HPCC Systems

Big Data pela perspectiva HPCC Systems

Alysson Oliveira

HPCC™ SYSTEMS

O Grupo RELX

RELX

RELX é um provedor global de análises baseadas em informações e ferramentas de decisão para clientes profissionais e empresariais. O Grupo atende clientes em mais de 180 países e possui escritórios em cerca de 40 países, com um total que supera 36 mil contribuidores.

Saiba mais em www.relx.com

Científico

ELSEVIER

Análise de risco

LexisNexis® RISK SOLUTIONS

Eventos

RX

Legal

LexisNexis® | Legal & Professional

HPCC SYSTEMS

HPCC Systems: Ativos e Clientes

- 12 petabytes de dados públicos e privados
- 270+ milhõe
- Clientes em
- 84% dos integrantes da Fortune 500

Unidade	Símbolo	Número de Bytes
Kilobyte	KB	2^10 = 1024 bytes
Megabyte	MB	2^20 = 1,048,576 bytes
Gigabyte	GB	2^30 = 1,073,741,824 bytes
Terabyte	TB	2^40 = 1,099,511,627,776 bytes
Petabyte	PB	2^50 = 1,125,899,906,842,624 bytes
Exabyte	EB	2^60 = 1,152,921,504,606,846,976 bytes
Zettabyte	ZB	2^70 = 1,180,591,620,717,411,303,424 bytes
Yottabyte	YB	2^80 = 1,208,925,819,614,629,174,706,176 bytes

- 9 dos 10 maiores bancos do mundo

s bancos

eguradoras do

- Mais de 7500 órgãos governamentais: locais, estaduais e federais

A LexisNexis Risk Solutions

Estrutura no Brasil

Total de 140 colaboradores

Área de atuação

Análise de dados para organizações que buscam gerenciar riscos, encontrar oportunidades e melhorar seus resultados. Sediada em Atlanta, Geórgia, a LexisNexis Risk Solutions tem mais de 11.000 funcionários ao redor do mundo.

