



**UNIVERSIDADE FEDERAL
DE SANTA CATARINA**

Aula Expositiva

Processamento e Análise de *Big Data* com HPCC Systems

Objetivo: Ao longo dessa aula serão apresentados aos alunos os conceitos essenciais de processamento e análise de dados em quantidades massivas (*Big Data*) fazendo uso da plataforma “*open source*” HPCC Systems.

Instrutores



- **Mauro D. Marques** – Engenheiro de SW na LexisNexis Risk Solutions
Engenheiro com pós-graduação nas áreas de TI e Educação
35 anos de atuação como engenheiro no setor automobilístico
11 anos de atuação como professor universitário nas áreas de Engenharia e Ciência da Computação
Mauro.Marques@lexisnexisrisk.com



- **Alysson R. Oliveira** – Engenheiro de SW na LexisNexis Risk Solutions
Engenheiro de Computação
03 anos como engenheiro de software e mentor de projetos acadêmicos
05 anos como instrutor de cursos técnicos na área de Engenharia de Computação
Alysson.Oliveira@lexisnexisrisk.com

Minicurso: Agenda

➤ **LexisNexis Risk Solutions: A Empresa**

- Quem somos nós?
- A nossa tecnologia: A evolução da plataforma HPCC Systems...

➤ **HPCC Systems: Visão Geral**

- Apresentação de conceitos;
- Aplicação de conhecimentos.

➤ **Próximos passos**

- Cursos online;
- Projetos de pesquisa;
- Oportunidades profissionais.

➤ **Considerações Finais**

LexisNexis Risk Solutions: A Empresa

- **Quem somos nós?**

A **LexisNexis Risk Solutions** é líder no fornecimento de informações essenciais que ajudam clientes de diversos setores e governos na avaliação, prevenção e gestão de riscos.

Fazemos parte do **LexisNexis Risk Solutions Group**, um portfólio de marcas que abrange vários setores que fornecem aos clientes tecnologias inovadoras, análises baseadas em informações e ferramentas de decisão e serviços de dados.

LexisNexis Risk Solutions Group



Saiba mais em: <https://risk.lexisnexis.com/group/our-brands>



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.

Saiba mais em www.relx.com

Científico



Eventos



Análise de risco



Legal



LexisNexis Risk Solutions: A Empresa

▪ A nossa tecnologia: A evolução da plataforma HPCC Systems...

2001



Primeira versão
da plataforma
HPCC é lançada

2011



Código aberto (licença
Apache e código no
GitHub)

