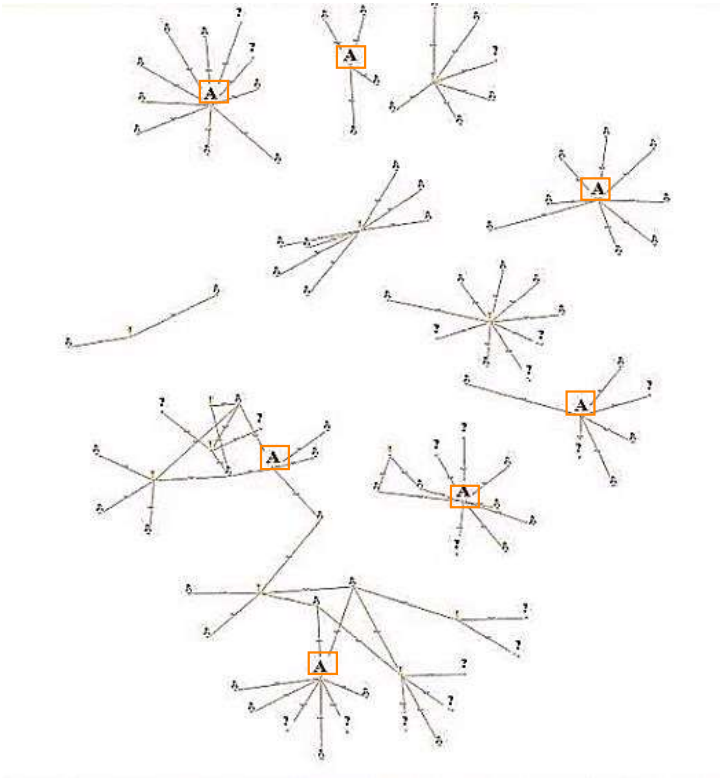
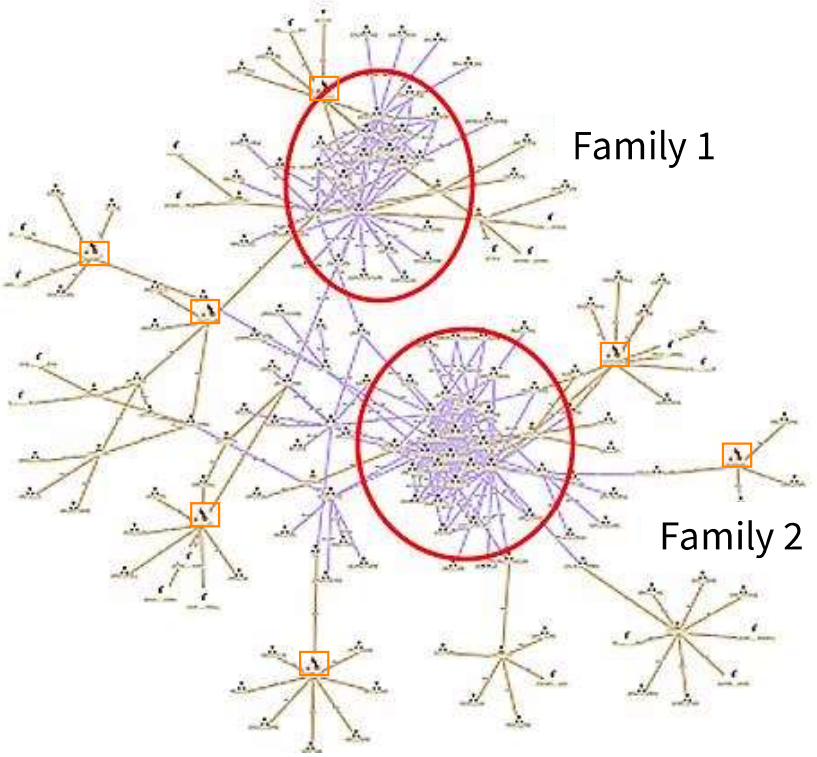

HPCC SYSTEMS
Internship Program
JANUARY 2023
Hugo Watanuki
LexisNexis Risk Solutions

