/

DePaul University. (2019). Master of Science 2018-2019 Information Systems, Database Administration Concentration, College of Computing and Digital Media. Retrieved from https://www.cdm.depaul.edu/academics/Pages/current/Requirements-MS-IS-Database-Administration.aspx

Furbush, J. (2018). Data engineering: A quick and simple definition [blog]. O'Reilly Media. Retrieved from https://www.oreilly.com/ideas/data-engineering-a-quick-and-simple-definition

Glassdoor (2019). 50 best jobs in America for 2019. Glassdoor, Inc. Retrieved from https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Glassdoor (n.d.) Data engineer job description. Glassdoor, Inc. Retrieved from https://www.glassdoor.com/Job-Descriptions/Data-Engineer.htm

Goldman, T. (n.d.) What is Data Engineering? What skills are needed to be successful in Data Engineering? [blog]. Infoworks Agile Data Engineering. Retrieved from https://www.infoworks.io/data-engineering-skills-data-engineer-need/

ICDE 34th IEEE International Conference on Data Engineering. (n.d.) Retrieved from https://icde2018.org/

University of Greenwich (2019). Big data and business intelligence MSc, Department of Computing & Information Systems. Retrieved from https://www.gre.ac.uk/postgraduate-courses/ach/cgbdbi

U.S. Bureau of Labor Statistics. (2019). Economic news release: County employment and wages summary. Retrieved from: https://www.bls.gov/news.release/cewqtr.nr0.htm

U.S. Bureau of Labor Statistics. (n.d.) Occupational Outlook Handbook: Computer and Information Technology Occupations. Retrieved from https://www.bls.gov/ooh/computer-and-information-technology/home.htm

White, S.K. (2018). What is a data engineer? An analytics role in high demand [blog]. CIO, IDG Communications, Inc. Retrieved from https://www.cio.com/article/3292983/what-is-a-data-engineer.html

Woodie, A. (2014). Rise of the big data engineer [blog]. Datanami, Tabor Communications, Inc. Retrieved from https://www.datanami.com/2014/09/08/rise-big-data-engineer/

Woodie, A. (2018). Why 2018 will be the year of the data engineer [blog]. Datanami, Tabor Communications, Inc. Retrieved from https://www.datanami.com/2018/02/05/2018-will-year-data-engineer/