Introduction to the Big Data Engineering Minitrack

Hong-Mei Chen
University of Hawaii at Manoa
hmchen@hawaii.edu

Ken Anderson
University of Colorado
kena@cs.colorado.edu

Wietske van Osch
Michigan State University
vanosch@msu.edu

Big Data Engineering (BDE) is a new field that has emerged to address the challenges and requirements of the 5 Vs (volume, velocity, variety, veracity, and value) of big data. The exponential growth of data over the last decade has fueled a new specialization for software technology: data-intensive, or big-data, software systems. Internet-born organizations such as Google and Amazon are on this revolution's cutting edge, collecting, managing, storing, and analyzing some of the largest data repositories ever constructed. Their pioneering efforts along with those of numerous other big-data innovators, have provided a variety of open source and commercial data management technologies that let any organization construct and operate massively scalable, highly available data repositories.

Addressing the challenges of engineering software for big-data systems requires careful design tradeoffs spanning the distributed software, data, and deployment architectures. It also requires extending traditional software architecture design knowledge to account for the tight coupling that exists in scalable systems. Highly parallel, distributed systems are required to provide the necessary velocity of processing and handle the ever-growing volumes of information in big data systems. Building such systems in an agile fashion to address evolving requirements requires judicious adoption of a range of off-the-shelf specialized data processing and management technologies that can provide the necessary system quality attributes at predictable cost.

This minitrack covers advances in the broad range of activities that are required to cost-effectively plan, design, build, evolve and manage big data systems. It aims at providing an outlet for researchers in various disciplines to exchange ideas and solutions.

This year we able to provide two papers that present research results in their advanced stages.

The first paper is entitled "Big Data Value Engineering for Business Model Innovation" by Hong-Mei Chen, Rick Kazman, Juan Garbajosa and Eloy Gonzalez. Big data value engineering for business model innovation requires a drastically different approach as compared with methods for engineering value under existing business models. This paper reports an exploratory multiple case study with 23 large enterprises to formulate the requirements for a method to aid in engineering value via innovation. A method, called Eco-ARCH (Eco-ARCHitecture), was developed for value discovery. This method is tightly integrated with the BDD (Big Data Design) method for value realization, to form a big data value engineering methodology for holistically addressing these requirements. The Eco-ARCH approach fills a methodological void for the big data value engineering where no central architecture pre-exists, system boundaries are fluid, requirements are ill-defined, many stakeholders are unknown, design goals are not provided, system behavior is non-deterministic and continuously evolving, and co-creation with consumers and prosumers is essential to achieve innovation goals. The method was empirically validated with a large IT service company in the Electric Power industry.

The second paper, "Batch to Real-time: Incremental Data Collection & Analytics platform" is authored by Ahmet Arif Aydin and Kenneth M. Anderson. It reports on real-time monitoring and querying of Twitter data while a mass emergency event is underway. To provide highly concurrent and efficient real-time analytics on streaming data at interactive speeds requires a well-designed software architecture that makes use of a carefully selected set of software frameworks. This paper presents the design and implementation of the Incremental Data Collection & Analytics Platform (IDCAP). The IDCAP provides incremental data collection and indexing in real-time of social media data; support for real-time analytics at interactive speeds; highly concurrent batch data processing supported by a novel data model; and a front-end web client that allows an analyst to manage IDCAP resources, to monitor incoming data in real-time, and to provide an interface that allows incremental queries to be performed on top of large Twitter datasets. Compared to a prior system, Epic Collect, as well as Apache Solr in handling this type of scenario, IDCAP is significantly faster and more efficient.

URI: http://hdl.handle.net/10125/41875
ISBN: 978-0-9981331-0-2