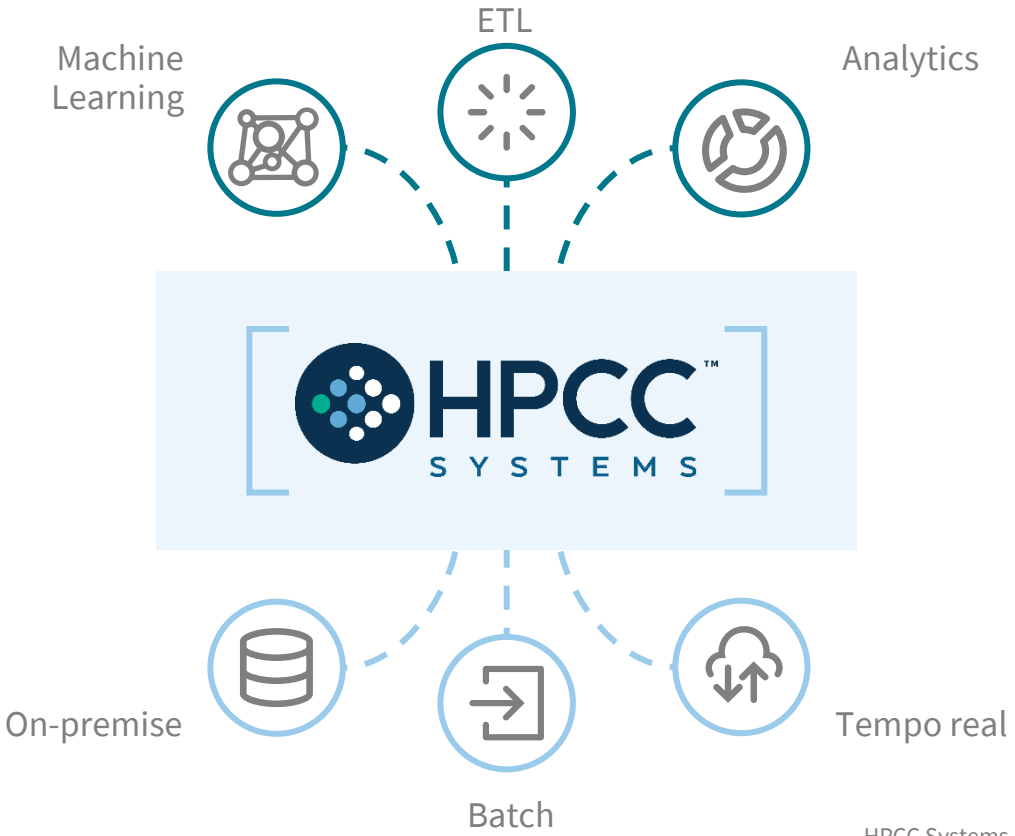
Tecnologia de código aberto

Plataforma de computação de Big Data de código aberto chamada HPCC Systems com vastos ativos de dados para proporcionar inteligência de decisão para clientes.

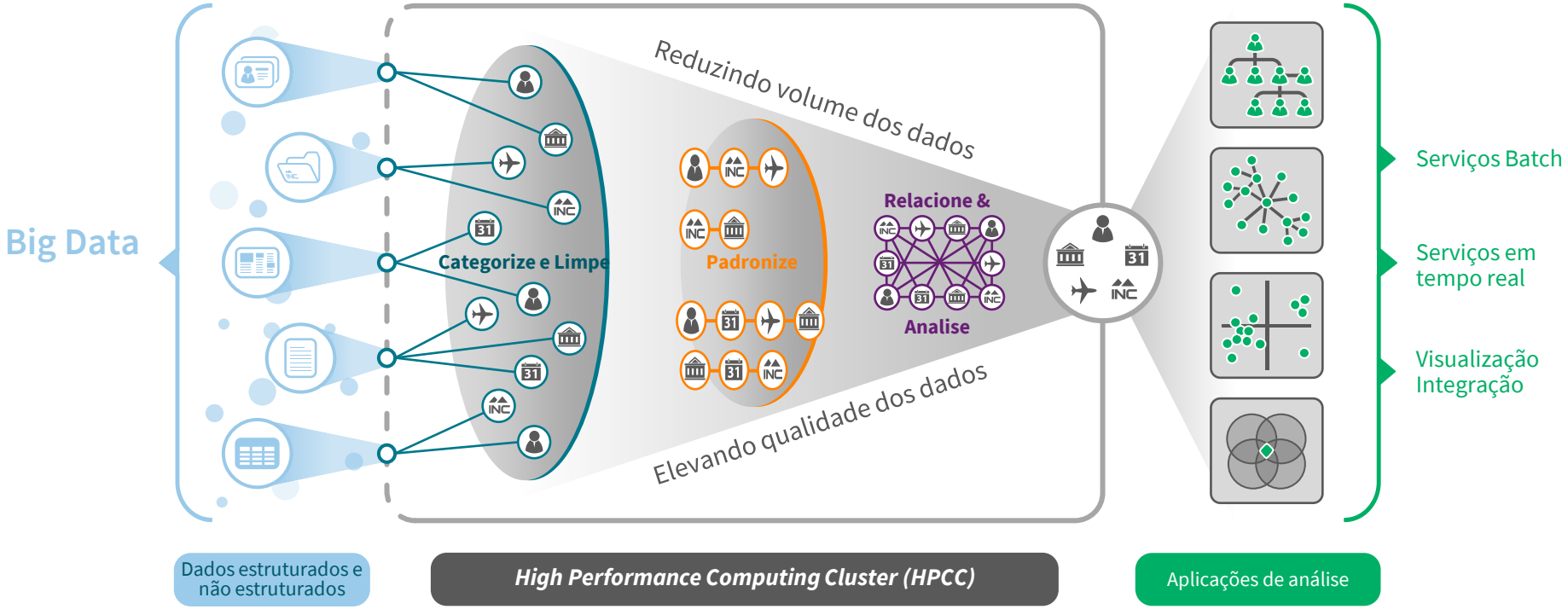
HPCC SYSTEMS

A plataforma HPCC Systems

- Stack para big data
- Processamento paralelo
- Dados distribuídos
- Código aberto
- Gratuita



