Presenting Concepts of Big Data in Information Technology Curricula

Ranjan K Sen

Kapgari.Inc, Washington, D.C.

Abstract

In the information technology industry today, there has been significant growth in rapid data mining and analytics applications that process very large volumes of data. This calls for adopting Big Data technology in the Information Technology curricula to prepare todays engineer for the industry. This requires aligning existing curriculum to cover topics such as distributed systems, parallel processing and coordination, fault-tolerance, scheduling and load balancing. In this paper we identify core concepts to be include in new courses. We also look at a typical Information Technology curriculum and present an approach of redesigning it to address the need. It identifies ways to extend, modify or replace existing course components to introduce these concepts in both four-year undergraduate and graduate programs.

Introduction

The convergences of intelligent devices, social networking, pervasive broadband communication, and analytics is ushering a new economic era that is redefining relationships among producers, distributors and consumers of goods and services. The key features of the data are its very large volume, high speed processing and the diversity. Recent articles such as [1] emphasizes the role of Big Data technologies in innovation that is leading us to an "intelligent" economy of smart cars, smart buildings, better healthcare, law enforcement and education, productivity gains in economy and new and efficient ways of interacting with customers.

The requirements for handling structured data with fixed schema, as well as unstructured data with no fixed schema, and theoretically unlimited volume and velocity of data has prompted the development of the so called Big Data technologies. Most of these are in the open source and has been spurring rapid adoption by the industry. Although their availability is easy these technologies are based on advanced Information Technology concepts such as service-oriented architecture, distributed parallel processing, cluster administration, performance monitoring and interoperability.

The goal of this paper is to identify the core and optional needs for the academic preparation for a graduate in Information Technology program, compare it with typical curriculum and suggest how any gap can be filled in. We begin with a brief highlight of the computing curricula 2008 from Association of Computing Machineries (ACM) for Information Technology, Information Science and Computer Science and the

Accreditation Board in Engineering and Technology (ABET). Then we identify topics in the ACM curriculum for Information Technology that are relevant to Big Data concepts and map them to a specific undergraduate program. We also present the courses in a masters level program in Information Technology and examine the relevance of these courses to Big Data education. present a short review of the curriculum of Information Technology in a typical four-year undergraduate and graduate levels today. We then identify courses that are relevant to concepts in Big Data. We present the current scope of the Big Data industry and the core concepts of Big Data technology. This is then followed by a plan for adoption of the Big Data concepts in to existing courses in Information Technology illustrated using the typical example used earlier. In general, we believe similar approach can be used for programs in Information Technology in other academic institutions.

Information Technology Curricula

Academic programs in Computer Science and Engineering, Information Systems, Information Technology share some core concepts. While the first focuses mainly in the design, development, maintenance and architecting digital computer systems, both hardware and software. The second deals with information systems composed of diverse information sub-systems including human operators and automatic processors such as digital computers or other devices to support business environments. They look at interfaces of computing sub-systems and emphasis the user and organizational aspects. The third is in-between the first and the second in their involvement in dealing with both business and the technology, although more emphasis is on development.

The special interest group on IT education (SIGITE) of ACM created accreditation criteria for the Accreditation Board for Engineering and Technology (ABET), which is the main US accreditation agency for programs in computing, engineering, engineering technology and applied sciences [2]. Also, they formed a curriculum committee, SIGITE Curriculum Committee, to develop a model curriculum for IT. The goal was to ensure that the curriculum can be accredited and it was developed in the context of the ACM Computing Curricula project [3]. Also, the curriculum must reflect the relationship of IT to other computing disciplines as they depend on materials covered in other computing disciplines. The curriculum must reflect the aspects that set IT apart from other computing disciplines. The overview report of the Computing Curricula 2005 was augmented as necessary and organized into a form acceptable to the Computing Curricula Series, which is a guideline for four-year undergraduate degree programs in IT from ACM and IEEE in 2008.

IT as an academic discipline is concerned with issues related to advocating for users and meeting their needs within an organizational and societal context through the selection, creation, application, integration and administration of computing technologies.

The program outcome in the ABET Course Advisory Committee (CAC) accreditation criteria is an important input in the development of the IT model curriculum. Among others, the CAC 2005 Overview Report lists the following:

- Ability to design, implement, and evaluate a computer-based system, process, component or program to meet desired needs
- Ability to analyze the impact of computing on individuals, organizations, and society, including ethical, legal, security and global policy issues
- Recognition of the need for an ability to engage in continuing professional development
- Ability to use current techniques, skills, and tools necessary for computing practices

Curriculum and course mapping

The following table identifies courses that are relevant to BigData.