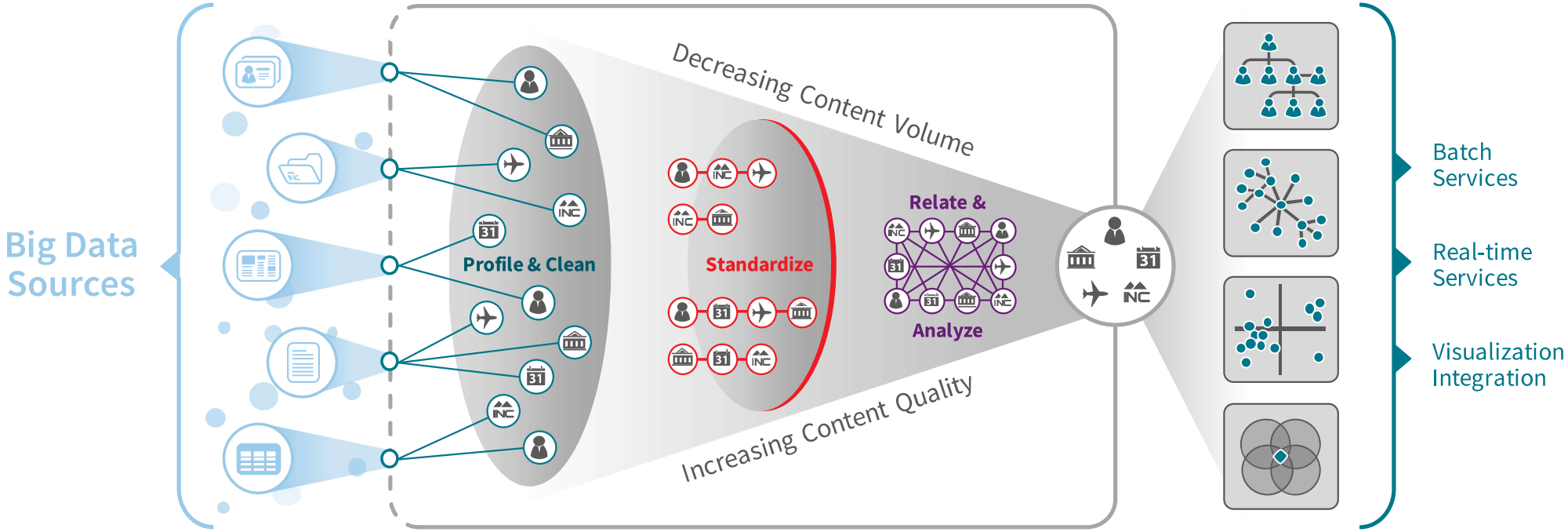
What We Do





HPCC SYSTEMS

Our Big Data Technology



HPCC SYSTEMS

HPCC Systems Internship Program

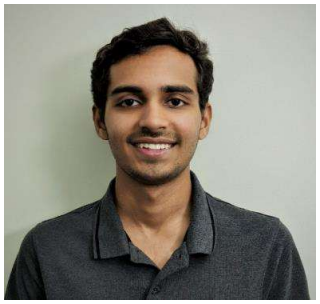
- **Are you interested in big data?**
- **Coding in an open language?**
- **Implementing machine learning algorithms?**
- **Or working on the architecture layer?**

If you answered yes to any of the above, our summer internship program might be a perfect fit!

- 12-week paid program over the summer
- Mentor-based and focused on open source HPCC Systems projects
- Onsite or remote based working options
- Open to undergraduate, Masters, and PhD students, as well as high school students
- Proposal-based application, either original or from our ideas list
- Proposal submission deadline is last week in March
- Scan the QR code to visit our Student **wiki**