“Funil” de dados no HPCC Systems



Cadeia de Big Data em HPCC Systems

Breve histórico do HPCC Systems

2001

Primeira versão da plataforma é lançada

2011

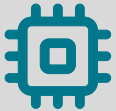
Código aberto (licença Apache e código no GitHub)

2012 – 16

Melhorias contínuas com **FOCO NA QUALIDADE**

Suporte e treinamento aprimorado

2017- Presente



Aprimoramentos de arquitetura (Cloud)

Desenvolvimentos em Machine Learning

Visão geral do stack

Cluster Thor

Extração, transformação e carregamento de dados

Cluster ROXIE

Entrega online de consultas em big data

Ferramentas para manipulação de dados

Perfilamento, limpeza, consolidação e linking de dados

Bibliotecas de Machine Learning

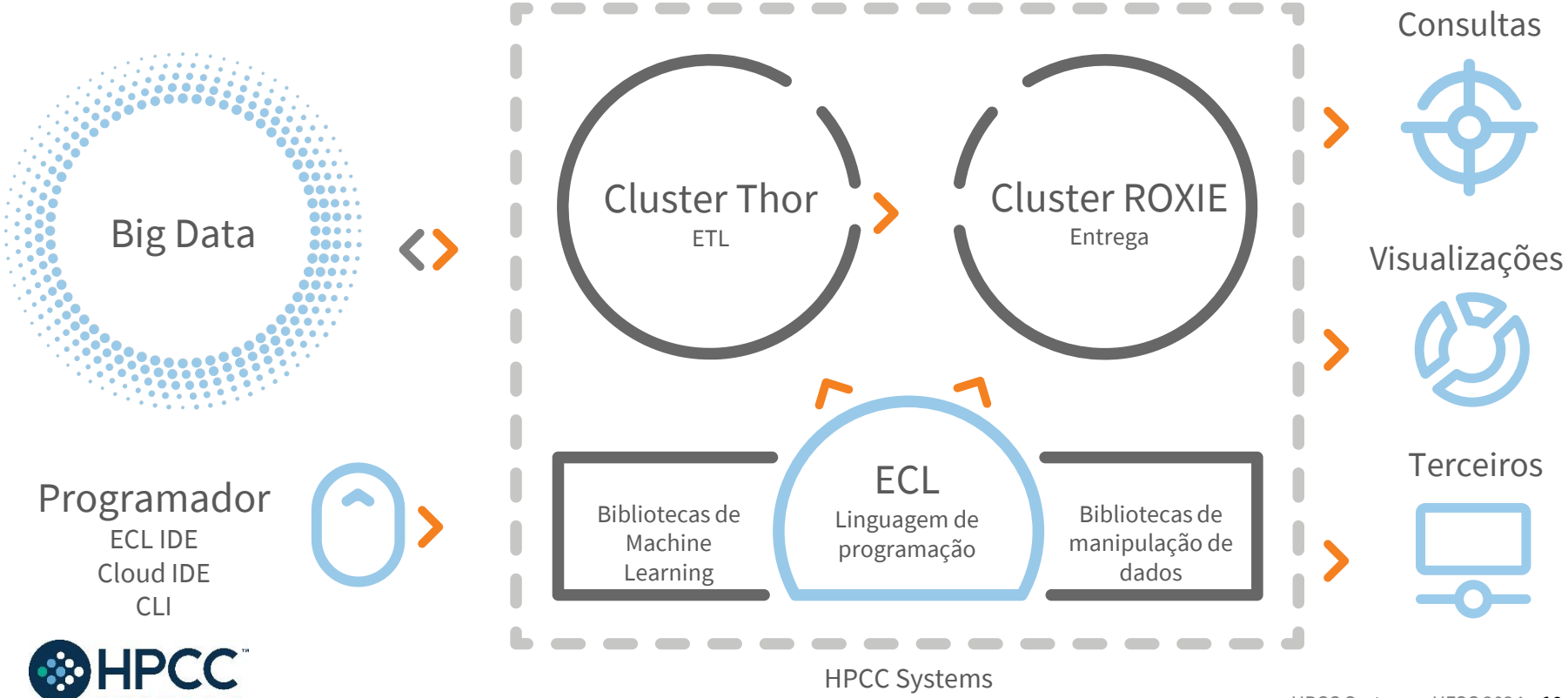
Supervisionado, não-supervisionado, aprendizagem profunda

