Introduction to Big Data and Analytics: Concepts, Techniques, Methods, and Applications Minitrack

Stephen H. Kaisler, D.Sc.
SHK & Associates
Laurel, MD
skaisler1@comcast.net

Frank J. Armour, Ph.D.
Kogod School of Business
American University
Washington, DC
farmour@american.edu

Alberto J. Espinosa, Ph.D.
Kogod School of Business
American University
Washington, DC
alberto@american.edu

Abstract

The Big Data minitrack features a number of papers addressing methods and techniques, issues and challenges, and organizational approaches to processing and managing Big Data within an organizational environment. This year …

The Big Data minitrack has been offered at HICSS for the past four years. This will be the fifth year in which interesting papers are being presented that address key and critical issues in Big Data and Analytics. This minitrack resulted from keynote speaker presentations at HICSS-44 and HICSS-45 which described the impact that Big Data was having and would continue to have on information systems and computer science. The co-chairs have given tutorials on Big Data at HICSS-46 through HICSS-49 and are presenting two distinct topic tutorials at HICSS-50. Attendance at these tutorials (approximately 50-90 people) and at the minitrack sessions (approximately 20-40 people) has reinforced our belief that HICSS is a major venue for the presentation of Big Data and Analytics research. The co-chairs have been heavily involved in planning degree programs as well as teaching courses in Big Data at their respective universities. Two papers on issues and challenges by the co-chairs which have been published at HICSS-46 and HICSS-47 have received numerous citations according to ResearchGate.

Introduction

The paper Value Oriented Big Data Strategy: Analysis and Caste Study by J. Arcondara, K. Jimmi, P. Guan and W. Zhou, addressed the question of whether big data is having or has had an impact on companies as measured by their success in the stock market. They examined data from the CAC40 to determine whether there was a relationship between stock performance and corporate usage of big data. Because data creates value for business organizations, it seemed logical to assume that such value would be reflected in the organizations stock performance. Companies were divided into four categories: Big Data competitors, overachievers, underachievers, and disadvantaged, e.g., those who have limited resources to expend on big data analysis. One measure they use was whether a company had a Chief Data Officer and/or a Chief Digital Officer. Companies having these positions outperformed the average stock price for all companies in the CAC40 using Big Data.

The authors examined the organizations Big Data strategy based on assessing Big Data Capability against Operational and Decision Dynamics. Four categories of strategy were examined: routine, excellence, integration, and strategic. Companies using Big Data moved from an integration usage to an excellence usage where Big Data drove many corporate decisions. A case study of the airline industry served to re-affirm their observations. They concluded that there is no link between underperforming companies and lack of data capability. Rather, performance was tied to failing to use available Big Data in dynamic decision making. They also concluded that Big Data can affect every part of the business decision-making process, but that the value it creates differ greatly from firm to firm.

The paper Data Systems Fault Coping for Real-time Big Data Analytics Required Architectural Crucibles by Stephen Cohen and William Money examined the role of unknown and unexpected faults introduced into real-time systems while processing Big Data. This is an area of research that has been neglected in the Big Data environment. Because many organizations now use Big Data on a continuing basis to make operational, tactical, and strategic business decisions, the impact of faults caused by failing to properly curate, cleanse, and transform Big Data can have significant effects on an organization's business success. The problem becomes particularly acute as more organizations utilize streaming data to make (near-) real-time business decisions. Dealing with fault – their

URI: http://hdl.handle.net/10125/41270
ISBN: 978-0-9981331-0-2

analysis, mitigation, and recovery – requires the creation of new architectural concepts in hardware, software and network topologies. Several cases are reviewed that reinforce the need for architectural responses to handling Big Data faults. The authors conclude that fault analysis and handling is an emerging critical problem that must be addressed in the design of systems dealing with Big Data.

The paper Service-Oriented Cost Allocation for Business Intelligence and Analytics: Who pays for BI&A? by R. Grytz and A. Krohn-Grimberghe addresses one of the key questions in the use of Big Data and Analytics: how to pay for the data preparation and analysis necessary to properly utilize Big Data in making key business decisions. The author's solution is to define a service-oriented model that can lead to a charging scheme for specific services used by the business operations. This transfers parts of the decision to use Big Data to business operations who can decide how much they want to allocate to BI&A along with other cost factors in their business operations. Developing a charging scheme in order to develop a cost allocation scheme can be difficult because BI&A has higher degrees of interdependencies and is more dynamic than typical IT schemes. The authors consider a BI&A service catalog with associated costs that allows business organizations to select the services based on their need for information and the decisions they need to make as well as their budgets. They propose a model that will be tested to determine its viability in address this critical area.

