2019 HPCC Systems® Community Day

Challenge Yourself – Challenge the Status Quo

HPCC SYSTEMS®

Wednesday October 16, 2019

Time	Topic	Presenter
8:30am - 8:35am	Welcome	
8:35am – 9:00am	**Opening Keynote** While Big Data processing has made tremendous progress over the last 10 years, the next generation of tools will provide more plug and play capabilities to perform data ingestion, curation and analysis. As the growth rate of raw data accelerates and novel applications impose more stringent real time requirements on data analytics, the need for easy to use tools will be in high demand. HPCC Systems is developing the next generation of toolsets to meet these needs. Tombolo, a tool for data curation and compliance tracking (GDPR, CCPA, HIPAA, DPPA etc.), ECL Cloud IDE to make coding complex analytics simple, and IoT Central to manage devices and ingestion pipelines are some of the tools planned for release this year. With the overall core HPCC Systems platform improvements, these additional tools will provide a comprehensive end to end open source stack that can better prepare the platform for a new generation of data scientists.	**Flavio Villanustre, LexisNexis Risk Solutions**
9:00am - 10:15am	**HPCC Systems in Industry** Featuring talks about a bio-technology use case and our own success stories.	**Track 1**
9:00am – 9:25am	**Leveraging HPCC Systems as Part of an Information Security, Privacy, and Compliance Framework** This presentation will describe how the Information Assurance and Data Protection Group (IADP), in collaboration with LexisNexis Risk Solutions, is leveraging HPCC Systems to support critical components of the RELX Group information security, privacy, and compliance framework. The goal of the IADP HPCC Systems program is to leverage the	**Andy Bayer & Marcus Mullins, RELX IADP & Naweed Mohammed, LexisNexis Risk Solutions**

	full capabilities of HPCC Systems and related technologies to ultimately improve the ability to respond to new threats more effectively and efficiently. There is also a strong reliance on complete and accurate data that is easily understood when it comes to ensuring efficient investigation and/or auditing processes. To achieve these goals, the HPCC Systems program is organized around four key areas: Data Ingestion; Advanced Search/Reporting; Fraud Detection/Alerts; and Workflow Integration.	
9:25am – 9:50am	**HPCC Systems Accelerating the Rapid Expansion of the China Big Data Market** China's economy has grown into the world's second biggest economy next to United States. Over the past decade, the mobile network has spread across the entire country with 800 million internet/mobile users and still growing, while United States has about 250 million users. Such a huge user base has generated an increasingly high demand on new products and services based on the mobile network, which have also produced a tremendous amount of data in recent years. The opportunity of leveraging the data generated by the mobile users to create analytical products and services is endless. LexisNexis Risk Solutions established its technology presence in China via the Genilex Joint Venture in 2015, which was acquired by RELX as a wholly owned subsidiary in early 2019. Yun will share what big data projects LexisNexis Risk Solutions has conducted leveraging HPCC Systems in China, as well as discuss current and future opportunities.	**Yun Chen, LexisNexis Risk Solutions**
9:50am – 10:15am	**The Power of HPCC Systems - a Bio-tech Industry Practice** This presentation is about a microbiota genetic analysis service for a customer in the bio-tech data processing space. Gut microbiota is closely related to numerous diseases such as obesity, diabetes metabolic syndrome, autoimmune diseases, ulcerative colitis and various mental diseases such as depression and Parkinson's disease (PD). In this solution, ECL is used to store the gene data mainly by HPCC Systems and fully leverage machine learning functions and the integration of HPCC Systems with other programming languages to provide a full-stack data-driven intelligent processing solution. With support from HPCC Systems, our project is working to achieve the following two conclusions: 1) gut microbiota signature/composition change in patients with mental diseases; and 2) use of this signature as a bio-marker for early diagnosis.	**Meng Han, Kennesaw State University**
10:15am - 10:30am	**Break**	
10:30am - 12:00pm	**HPCC Systems in Academia** Presentations include research findings and project outcomes from our HPCC Systems Academic Community.	**Track 2**
10:30am – 10:55am	**Using High Dimensional Representation of Words (CBOW) to Find Domain Based Common Words** Text cleaning is becoming an essential step in text classification. Stop word removal is a	**Farah Alshanik, Clemson University**

