Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas (nível intermediário)

Aula 6

LexisNexis® RISK SOLUTIONS

RELX Group

PRO

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

USP UNIVERSIDADE DE SÃO PAULO

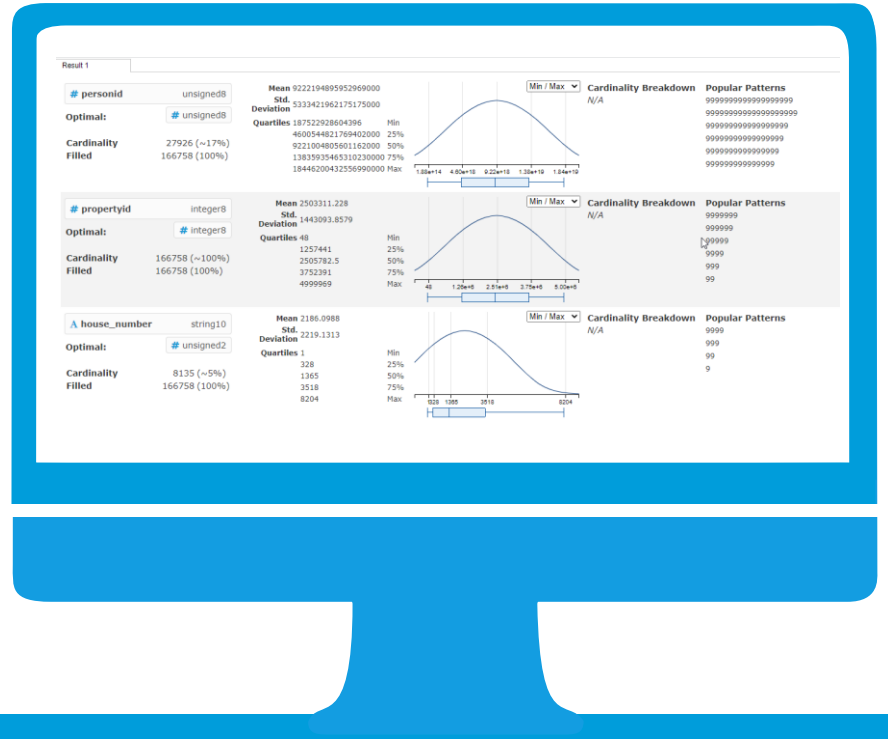
Bem-vindo! – Agenda da aula 6

- ✓Desafio Lending Club
- ✓Visualização de dados
- ✓Intervalo
- ✓Próximos passos

Exercício prático:

Treine e avalie um modelo de análise de risco de pedido de empréstimo

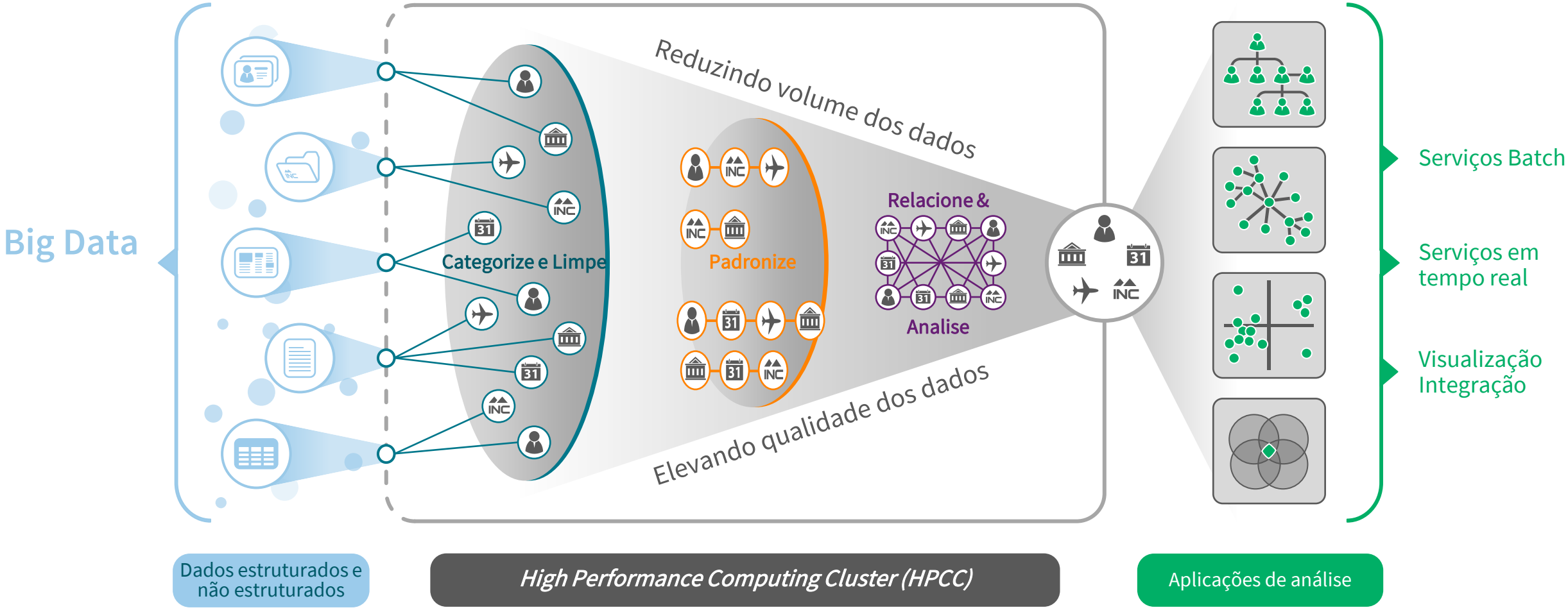
- Considere a aplicação de aprendizagem supervisionada



