SKILLS AND KNOWLEDGE

Skills in:

Analytical Thinking
Applying Statistical Methods
Computational Thinking
Computer Programming (e.g., R, Python)
Critical Thinking
Data Decoding (e.g., UTF, ASCII)
Data Management
Data Manipulation
Data Security
Database Administration
Database Programming (e.g., DB, Query data tables)
Educating Clients and Customers
Internet Search Strategies
Intra-company Communications
Machine Learning
Manipulating Data Tables
Parallel Programming (e.g., MPI, Hadoop, MapReduce)
Problem Solving
Project Management
Relational Databases (e.g., Oracle, SQL)
Research Methods
Scanning Technical Literature
Scripting
Statistical Methods
Synthetic Thinking (Big Picture)
Time Management
Troubleshooting
Visualization Design
Working with Spreadsheets
Writing

Knowledge of:

Algorithms (e.g., machine learning, statistics)
Analytic Thinking
Best Practices
Brute Force Analytics
Communication
Computation
Concurrency
Critical Thinking
Data Modeling
Data Practices (e.g., HIPAA, SOX)
Data Security and Privacy
Data Standards
Data Structures
Databases (e.g., SQL, NoSQL)
Discrete Logic
Distributed Systems
Distributed Computing Methods (e.g., Hadoop, HANA)
Domain/Field Knowledge (i.e., deep & broad)
Math
Metadata Standards
Numerical Methods
Performance Metrics
Programming
Proper Use of Data (e.g., governance)
Rapidly Evolving Technology Landscape
Relational Algebra
Research Methodology
Resource Allocation
Scientific Method
Statistics
Unstructured Data (e.g., images, text)
Visualization

EQUIPMENT/TOOLS/SUPPLIES

Article Server/Search System (e.g., Google Scholar, Web of Knowledge)
Big Data Hardware (e.g., Clusters/Servers, Networking {Infininode, Fiberlink}, Cloud {AWS, Azure, etc.})
Collaborative Tools
Data Mining Tools
Data Security Software and Appliances
Data Warehouse (e.g., ETL Tools)
Databases (e.g., SQL, NoSQL)
Desktop Productivity (e.g., Word proc., spreadsheet, slide prog., e-mail)
File Systems (e.g., HDFS, GPFS)
Job Scheduler (e.g., HTCondor, GridEngine)
Knowledge Management Tools
Knowledge Networks
MapReduce (e.g., Hadoop, Sparc, YARN, KEPLER)
Operating Systems
Personal Hardware (e.g., Desktop PC, Laptop, Smartphone, tablet)
Programming Packages (e.g., Python, C#, Java)
Project Management Tools
Simulation Packages
Skype/GoToMeeting
Source Control Systems (e.g., SUN, Git)
Statistics Packages (e.g., R, Matlab, SAS)
Visualization and Analytics Software (e.g., D3, Tableu, Ayasdi)
Workflow Tools (e.g., Proficiency Builder Editor)

BEHAVIORS

A successful big-data-enabled specialist is…

A choreographer
A connector of domains/ideas
A data lover
A lifelong learner
A mentor
A multi-tasker
A problem solver
A risk taker
A seeker of patterns
A storyteller
A strategic thinker
A tinkerer
Collaborative
Curious
Detail oriented
Ethical
Flexible
Inclusive
Logical
Open-minded
Organized
Passionate
Patient
Respectful
Self-directed
Skeptical
Socially aware
Willing to question

TRENDS/CONCERNS

Accelerating data growth leads to fragmentation of ad hoc solutions
Big data field evolving from individual disciplines to trans-disciplinary melting pot
Demand for big-data-enabled specialists is rapidly increasing, while supply of individuals with those skills is not
Difficulty in discovering poorly collected data
Exponential growth rate of data
Fragmentation of practices and tools exceeds the capacity of training programs and workforce professional development
Growth of government involvement in organizational data practices
Increased need for real-time analytics for streaming data
Increased risk to data security due to security breaches
Industry tools stand in contrast to workforce skill needs
Insufficient bandwidth to curate and clean data
Insufficient workforce skilled in big data
Lack of access to electric power to run data centers
More complex statistical results/visualizations are increasingly present in media
Need for ethical safe harbor for data sharing
Proliferation of computing in developing nations creates new challenges
Proliferation of diverse policies on governing data security
Proliferation of practices and internal tools exceeds the capacity of training programs and workforce professional development
Public interest in data literacy is growing
Public understanding of data remains low
Rapid drop in cost, along with rapid rise in accountability and ubiquity of cloud computing
Rapid obsolescence of technology and tools
The big-data-enabled specialist is transitioning from a technical role to a business-driven role
The Internet of things creates more data than existing capacity can process
The role of the big-data-enabled specialist is not well defined in organizational culture

Five years from now …

Client base will move to smaller organizations using larger data sets to solve more sophisticated problems
Compute availability will be on the evening news
Continuous increase in data but deflation of the big data hype with a much greater focus on impact and ROI
Data will be collected at even greater scales, yet software/tools/methods will still lag behind
Data and analysis will be provided more efficiently and transparently using new technologies and methods
Development of global data retention standards (e.g., safe harbor, templates)
Increase in data-driven decision making
Increase of data will increase solvability of crimes
Less hype and frenzy, and more productivity
Shift from documents/PDF to interactive data methods and visualizations to ensure reproducibility
Using big data modeling and capture to change the mode of global tectonic studies from local cases to global monitoring

