HPCC Systems Intern Program

Lorraine Chapman
Manager Business Analyst
HPCC Systems Intern Program Manager
LexisNexis Risk Solutions

HPCC SYSTEMS®

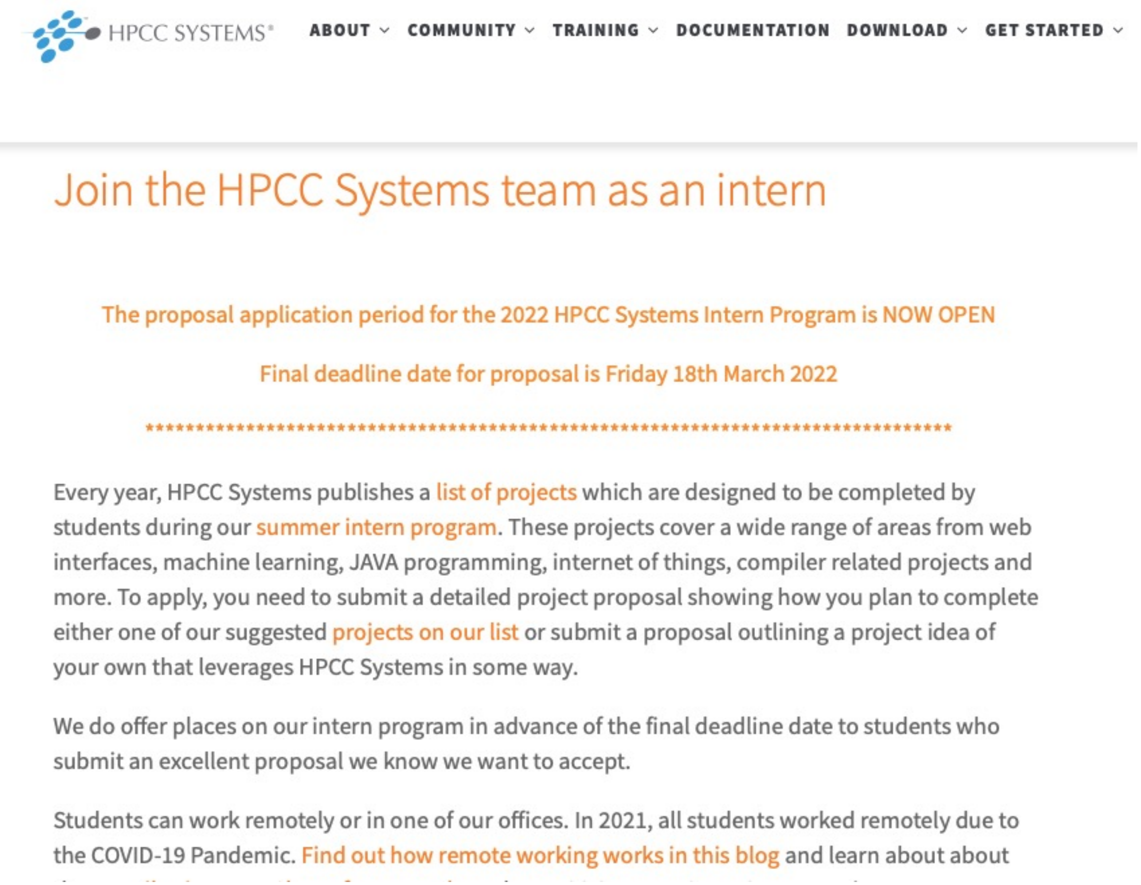
What is HPCC Systems?

- End to end big data analytics solution – but it doesn’t have to be
- First created in 2000, went open source in 2011 – hpccsystems.com
- Underpins a number of our own data driven business initiatives
- Used in a wide variety of ways – IoT, academic research, business
- Thor Cluster – Data Refinery
- ROXIE Cluster – Data Delivery
- Queries coded in ECL (Enterprise Control Language) training available
- Contribute to the platform / machine learning library /use case
- Cloud native platform - https://wiki.hpccsystems.com/x/FYD0B

HPCC Systems Intern Program

- 12 weeks paid program
- High school through to PhD
- Global – Asia, US and Europe
- May through August
- Remote or LN office based (Alpharetta, GA or Boca Raton, FL)
- Proposal period deadline: 18th March 2022
- Review panel decides
- Mentoring
- Community involvement

Read my blog hpccsystems.com/intern



Join the HPCC Systems team as an intern

The proposal application period for the 2022 HPCC Systems Intern Program is NOW OPEN

Final deadline date for proposal is Friday 18th March 2022

****

Every year, HPCC Systems publishes a list of projects which are designed to be completed by students during our summer intern program. These projects cover a wide range of areas from web interfaces, machine learning, JAVA programming, internet of things, compiler related projects and more. To apply, you need to submit a detailed project proposal showing how you plan to complete either one of our suggested projects on our list or submit a proposal outlining a project idea of your own that leverages HPCC Systems in some way.

We do offer places on our intern program in advance of the final deadline date to students who submit an excellent proposal we know we want to accept.