	crucial space-saving technique in text cleaning which saves huge amounts of space in text indexing. There are many domain-based common words which differ from one domain to another and have no value within a particular domain. Eliminating these words will reduce the size of the corpus and enhance the performance of text mining. This talk will discuss how the text vectors bundle, (CBOW), in HPCC Systems was used to find the domain based common words, and the methodology applied to enhance the performance of classification methods.	
10:55am – 11:20am	**Athlete 360: Leveraging HPCC Systems for Maximizing Player Performance** The ultimate goal in any sport is to win. This requires athletes to compete at a high level, which in turn requires coaches and athletes to strive towards improving performance. Although in team sports, there are many external variables that cannot be controlled. This makes the process of gauging performance of individual athletes difficult. The field of sport science is still relatively young and is rapidly evolving. This state of growth is important because the more data that can be collected, the better the understanding of what an athlete does and how their body responds. Within collegiate athletics, and sports in general, there is a struggle to be able to interpret data from different streams together in a single report. Furthermore, streamlined data collection can better provide an understanding of what an athlete does and how their body responds. This involves data from all aspects of an athlete's day including wellness questionnaires, practice training loads, weight room training loads, and weight room assessments of strength, power, and fatigue. In this talk, learn how the NC State University Strength and Conditioning Coach team is addressing this challenge by using HPCC Systems for collecting and analyzing all the various data streams for creating a 360° view of an athlete's wellbeing to ensure they are performing at their highest potential.	**Vincent Freeh & Chris Connelly, NC State University**
11:20am – 11:40am	**Beyond the Spectrum – Creating an Environment of Diversity and Empowerment with HPCC Systems** Hear how the Florida Atlantic University Center for Autism and Related Disabilities has partnered with the HPCC Systems community to provide young people with autism both the technology and professional skills needed to compete in today's workplace. Mentoring and hands-on coding through ECL workshops have positively impacted students, opening doors to new opportunities for both students and employers.	**Darius Murray, Florida Atlantic University Center for Autism and Related Disabilities**
11:40am – 12:00pm	**A Success Story of Challenging the Status Quo: Gadget Girls and the Inclusion of Women in STEM Teaching** Join NSU University School student and program leader for girls in robotics, Ronnie Shashoua, as she talks about Gadget Girls - a project in collaboration with the NSU Alvin Sherman Library, NSU University School, the South Florida Girl Scouts, and sponsorship	**Ronnie Shashoua, NSU University School**

	from the HPCC Systems Academic Program. Gadget Girls is a program aimed at encouraging girls in fourth and fifth grade to explore their interests in and love for STEM, especially robotics and engineering. Shashoua will discuss the underrepresentation of girls in the Florida Vex Robotics circuit, such as how it demonstrates a larger trend of low numbers of women undertaking STEM educational and career paths and the role it played in inspiring the creation of Gadget Girls.	
12:00pm - 12:15pm	**Community Awards Ceremony** Let's congratulate the winners of the 2019 HPCC Systems Poster Competition and other award recipients.	**Trish McCall & Flavio Villanustre, LexisNexis Risk Solutions**
12:15pm - 1:00pm	**Lunch**	
1:00pm - 1:30pm	**Interactive Expo** New this year! Robotics showcase featuring American Heritage School and NSU University School, hands-on demos, and Q&A with our HPCC Systems experts in our first ever Interactive Expo.	
1:30pm - 3:00pm	**HPCC Systems Breakouts** Sessions include a deeper dive into specific technical topics and components of the HPCC Systems platform.	**Track 3**
1:30pm – 2:10pm	**HPCC Systems Breakout Rotation 1**	
	Theme 1: System Enhancements	
Chairperson: **Kunal Aswani**	**Workunit Analysis Tool** The Workunit Analyser examines the entire workunit to produce advice that both novices and experienced ECL developers should find useful. The Workunit Analyser is a post-execution analyser that identifies potential issues and assists users in writing better ECL.	**Shamser Ahmed, LexisNexis Risk Solutions**
Chairperson: **Ken Rowland**	**Leveraging the Spark-HPCC Ecosystem** Join us for an introductory walk-through of using the Spark-HPCC Systems ecosystem to analyze your HPCC Systems data using a collaborative Apache Zeppelin notebook environment.	**James McMullan, LexisNexis Risk Solutions**
	Theme 2: Usability Improvements	
Chairperson: **Trish McCall**	**Dapper Tool - A Bundle to Make your ECL Neater** Have you ever written a long project for a simple column rename and thought, this should be easier? What about nicely named output statements? Yeah they bother me too. Oh, and DEDUP(SORT(DISTINCT()))? There is a better way! Learn how Dapper can help!	**Rob Mansfield, Proagrica**

