HPCC SYSTEMS

Summer Internship Program

JANUARY 2024

Hugo Watanuki
LexisNexis Risk Solutions

Contents

HPCC SYSTEMS

The Business

LexisNexis Risk Solutions is Part of RELX

RELX

RELX is a global provider of information-based analytics and decision tools for professional and business customers, enabling them to make better decisions, get better results and be more productive. The Group serves customers in more than 180 countries and has offices in about 40 countries.

It employs over 33,000 people, of whom almost half are in North America.

Learn more at www.relx.com

RELX operates in four major market segments:

Scientific, Technical & Medical

ELSEVIER

Risk

LexisNexis® RISK SOLUTIONS

Exhibitions

RX

Legal

LexisNexis® | Legal & Professional

HPCC SYSTEMS

What we do

We leverage five main capabilities to provide end-to-end solutions that help customers assess risk and opportunity.

Vast Data Resources

We maintain over 12 petabytes of content comprising billions of public and proprietary records.

Big Data Technology

We designed our massively-scalable super-computing platform, HPCC Systems®, enabling us to process at very high speeds – over 270 million transactions per hour.

Advanced Linking

We use our own unique identifier, LexID®, together with a proprietary linking technology. Our patented linking and clustering method is the engine behind many of our products.

Sophisticated Analytics & Insight Capability

We apply data science and leverage patented algorithms, predictive modeling, machine learning and AI to provide data driven solutions and better decision intelligence.

Industry-Specific Expertise & Delivery

The people in our businesses have deep industry experience and expertise – we employ professionals that worked in the industries we serve, so they have walked in the shoes of our customers.

Customer-Focused Solutions

We connect the dots between public records and transactions, resulting in actionable information our customers use to advance their goals.

HPCC SYSTEMS

Our Customers

We work with Fortune 1000 and mid-market clients globally across industries, and federal and state governments.

- **Customers in more than 180 countries**
- **9 out of 10 of the world's top 10 banks**
- **78% of the Fortune 500 companies**
- **98 of the top 100 personal lines insurance carriers**
- **More than 7,500 federal, state and local government agencies**

HPCC SYSTEMS

The Technology

Open Source Technology for Big Data Processing

HPCC Systems®

Born from the deep data analysis experience of LexisNexis Risk Solutions, HPCC Systems is a comprehensive, dedicated cloud-native platform that makes combining data stored in massive, mixed schema data lakes easier and faster. The platform scales very quickly as your data needs grow enabling companies of all sizes to save time and money, now and in the future.

www.hpccsystems.com video

Single, Complete Platform

Two data engines (query and refinery) operating at a high level of speed and accuracy

Cloud-Native Architecture

Automation of Kubernetes makes it easy to set-up, manage and scale your data

Parallel Programming Language (ECL)