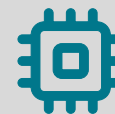
2012 – 16



Melhorias contínuas
com **FOCO NA
QUALIDADE**

Suporte e treinamento
aprimorado

2017-Presente

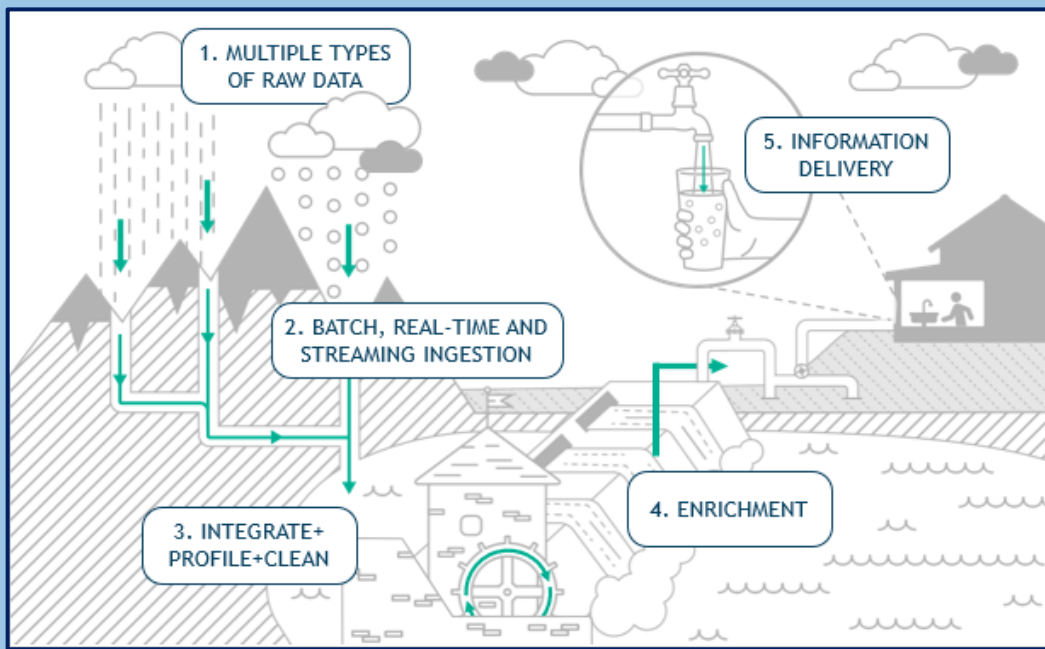


Aprimoramentos de
Arquitetura (*Cloud*)