Conectividade

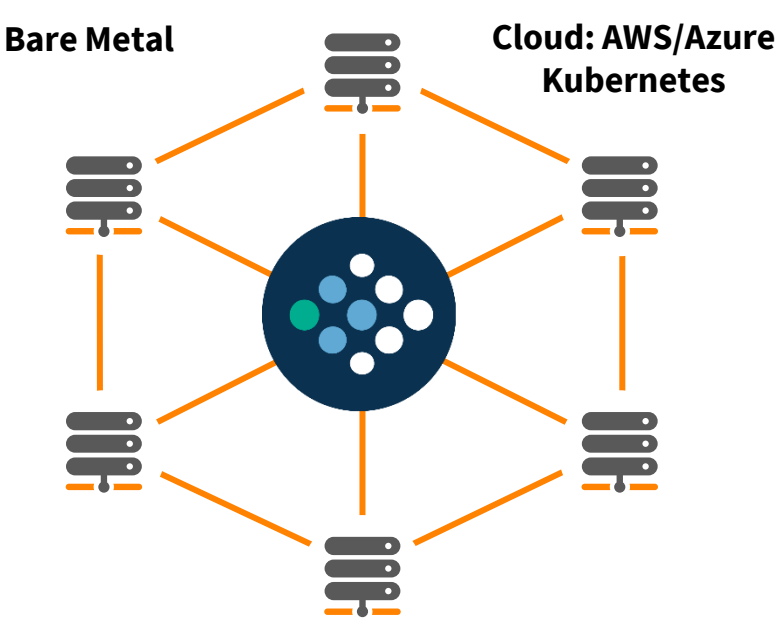
Plugins de integração com outros sistemas