High level, data-centric declarative programming language

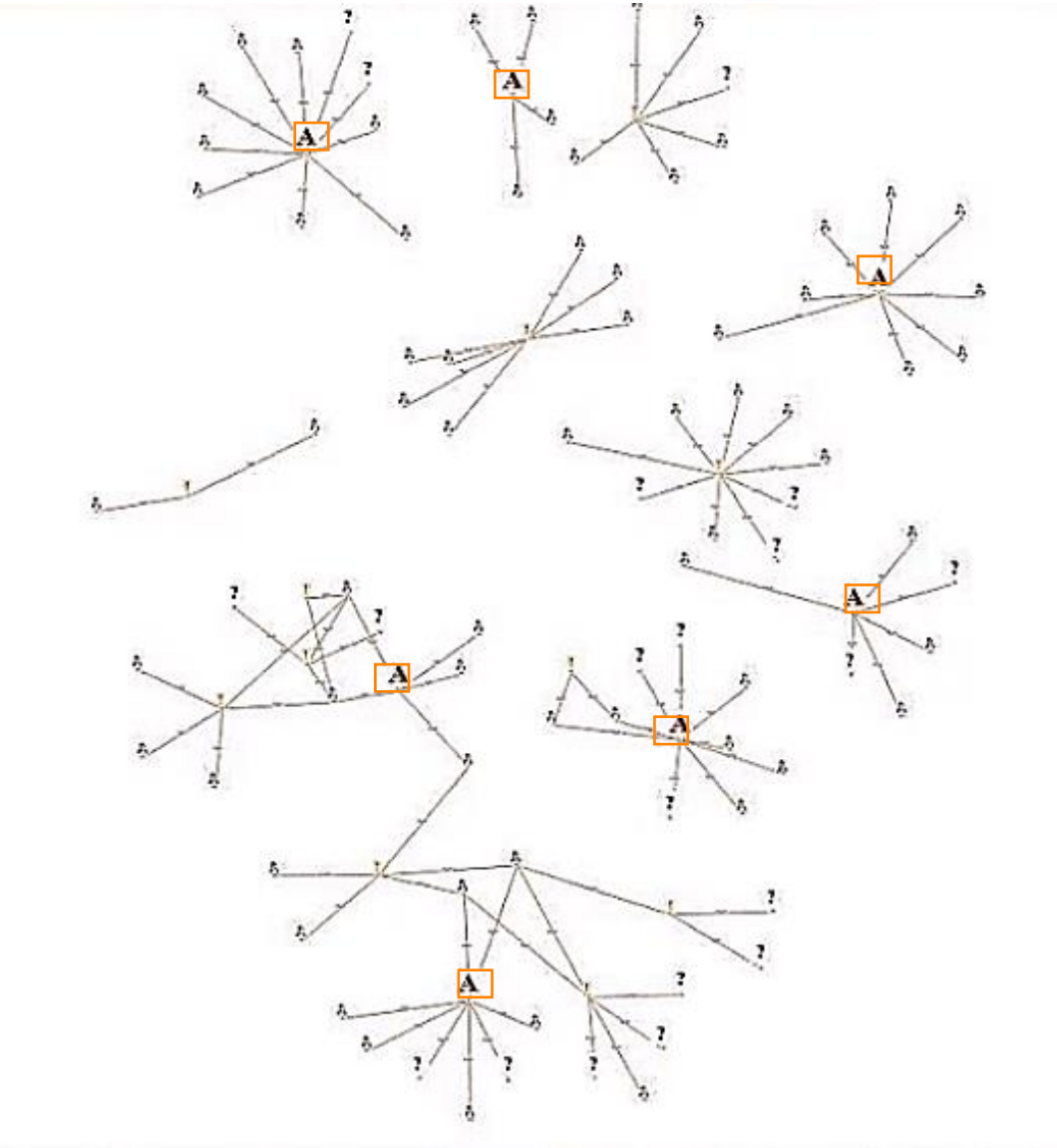
Data Management Tools

Data is easier to manage with robust profiling, cleansing, update and consolidation tools