Desenvolvimentos em
Machine Learning

HPCC Systems: Visão Geral

■ Apresentação de conceitos



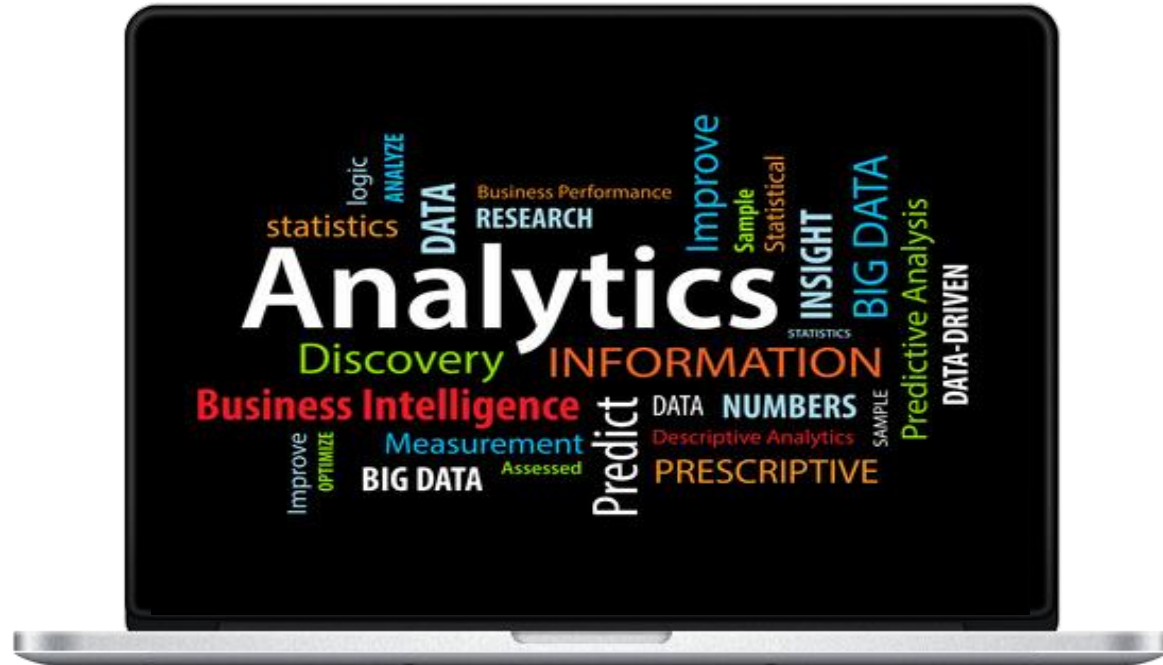
HPCC Systems

Processamento e Análise de *Big Data*

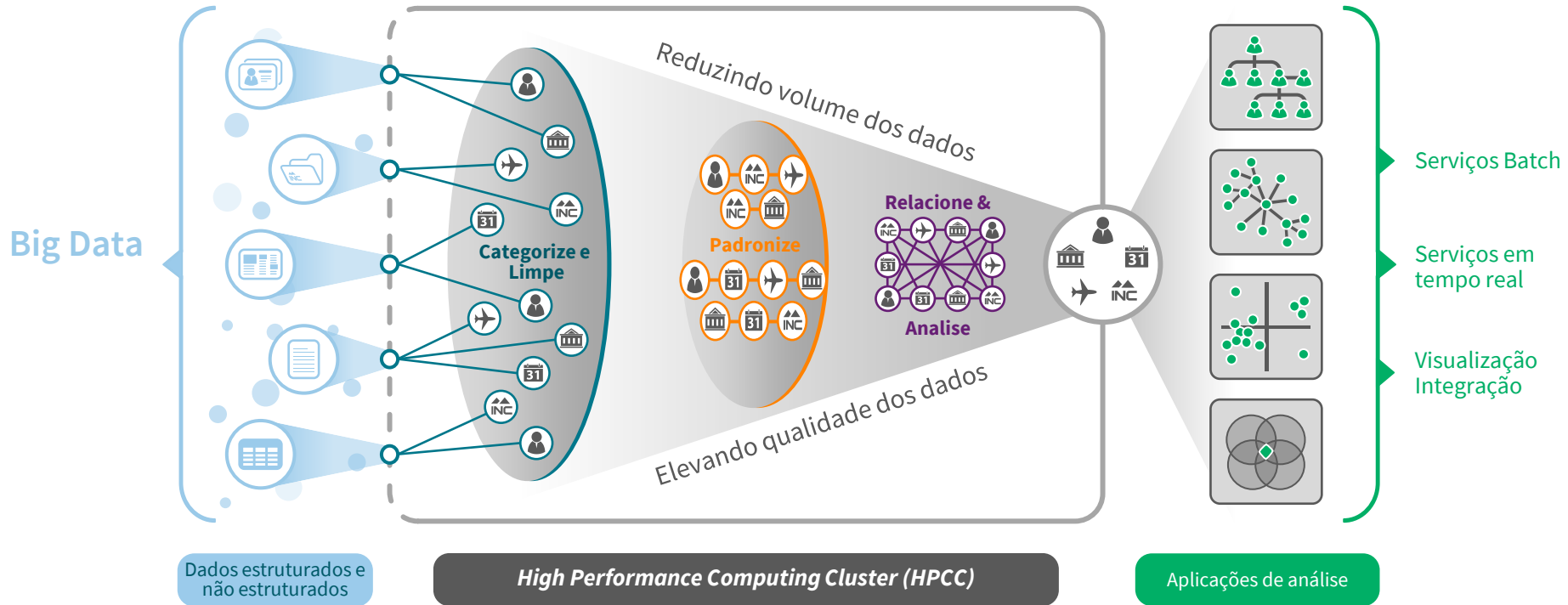


Apresentação de Conceitos

- Os cinco **V's**:
 - **V**olume
 - **V**ariiedade
 - **V**elocidade
 - **V**eracidade
 - **V**alor