HPCC SYSTEMS

Student Projects



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

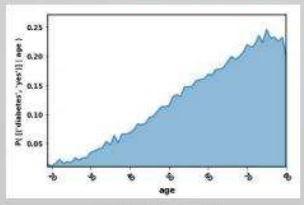
Analysis Steps

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

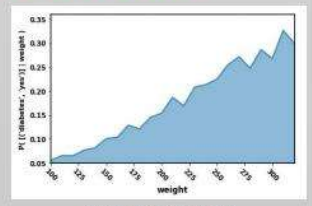
Causal Hypothesis Question

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

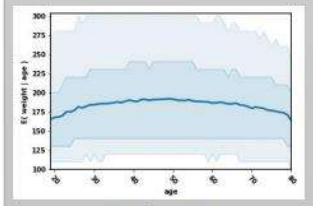
Analysis LLCP-CDC Dataset



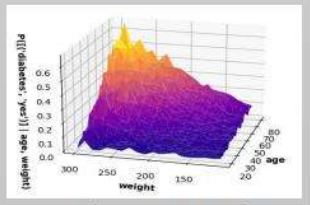
Diabetes vs Age



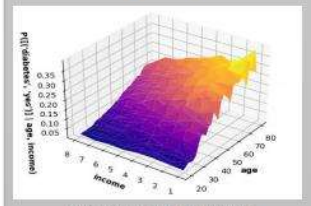
Diabetes vs Weight



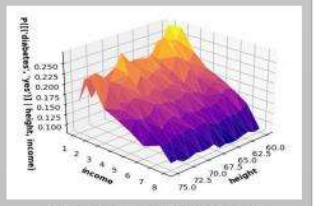
Weight vs Age



Diabetes vs Age, Weight

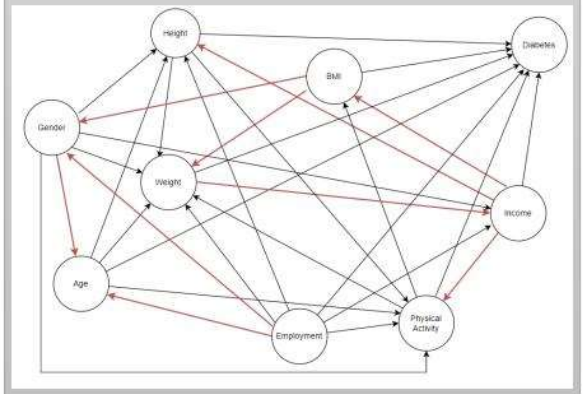


Diabetes vs Age, Income



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

Student Projects (cont.)

Nepali NLP Initiative and Translation

Initiative: Don't Let Nepali Language Die in the 21st Century

Ananya Gupta, David de Hilster

Problem Statement:

- Nepali language is **underrepresented** in digital world (about 16,700 out of 100,000 words (~**16.7%**) are listed in Nepali Wiktionary)
- Existing words do **not** give **complete** linguistic information and are **not** in **proper format**
  - Not useful to perform Natural Language Processing in Nepali language

Solution:

- Developed a framework to populate Nepali words in Wiktionary,
- Built a Nepali dictionary using NLP++ and HPCC Systems

Step 1: Developed a standard Word Entry Template for Nepali Wiktionary

```
==नेपाली==
==उच्चारण==
==पद वर्ग १==
===अर्थ १===
====उदाहरण====
# उदाहरण
# उदाहरण
===अर्थ २===
====उदाहरण====
# उदाहरण
# उदाहरण
===अर्थ ३===
==पद वर्ग २==
===अर्थ १===
====उदाहरण====
# उदाहरण
===अर्थ २===
====उदाहरण====
# उदाहरण
# उदाहरण
==समानार्थी शब्द==
==व्युत्पन्न सर्तहरू==
==अनुवाद==
अङ्ग्रेजी:
```

```
==Nepali==
==Pronunciation==
==part of speech 1==
===Meaning 1===
====Example====
# Example 1
# Example 2
===Meaning 2===
====Example====
# Example 1
# Example 2
===Meaning 3===
====Example====
# Example 1
# Example 2
==Part of speech 2==
===Meaning 1===
====Example====
# Example 1
# Example 2
===Meaning 2===
====Example====
# Example 1
# Example 2
===Meaning 3===
====Example====
# Example 1
# Example 2
==Synonyms==
==Derived terms==
==Translation==
English
```