Machine Learning Library

Predictive modeling functionality for machine learning performance

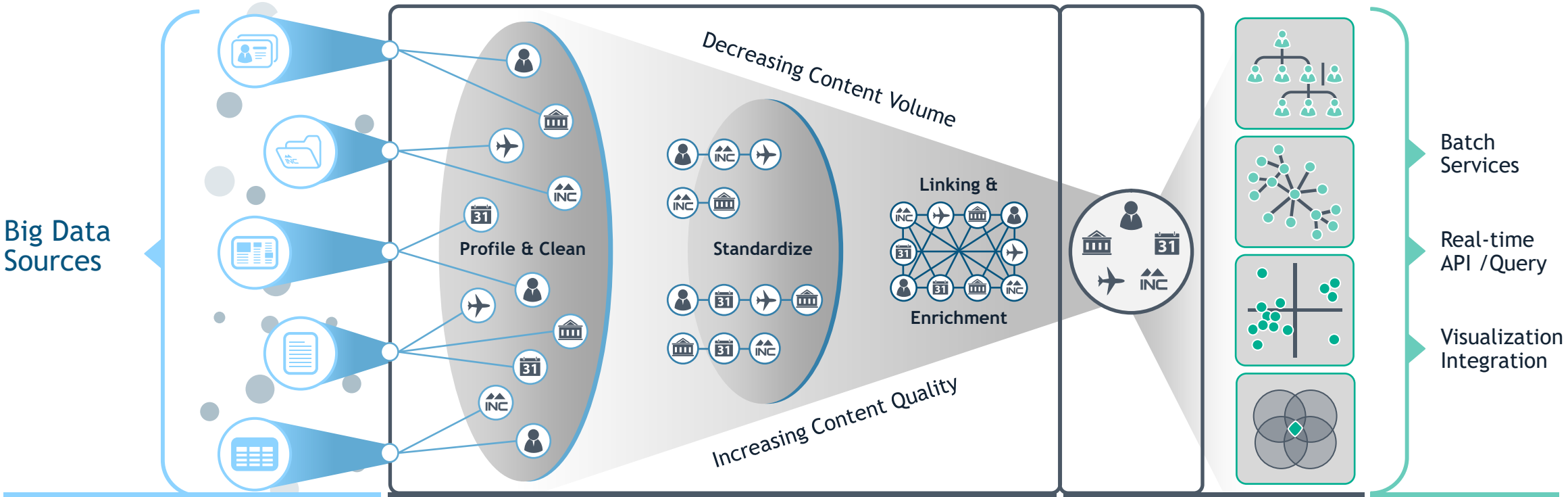
Insurance Fraud Use Case



HPCC SYSTEMS

HPCC Systems Data Enrichment Pipeline

Integrate › Profile › Clean & Standardize › Linking & Enrichment › Index Build › API Publish › Deliver



Unstructured and Structured Content

- Over 12 petabytes of content
- More than 10,000 sources
- 7.5 billion unique name and address combinations
- 7.3 billion motor vehicle records

Thor HPCC

- Distributed Computing
- Data-centric language (ECL)
- An integrated delivery system that offers data plus analytics
- 86 billion records processed through Thor

