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](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

3
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
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:

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

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 other languages
  • 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
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

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
Get in Touch

Lorraine.Chapman@lexisnexisrisk.com