Tech topics cited in the ACM CC 2008	**Relevance in BigData context**	**Course Map (90:_)** (see Table 2 for titles of course numbers)
The World Wide Web and its applications	Common Web GUI for BigData service management portal	231, 238, 247, 248, 291, 302
Networking technologies, particularly those based on TCP/IP	Networking in clusters, socket level, application level, load estimation, supporting distributed services	461, 462, 464
Systems administration and maintenance	Cluster administration, service administration and management, troubleshooting	311,312, 313, 319, 321
Graphics and multimedia	Understanding distributed graphics and multimedia application in BigData environment	230, 267, 268, 311, 313, *
Web systems and technologies	Techniques used in Web server systems useful in design, development of BigData services that are efficient	91.113, 250
Service-oriented architecture	Recognizing service-oriented architectural concepts in understanding BigData products	311, 250, 91.113, 303, 305, *
E-commerce technologies	Understanding capability scope of BigData	250
Client-server technologies	Understanding role in design of BigData products	311, 313
Interoperability	Understanding role of message level interoperability, products such as Apache Thrift, Gueva(?) and their role in BigData	303, 311, 313, *
Technology integration and deployment	Recognizing role of integration and deployment of BigData services	303, 311, 313, 450, 453, 455, *
Object-oriented even-driven programming	Recognizing role of object-oriented and event driven programming in the context of BigData	225, 268, 301, 303
Sophisticated application programmer interfaces (APIs)	Recognizing APIs used in developing applications that uses BigData framework and products	301, 303, 360, *
Human-computer interaction	Role of debug and application logs, system and service performance monitoring tools, tools for administration and management of services in the BigData context	311, 313, 303, 360, *
Security	Issues of security at different levels and their benefits/limitations in BigData context	385, 457

Application domains	Understanding of distributed parallel programming (MapReduce, Distributed File Systems, Index structures, memory resident objects, load balancing, workload distribution, coordinations and scheduling.	220, 224, 225, 306, 308, 247, 268, 269, 270, 271, 297, 301, 302, 303, 305, 459

Table 1: Relevance to Big Data

Table 2 is the list of undergraduate courses in Information Technology offered by the University of Massassuchetts, Lowell http://www.continuinged.uml.edu/degrees/bs_informationtechnology.cfm). The third column in Table 1 gives a rough mapping of these courses to the topics. The following table lists the courses.

Elective course (10 to choose)	
90.220 Visual Basic	90.306 Introduction to XML
90.224 Advanced Visual Basic	90.308 Agile Software Development with Java
90.225 Survey of Programming Languages	90.311 Introduction to the Linux/UNIX Operating System
90.230 Introduction to Multimedia	90.312 Shell Scripting
90.231 Graphics for Multimedia and the World Wide Web	90.313 Linux/Unix Internals Overview
90.232 Desktop Video Production	90.319 Introduction to Linux
90.238 Website Development: Adobe Dreamweaver	90.321 Linux/UNIX System Administration
90.247 Web Authoring: Flash	90.346 Digital Media Delivery
90.248 Website Database Implementation	90.360 Introduction to Data Structures
90.250 E-Commerce on the Web	90.364 Problem Solving with C
90.267 C Programming	90.385 Introduction to Information Security
90.268 C++ Programming	90.450 Database Administration I: Introduction to Oracle 11g
90269 Advanced C++	90.453 Database Administration II: Advanced Oracle 11g
90.270 Visual C++.NET	90.455 Database Administration III: Oracle 11g projects
90.271 C# Programming	90.457 Network Security
90.291 Introduction to DHTML	90.459 PL/SQL I: Introduction to Oracle 11g PL/SQL
90.297 Introduction to Java Programming	90.461 LAN/WAN Technologies
90.301 JAVA Programming	90.462 TCP/IP and Network Architecture
90.302 JavaScript	90.464 Network Management
90.303 Advanced Java Programming	90.467 Relational Database Concepts
90.305 Survey of Perl/Python/PHP	91.113 Exploring the Internet

Table 2: Courses at UM Lowell

The University of Massachusetts Lowell Bachelor of Science degree in Information Technology is available as online, in campus or as a mix of online and in campus courses.

Masters level degree programs

The http://www.thebestschools.org/blog/2012/10/29/20-online-master-information-technology-it-degree-programs/ lists universities that offer online masters level degree programs in IT. Generally, computing or business schools offer these programs. We examine the Masters program in IT at New Jersey Institute of Technology. To understand the flavor of the masters program first observe the range of specializations offered at the BS level by the department of information technology http://it.njit.edu/academics/graduate/BSITBrochure.pdf . These are "Systems integration, administration, design, deployment and management of computing and telecommunication resources and services", "Security and information assurance, ethical hacking, intrusion detection", "Multi-media, graphics design, entertainment technology, and animation", "Game development including options in programming, art and design", "Management information systems and accounting", "Criminal justice and law, computer forensics". They offer the MS degree in IT Administration and Security.