ROXIE HPCC

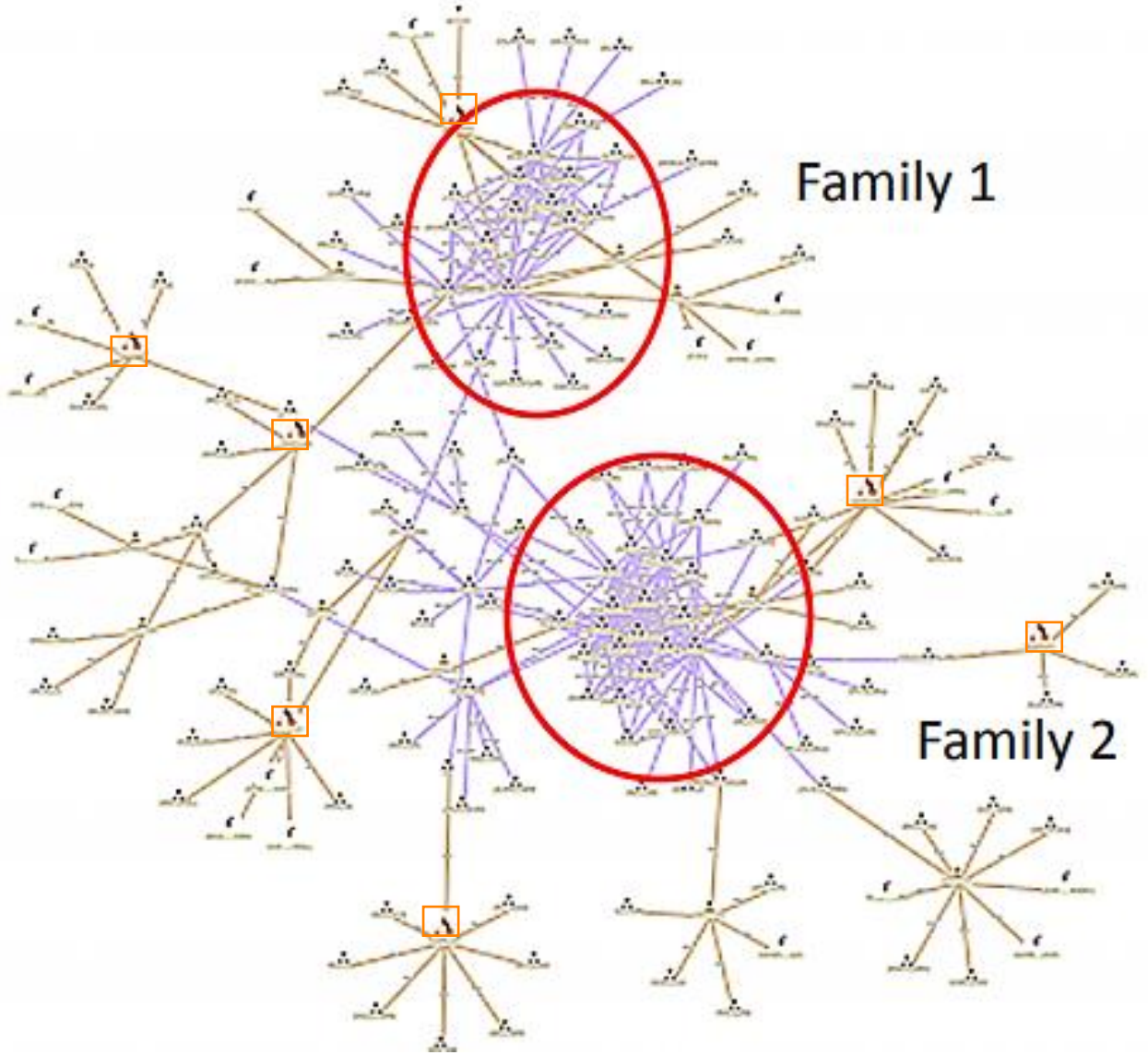
- Complex Searching and Matching
- Approximately 400+ models deployed
- Deliver 270 million ROXIE calls every hour

Products

- LexisNexis Risk Products

HPCC SYSTEMS

Insurance Fraud Use Case



HPCC SYSTEMS

The Internship

HPCC Systems Internship

Are you interested in big data?

Coding in an open language?

Working with machine learning algorithms?

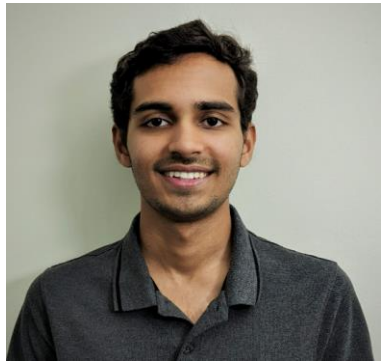
Or contributing to a cloud native platform?

- 12-week paid and remote program over the summer
- Mentor-based and focused on open source HPCC Systems projects
- Open to students from high school to PhD levels that are authorized to work
- Proposal-based application, either original or from our list
- Visit our student wiki: hpccsystems.com/student-wiki

Questions can be emailed to **students@hpccsystems.com**

HPCC™ SYSTEMS

MSc Student Project



NC STATE UNIVERSITY

Applying Causality Toolkit to Real-world Datasets

Arun Gaonkar
Mentor : Roger Dev

HPCC SYSTEMS
LexisNexis RISK SOLUTIONS

Introduction

Everything in this universe happens for a reason and every action has a reaction.
Analyzing causality can help in medical diagnostic analysis, time series analysis, and strategic planning
In real-world datasets, variables are inter-related, implying subtle correlations, which makes causal analysis difficult.