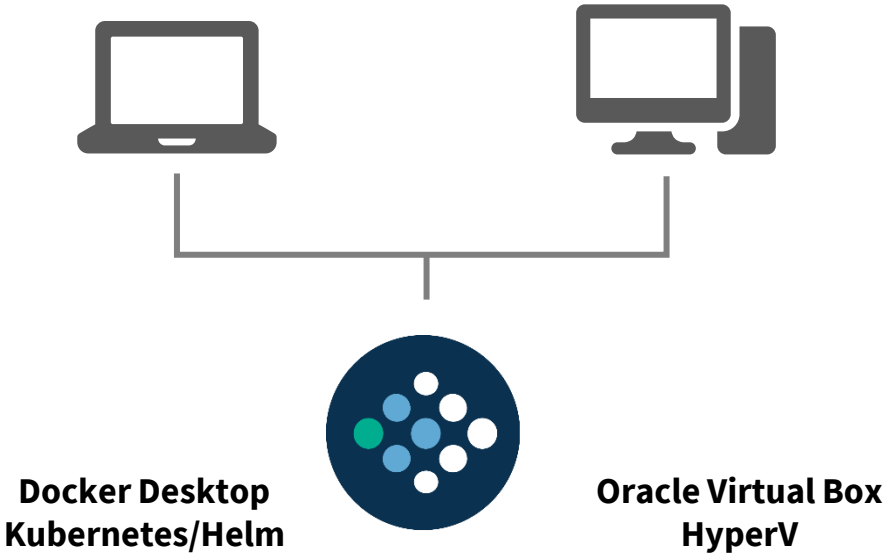
HPCC SYSTEMS

Os componentes da plataforma

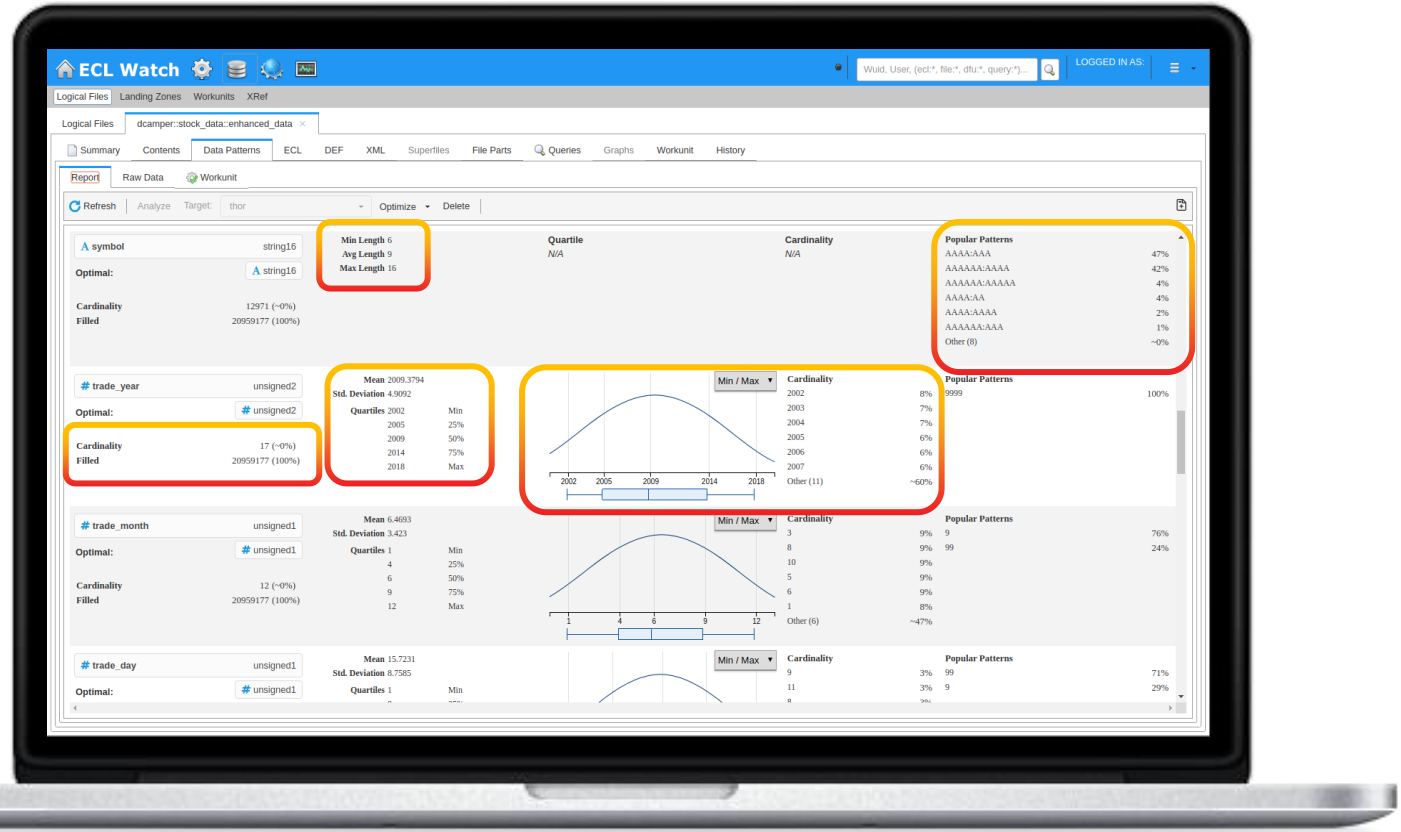


Jornada em direção à nuvem





Bibliotecas de perfilamento de dados



Bibliotecas de machine learning

Não supervisionado

Clusterização

DBSCAN

K-Means

PLN

Text Vectors

Levenshtein Deletion Neighborhood

Redução de Dimensão

PCA

Supervisionado

Classificação

SVM

Árvores de decisão

Regression logística

Classification Forest

Alocação Latente de Dirichlet (Topic Modeling)