Fluxo de Dados no HPCC Systems: 'Funil' de Dados



“Stack” tecnológico da plataforma HPCC Systems



Cluster Thor

ETL: Extração, Transformação e Carregamento de dados



Cluster ROXIE

Entrega online de consultas em *Big Data*



Ferramentas para manipulação de dados

Perfilamento, limpeza, consolidação de dados



Bibliotecas de *Machine Learning*

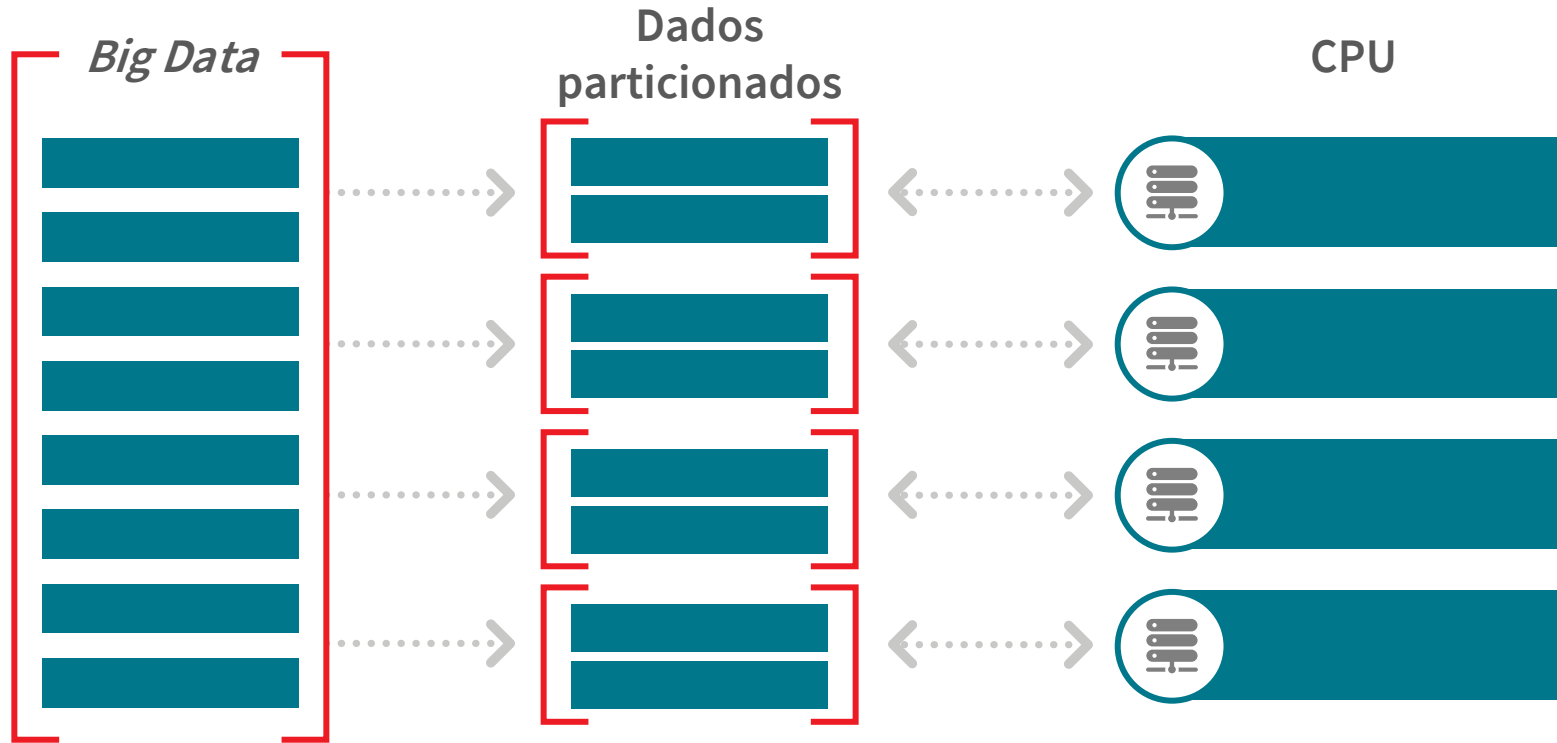
Supervisionado, não-supervisionado, aprendizagem profunda