Students can work remotely or in one of our offices. In 2021, all students worked remotely due to the COVID-19 Pandemic. Find out how remote working works in this blog and learn about about

Class of 2021 Projects - Use Cases

- Ingestion/Analysis of collegiate women’s basketball GPS data using HPCC Systems and RealBI
- COVID-19 Tracker and Global Map Improvements
- Processing Robotics Data with and HPCC Systems Cluster on Kubernetes
- Improvements to the HPCC Systems Structured Query Language (HSQL)

Class of 2021 Projects – Cloud Related

- Using Azure Spot Instances
- Ingress Configuration
- Apply Docker Image Build and Kubernetes Security Principles

Class of 2021 Projects – Machine Learning

- Implement a PMML Processor
- Toxicity Detection Machine Learning Project
- Causality – Probabilities and Conditional Probabilities
- Causality– Independence, Conditional Independence and Directionality
- Causality – Counterfactual and Interventional Layers
- Causality 2021 Project Details

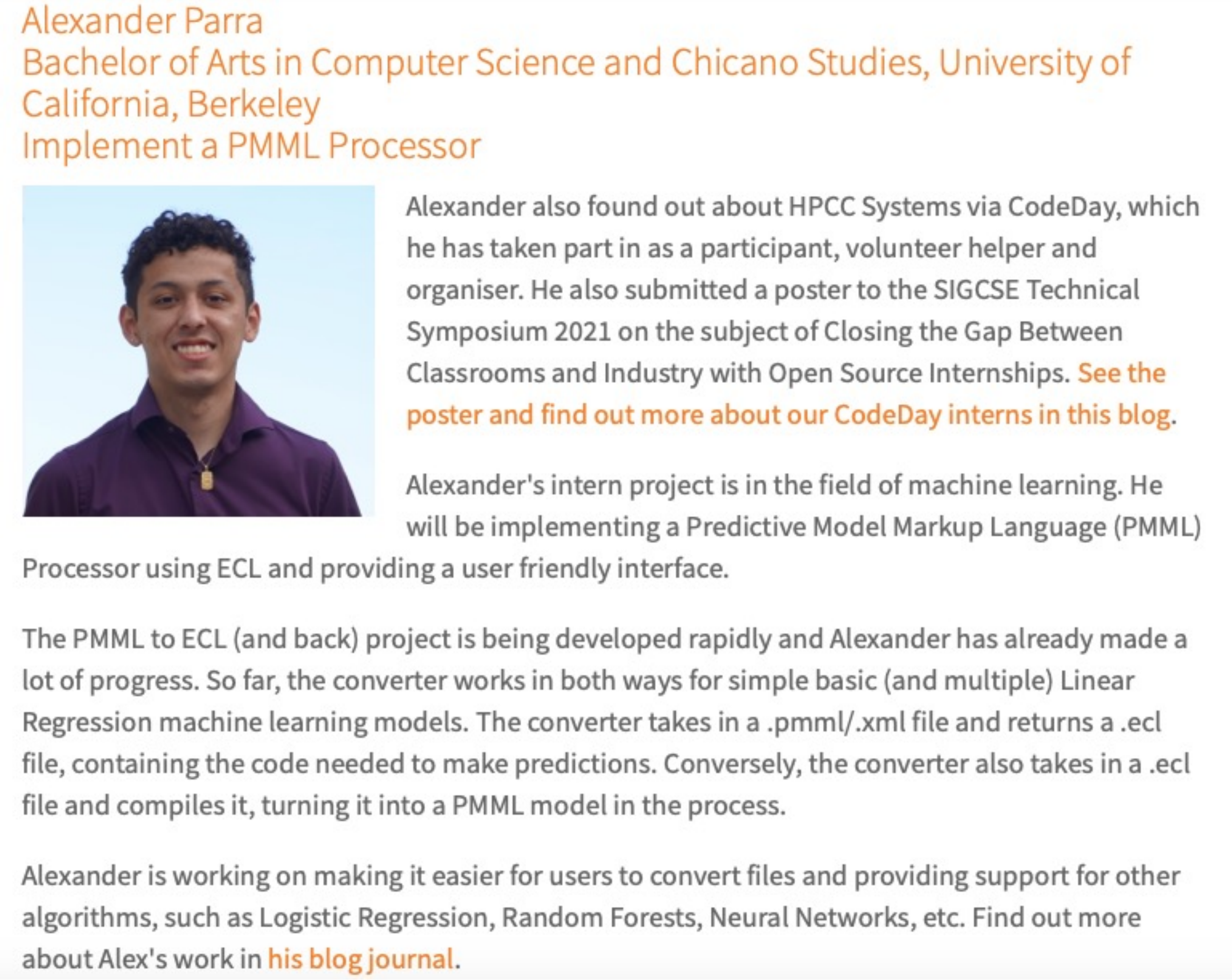
  https://hpccsystems.com/blog/causality-2021

Class of 2021 Projects – Machine Learning

- Implement a PMML Processor
- Toxicity Detection Machine Learning Project
- Causality – Probabilities and Conditional Probabilities
- Causality– Independence, Conditional Independence and Directionality
- Causality – Counterfactual and Interventional Layers
- Ingestion/Analysis of collegiate women’s basketball GPS data using HPCC Systems and RealBI
- COVID-19 Tracker and Global Map Improvements
- Processing Robotics Data with and HPCC Systems Cluster on Kubernetes

Learn more about the 12 students who joined our program in 2021

https://hpccsystems.com/blog/intern-intro-2021



Alexander Parra
Bachelor of Arts in Computer Science and Chicano Studies, University of California, Berkeley
Implement a PMML Processor

Alexander also found out about HPCC Systems via CodeDay, which he has taken part in as a participant, volunteer helper and organiser. He also submitted a poster to the SIGCSE Technical Symposium 2021 on the subject of Closing the Gap Between Classrooms and Industry with Open Source Internships. See the poster and find out more about our CodeDay interns in this blog.