Step 2: Added words to Nepali Wiktionary

Step 3: Participant's Recruitment, Networking & Website Development

Step 4: Analyzer and KB Development

Step 5: Building Nepali Dictionary using HPCC Systems

ECL Playground

```
Parsed ParseIt(WordsConverted l, INTEGER c) := TRANSFORM
  SELF.parsed := (>UNICODE<)nlp.UnicodeAnalyzeText('Nenlktionary',l.text);
  SELF.wordid := c;
  SELF := l;
END;

WordsParsed := PROJECT(WordsConverted, ParseIt(LEFT,COUNTER));
//output(WordsParsed);

XMLrecbeg := '<File><filename>';
XMLrecid := '</filename>';
XMLrecend := '</File>\n';

XMLrec := RECORD
  INTEGER wordid;
  UTF8 myXML;
END;
P := PROJECT(WordsParsed,
             TRANSFORM(XMLrec,
                       SELF.wordid := LEFT.wordid,
```

Final Output: Record Structure of Nepali Dictionary built on HPCC Systems

HPCC SYSTEMS

CLEMSON School of COMPUTING

Nepali NLP

LexisNexis RISK SOLUTIONS

References:

- Wiktionary. (n.d.). Retrieved July 29, 2022, from http://wiki.nepalinlp.org
- https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries
- http://nepalinlp.org/word-entry-template-and-examples/
- https://visualtext.org/nlp/
- https://hpccsystems.com/training/documentation/learning-ecl

HPCC SYSTEMS

Student Projects (cont.)

Leveraging and Evaluating Kubernetes support for HPCC Systems on Azure

Yash Mishra | Advisor: Dr. Amy Apon | Mentor: Dr. Dan Camper

HPCC SYSTEMS · LexisNexis RISK SOLUTIONS · CLEMSON School of COMPUTING

Introduction

Deployment of HPCC Systems to commercial clouds can be done in multiple ways, such as Lift-and-Shift or Containerization, depending on various business needs. This project utilizes the containerized version of HPCC Systems and orchestrates the new environment via Kubernetes, targeting Microsoft Azure. In the new Kubernetes orchestration of HPCC Systems, several things appear to be different than in the legacy version. HPCC Systems components are converted to pods, completely decoupling it from the node-level dependencies. Pods run the system processes that communicate with other pods in the cluster. Moreover, the storage handling and scaling also changes. The project explores these options to understand the operation of HPCC Systems in cloud-native environment.

Kubernetes considerations on Azure

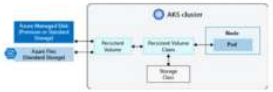
Subscription
Resource Group| Deployment Region
Primary Node Pool - Number and size of nodes in the cluster along with node type
Authentication - Service Principal or System-assigned Managed Identity
Helm Manifest Configuration

Storage Option Considerations

- **Azure File** - Offers SMB access to file shares. This meets the shared data requirement for dllStorage and dataStorage classes.
- **Azure Disk:** Mounted as *ReadWriteOnce*, so it is only available to a single pod. This does not meet the shared storage requirement for dataStorage and dllStorage classes in a multi-node cluster
- Orchestrated via Persistent Volumes and referenced by Persistent Volume Claims

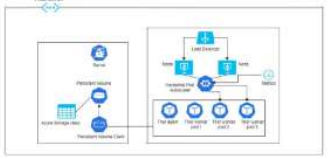
Persistent Volumes

The following diagram illustrates storage architecture in Kubernetes [1]



Network Topology

Pod Scaling and Shared Storage



Cloud Costs

Cloud costs vary by region. For example, Standard_D2s_v2 instance type costs $0.146/hour in the US East region, and costs $0.14/hr in West US region. Choosing a different region may be cheaper, but it might impact the latency.

Challenges and Future Work

- Getting data in and out of the cluster
- Persisting data longer than helm charts
- Exploring alternate storage options - Azure Blob | Azure Data Lake

References:

[1] *Concepts - Storage in Azure Kubernetes Services (AKS).* https://docs.microsoft.com/en-us/azure/aks/concept-storage
[2] *Setting up a Default HPCC Systems Cluster on Microsoft Azure Cloud Using HPCC Systems 7.8.x and Kubernetes*, Jake Smith | HPCC Systems. https://hpccsystems.com/blog/default-azure-setup
[3] *Persisting Data in an HPCC Systems Cloud Native Environment*, Gavin Halliday | HPCC Systems. https://hpccsystems.com/blog/persisting-data-cloud.