Fundação Vanzolini

Visualização de dados

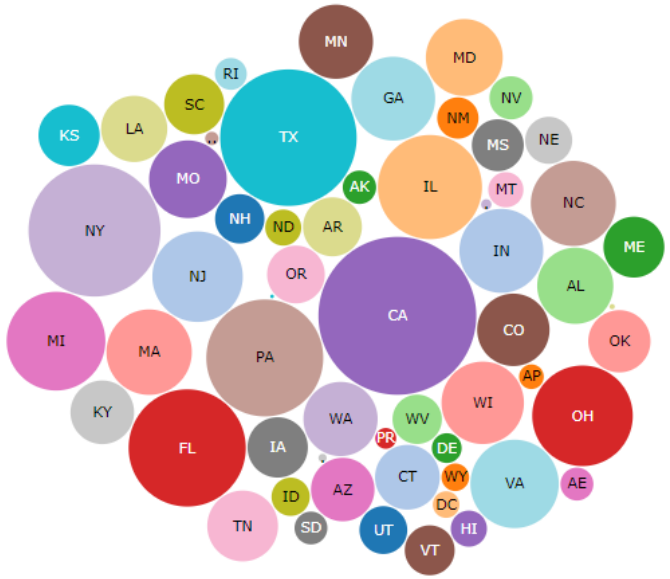
Extract, Transform, Load

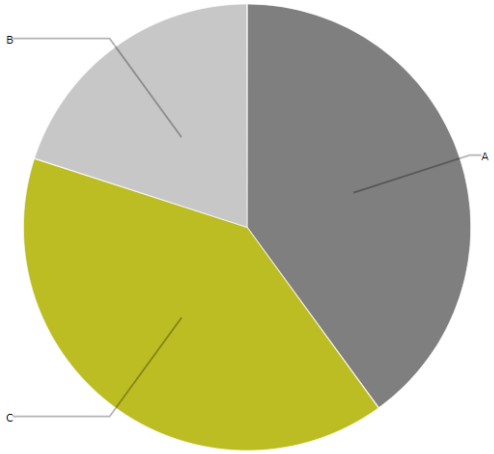


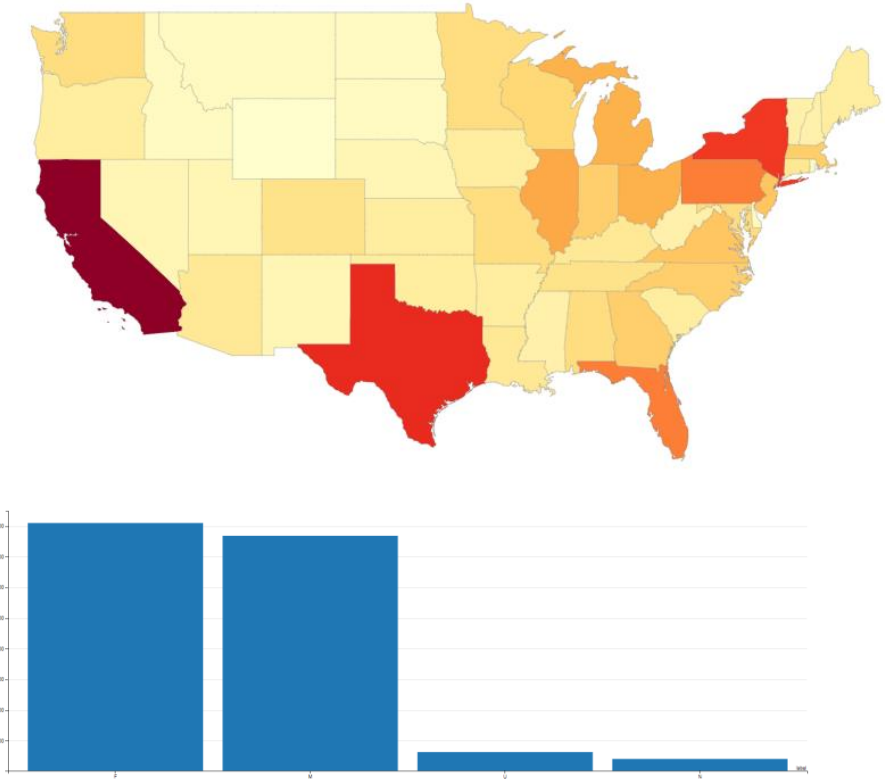
Fundação Vanzolini

Ferramentas de visualização

A plataforma HPCC Systems disponibiliza ferramentas de visualização para dados de saída via gráficos e mapas.







Fundação Vanzolini

Ferramentas de visualização (cont.)

Os dados podem ser visualizados por meio de três métodos:

- Via a ferramenta de visualização do Playground.
- Via a aba “Visualize” em qualquer workunit.
- Via a aba “Resources” em conjunto com o pacote de Visualização ECL.

Instalação:

```
ecl bundle install https://github.com/hpcc-systems/Visualizer.git
```

Gerando a visualização