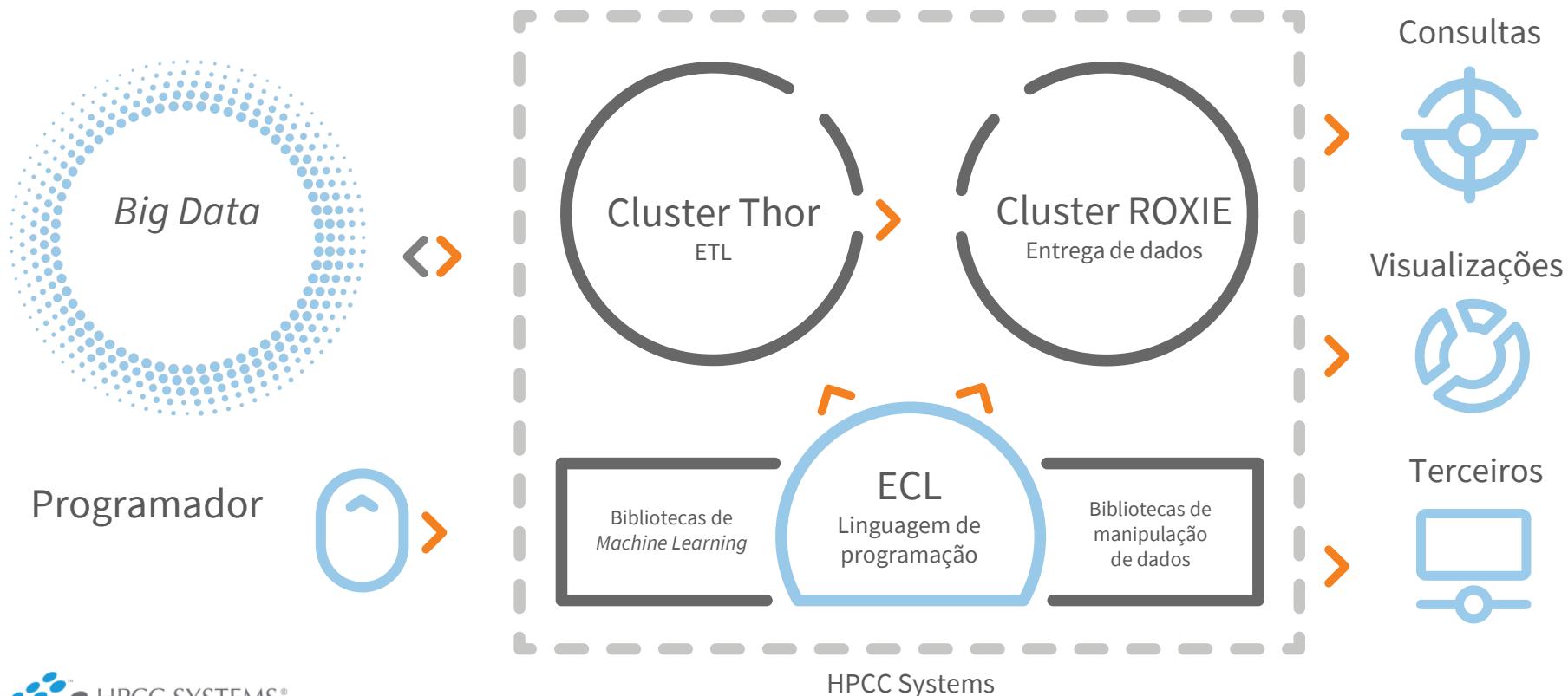
Conectividade

Plugins de
integração com
outros sistemas

Dados Distribuídos e Processamento Paralelo



Arquitetura da plataforma HPCC Systems



Enterprise Control Language (ECL)

Linguagem de programação centrada em dados (*Data Flow*)

- Declarativa e não-procedural
- Códigos menores e reutilizáveis
- Biblioteca para manipulação de dados

Compilador

- Gera código otimizado (C++)
- Lógica para processamento paralelo e distribuído

Como fazer



vs.