The curriculum is given in http://it.njit.edu/academics/graduate/index.php. The objective of this program to prepare students for database, network, security and web services administrators, enterprise application administrators and as IT administration managers. The students are expected to be a mix of working professionals and graduates of their undergraduate computing programs most of who take the network and security concentrations.

The required (in bold) and elective courses in the MS AS course at NJIT are as follows:

IT 620 Wireless Network Security & Administration	CS 633 Distributed Systems
IT 635 Database Administration	CS 652 Computer Networks- Architectures, Protocols and Standards
IT 610 System Administration	IS 631 Enterprise Database Management
CS 656 Internet and Higher Layer Protocols	IS 677 Information System Principles
CS 696 Network Management and Security	IS 679 Management of Computer and Information Systems
CS 640 Web and Domain Server Administration	IS 680 Information Systems Auditing
CS 631 Data Management System Design	IS 681 Information Systems Auditing
CS 632 Advanced Database System Design	ECE 645 Wireless Networks
CS 697 Principles of Broadband Networks	HRM 601 Organizational Behavior

Table 3: NJIT Masters courses

The Big Data

Big Data means processing of large volumes of data, often in semi-real time, where data can be of different types. Over two billion people have access to the Internet today. This results in the accumulation of very large amount of data, which can be a gold mine of valuable information. With large volume of data one is interested in scalability or the capability of linearly proportional processing time with data size.

Big Data industry

A brief survey of the tremendous value in processing of ever-increasing volumes of data at high speed and unstructured format can be found in [4]. There are many other literatures on the role of data analytics in the context of Volume, Velocity and Variety [5, 6].

Essential concepts in Big Data

Todays Big Data industry is credited to the work of Google, Yahoo and others during 2001-3. It derived from the idea of high performance computing using cluster of commodity machines. The technical foundations for this was multi-processing on a distributed computing environment built using cluster of commodity machines. This is often called – horizontal scaling. Scaling an application is the ability of maintaining processing time to be more or less constant with increasing data volume as long as one is allowed to increase the number of commodity machines proportionately.

The applications that were of interest were processing ever-increasing log files, mostly generated by automated interactions of software systems such as search, web logs, web documents. These applications proliferated into the area of data mining and machine learning for building intelligent alert, notification, and business intelligence and decision support systems. The latter application scenarios and use cases nurtured the Big Data ecosystem and cloud computing. Initiated by companies such as Amazon, eBay and others this just spread in multiple ways.

In addition to cluster of computers and distributed parallel processing with a service oriented approach data processing support for structured and unstructured (with no fixed schema) were core technologies associated with data mining and analytics support. Today, data science application is closely associated with Big Data processing.

The core topics related to Big Data are the following:

1. Cluster of commodity machines
2. Distributed parallel processing – speed and scalability
3. Structured and unstructured data
4. Distributed coordination service
5. Redundancy and fault-tolerance

One approach can be to revise and adopt existing courses in Information Technology to include Big Data concepts. Another approach can be to replace and add one or more new course, depending on the goal of the level of preparation of students, that covers the topics listed above.

New Course(s)

The contents of one or more new courses are based on the need of covering the cores concepts presented above. The contents are described below.
Topic 1: Computer Cluster

The idea of using multiple commodity computers, connected via local area network, originated with the work of building a cheap super-computer. The essential goal was to set up an affordable but useful hardware/software platform for distributed parallel processing that can support high performance computing based on distributed memory parallel processing. It was clear that a key bottleneck in such a set up is the inter-processor communication capability.

Topic 2: Parallel and Distributed Processing – speed and scalability

The Big Data uses a highly popular and simple model for parallel processing. This is known as Map Reduce [7]. A basic concept is functional programming, in which data is available as a key-value pair. The map and reduce functions transform key-values into key-values. The input data (considered a key-value pair) is converted to a new key-value pair by a map function. The reduce function can aggregate key-value pairs by the key value.

MapReduce (MR) framework was designed to address the challenges of large-scale computing in the context of long-running batch jobs. The success of MR led to a wide range of third-party implementations, notably the open-source Hadoop [8] and a number of hybrid systems that combine parallel DBMS with MR, offered by vendors like Aster, Cloudera, Greenplum and Vertica. Hadoop 2.x contains Yarn, which is a general-purpose resource management system for the underlying cluster of commodity system.

