Big Data, Analytics and Business Intelligence: Recruiting for the New Age

Larry Gale, MS, CPHIMS, CISSP

What is meant by “Big Data”?

From Wikipedia:

- Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.
- The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

How does this impact Data Warehouses?

The TDWI report “Data Warehouse Modernization in the Age of Big Data Analytics” identifies some of the impacts of Data Warehouse Modernization:

- *server upgrades and tweaks for data models*
- *adding new platforms into the extended data warehouse environment (DWE)*
- *replacing the primary DW platform.*
- *using features previously untapped, such as:*
  - *in-memory and/or columnar databases*
  - *in-database analytics*
  - *real-time functions*
  - *data federation or virtualization.*

How does this impact Data Warehouses? cont'd.

Analytics, reporting, and data integration are also modernizing, and the DW is under pressure to provision data in ways that enable modern end-user practices such as:

- Visualization
- Advanced Analytics
- Natural Language Processing
- Data Preparation and Data Quality
- Self Service Data Access

The arrival of big data has made such provisioning more business critical—and more difficult.

Why does this create staffing problems?

TDWI surveyed Data Warehouse professionals, and found that most DW professionals have worked primarily with data that is relational or otherwise structured. Their skills and tool portfolios—very much tuned to relational data and technologies (e.g., SQL)—are currently being challenged by:

- *The diversification of data types and formats (non-relational, unstructured, social)*
- *The diversification of data sources (sensors, machines, GPS)*
- *The arrival of streaming data*

In organizations that are experiencing these forms of new big data, the data's unusual formats and sources are driving them to update both skills and portfolios of tools and data platforms.

How big is the problem?

The Aberdeen Group, in their report “The Modern Data Warehouse: Hybrid, Fluid and Formidable” states that as new data sources come online there arise challenges.

- *Companies are eager to exploit different forms of data (e.g. unstructured social media data, location-based data) but the infrastructure is often ill-equipped to house these disparate types and make them accessible.*
- *Along with new data sources comes human error and issues with quality and usability of data.*
- ***Looming large over all of these issues is the fundamental lack of IT manpower and experience to tackle all this data complexity***

What are the “new” job titles?

A variety of new job titles are emerging in industry in response to the rise of Big Data, Analytics, and Business Intelligence as strategic areas for the enterprise. Staffing firm Robert Half Technology provides job descriptions for a number of these new positions, including:

- Big Data Engineer
- Data Architect
- Data Modeler
- Data Scientist
- Business Intelligence Analyst

The following slides provide the descriptions and duties of those positions.

Big Data Engineer

Big data engineers communicate with business users and data scientists to understand the business objectives and translate those objectives into data-processing workflows. Big data engineers should have strong knowledge of statistics, extensive programming experience, ideally in Python or Java, and the ability to design and implement solutions for big data challenges. Knowledge and experience in data mining, processing large amounts of raw data, and designing and maintaining relational databases for storage and data acquisition are desired. Experience with NoSQL is preferred. This individual communicates directly with business users and data scientists to understand objectives and create data-processing workflows. Employers often require a bachelor's degree in a related field and four to six years of experience.

Typical duties include:

- Gathering and processing raw data and translating analyses
- Evaluating new data sources for acquisition and integration
- Designing and implementing relational databases for storage and processing
- Working directly with the engineering team to integrate data processing and business objectives

Data Architect

Candidates for data architect positions require a high level of analytical and creative skills, along with in-depth knowledge of data systems and database methodology, design and modeling. They must be able to communicate effectively in order to plan and coordinate data resources. Working knowledge of network management, distributed databases and processing, application architecture, and performance management is highly valued. Employers generally seek a bachelor's degree in computer science or a related field, as well as experience with Oracle, Microsoft SQL Server or other databases in various operating system environments such as Unix, Linux, Solaris and Microsoft Windows.

Typical duties include:

- Understanding and evaluating business requirements and translating them into specific database solutions
- Creating data design models, database architecture and data repository design
- Working with the systems and database administration staff to implement, coordinate and maintain enterprisewide data architecture
- Providing leadership in establishing and documenting data standards
- Creating and testing database prototypes

Data Modeler

Data modelers must possess excellent data analysis and problem-solving skills, and be able to both communicate effectively and work as part of a team. Employers normally request a bachelor's degree in computer science, IT or mathematics, in addition to several years of relevant data management experience. Candidates should be familiar with data modeling tools and methodologies and be knowledgeable in database system applications, stored procedures and data warehousing.