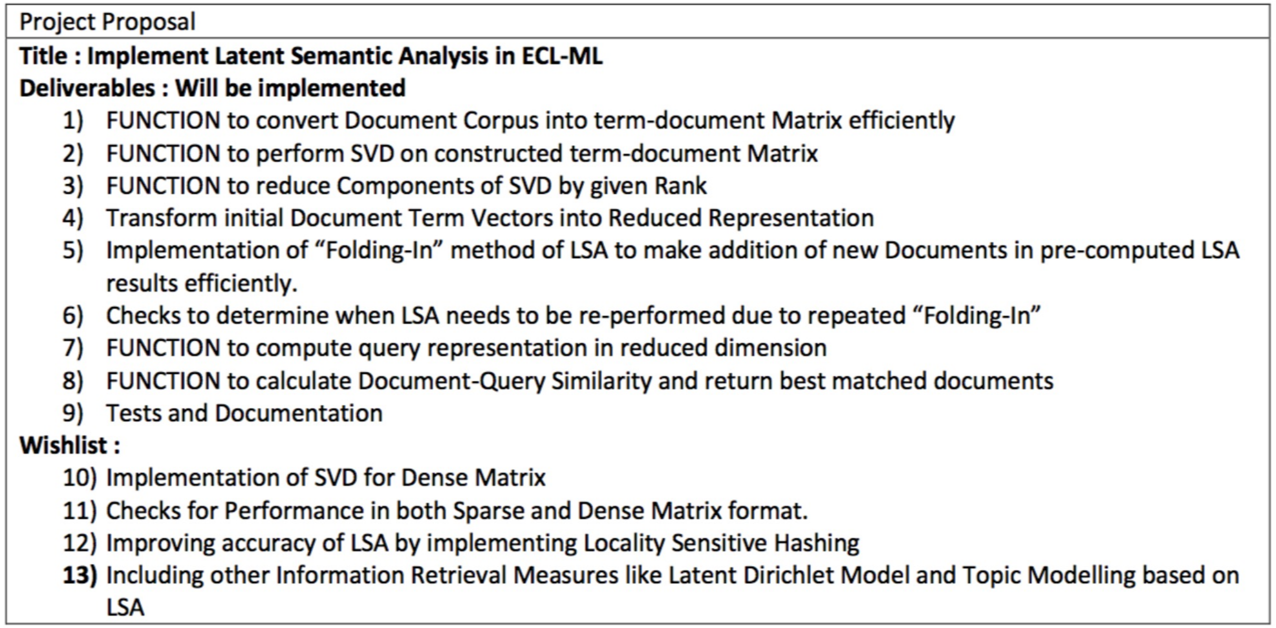
Alexander's intern project is in the field of machine learning. He will be implementing a Predictive Model Markup Language (PMML) Processor using ECL and providing a user friendly interface.

The PMML to ECL (and back) project is being developed rapidly and Alexander has already made a lot of progress. So far, the converter works in both ways for simple basic (and multiple) Linear Regression machine learning models. The converter takes in a .pmml/.xml file and returns a .ecl file, containing the code needed to make predictions. Conversely, the converter also takes in a .ecl file and compiles it, turning it into a PMML model in the process.

Alexander is working on making it easier for users to convert files and providing support for other algorithms, such as Logistic Regression, Random Forests, Neural Networks, etc. Find out more about Alex's work in his blog journal.

Project Proposal

- Find out more:
  hpccsystems.com/student-wiki
- Available Projects List
  hpccsystems.com/ideas-list
- Highlight the main tasks
- Timeline of work for each week
- Challenges and solutions
- Liaise with the project mentor