Topic 3: Structured and Unstructured data

The very nature of log data is unstructured and varied. Consequently one requirement for Big Data has been the ability to process data in different formats, which may not be known beforehand. The key-value pair data structure is amenable to flexible formatting. Structured data is typically processed as tables with schema in relational database systems (RDBMS). High-speed data operations can be performed in parallel RDBMS. In distributed systems programming languages and messages use a nested representation (such as a tree). Normalizing or recombining these at Web scale is usually very costly. Note the complementary nature of MR and parallel DBMSs [9].

In processing unstructured data a flexible scheme is needed, as the format of data is not known beforehand. The granular key-value data format is used to create dynamic schema on the fly. This is supported via a column oriented table data structure known as Big Table, which is also amenable to distributed data allocation and processing. Examples of usage scenarios include – analysis of crawled web documents, spam analysis, debugging of map tiles on Google Maps, tablet migrations in managed Big Table instances, results of tests run on distributed build system, disk I/O statistics for hundreds of thousands of disks, resource monitoring for jobs run in data centers etc. at Google. The column oriented data structure can be efficiently partitioned and redistributed across the nodes of a cluster. Columnar storage format is supported by many data processing tools such as MR, Sawzall [10], FlumeJava [11].

Data in a Big Data framework resides in a distributed file system modeled after Google File System developed at Google [12] and Big Table. Data is structured similar to a serving tree of a distributed search engine. The machines in a cluster hosts sub trees in different machines. A query gets pushed down the tree to the machines hosting the sub trees for computing the results. The results are then aggregated in parallel. This is a well-known parallel processing paradigm.

Dremel [13] can execute many queries that run faster than running sequences of Map Reduce (MR) jobs. Dremel is used in conjunction with MR to analyze outputs of MR pipelines or rapidly prototype larger computations. Unlike Pig [14] or Hive [15] it does not execute queries as MR jobs.

In Big Data storage technology key design goal is to meet reliability and scaling. The Amazon Simple Storage System or S3 is based on Dynamo [16] with a tight control over the tradeoffs between availability, consistency, cost-effectiveness and performance. Actual production performance of Dynamo supports 3 million checkouts of purchase portal of Amazon. Different techniques have been combined to provide a single highly available system. It is an eventually consistent storage system that can support demanding applications.

Topic 4: Distributed Coordination Service

This distributed service is fault-tolerant and allows a registry service where state information can be preserved. This allows a service to restart and reinitialized by accessing the state information stored in the registry. The content should cover elements of distributed coordination services that include leader election using quorums, use of technology such as zookeeper [17] and related matters.

Topic 5: Redundancy and fault-tolerance

Scalability, fault-tolarence and availability in Dynamo is implemented through data partitions and replications using consistent hashing, object versioning. Consistency among replicas during updates is maintained by a quorum-like technique and a decentralized replica synchronization protocol. Gossip based distributed failure detection and membership protocol.

Curriculum adoption

A number of companies such as Cloudera, Hortonworks, MapR and others offer specialized trainings on various Big Data products, most of which are open source, and technologies. Scores of online documentations and tutorials are available as well. Unfortunately these materials are not suitable for inclusion in a curriculum directly. Utmost care needs to be given to select appropriate context and material when existing courses are modified or extended to include such materials. Below we identify a set technology, which in our opinion forms the base that demonstrates the core concepts of Big Data.

Concepts/Technology	Possible target course for modification
Distributed File System	90.360 Introduction to Data Structures 90.311 Introduction to the Linux/UNIX Operating System CS 633 Distributed Systems
Parallel processing (Map Reduce, Query processing on column-stripped nested data)	90.360 Introduction to Data Structures 90.250 E-Commerce on the Web 90.303 Advanced Java Programming
Big Data eco-system (Hadoop, Pig, Hive…)	90.467 Relational Database Concepts

Distributed Coordination	90.360 Introduction to Data Structures 90.303 Advanced Java Programming
Cloud computing technology (SaaS, PaaS, IaaS)	90.250 E-Commerce on the Web 90.311 Introduction to the Linux/UNIX Operating System
Cluster administration (chef, puppet), monitoring and tuning	90.311 Introduction to the Linux/UNIX Operating System 90.461 LAN/WAN Technologies 90.462 TCP/IP and Network Architecture 90.464 Network Management **IT 610 System Administration** **CS 696 Network Management and Security** CS 632 Advanced Database System Design
Big Table and column oriented data structure processing	90.360 Introduction to Data Structures 90.467 Relational Database Concepts 90.303 Advanced Java Programming
Object oriented database for indexed key-value storage	90.360 Introduction to Data Structures 90.467 Relational Database Concepts
Service oriented design, installation and configuration, interoperability	90.250 E-Commerce on the Web 90.311 Introduction to the Linux/UNIX Operating System 90.462 TCP/IP and Network Architecture 90.303 Advanced Java Programming