Panel

Kirk Borne
Professor of Astrophysics and Computational Science
George Mason University
Fairfax, Virginia

Randy Bucciarelli
Programmer/Analyst
Scripps Institution of Oceanography
UC San Diego
La Jolla, California

Tim Chadwick
Principal Engineer
Dynamic Network Services, Inc.
Manchester, New Hampshire

Benjamin Davison
Quantitative User Experience Researcher
Google
Boston, Massachusetts

Lucy Drotning
Associate Provost of Planning and Institutional Research
Columbia University
New York, New York

Ryan Kapaun
Law Enforcement Analyst
Eden Prairie Police Department
Eden Prairie, Minnesota

Juan Miguel Lavista Ferres
Principal Data Scientist
Bing/Microsoft
Seattle, Washington

Shannon McWeeney
Head of Division of Bioinformatics and Computational Biology
Oregon Health & Science University
Portland, Oregon

Jay Parker
Earth Scientist
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Steve Ross
Consultant on Data Quality Control
Corporate Editor
Broadband Communities Magazine
Revere, Massachusetts

Kartik Shah
Principal Consultant
Strategix Solutions
Toronto, Canada

Oceans of Data Institute

Ruth Krumhansl
Director

Profile Facilitators

Joseph Ippolito

Joyce Malyn-Smith

Suggested Citation:
Oceans of Data Institute. (2014). *Profile of a big-data-enabled specialist*. Waltham, MA: Education Development Center, Inc.

Profile of a Big-Data-Enabled Specialist

EDC OCEANS of DATA INSTITUTE

http://oceansofdata.org | Email: oceansofdata@edc.org

Learning Occupation: The big-data-enabled specialist is an individual who wrangles and analyzes large and/or complex data sets to enable new capabilities including discovery, decision support, and improved outcomes.

DUTIES	TASKS											
1. Defines the Problem	1A. Identifies stakeholders	1B. Determines stakeholders' needs	1C. Articulates question	1D. Aligns study to organizational goals and objectives	1E. Translates question into research plan	1F. Designs experiment	1G. Develops deep domain knowledge of data source	1H. Discerns whether big data is needed to solve problem	1I. Identifies resources (e.g., experts, software)	1J. Performs gap analysis	1K. Assesses risk and bias involved in conducting study/project	1L. Communicates cost/risks of study to stakeholders
	1M. Negotiates plan, including deadlines and budgets	1N. Creates requirement document (sign-off)										
2. Wrangles Data	2A. Performs data exploration	2B. Identifies data	2C. Creates data dictionary	2D. Collects data	2E. Assesses the extent/methods to clean the data	2F. Maps data across heterogeneous sources	2G. Identifies outliers and anomalies	2H. Cleans data	2I. Transforms data	2J. Synthesizes data	2K. Defines new metrics/ attributes based on accessible data	2L. Performs data visualization
	2M. Writes software to automate tasks	2N. Documents process										
3. Manages Data Resources	3A. Manages data life cycle	3B. Conducts capacity planning of resources	3C. Complies with legal obligations	3D. Applies ethical standards	3E. Identifies tools that may be needed for purchase or modification	3F. Protects data and results	3G. Determines access to data	3H. Designs ETL workflow	3I. Implements ETL workflow	3J. Stores data	3K. Upserts data sources	
4. Develops Methods and Tools	4A. Researches current methods/ models	4B. Extends existing methods/ models, if possible	4C. Selects tools/ software/ programming environment	4D. Develops new methods/models	4E. Runs simulations	4F. Iterates correctness and scalability of methods/models	4G. Validates methods/ models with test cases	4H. Disseminates methods/models for peer review	4I. Documents methods/models			
5. Analyzes Data	5A. Develops analysis plan	5B. Applies methods and tools	5C. Conducts exploratory analysis (e.g., identifies anomalies, outliers, bias in sampling; visualizes)	5D. Evaluates results of the analysis (e.g., significance, effect, size)	5E. Estimates precision and accuracy of answer	5F. Determines level of confidence in results	5G. Compares results with other findings	5H. Answers the question (e.g., insights drawn from results)	5I. Submits preliminary findings for peer review	5J. Documents preliminary findings		
6. Communicates Findings	6A. Selects documentation media (e.g., dashboard, PowerPoint, e-mail)	6B. Compiles report	6C. Describes problem, method, and analysis	6D. Identifies limitations (e.g., data use, data application methods)	6E. Scopes data narrative based on time, depth, and method	6F. Prepares visualizations	6G. Guides interpretation	6H. Articulates conclusions	6I. Contrasts alternative approaches and past results	6J. Provides recommenda-tions based on results	6K. Tells "data story" to convey insight (e.g., talks to CEO)	
7. Engages in Professional Development	7A. Seeks out mentors	7B. Stays current on emerging technologies, data types, and methods	7C. Attends relevant big data conferences	7D. Contributes new knowledge to the field	7E. Maintains professional library	7F. Participates in professional organizations	7G. Mentors others	7H. Engages in cross-discipline training	7I. Articulates value of big data activities to other departments/ functions of organization	7J. Articulates evolving role of big data in supporting organizational goals		