Project Proposal

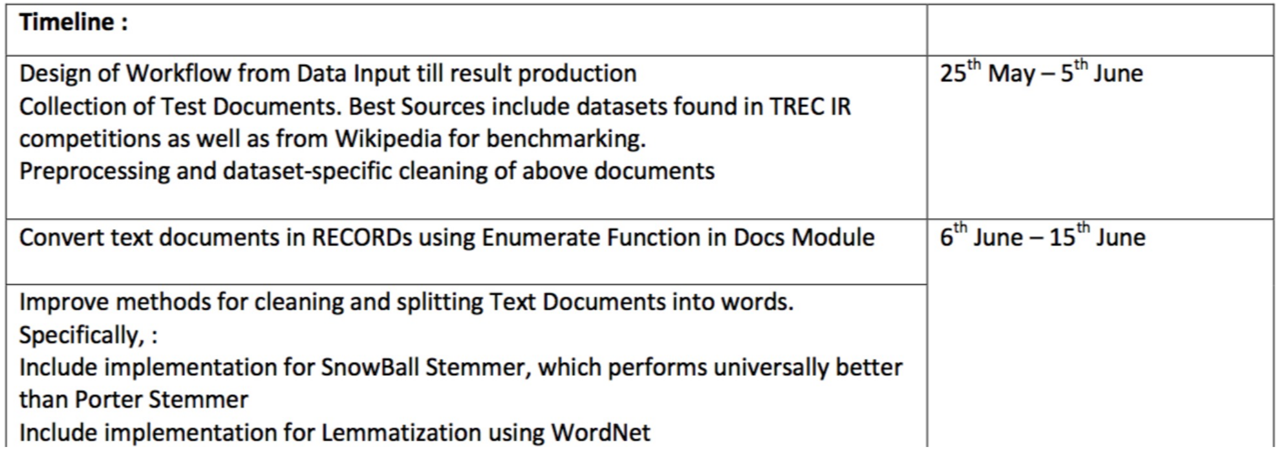
Title : Implement Latent Semantic Analysis in ECL-ML

Deliverables : Will be implemented

1) FUNCTION to convert Document Corpus into term-document Matrix efficiently
2) FUNCTION to perform SVD on constructed term-document Matrix
3) FUNCTION to reduce Components of SVD by given Rank
4) Transform initial Document Term Vectors into Reduced Representation
5) Implementation of “Folding-In” method of LSA to make addition of new Documents in pre-computed LSA results efficiently.
6) Checks to determine when LSA needs to be re-performed due to repeated “Folding-In”
7) FUNCTION to compute query representation in reduced dimension
8) FUNCTION to calculate Document-Query Similarity and return best matched documents
9) Tests and Documentation

Wishlist :

10) Implementation of SVD for Dense Matrix
11) Checks for Performance in both Sparse and Dense Matrix format.
12) Improving accuracy of LSA by implementing Locality Sensitive Hashing
13) Including other Information Retrieval Measures like Latent Dirichlet Model and Topic Modelling based on LSA

