## A Certificate Program in Data-Intensive Computing*

**Introduction:** Data-intensive computing deals with computational methods and architectures to analyze and discover intelligence in huge volumes of data generated in many application domains. This undergraduate-level certificate program addresses the increasing need for workforce personnel who are competent in data-intensive computing and other closely related technologies such as grid computing, parallel computing, and cloud computing. The program consists of a total of five courses: three required CSE courses, one elective course from the major discipline of the student and a capstone project course that applies the concepts in the earlier courses of the certificate. See below for details.

**Motivation**: Data-intensive computing has been receiving much attention as a collective solution to address the data deluge that has been brought about by tremendous advances in distributed systems and Internet-based computing. The need to deal with huge volumes of data is being felt across industry, government, and academia. This includes the analysis needs of intelligence agencies for data from satellite imagery, signal intercepts, sensors, etc.; medical agencies analyzing images, genetic and other diagnostic data; environmental agencies with sensor-collected data on air, water, and meteorological conditions; retail companies such as Wal-Mart and amazon.com analyzing sales data; and organizations in general needing to exploit the massive document databases on the world wide web (several hundred trillion byes of text); among others.

An innovative programming model called MapReduce and a large-scale distributed file system to support it have revolutionized and fundamentally changed approaches to large scale data storage and processing. These data-intensive computing approaches are expected to have a profound impact on any application domain that deals with large scale data, from healthcare delivery to military intelligence.

A new forum called the *big-data computing group* has been formed by a consortium of industrial stakeholders and agencies including the NSF (National Science Foundation) and CRA (Computing Research Associates) to promote wider dissemination of big-data solutions and transform mainstream applications. Given the omnipresent nature of large scale data and the tremendous impact they have on a wide variety of application domains, it is imperative that we prepare our workforce to face the challenges in this area. The timely introduction of these concepts to our undergraduate students is important for them to remain competitive in a fast-moving global environment. Furthermore, our in-place industry workforce needs to achieve and maintain their currency and competency in the data-intensive computing areas to meet the big-data challenges. Industry personnel need continuing education to upgrade their skills.

In response to the demand for knowledge and expertise in this area and with the support of a NSF CCLI Phase 2 grant, the Computer Science and Engineering Department has developed an undergraduate-level certificate program designed to educate students in data-intensive computing and related technologies. This certificate program is also open to any non-CSE undergraduate student who satisfies the prerequisites needed for the courses.

**Intended Audience:** CSE undergraduates are the primary target audience for the certificate program. CSE majors can receive the certificate by appropriately choosing their electives and capstone project area. This program is also suitable for non-CSE engineering majors, biomedical

* This program is partially supported by NSF grant DUE-CCLI-0920335

majors, bioinformatics majors and any student wishing to pursue a minor in CSE. This program will also serve industrial workforce personnel who are considering retraining or updating their skills, and those who are interested in focused exposure to data-intensive computing and technologies.

**Program Outline:**

Required Courses (3):
CSE250: Data Structures and Algorithms, or equivalent
CSE486: Distributed Systems
CSE487: Data-Intensive Computing

Elective Courses (2):
4XX/3XX: any course with data-intensive problems in the major area of the student
4XX: Capstone project in the major area of the student

**Sample Courses for the certificate:**

Sample courses for CSE majors:

1. CSE250, CSE486, CSE487, CSE435 (Information Retrieval), CSE494 (Senior capstone course)
2. CSE250, {CSE486, CSE442} , {CSE487, CSE453} (CSE486 and CSE442 can be taken at the same time, CSE487, CSE453 can be taken at the same time)

Sample courses for Bioinformatics majors:
CSE250, CSE486, CSE487, BIO400 (Bioinformatics), BIO499 (Capstone Independent Study)

**Financial incentives for students**: A grant of $12000 ($4000 X 3 years) is available from NSF for attracting under-represented minorities and industrial workforce to enroll and complete the certificate program.

**Interdisciplinary support**: Chairs of the Math Department and Biological Sciences support this certificate program and will encourage their undergraduates to enroll in this program.

For additional information send email to bina@buffalo.edu, vipin@buffalo.edu

Certificate application is in progress.