HPCC SYSTEMS

Internship Project Life Cycle

Onboarding and welcome	Development phase	Completion phase	HPCC Systems Community Day
• Couple of weeks before the kickoff • Setup infrastructure and accesses • Training • Welcome meeting	• Daily meetings • Code development • Troubleshooting • 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:

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

3) Submit the proposal and your CV:

- Email: students@hpccsystems.com

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

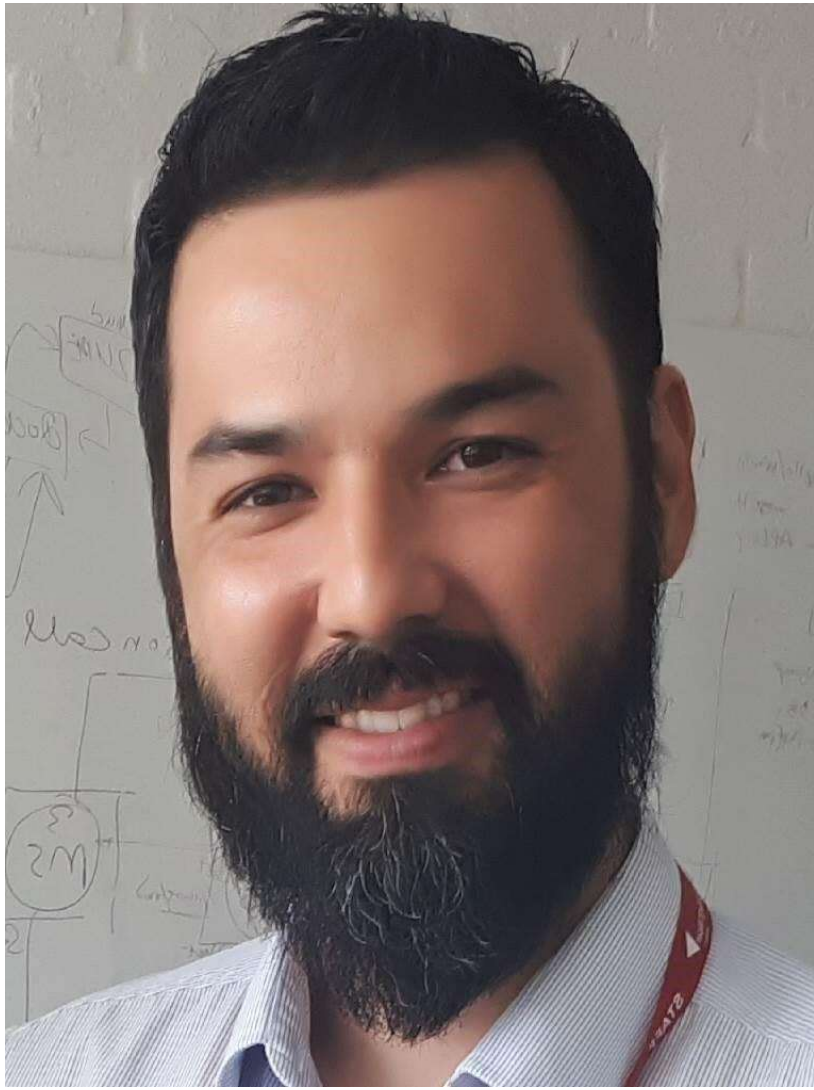
Timeline :	
Design of Workflow from Data Input till result production Collection of Test Documents. Best Sources include datasets found in TREC IR competitions as well as from Wikipedia for benchmarking. Preprocessing and dataset-specific cleaning of above documents	25th May – 5th June
Convert text documents 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	

Be Part of the Team and Make Your Mark

HPCC SYSTEMS

Get in Touch

Hugo.Watanuki@lexisnexisrisk.com



HPCC SYSTEMS