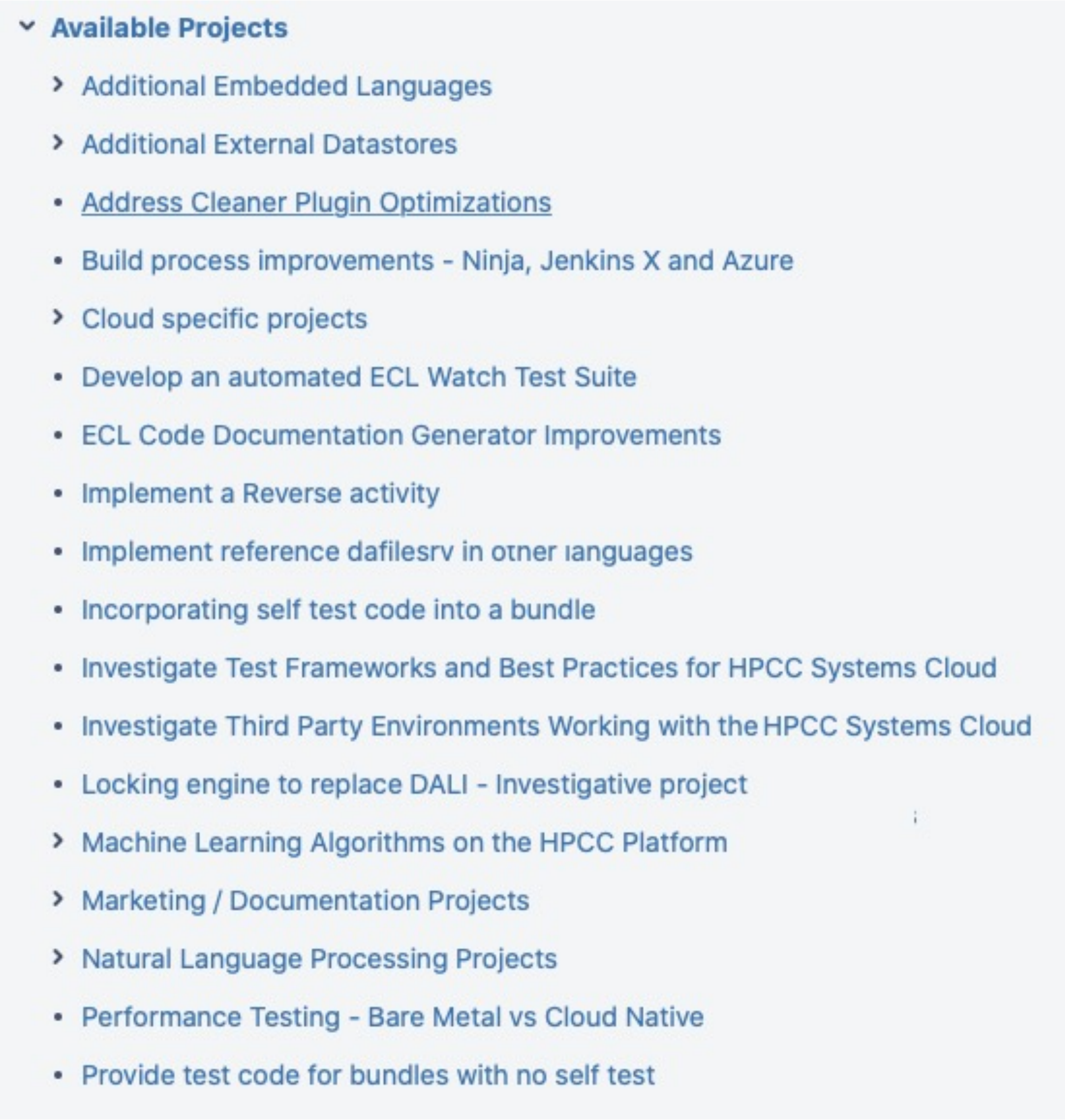


Timeline :	
Design of Workflow from Data Input till result production Collection of Test Documents. Best Sources include datasets found in TREC IR competitions as well as from Wikipedia for benchmarking. Preprocessing and dataset-specific cleaning of above documents	25th May – 5th June
Convert text documents in RECORDs using Enumerate Function in Docs Module	6th June – 15th June
Improve methods for cleaning and splitting Text Documents into words. Specifically, : Include implementation for SnowBall Stemmer, which performs universally better than Porter Stemmer Include implementation for Lemmatization using WordNet	

Research and Development can be Unpredictable

- We don’t expect you to know everything
- We don’t know everything
- The only stupid question is the one you don’t ask!
- Expect the unexpected
- Support from great mentors
- You can do it!

HPCC SYSTEMS®



Available Projects

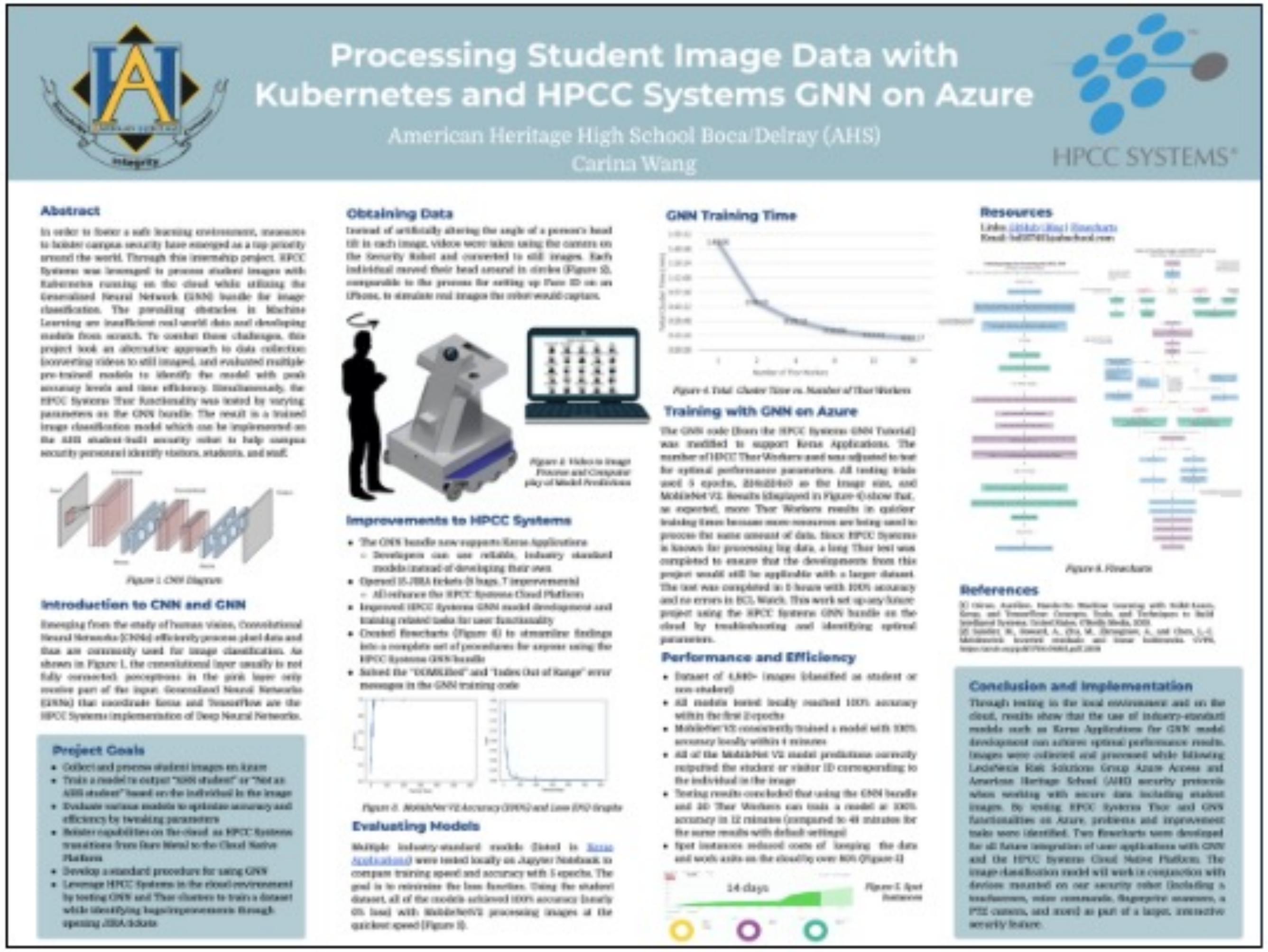
- Additional Embedded Languages
- Additional External Datastores
- Address Cleaner Plugin Optimizations
- Build process improvements - Ninja, Jenkins X and Azure
- Cloud specific projects
- Develop an automated ECL Watch Test Suite
- ECL Code Documentation Generator Improvements
- Implement a Reverse activity
- Implement reference dafilesrv in otner ıanguages
- Incorporating self test code into a bundle
- Investigate Test Frameworks and Best Practices for HPCC Systems Cloud
- Investigate Third Party Environments Working with the HPCC Systems Cloud
- Locking engine to replace DALI - Investigative project
- Machine Learning Algorithms on the HPCC Platform
- Marketing / Documentation Projects
- Natural Language Processing Projects
- Performance Testing - Bare Metal vs Cloud Native
- Provide test code for bundles with no self test