Typical duties include:

- Analyzing organizational data requirements and creating logical and physical models of data flow
- Interviewing key project stakeholders, documenting findings and making detailed recommendations
- Working with database administrators and reporting teams to ensure the availability of standard and ad hoc data reporting in a production environment
- Addressing data quality issues with clients and management

Data Scientist

Data scientists must have a range of mathematical and analytical skills, as well as business acumen. Big data scientists analyze and integrate multiple data sets and make recommendations based on their findings. Experience in programming languages – commonly Python or Java – is often required, as is a Ph.D.

Typical duties include:

- Gathering and processing raw data
- Providing analysis to leaders in order to support business decisions
- Developing metrics and prototypes that can be used to drive business decisions
- Identifying emerging trends and opportunities for business growth

Business Intelligence Analyst

Candidates for business intelligence analyst positions need a strong background in all aspects of database technology, with an emphasis on the use of analytical and reporting tools. Employers seek a bachelor's degree in computer science, information systems or engineering, as well as several years of experience with database queries, stored procedure writing, Online Analytical Processing (OLAP) and data cube technology. Excellent written and oral communication skills are a must.

Typical duties include:

- Designing and developing enterprise-wide data analysis and reporting solutions
- Reviewing and analyzing data from multiple internal and external sources
- Communicating analysis results and making recommendations to senior management
- Developing data cleansing rules

What do these jobs pay?

Robert Half Technology 2017 predicted national pay ranges:

Job Title	2017 Salary Range
Big Data Engineer	$135,000 - $ 196,000
Database Manager	$122,250 - $ 177,000
Database Developer	$108,000 - $ 161,500
Database Administrator	$ 98,500 - $ 148,500
Data Analyst/Report Writer	$ 77,500 - $ 118,750
Data Architect	$131,250 - $ 184,000
Data Modeler	$111,000 - $ 161,500
Data Scientist	$116,000 - $ 163,500
Data Warehouse Manager	$129,000 - $ 179,000
Data Warehouse Analyst	$107,500 - $ 155,750
Business Intelligence Analyst	$118,000 - $ 171,750

Where Do We Find Qualified Candidates?

None of these positions resemble "entry level" positions. They all require both very specific technical skills and/or education.

- "a bachelor's degree in computer science, information systems or engineering"
- "strong knowledge of statistics, extensive programming experience, ideally in Python or Java"
- "Experience in programming languages – commonly Python or Java – is required, as is a Ph.D"
- "Experience with NoSQL"

Where Do We Find Qualified Candidates?, cont'd.

They also require specific, relevant job experience:

- "Candidates should be familiar with data modeling tools and methodologies and be knowledgeable in database system applications, stored procedures and data warehousing"
- "Several years of experience with database queries, stored procedure writing, Online Analytical Processing (OLAP) and data cube technology"

Colleges & Universities

Most S.C. colleges and universities offer “traditional” computer science majors. I surveyed offerings at Clemson, USC, College of Charleston, the Citadel. Their offerings are all similar, including:

- A “Computer Science” and/or “Computer Engineering” major (which is very technically focused)
- “Computer Information Systems”, which requires less high-level math, and is more focused on business applications of computers.

Graduates of these programs should be qualified for entry-level positions in the Information Technology field, but generally lack the requisite specialized experience required for the positions we’re discussing.

Colleges & Universities, cont'd.

The College of Charleston has something unique: the first undergraduate "Data Science" major in the nation:

"At the College of Charleston, data science students are taught to scrape, process, organize, and analyze large data sets to identify patterns and trends. They learn to use the tools and problem-solving skills of mathematics and computer science to gather information from large, multidimensional data sets, data streams and complex systems."

The Bachelor of Science in Data Science is intended for students who wish to pursue professional positions in data mining, business intelligence, national security, genomics, drug informatics, and any other data-intensive field in which new knowledge is discovered through patterns in data.

Colleges & Universities, cont'd.