A utilização do pacote de visualização envolve basicamente quatro passos.

1. Faça OUTPUT dos seus dados usando o atributo NAMED.
2. Defina os campos a serem visualizados.
3. Customize a aparência do gráfico.
4. Compartilhe o gráfico.

```
IMPORT $, Visualizer;
GenderDS := DATASET([{'Female', 404988},
                     {'Male', 384182},
                     {'Neutral', 20508},
                     {'Unknown', 70722}],{STRING Label,UNSIGNED4 Value});
OUTPUT(GenderDS,NAMED('VizPie'));
Visualizer.TwoD.Pie('Pie',,'VizPie');
```

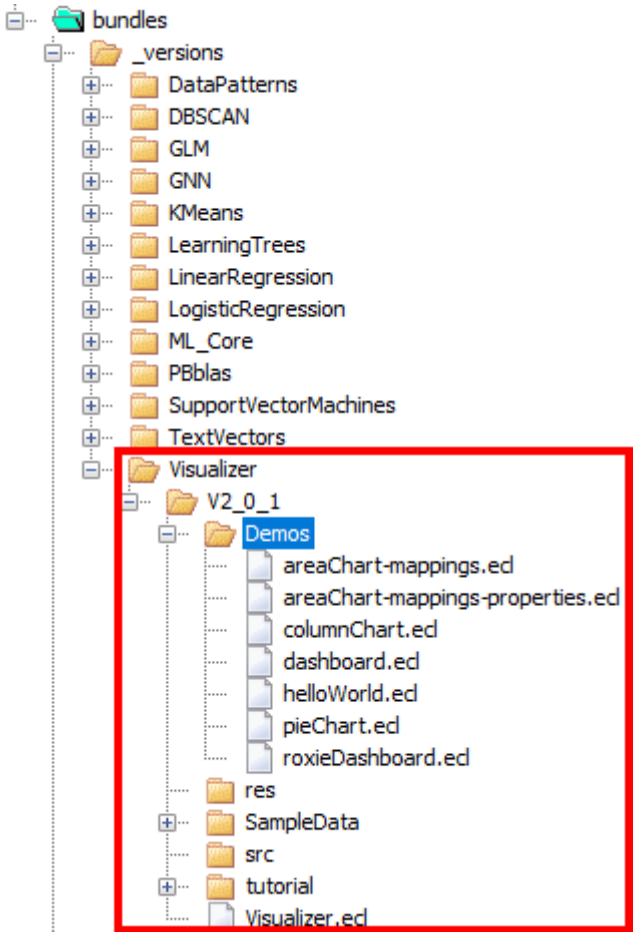
https://cdn.hpccsystems.com/releases/CE-Candidate-8.4.14/docs/EN_US/VisualizingECL_EN_US-8.4.14-1.pdf

https://hpccsystems.com/blog/visualizing-ecl-and-sharing-your-results-hpcc-systems-visualizer