O que fazer

HPCC Systems: Visão Geral

■ Aplicação de conhecimentos

The screenshot displays the ECL Watch application interface. The main window is titled "hthor" and shows a "Produto: Serviço de Consulta de Dados Pessoais" (Product: Personal Data Query Service). The interface includes a "fetch_persons_mdm" function and a "Dynamic Form" dropdown. Below this, there is a section titled "FETCH_PERSONS_MDMREQUEST" with a checkbox. The form contains three input fields: "firstname_value:", "lastname_value:", and "state_value:". At the bottom of the form, there are buttons for "Output Tables", "FORM POST", "Submit", and "Clear All". The left sidebar shows a list of tasks under "Tarefas" and "Área de Recreio", including a list of WUIDs. The bottom status bar indicates "1 - 50 de 1252 resultados" and shows various error and warning counts.

2. Enterprise Control Language (ECL)

■ Conceitos básicos de ECL:

- Paradigma declarativo (não-procedural)
- ECL não é sensível a caixa alta/baixa
- Espaço em branco é ignorado para melhor leitura
- Comentários em linha (//) e em bloco (/* e */)
- ECL utiliza sintaxe **objeto.propriedade**

Dataset.Campo

// referencia um campo em um dataset

NomedoDiretorio.Definicao

// referencia uma definição em outro módulo

2. Enterprise Control Language (ECL)

- **Conceitos básicos de ECL:**

- O código ECL é constituído de:

- ❖ **Definições** estabelecem o que as coisas são (arquivos de definição ECL)

mydef := 'People'; // não inicia uma **WU**

- ❖ **Ações** resultam em compilação e execução (arquivos **BWR** – *Builder Window Runnable*)

OUTPUT('People'); // inicia uma **WU**

OUTPUT(mydef); // inicia uma **WU**

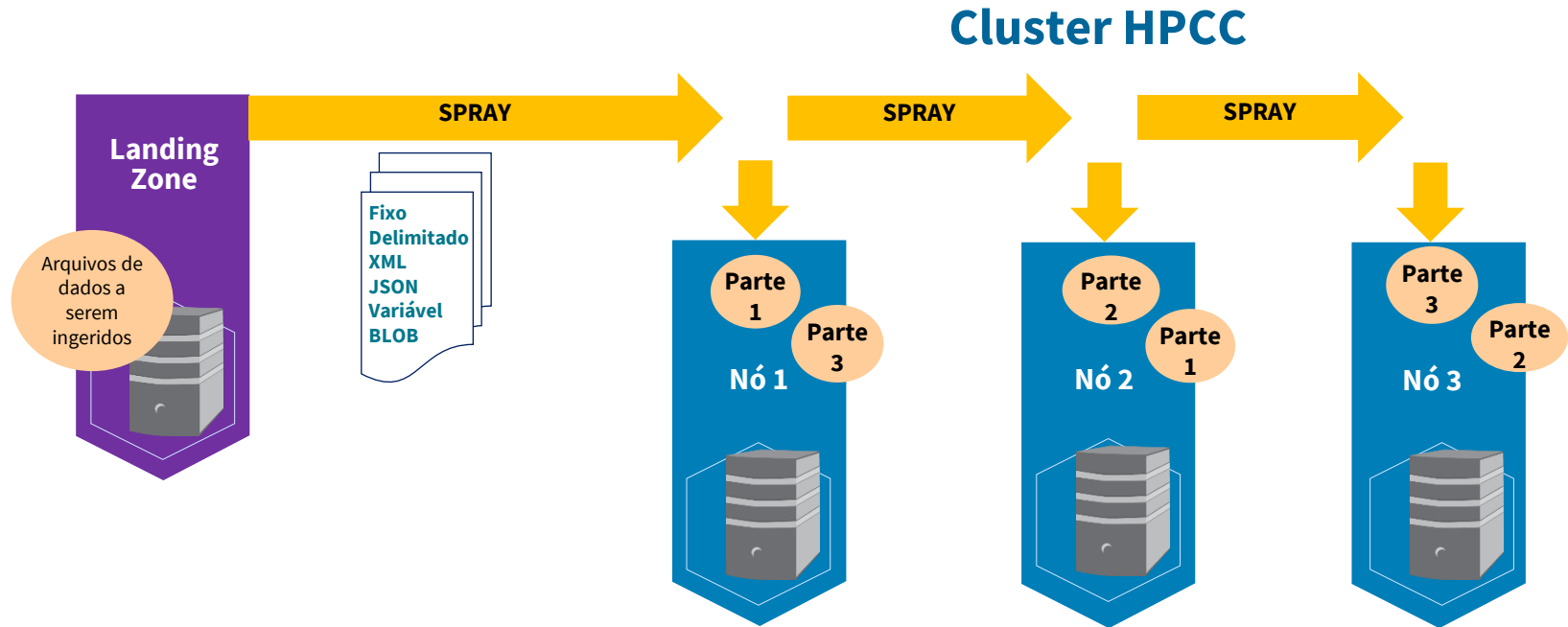
2. Enterprise Control Language (ECL)