The paper A Correlation Network Model for Structural Health Monitoring and Analyzing Safety Issues in Civil Infrastructures by A. Fuchsberger and H. Ali addresses the key problem in civil infrastructures that has been identified as a multi-trillion dollar problem for the foreseeable future. The authors note that the Federal Highway Administration inspects over 600,000 bridges and other structures every two years no matter what their status. But, this manual inspection often leads to erroneous data. Although new types of sensors are becoming available (e.g., acoustics, xrays, etc.), they are not widely distributed. The authors focus on analyzing this data using graph analytics techniques to identify problems based on similarities among different types of structures with similar attribute values (age, construction type, etc.). Their analysis has shown that current monitoring is based on anomaly detection, but often signs of damage are not rare. Being able to assess their severity can lead to predictions about the status and reliability of the infrastructures. The author's approach may offer a new tool for determining the status of civil infrastructures, increasing the frequency of inspections – with or without new sensors, and the ability to predict and then alert authorities as to the status of critical infrastructures.

The paper Introducing Data Science to Undergraduates through Big Data: Answering Questions by Wrangling and Profiling a Yelp Dataset by S. Jensen addresses another area that has been somewhat neglected in the rush to Big Data: How do we train the next generations of data scientists at the undergraduate level in our colleges and universities? Most data science work is focused on cleansing, curating, transforming and wrangling data. This is not necessarily exciting work (as one of the co-chairs can attest). Getting undergraduate students excited about the prospects for data science has to go beyond the data preparation phase to the actual analysis phase using a variety of tools. This paper focuses somewhat on data preparation with the idea of trying to cast it as a challenging problem (which it is) and how to show undergraduates that out of the wrangling process can arise interesting business intelligence questions. The author wanted to provide some insight into data analysis tools and to determine whether differences exist between male and female students and how best to serve each group's needs for understanding the concepts. The author set up an analysis system using Hadoop, Hive and Tableau using a real social media dataset. He concludes that this is both a practical and effective way to get students to understand how to use the tools and how to frame/pose data analysis questions for which they could use the tools at hand.

The paper An Introduction to the MISD Technology by A. Popov focuses on the use of multiple instruction, single data stream (MISD) hardware to apply multiple analytic techniques to data streams. Heretofore, MISD had been dismissed because few examples existed, such as CMU's Systolic Array processors, but recent hardware architectural efforts have resulted in new systems emerging. The author describes the Structure Processing Unit (SPU), a new implementation of MISD technology. The author describes the principles of design and the programming model for the SPU. The SPU was implemented using the Virtex FPGA with an on-ship PowerPC 405 CPU and benchmarking tests were performed. The basic idea behind using MISD technology is that an algorithm is divided into several parts each of which can be performed concurrently. The author concludes that this implementation of MISD technology offers a new processing model for Big Data.

The paper Comparing Data Science Project Management Methodologies via a Controlled Experiment by J. Saltz, I. Shamshurin, and K Crowston addresses another critical problem in Big Data and Analytics: are specific technologies and analytic methods better than others and how do we determine which is which? The authors discuss an experiment to compare four different data science project methodologies. They define a general model and then map the four methodologies to the general model. The four methodologies were Agile Scrum, Agile Kanban, CRISP, and Baseline (e.g.,, no methodology). Each team was instructed in data science, given access to a business expert, and instructed in the particular methodology. CRISP was reported as the most effective methodology, followed by Kanban, Baseline, and then Scrum. Although Kanban came in second to CRISP in the expert's evaluation, the student's review rated it the more effective method based on the survey instrument. That Scrum cam in dead last is perhaps testament to their leap into doing analytics immediately without focusing on requirements. CRISP was effective because it seemed to be a natural way to execute a project through stages and, yet, supported iterations as necessary. The authors plan further research to refine these results.