The College of Charleston Data Science core classes include:

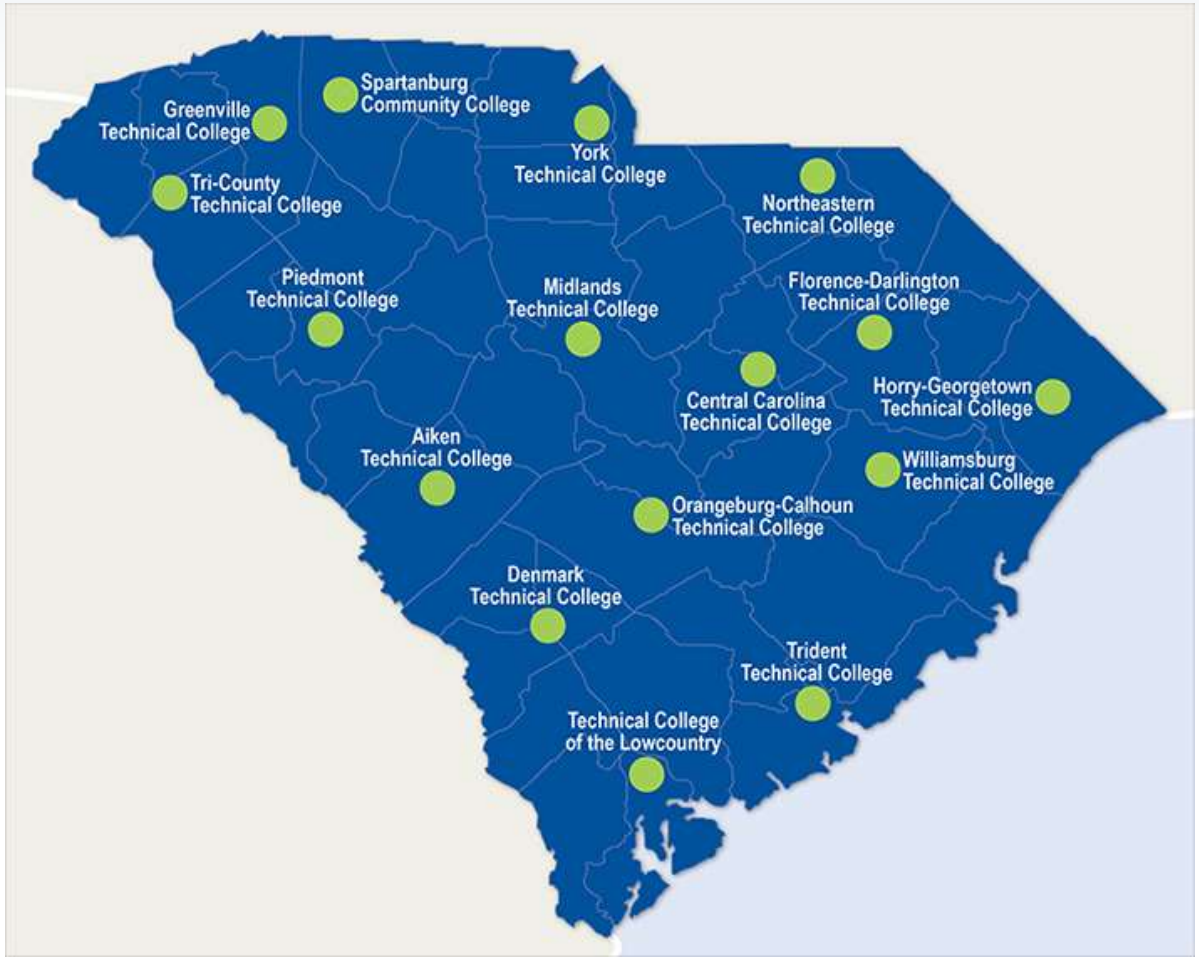
- Computer Programming
- Data Mining
- Artificial Intelligence
- Data Science
- Math (including Calculus and Linear Algebra)
- Statistics (9-10 hours).

Colleges & Universities, cont'd.

Students then choose one of 14 subject areas to specialize in, so when they graduate, they have actual subject area experience in data mining and data science for that area. These include:

- Accounting
- Biomechanics
- CRM
- e-Commerce
- Economics
- Exercise Physiology
- Finance
- Geoinformatics
- Molecular Biology
- Organismal Biology
- Physics and Astronomy
- Psychology
- Sociology
- Supply Chain Management

Technical Colleges



Technical Colleges, cont'd.