Table 4: Existing Courses to modify to adoption of Big Data Concepts

For distributed file system, we can introduce topics on data structures that allow concurrent operations on data. Benchmarks can be introduced in this context. Replication to preserve data quality despite faulty hardware and achieve fast response times in presence of stragglers. Resource scheduling for CPU, rescheduling slow or faulty process and load balance are some related topics.
For a master level course, the subject titled distributed systems is often part of existing concurrency. This course can provide the context of most basic concepts pertinent to the Big Data technology. In masters level we can cover the topic of data analytics at scale that needs a high degree of parallelism.

Parallel processing can be introduced in the data structure course as well as in advanced Java programming course. Functional programming, data parallel programming and Map Reduce as functional programming paradigm, parallel query processing where sub-queries are evaluated in parallel followed by aggregation by combining the result are some of the topics that can be covered. Data structure can be a complementary course in which column stripping and data splitting to support scalability and data parallel mode of computation. Construction of query processing engines for nested data records in columnar representation of nested data and its platform neutral nature together with serialization is a related topic that can be part of the data structure course.

As part of the combination of programming and data structure courses we can include the topic of code generation tools, bindings for programming languages such as C++ or Java. In addition, topics such as cross-language interoperability – using a standard binary on-the-wire

representation of records in which field values are laid out sequentially as they occur in the record can be included. Also, assembling columnar representation to record for interoperation with MR and other data processing tools can be included.

Distributed coordination system can be included in advanced programming course in the context of concurrency. Products such as Zookeeper [17] can be included via lab exercises together with Java based programming. Cloud computing topics can be included in the E-commerce on the Web from a use case scenario. In this context both SaaS and PaaS can be introduced. IaaS can be introduced in the context of Linux/UNIX operating system. Cluster administration topics can be added in Linux/UNIX operating system. Associated topics can be covered in subjects on LAN/WAN, TCP/IP and Network Architecture and Network Management. Big Table topics can be included in Data Structures, relational database and advanced programming. Object oriented database can be covered in this context. Service oriented design; installation, configuration and interoperation can be covered in advanced programming, Linux/UNIX operating system

Bibliography

1. Unlocking BigData – Foundation for Innovation. https://www.unlockingbigdata.com/
2. Joseph J. Ekstrom, The Information Technology Model Curriculum, Journal of Information Technology Education, Vol. 5, 2006, pp. 343 -361.
3. ACM Computing Curricula project www.acm.org/education/
4. Big Data's Big Impact Across Industries http://www.forbes.com/sites/howardbaldwin/2014/03/28/big-datas-big-impact-across-industries/.
5. Defining Big Data: Volume, Velocity and Variety, http://www.forbes.com/sites/howardbaldwin/2014/03/28/big-datas-big-impact-across-industries/
6. Big Data http://en.wikipedia.org/wiki/Big_data
7. J.Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In OSDI, 2004
8. Apache Hadoop. hadoop.apache.org/
9. D. J. Abadi, P. A. Boncz, and S. Harizopoulos. Column-Oriented Database Systems. *VLDB*, 2(2), 2009.
10. R. Pike, S. Dorward, R. Griesemer, and S. Quinlan. Interpreting the Data: Parallel Analysis with Sawzall. *Scientific Programming*, 13(4), 2005.
11. C. Chambers, A. Raniwala, F. Perry, S. Adams, R. Henry, R. Bradshaw, and N. Weizenbaum. FlumeJava: Easy, Efficient Data-Parallel Pipelines. In *PLDI*, 2010.
12. S. Ghemawat, H. Gobioff, and S.-T. Leung. The Google File System. In *SOSP*, 2003.
13. S. Melnik et al. Dremel: Interactive Analysis of Web-Scale Datasets, *Proceedings of the VLDB Endowment,* Vol. 3, No. 1
14. C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig Latin: a Not-so-Foreign Language for Data Processing. In *SIGMOD*, 2008.
15. Hive. http://wiki.apache.org/hadoop/Hive, 2009.
16. G. DeCandia et al. Dynamo: Amazon's Highly Available Key-value Store, SOSP 07, October 14-17, 2007, Stevenson, Washington, USA
17. Zookeeper project zookeeper.apache.org