Mentoring and Support

- Experienced RELX Colleagues, School Teachers and Professors
- Subject Matter Experts and Wellbeing
- Evaluations – Mid Term and Final
- Progress Reports - Weekly
- Daily stand-up call – First month (and beyond)
- Email, Gitter, Teams/Zoom etc
- Intern Chat and Share

Opportunities to Share Your Work

- Personal Blog Journal
- HPCC Systems Blog
- Presentations
- Poster Contest
- Community Day
- Social Media Channels
- HPCC Systems Website
- GitHub Repository



Processing Student Image Data with Kubernetes and HPCC Systems GNN on Azure

American Heritage High School Boca/Delray (AHS)
Carina Wang

Abstract

In order to foster a safe learning environment, measures to bolster campus security have emerged as a top priority around the world. Through this internship project, HPCC Systems was leveraged to process student images with Kubernetes running on the cloud while utilizing the Generalized Neural Network (GNN) bundle for image classification. The prevailing obstacles in Machine Learning are insufficient real-world data and developing models from scratch. To combat these challenges, this project took an alternative approach to data collection (converting videos to still imaged), and evaluated multiple pre-trained models to identify the model with peak accuracy levels and time efficiency. Simultaneously, the HPCC Systems Thor functionality was tested by varying parameters on the GNN bundle. The result is a trained image classification model which can be implemented on the AHS student-built security robot to help campus security personnel identify visitors, students, and staff.

Figure 1. CNN Diagram

Introduction to CNN and GNN

Emerging from the study of human vision, Convolutional Neural Networks (CNNs) efficiently process pixel data and thus are commonly used for image classification. As shown in Figure 1, the convolutional layer usually is not fully connected; perceptrons in the pink layer only receive part of the input. Generalized Neural Networks (GNNs) that coordinate Keras and TensorFlow are the HPCC Systems implementation of Deep Neural Networks.

Project Goals

- Collect and process student images on Azure
- Train a model to output "AHS student" or "Not an AHS student" based on the individual in the image
- Evaluate various models to optimize accuracy and efficiency by tweaking parameters
- Bolster capabilities on the cloud as HPCC Systems transitions from Bare Metal to the Cloud Native Platform
- Develop a standard procedure for using GNN
- Leverage HPCC Systems in the cloud environment by testing GNN and Thor clusters to train a dataset while identifying bugs/improvements through opening JIRA tickets

Obtaining Data

Instead of artificially altering the angle of a person's head tilt in each image, videos were taken using the camera on the Security Robot and converted to still images. Each individual moved their head around in circles (Figure 2), comparable to the process for setting up Face ID on an iPhone, to simulate real images the robot would capture.

Figure 2. Video to Image Process and Computer play of Model Predictions

Improvements to HPCC Systems

- The GNN bundle now supports Keras Applications
  - Developers can use reliable, industry standard models instead of developing their own
- Opened 15 JIRA tickets (8 bugs, 7 improvements)
  - All enhance the HPCC Systems Cloud Platform
- Improved HPCC Systems GNN model development and training related tasks for user functionality
- Created flowcharts (Figure 6) to streamline findings into a complete set of procedures for anyone using the HPCC Systems GNN bundle
- Solved the "DOMLEfied" and "Index Out of Range" error messages in the GNN training code

Figure 3. MobileNet V2 Accuracy (100%) and Loss (0%) Graphs

Evaluating Models

Multiple industry-standard models (listed in Keras Applications) were tested locally on Jupyter Notebook to compare training speed and accuracy with 5 epochs. The goal is to minimize the loss function. Using the student dataset, all of the models achieved 100% accuracy (nearly 0% loss) with MobileNetV2 processing images at the quickest speed (Figure 3).