Causal Toolkits

1. HPCC_Causality
2. Because
   - Visualization bundle
   - Dependence & Independence tests
   - Causal Direction Tests

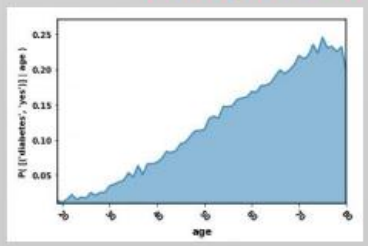
Analysis Steps

1. Finding & Analyzing Dataset
2. Pre-processing the dataset
3. Propose a Causal Hypothesis
4. Applying causal toolkits & analyzing
5. Interpretation & Causal Model
6. Hypothesis & Model Verification

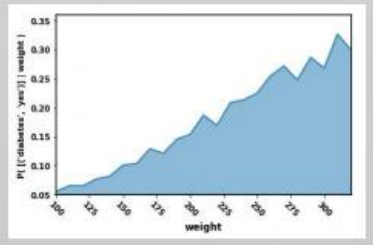
Causal Hypothesis Question

Proposed causal hypothesis question:
"What factors can influence the likelihood of a person having Diabetes?"

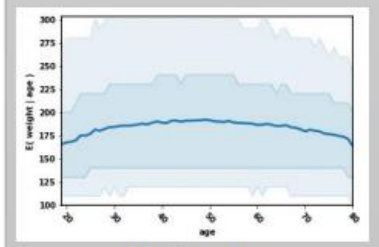
Analysis LLCP-CDC Dataset



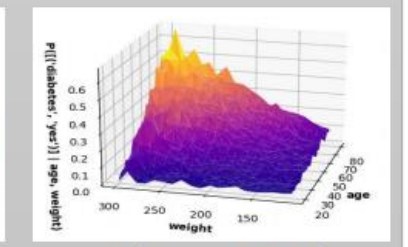
Diabetes vs Age



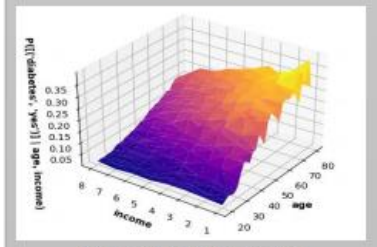
Diabetes vs Weight



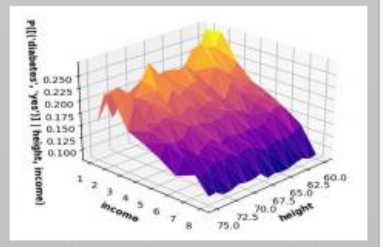
Weight vs Age



Diabetes vs Age, Weight

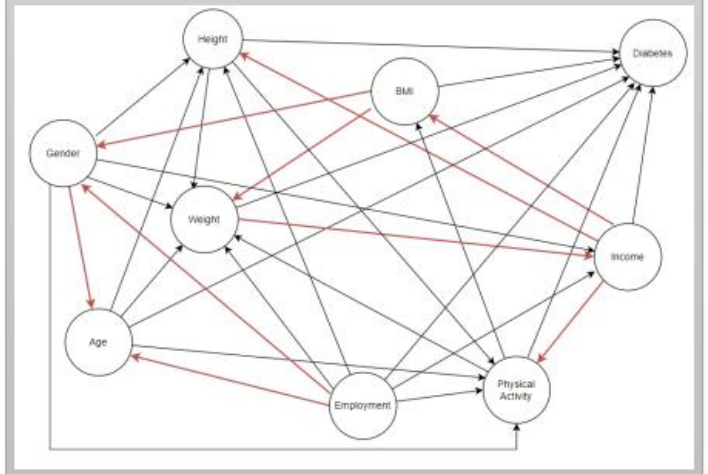


Diabetes vs Age, Income



Diabetes vs Height, Income

Causal Model



Conclusion

Factors like Age, Height, Weight, Income, type of Employment, Physical Activity, and gender have their effect on Diabetes.