Tutorial de visualização de dados

Bundle Visualizer

- Biblioteca de visualização de resultados gerados por código ECL
- Gráficos bidimensionais
  - `Visualizer.TwoD.__test e Visualizer.TwoDLinear.__test`
- Gráficos multidimensionais
  - `Visualizer.MultiD.__test`
- Gráficos geoespaciais
  - `Visualizer.Choropleth.__test`
- Gráficos relacionais
  - `Visualizer.Relational.__test`
- Tabelas genéricas
  - `Visualizer.Any.__test`



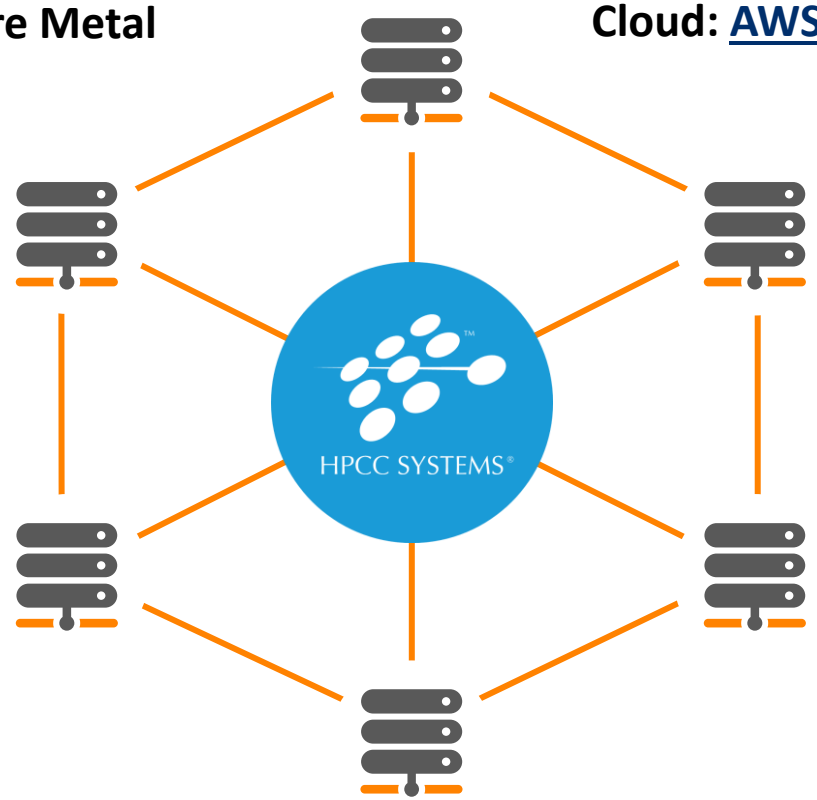
Próximos passos

Próximos passos

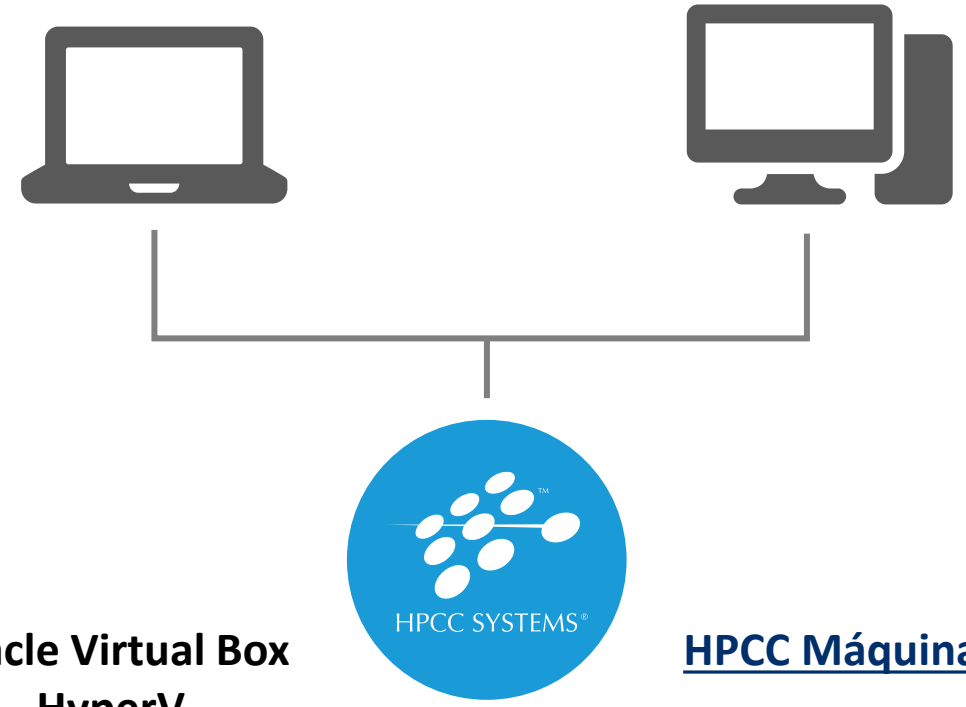
- ✓Enviar badges de certificação e exemplos de Código
- ✓Certificado USP
- ✓Playground / Treinamento online / documentação / fórum
- ✓Conferência HPCC Systems (Outubro/22)
  - ✓https://hpccsystems.com/community/events/hpcc-systems-summit-2022

Opções de uso: play.hpccsystems.com

Bare Metal



Cloud: AWS/Azure



Oracle Virtual Box
HyperV
Docker
GitPod

HPCC Máquina Virtual

✓https://hpccsystems.com/pt-br/try-now

Fundação Vanzolini

Cursos online: +170 aulas (learn.lexisnexis.com/hpcc)

- Introdução ao ECL (parte 1)
  - Conceitos e consultas
- Introdução ao ECL (parte 2)
  - ETL com ECL
- ECL Avançado (parte 1)
  - Dados relacionais
- ECL Avançado (parte 2)
  - Superarquivos, XML/JSON e PLN
- ECL Aplicado
  - Geração e automação de código ECL

ROXIE ECL (parte 1)
- Índices e consultas

ROXIE ECL (parte 2)
- Otimização de consultas

Machine Learning com HPCC Systems
- Fundamentos para uso dos plugins

Administração de Sistemas
- Conceitos e operação básica

HPCC para gestores
- Visão geral e aplicações da plataforma

Links úteis

- Site principal: hpccsystems.com
- Primeiros passos: hpccsystems.com/Why-HPCC-Systems
- Canal do youtube: youtube.com/user/HPCCSystems
- Fórum da Comunidade: hpccsystems.com/forums