Regressão

Regressão linear

GLM

Regression Forest

Redes neurais & Deep Learning

Autoencoders

Redes neurais convolucionais

Redes neurais recorrentes

Perceptrons

Métodos ensemble

Random Forest

Gradient Boosted Forest

Gradient Boosted Trees

Plugins para conectividade

WsSQL

TOMBOLO

SPARK

JDBC/ODBC Driver

KAFKA

HPCC

PENTAHO

Couchbase

Tableau

SQS

Java API

MEMCACHED

REDIS

Linguagens suportadas

- C++
- R
- Python
- Java
- Cassandra
- SQL/SqLite

CODE: SELECT ALL

```
IMPORT python;
r := RECORD
  STRING word;
  UTF8 tags;
END;
DATASET(R) tag(STRING text) := IMPORT(python, './ex2.tag');
tag('Once upon a time there was a boy called Richard');
```

CODE: SELECT ALL

```
IMPORT java;
STRING jcat(STRING a, STRING b) :=
   IMPORT(java,
          'JavaCat.cat:(Ljava/lang/String;Ljava/lang/String;)Ljava/lang/String;' :
classpath('/opt/HPCCSystems/classes'));

jcat('Hello ', 'world!');
```

CODE: SELECT ALL

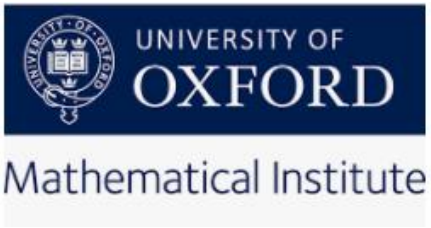
```
IMPORT python;
SET OF STRING split(STRING text) := EMBED(python)
  return text.split()
ENDEMBED;
split('Once upon a time');
```

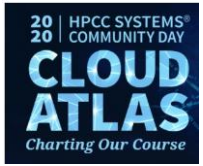
CODE: SELECT ALL

```
IMPORT MySQL;
stringrec := RECORD
   string name
END;
sqlrec := RECORD
   string ssn;
   string address;
END;
DATASET(sqlrec) MySQLJoin(dataset(stringrec) inrecs) := EMBED(mysql)
  SELECT * from tbl1 where name = ?;
ENDEMBED;
MySQLJoin(indata);
```

Relacionamento com Academia

Universidade de São Paulo
Brasil
USP





HPCC SYSTEMS

Universidades Brasileiras

Universidade de São Paulo
Brasil

- Disciplina Optativa na Poli/USP (Link para a disciplina)
- Curso de Difusão (Fundação Vanzolini)
- Co-orientação de IC's (PIBIC)
- Co-Orientação de TCC's

UNIVERSIDADE FEDERAL
DE SANTA CATARINA

- Co-Orientação de IC's
- Co-Orientação de TCC's
  - Artigos publicados (ERAD/RS, CotB, etc)
  - Apresentações no HPCC Summit
- Co-Orientação de Mestrado
- Compra de equipamentos

HPCC SYSTEMS

Universidades Estrangeiras

Imperial College London

- Pesquisas de Doutorado
  - Deep Learning, Machine Learning, Text Mining, Natural Language Processing