South Carolina's Technical College system is very responsive to requests from industry to develop training programs for high need areas.

- They focus on practical knowledge and hands-on training, rather than just academic study.
- They offer degree programs, such as Associate's Degrees, as well as non-degree Certificate programs.
- These Certificate programs may be valuable for helping existing employees keep their skills current and relevant.

Technical Colleges, cont'd.

Trident Technical College, in Charleston, for example, offers associate degrees in:

- Computer Programming
- Information Systems
- Network Systems Management
- Management – Business Information Systems

Technical Colleges, cont'd.

Trident Tech also offers Certificate programs in:

- Database Administration
- Internet Programming
- e-Commerce
- Cybersecurity
- Linux Systems Administration
- Virtualization and Cloud Computing
- Network Security
- Mobile Application Programming

Technical Colleges, cont'd.

The Technical Colleges are generally very responsive to the needs of local industry, so SC HIMSS could start to establish working relationships with those institutions in order to help develop courses of study that would allow both degree program students and individuals pursuing certificates and continuing education tracks to be prepared for these new and evolving positions.

What's missing from these degree programs?

The baccalaureate programs are relatively unchanged since the 1990's; at least, the catalogs read very similar to what I remember was being taught at that time:

- Introduction to Computer Science
- Algorithms and Data Structures
- Introduction to Machine Organization
- Foundations of Software Development
- Operating Systems
- Computer Networks

What's missing from these degree programs? Cont'd.

They occasionally offer "special topics" classes that address areas of interest, such as:

- Computer Forensics
- Bioinformatics
- Geographic Information Systems
- Databases
- Cybersecurity
- Natural Language Processing
- Artificial Intelligence

What's missing from these degree programs? Cont'd.

In order to help cultivate graduates who have at least some familiarity with these new and emerging disciplines, it may be worth having SC HIMSS representatives meet with these institutions to recommend, develop, or even offer to teach special topics classes in areas like:

- Big Data
- Hadoop
- NoSQL and Non-Relational Database Systems
- Advanced Analytics
- Data Mining

What about existing staff?

Current IT staff have many of the necessary technical skills and domain knowledge to work in these areas, but may need additional specialized course work or professional development opportunities to move into these positions.

Staff could be encouraged to pursue specialized coursework from colleges, as well as through private sector training entities.

Affordable options such as MOOCs (Massive Open Online Course) may also be very good options for existing employees.

MOOCs?

From Wikipedia:

A Massive Open Online Course (MOOC /muːk/) is an online course aimed at unlimited participation and open access via the web. In addition to traditional course materials such as filmed lectures, readings, and problem sets, many MOOCs provide interactive user forums to support community interactions among students, professors, and teaching assistants (TAs). MOOCs are a recent and widely researched development in distance education which were first introduced in 2008 and emerged as a popular mode of learning in 2012.

MOOCs, cont'd.

Generally flexible scheduling of on-line offerings, frequently very inexpensive, even free. Some high quality content providers participate. For example, coursera.org offers:

- A Big Data Specialization from UC San Diego
- A Data Science Specialization from Johns Hopkins
- Relational Database Support for Data Warehouse from University of Colorado System
- Data Visualization with Tableau from UC Davis

as well as many other course offerings that would allow current IT workers to enhance their skills in order to fill some of these new positions.

Final thoughts…

Data Warehouse Environments (DWEs) are rapidly evolving from traditional relational database technologies to environments incorporating non-structured data, using technologies such as Hadoop and NoSQL.

Analytics tools and techniques are changing to accommodate these new data types.

The technical skills required to support these technologies are substantial, and individuals with those skills are in short supply.

Final thoughts…

To fill these positions, you can:

- Use Contractors with the requisite technical skills and experience
- Hire new employees with the requisite technical skills and experience
- Hire new graduates and provide on-the-job training to bring them up to speed
- Leverage your current IT staff by providing the specific technical training to help them learn the new technologies

Final thoughts…

Additionally, relationships with institutions of higher learning should be leveraged to insure that new graduates are being produced with the necessary skill sets to eventually work in these interesting and necessary positions.

Questions?

Larry Gale, MS, CPHIMS, CISSP
Data Architect & Senior Systems Engineer
Office of the Chief Information Officer
Medical University of South Carolina
(843) 792-4806
galel@musc.edu