Faça parte da Comunidade

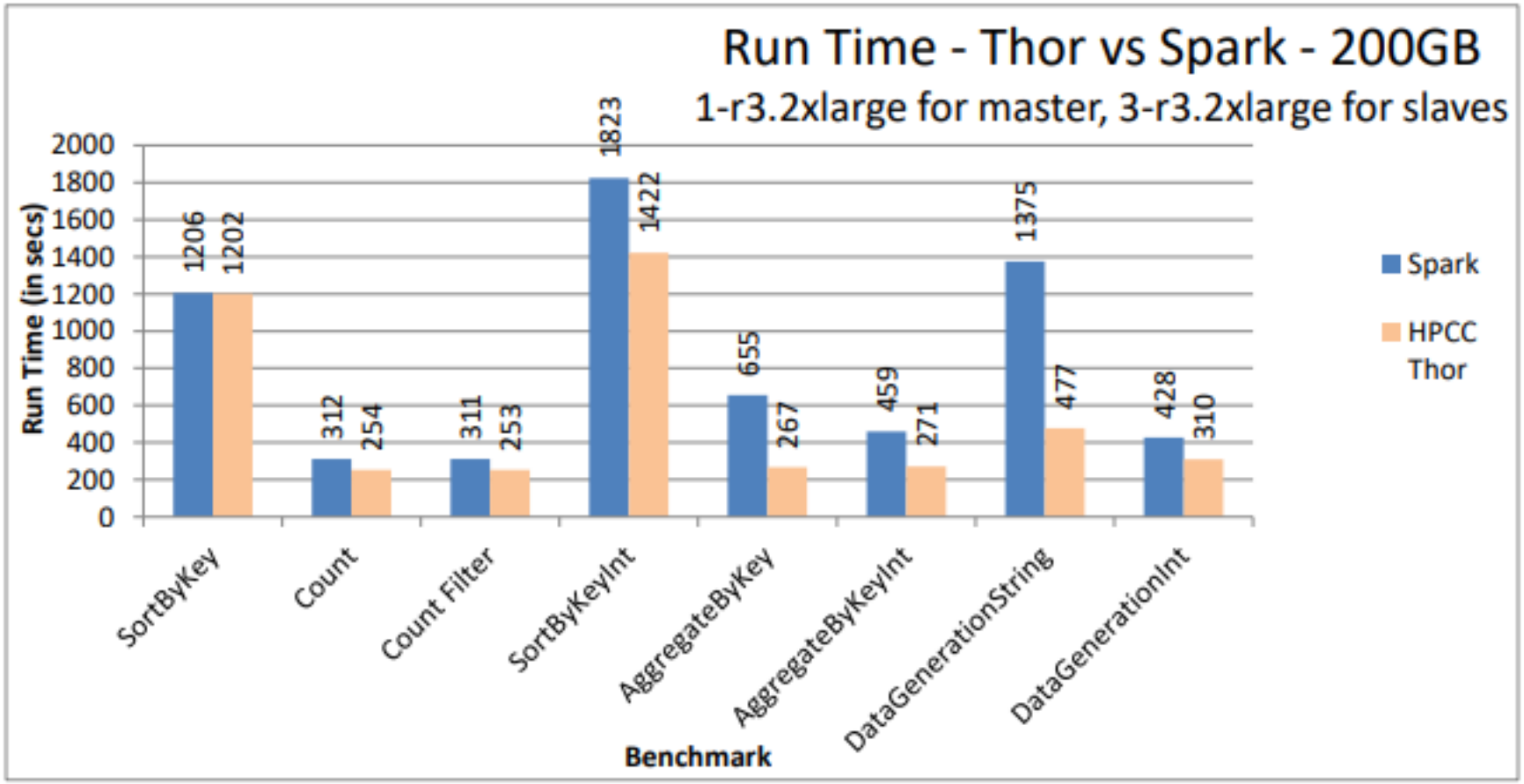
Registre-se em hpccsystems.com

Benchmark

Table 1. HPCC vs Hadoop vs Spark

Topic	HPCC	Hadoop	Spark
Parallelism Paradigm	Dataflow Three parallel execution modes: • **Data:** Data partitioned across nodes; Compute occurs on each node in parallel • **Pipeline:** Consecutive operations on the same dataset at the same time; Data processed by one operation immediately passed to the next • **System:** Independent operations try to execute in parallel	MapReduce **Data** parallelism only, and only in the Map phase.	RDD (Resilient Distributed Dataset) **Data** parallelism only
Compilation	Yes. The C++ generated by the ECL Compiler is compiled for execution	No. JVM-based	No. JVM-based
Built-in End User Query Support	Yes. Roxie clusters deliver thousands of concurrent end-user transactions per second (actual numbers dependent on the number of nodes in the cluster and the complexity of the queries themselves)	No. Third party tools required.	No. Third party tools required.
Production Monitoring	Yes. Ganglia and Nagios included as part of the platform.	No. Third party tools required.	No. Third party tools required.
Language(s) Supported	ECL built in with any other language embeddable inline. C++, Java, Javascript, Python, SQL, and R currently supported. More embed languages can be added by the community	Java, Hive, Pig	API allows JVM-based language programming (like Java, Python, Scala, and R)

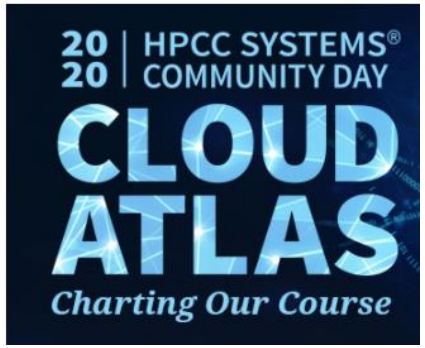
Desempenho comparativo



https://cdn.hpccsystems.com/whitepapers/hpccsystems_thor_spark.pdf

Relacionamento com Academia

https://hpccsystems.com/community/academics



UNIVERSIDADE FEDERAL
DE SANTA CATARINA

Fundação Vanzolini

Projetos de Pesquisa

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

Até o próximo curso!!!