GNN Training Time

Figure 4. Total Cluster Time vs. Number of Thor Workers

Training with GNN on Azure

The GNN code (from the HPCC Systems GNN Tutorial) was modified to support Keras Applications. The number of HPCC Thor Workers used was adjusted to test for optimal performance parameters. All testing trials used 5 epochs, 224x224x3 as the image size, and MobileNet V2. Results (displayed in Figure 4) show that, as expected, more Thor Workers results in quicker training times because more resources are being used to process the same amount of data. Since HPCC Systems is known for processing big data, a long Thor test was completed to ensure that the developments from this project would still be applicable with a larger dataset. The test was completed in 8 hours with 100% accuracy and no errors in ECL Watch. This work set up any future project using the HPCC Systems GNN bundle on the cloud by troubleshooting and identifying optimal parameters.

Performance and Efficiency

- Dataset of 4,880+ images (classified as student or non-student)
- All models tested locally reached 100% accuracy within the first 2 epochs
- MobileNet V2 consistently trained a model with 100% accuracy locally within 4 minutes
- All of the MobileNet V2 model predictions correctly outputted the student or visitor ID corresponding to the individual in the image
- Testing results concluded that using the GNN bundle and 20 Thor Workers can train a model at 100% accuracy in 12 minutes (compared to 48 minutes for the same results with default settings)
- Spot instances reduced costs of keeping the data and work units on the cloud by over 80% (Figure 5)

Figure 5. Spot Instances

Resources

Links: GitHub | Blog | Flowcharts
Email: [illegible]

Figure 6. Flowcharts

References

[1] Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. United States, O'Reilly Media, 2019.
[2] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. CVPR, [illegible], 2018

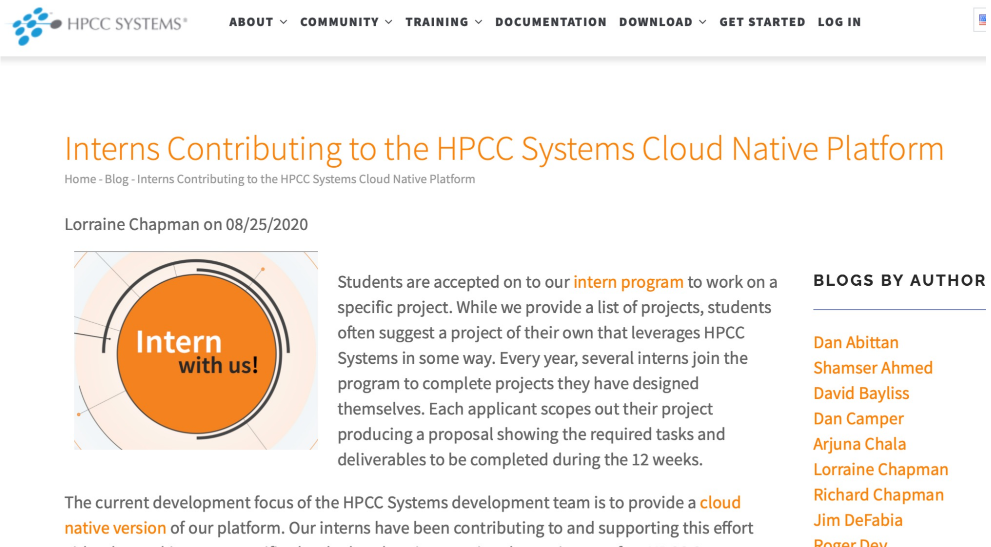
Conclusion and Implementation

Through testing in the local environment and on the cloud, results show that the use of industry-standard models such as Keras Applications for GNN model development can achieve optimal performance results. Images were collected and processed while following LexisNexis Risk Solutions Group Azure Access and American Heritage School (AHS) security protocols when working with secure data including student images. By testing HPCC Systems Thor and GNN functionalities on Azure, problems and improvement tasks were identified. Two flowcharts were developed for all future integration of user applications with GNN and the HPCC Systems Cloud Native Platform. The image classification model will work in conjunction with devices mounted on our security robot (including a touchscreen, voice commands, fingerprint scanners, a PTZ camera, and more) as part of a larger, interactive security feature.

Personal Development Experience

- Office/teamworking experience
- Project management
- Communication skills
- Research and development real world experience
- Bleeding edge development
- Confidence in decision making
- Use of developer tools and processes – GitHub, JIRA, code editors, testing, checking in code etc
- Contribute to an active open-source community

Be Part of the Team and Make Your Mark



Interns Contributing to the HPCC Systems Cloud Native Platform

Home - Blog - Interns Contributing to the HPCC Systems Cloud Native Platform

Lorraine Chapman on 08/25/2020

Students are accepted on to our intern program to work on a specific project. While we provide a list of projects, students often suggest a project of their own that leverages HPCC Systems in some way. Every year, several interns join the program to complete projects they have designed themselves. Each applicant scopes out their project producing a proposal showing the required tasks and deliverables to be completed during the 12 weeks.