- Most of the relations are practically and analytically correct.
- Some relations are unexpected, but probable valid proof can be generated.
- For some other relations, any explanation is rationally invalid.

Causality toolkit can be applied to analyze real-world datasets. But the cause-effect of latent variables cannot be incorporated into the causal model.

HPCC SYSTEMS

Undergraduate Student Project

TEST SUITE FOR A ROXIE CLUSTER ON KUBERNETES

NIVEDHA SIVAKUMAR | GEORGIA STATE UNIVERSITY |
MENTORS: KRISHNA TURLAPATHI & ATTILA VAMOS

ABSTRACT

HPCC Systems is an open-source platform offering other businesses high-performance data processing and analytics.

- Thor cluster is responsible for manipulating massive amounts of data
- Roxie cluster supports high-performance data delivery applications using indexed data files.

The motivation of my project is to create a test suite for Roxie designed to provide a more in-depth understanding of how different query, cluster, and infrastructure configurations can affect the functionality and performance of Roxie in the cloud.

PROJECT GOALS

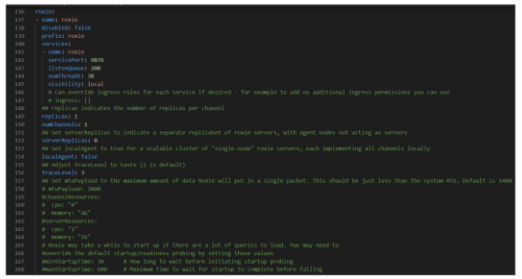
The primary goal for the project is to develop a test suite for Roxie in Kubernetes

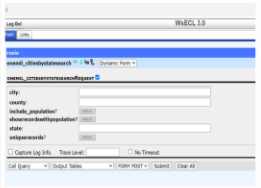
The results provided by the test suite can give indications or guidelines as to what configuration will be suitable for each use case of Roxie in the future. The test cases will be executed to simulate different usage patterns and can help identify potential problems or areas for improvement within the cluster which can be particularly useful during new releases of HPCC Systems or changes in the infrastructure

METHODOLOGY

Test Execution

- Cloned terraform repo made by Godji, made changes in the esp.yaml file. This portion of the Roxie configuration needs to be added to the YAML file and used the default Roxie configuration which was 4 CPUs, and 3 nodes Roxie.
- Started with running the HPCC System cluster on AKS with two different storage types (Premium & Standard).
- Execute ***kubectl get svc*** command to obtain the ECL Watch and the ECL Queries page info
- ***ecl.exe publish --target=roxie --server=<eclqueries external IP> --no-reload --no-files --wait=7200000 -v "<query file directory>" --name=<name of the file> --comment="Test" -l "<query folder directory>".***





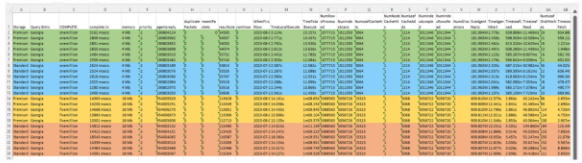
Log File Information

- Once the results are displayed, I will then execute ***kubectl logs (podname) > filename.log*** command
- But here's the problem, we were not able to capture the proper information we needed.
- As a result, we had to develop our own Python script that would extract only the information we needed and format them properly in Excel.