Chairperson: **Becky Champion**	**DataPatterns - Profiling in ECL Watch** DataPatterns.Profile() has been evolving since the last time you may have seen it. It does more. It looks better. It has been integrated into the ECL standard library and into ECL Watch. Learn what this data profiler can do for you and how its built-in visualization easily summarizes the results.	**Dan Camper, LexisNexis Risk Solutions**
2:10pm - 2:20pm	**Break**	
2:20pm - 3:00pm	**HPCC Systems Breakout Rotation 2**	
	Theme 1: Novel Applications	
Chairperson: **Richard Taylor**	**Leveraging Intra-Node Parallelization in HPCC Systems** HPCC Systems offers parallelization by assuming data-independent tasks. For operations having complex predicates, such as the set similarity join (SSJ) predicates, this assumption might create a tradeoff. On the one hand, one may choose a data replication strategy with large intermediate data groups assuring that all intermediate results fit into main memory. However, this can lead to an underutilization of today's massively parallel CPUs. On the other hand, you may choose a higher degree of replication for better CPU utilization. Such a choice may lead to an overutilization of main memory on the compute nodes. Our research focuses on the parallel implementation of the set similarity join (SSJ) operator. This operator finds all pairs of records which have a similarity above a defined threshold using a similarity measure such as Jaccard. Our goal is a robust approach for executing SSJ that does not over-utilize memory and exploits CPU parallelization as much as possible. This approach requires data sharing between tasks/threads which is not foreseen in HPCC Systems so far. In this talk, we describe how we implemented multi-threaded user-defined functions for the SSJ operator. To be able to control NUMA-specific parallelization conditions, we implemented a C++ plugin for HPCC Systems. Furthermore, we show how we visualized the relevant system parameters (CPU and memory usage). This talk is intended for anyone interested in extending HPCC Systems by plugins and monitoring distributed program execution.	**Fabian Fier, Humboldt University Berlin**
Chairperson: **Bob Foreman**	**Expanding HPCC Systems Deep Neural Network Capabilities** The training process for modern deep neural networks requires big data and large computational power. Though HPCC Systems excels at both of these, HPCC Systems is limited to utilizing the CPU only. It has been shown that GPU acceleration vastly improves Deep Learning training time. In this talk, Robert will go into detail on the first GPU accelerated library for HPCC Systems and how it greatly expands its deep neural network capabilities.	**Robert Kennedy, Florida Atlantic University**
	Theme 2: Cloud Enablement	

Chairperson: Sarah Fabius	**Docker Support** Learn how to package the HPCC Systems Platform in a Docker container and deploy it locally, and build an HPCC Systems Platform AMI followed by an AWS deployment.	**Xiaoming Wang & Godson Fortil, LexisNexis Risk Solutions**
Chairperson: Helen Graham	**Progress Towards the Cloud** General discussion of the datacenter migration toward the public cloud. Includes a high-level overview on decisions, planning, and methodology along with success stories and patterns.	**Jon Burger, LexisNexis Risk Solutions**
3:00pm - 3:15pm	**Break**	
3:15pm - 4:40pm	**HPCC Systems Roadmap Tech Talks** Our platform team will share an update on the latest features included on the roadmap.	**Track 4**
3:15pm - 3:30pm	**Advancements in HPCC Systems Machine Learning** This presentation will provide an overview of the latest advancements in Machine Learning modules over the past year, including Clustering, Natural Language Processing, Deep Learning, and the Expanded Model Evaluation Metrics.	**Roger Dev, LexisNexis Risk Solutions**
3:30pm - 3:45pm	**Clustering Methods of the HPCC Systems Machine Learning Library** The clustering method is an important part of unsupervised learning. To gain the unsupervised learning capability, two widely applied clustering methods, KMeans and DBSCAN are adopted to the current Machine Learning library. This presentation will introduce the newly developed clustering algorithms and the evaluation methods.	**Lili Xu, LexisNexis Risk Solutions**
3:45pm - 4:05pm	**Geohashing with Uber's H3 Geospatial Index** An introductory look at the ECL H3 Plugin (available since v7.2.0) - a journey from lat/long to ROXIE Service driven visualizations.	**Gordon Smith, LexisNexis Risk Solutions**
4:05pm - 4:25pm	**Release Cycle Changes** This talk will explain the reasoning behind the release cycle changes, and how overcoming the challenges faced in the previous practice of automated testing has introduced new benefits and wider acceptance from the wider community.	**Attila Vamos, LexisNexis Risk Solutions**
4:25pm - 4:40pm	**Path to 8.0** Come hear a brief overview on the direction the HPCC Systems platform is heading, and get a glimpse into some of the likely highlights included in the next minor and major versions.	**Gavin Halliday, LexisNexis Risk Solutions**
4:40pm - 4:55pm	**Community Website: Virtual Ribbon Cutting** The reveal of how HPCC Systems is improving and expanding around the world for building a global community.	**Jessica Lorti, LexisNexis Risk Solutions**
4:55pm - 5:00pm	**Closing / Adjourn**	

#HPCCSummit
Event Details: hpccsystems.com/hpccsummit2019
Livestream: http://bit.ly/2019hpccsummit

Sponsored by:

DELL Technologies

Infosys®
Navigate your next

Platinum

Gold

Datum
SOFTWARE

KFORCE®
TECHNOLOGY STAFFING

Bronze

Bronze

#HPCCSummit
Event Details: hpccsystems.com/hpccsummit2019
Livestream: http://bit.ly/2019hpccsummit