The current development focus of the HPCC Systems development team is to provide a cloud native version of our platform. Our interns have been contributing to and supporting this effort either by wor

cluster on a s

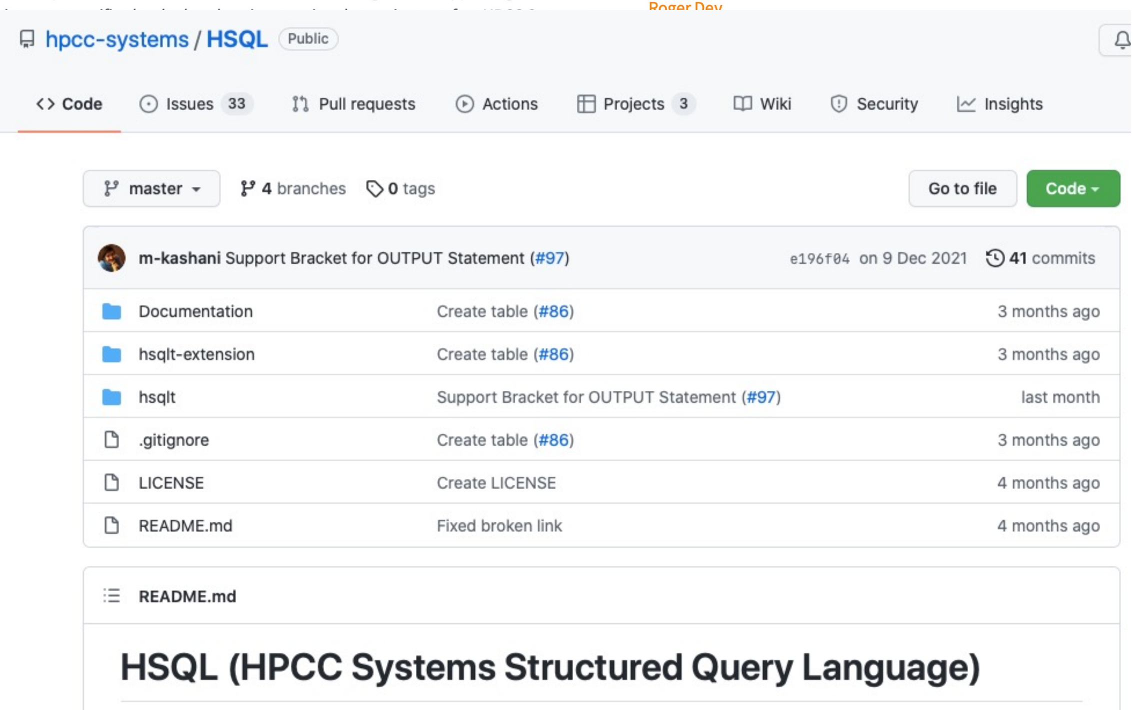
environment

BLOGS BY AUTHOR

Dan Abittan
Shamser Ahmed
David Bayliss
Dan Camper
Arjuna Chala
Lorraine Chapman
Richard Chapman
Jim DeFabia
Roger Dev

Testing

Since our nev



m-kashani Support Bracket for OUTPUT Statement (#97) e196f04 on 9 Dec 2021 41 commits

Documentation	Create table (#86)	3 months ago
hsqlt-extension	Create table (#86)	3 months ago
hsqlt	Support Bracket for OUTPUT Statement (#97)	last month
.gitignore	Create table (#86)	3 months ago
LICENSE	Create LICENSE	4 months ago
README.md	Fixed broken link	4 months ago

README.md

HSQL (HPCC Systems Structured Query Language)

Configuring a HashiCorp Vault as a Certificate Manager on HPCC Systems

Nikita Jha on 08/03/2021

Nikita Jha is a high school student, who joined the 2021 HPCC Systems Intern Program to complete a project focusing on applying docker image build and Kubernetes security principles to our new Cloud Native platform. This tutorial style blog covers the importance of certificate management and provides instructions for setting up and configuring a Hashicorp Vault.

Future Employment Opportunities

- Open positions at LexisNexis Risk Solutions Group: https://risk.lexisnexis.com/group/careers
- Previous interns have a head start and we can vouch for you
- Hired 2 interns from the 2021 program
- Not all interns are ready for immediate employment
- Intern for a second time or a third…

Diversity and Inclusion at LexisNexis Risk Solutions Group

- Priority for LN RSG: https://risk.lexisnexis.com/group/diversity
- Reduce disparities in employment, compensation and progression
- Unconscious bias training
- Employee Resource Groups – Forums for women, pride, disability, mental health and many more…
- Listed in the Bloomberg Gender Equality Index - RELX PLC

Learn more…

- Student wiki - hpccsystems.com/student-wiki
- Available Projects List - hpccsystems.com/ideas-list
- Student journals and presentations https://wiki.hpccsystems.com/x/GQFdB
- Read about the program - hpccsystems.com/interns/
- Student Testimonials https://wiki.hpccsystems.com/x/KoNc
- HPCC Systems GitHub Repository https://github.com/hpcc-systems
- Student Posters https://wiki.hpccsystems.com/display/hpcc/HPCC+Systems+Technical+Presentations

HPCC SYSTEMS®

Get in Touch

Lorraine.Chapman@lexisnexisrisk.com