CLEMSON UNIVERSITY

- Estagiários
  - Machine Learning

HPCC SYSTEMS

Projetos de Pesquisa

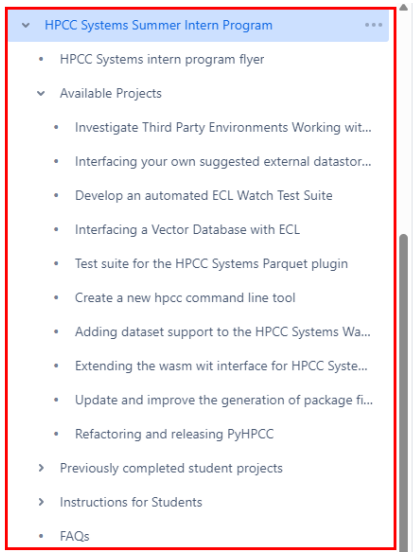
Site: https://hpccsystems.com/community/academics

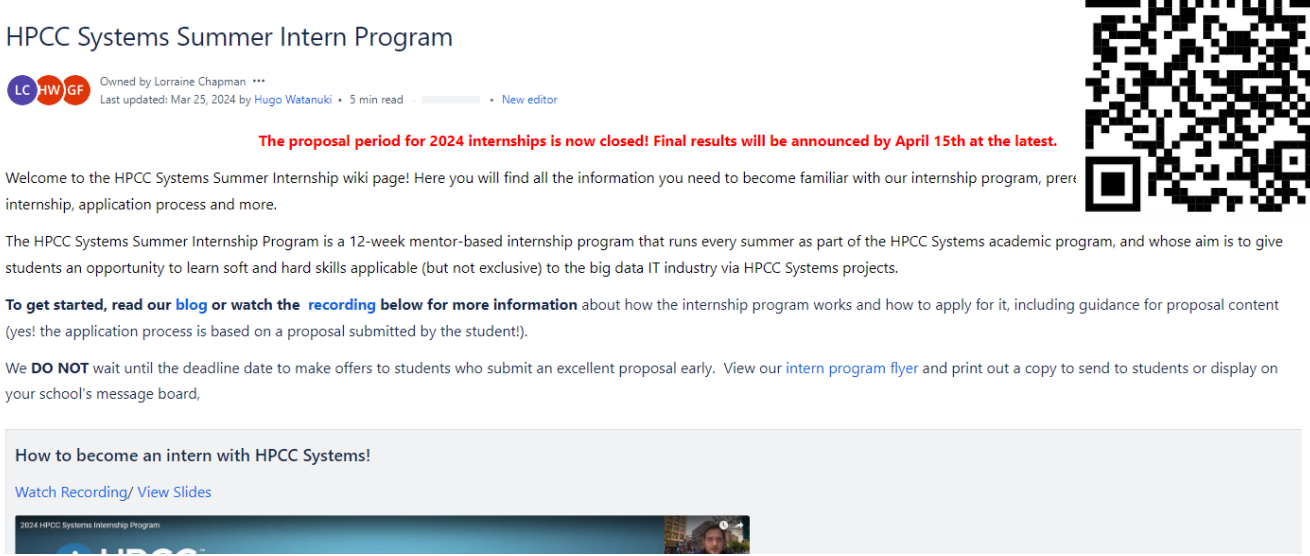
- Programa de Estágio
  - Verão do Hemisfério Norte (Summer Intern Program)
  - Mentoria
  - Bolsas de Estudo
- Publicações Acadêmicas
- Treinamentos