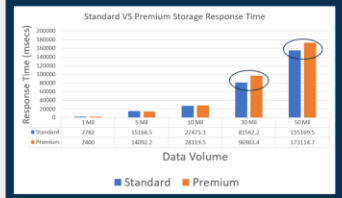
DATA COLLECTION

- One challenge was making sure some numbers were consistent throughout the test cases to compare the results.
- If the numbers didn't match up, we had to go back and execute the test cases again, or try to understand why the difference is occurring.
- During our process, of data analysis, we mainly focus on the metrics response time, and local time execute

GitHub Repo
https://github.com/Nivedha2/Test_Cases/tree/master



RESULTS



The graph displays the results of response time for each query in msecs. You may notice premium was slightly better than standard with smaller datasets, but had higher response time in larger datasets. As of right now, we are not sure why it is happening, but it will be discussed further with Microsoft.

CONCLUSION & NEXT STEPS

•Conclusion: Based on the tests performed so far, there has been no evidence that the premium storage offered by Azure provides enhanced performance for Roxie queries.

•Next steps:

•An opportunity exists to further enhance the set of test cases to cover more diversified types of Roxie queries (Such as involving different index types, ECL functions, data manipulation strategies, etc)

•Additional testing of different sizes of datasets is required to collaborate the current results with the next set of test results

ACKNOWLEDGMENTS

I would like to thank my mentors Krishna Turlapathi, Attila Vamos, Hugo Watanuki, and the HPCC Systems Team for their support and guidance throughout my internship project.

High School Student Project

HPCC Systems Storage Support with Container Storage Interface (CSI)

American Heritage School

Ryan Rao, HPCC Systems intern, American Heritage Palm Beach | Mentors: Xiaoming Wang, Godji Fortil

HPCC SYSTEMS

LexisNexis RISK SOLUTIONS

Introduction

AWS EFS and AWS FSx for Lustre are two powerful storage services that offer different use cases. EFS is similar to a standard NFS and has a more general usage, while FSx for Lustre is designed specifically for HPC applications due to its fast speed. As it is important to provide users a versatile range of storage options when deploying HPCC Systems, my project focused on implementing these two storages through CSI to allow HPCC Systems to utilize their different features. In the first phase of my project, I focused on developing a storage lifecycle beyond Kubernetes for EFS, and in the second phase, I focused on developing a storage lifecycle within Kubernetes for FSx for Lustre.

Project Goals

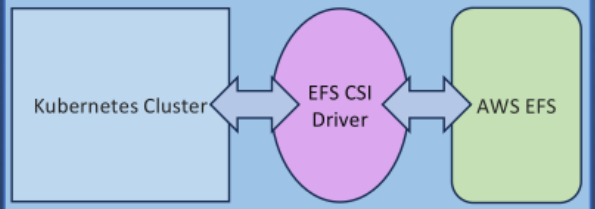
- Investigate and understand the three storage lifecycles
- Design and build the helm chart for the EFS lifecycle beyond Kubernetes
- Explore FSx for Lustre and any limitations
- Design and build the helm chart for the FSx for Lustre lifecycle within Kubernetes

Three Storage Lifecycles

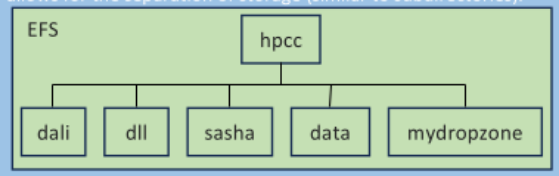
1. Storage lives on the HPCC Systems cluster level. It depends on the HPCC Systems cluster.
2. Storage lives within the Kubernetes cluster. It is independent of the HPCC Systems cluster.
3. Storage lives beyond the Kubernetes cluster. It is independent of the Kubernetes cluster.

Phase 1: EFS

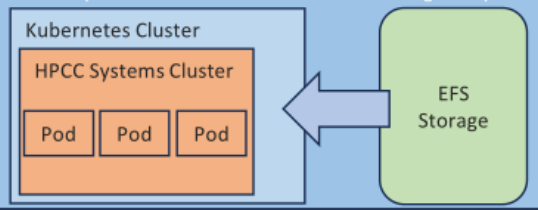
The user installs the **EFS CSI driver** by running a script. The EFS CSI driver allows the Kubernetes cluster communicate and interact with AWS resources, in this case, EFS.