HPCC SYSTEMS

Projetos de Pesquisa

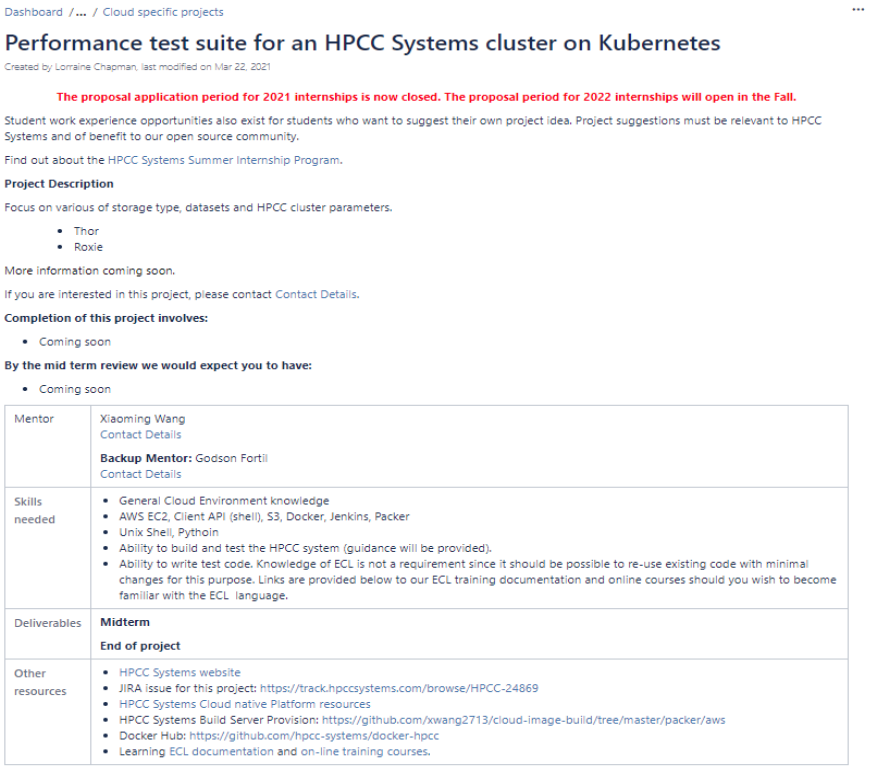
https://wiki.hpccsystems.com/display/hpcc/HPCC+Systems+Summer+Intern+Program





Projetos de Pesquisa

https://wiki.hpccsystems.com/display/hpcc/Available+Projects



Dashboard / ... / Cloud specific projects

Performance test suite for an HPCC Systems cluster on Kubernetes

Created by Lorraine Chapman, last modified on Mar 22, 2021

The proposal application period for 2021 internships is now closed. The proposal period for 2022 internships will open in the Fall.

Student work experience opportunities also exist for students who want to suggest their own project idea. Project suggestions must be relevant to HPCC Systems and of benefit to our open source community.

Find out about the HPCC Systems Summer Internship Program.

Project Description

Focus on various of storage type, datasets and HPCC cluster parameters.

- Thor
- Roxie

More information coming soon.

If you are interested in this project, please contact Contact Details.

Completion of this project involves:

- Coming soon

By the mid term review we would expect you to have:

- Coming soon

Mentor	Xiaoming Wang Contact Details **Backup Mentor:** Godson Fortil Contact Details
Skills needed	• General Cloud Environment knowledge • AWS EC2, Client API (shell), S3, Docker, Jenkins, Packer • Unix Shell, Pythoin • Ability to build and test the HPCC system (guidance will be provided). • Ability to write test code. Knowledge of ECL is not a requirement since it should be possible to re-use existing code with minimal changes for this purpose. Links are provided below to our ECL training documentation and online courses should you wish to become familiar with the ECL language.
Deliverables	**Midterm** **End of project**
Other resources	• HPCC Systems website • JIRA issue for this project: https://track.hpccsystems.com/browse/HPCC-24869 • HPCC Systems Cloud native Platform resources • HPCC Systems Build Server Provision: https://github.com/xwang2713/cloud-image-build/tree/master/packer/aws • Docker Hub: https://github.com/hpcc-systems/docker-hpcc • Learning ECL documentation and on-line training courses.

Código Aberto

Github: https://github.com/hpcc-systems

- Linguagem: C++
- Repositório bastante ativo
  - 170+ Commits nos últimos 30 dias
- Documentação
  - Arquivos README.md dentro do repositório
  - Site do HPCC (https://hpccsystems.com/training/documentation)
- Tickets
  - https://track.hpccsystems.com/secure/Dashboard.jspa

Considerações Finais & Perguntas

Contato

- Alysson.Oliveira@lexisnexisrisk.com

- Mauro.marques@lexisnexisrisk.com