The pre-created EFS file system uses **five manually created access points**, which act as isolated entry points into the file system. This allows for the separation of storage (similar to subdirectories).

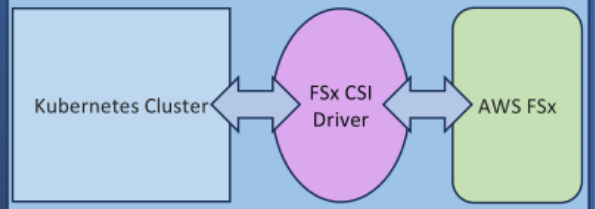


The **hpcc-efs-static-pv helm chart** deploys the required Kubernetes components, connecting the external EFS storage to the cluster. The HPCC Systems cluster can now use the 3rd storage lifecycle.

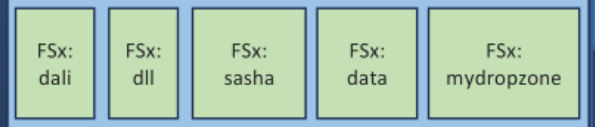


Phase 2: FSx for Lustre

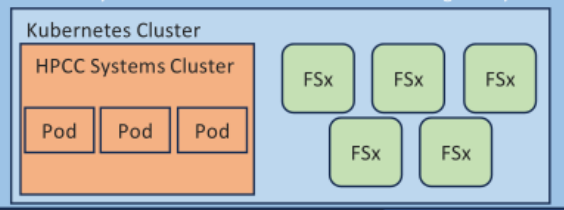
The user installs the **FSx CSI driver** by running a script. The FSx CSI driver allows the Kubernetes cluster to communicate and interact with AWS resources, in this case, FSx.



FSx doesn't offer an access point feature like EFS, so there is no way to have subdirectories within a single FSx file system. Instead, **five separate FSx file systems** are created, allowing for the separation of storage.



The **hpcc-fsx-dynamic-pv helm chart** deploys the required Kubernetes components to set up the FSx storage in Kubernetes. The HPCC Systems cluster can now use the 2nd storage lifecycle.



Results

- My PR for the EFS storage implementation has been approved for merging. It includes a new helm chart, improved documentation, and simplified code.
- In the next few weeks, I plan to create a PR for the FSx storage implementation, including a new helm chart, documentation, and various other files.
- FSx integrates with S3. After the data is imported from FSx to S3, the FSx volume can be temporarily shut down. This is something to explore in the future.

Conclusion

As new storage services emerge, it is important for HPCC Systems to integrate with these storage solutions, offering its users a range of choices. To this end, I hope my work with EFS and FSx for Lustre and any new features I have added serve to enhance the platform's capabilities.

Resources

Blog | EFS | FSx for Lustre

HPCC SYSTEMS

Internship Project Life Cycle

Onboarding and welcome	Development phase	Completion phase	HPCC Systems Community Day
• Welcome meeting • Infrastructure and access • Training	• Daily meetings • Code development • Testing • Blog reports	• Check-in code • Documentation • Team presentations • Poster submission	• Annual conference • Poster competition • Talk (encouraged!)

Application Process

1) Select your individual project:

- Available projects list
- Suggest your own project

2) Write a proposal:

- Liaise with a mentor
- Highlight the deliverables
- Timeline of work for each week

3) Submit the proposal and your CV:

- Apply online: RELX careers website

HPCC SYSTEMS

Important Dates

- Deadline for applications: March 22, 2024
- Notification of acceptance: up to mid-April, 2024
- Hiring and onboarding: late April-early May, 2024
- Start date: May 20, 2024
- End date: Aug 9, 2024

HPCC SYSTEMS

Get in Touch

Hugo.Watanuki@lexisnexisrisk.com

linkedin.com/in/hugowatanuki

HPCC SYSTEMS