- **“Inline” Datasets utilizados durante a aula expositiva:**

Firstname	Lastname	Gender	Age	Nationality	Occupation	Account	Balance	Income
Isaac	Newton	M	84	english	scientist	cc100	100	3500.00
Albert	Einstein	M	76	german	scientist	cc200	-100	4000.30
Marie	Curie	F	66	polish	scientist	cc300	200	3640.10
Victor	Hugo	M	83	french	writer	cc400	150	1900.00
Jane	Austen	F	41	english	writer	cc500	180	2000.00
Emily	Bronte	F	30	english	writer	cc600	120	1800.00
Jane	Doe		25	brazilian	unemployed	cc700	-500	0.00
John	Doe	U	65	american	retired	cc800	750	3211.11

Firstname	Lastname	Email	Phone
ISAAC	NEWTON	isaac.newton@cam.ac.uk	16431727
ALBERT	EINSTEIN	albert.einstein@princeton.edu	18791955
MARIE	CURIE	marie.curie@sorbonne.fr	18671934
VICTOR	HUGO	victor.hugo@lacroix.fr	18021885
JANE	AUSTEN	jane.austen@hampshire.uk	17751817
EMILY	BRONTE	emily.bronte@thornton.uk	18181848
JANE	DOE	jane.doe@hotmail.com	
JOHN	WAYNE	john.wayne@paramount.com	12345678

3. Extração dos dados



As partes do arquivo são referenciadas em ECL como um único arquivo lógico...

3. Extração dos dados

- **Escopo e Nomes de arquivos lógicos**

Os nomes de arquivos lógicos sempre começam com um escopo (estrutura de diretórios) e termina com o nome do arquivo.

O HPCC busca por **arquivos cujos nomes começam com um escopo padrão (THOR):**

'DIR1::DIR2::NomeArquivo' //dado isso, HPCC procura por:

'**THOR**::DIR1::DIR2::NomeArquivo' //esse arquivo

O sinal de “til” (~) indica a **supressão do escopo padrão:**

'~DIR1::DIR2::NomeArquivo' //dado isso, HPCC procura por:

'DIR1::DIR2:: NomeArquivo' //esse arquivo

3. Extração dos dados

- **Upload do “raw” Dataset para a ‘Zona de Entrada de Arquivos’:**
 - Importação dos Dados Brutos (Upload = *Landing Zone*)

Uploader(Enviador) de Arquivo

Zona de Entrada de Arquivos:

mydropzone

Máquinas:

10.0.0.91

Pasta:

/UFSC

#	Type	File Name	Size
1	CSV	<u>DatasetInline01.csv</u>	495

Sobrescrever

Começar

Fechar

3. Extração dos dados

- **Spray para a distribuição do arquivo entre os 'nós' do *Cluster*:**
 - Spray do arquivo (Spray = *Logical Files*)

Spray (Distribuir aos Nós): Fixo ▾ Delimitado ▾ XML ▾ JSON ▾ Variável ▾ BLOB ▾

Destino

Grupo: mythor ▾

Fila: dfusever_queue ▾

Escopo de Destino: CLASS::MDM::DEMO::

Nome do Destino

Datasetline01

Opções

Formato: ASCII ▾

Maximo tamanho do registro: 8192

Seperadores: \,

Omitir Separador: ☐

Escapar:

Terminador de linhas: \n, \r, \n

Aspas: "

Sobrescrever: ☐

Sem Separador: ☐

Compactar: ☐

Estrutura de registro disponivel: ☐

Expira em (dias):

Replicar: ☐

Incomum: ☒

Falha em caso de arquivo sem fonte: ☐

Terminador em Aspas: ☐

Replicação atrasada: ☒

Spray (Distribuir aos Nós)

Próximos passos

▪ **Cursos online: +170 aulas** (<https://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

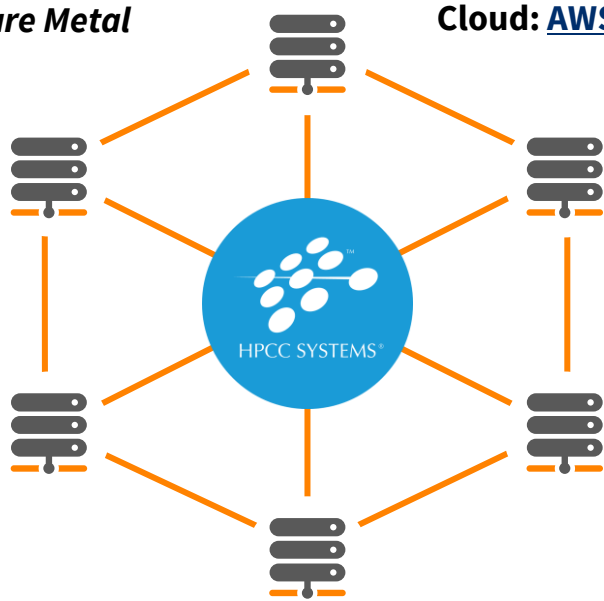
- Visão geral e aplicações da plataforma

Próximos passos

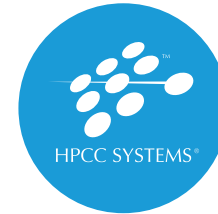
- Opções de utilização: **Playground** (<http://play.hpccsystems.com:8010/>)

Bare Metal

Cloud: [AWS](#)/[Azure](#)



Oracle Virtual Box
HyperV
[Docker](#)
Gitpod

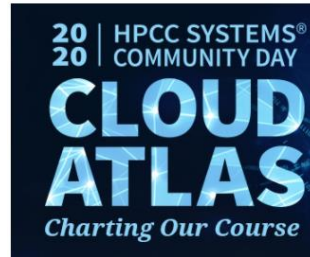


[HPCC Máquina Virtual](#)

Saiba mais em: <https://hpccsystems.com/getting-started/>

Próximos passos

- **Relacionamento com a Comunidade Acadêmica:**
<https://hpccsystems.com/community/academics>



Universidade de São Paulo
Brasil



Próximos passos

■ Universidades Brasileiras

Universidade de São Paulo
Brasil



- Disciplina Optativa na Poli/USP ([Link](#))
- Cursos de extensão ([Link](#))
- Co-Orientação de IC's (PIBIC [Link1](#) [Link2](#) [Link3](#))



UNIVERSIDADE FEDERAL
DE SANTA CATARINA

- Co-Orientação de TCC's ([Link1](#) [Link2](#))
- Co-Orientação de IC's ([Link](#))
- Co-Autoria de artigos científicos ([Link](#))
- Auxílio para aquisição de equipamentos

Próximos passos

- **Projetos de pesquisa**



Saiba mais em: <https://wiki.hpccsystems.com/display/hpcc/Available+Projects>

Próximos passos

■ Oportunidades profissionais

#ExploreMore


<https://risk.lexisnexis.com/about-us/careers>
<https://www.linkedin.com/company/lexisnexis-risk-solutions/>
<https://www.vagas.com.br/v2273659>

#Contato

Ana Cristina Vieira
Senior Talent Acquisition
LexisNexis Risk Solutions Group
☎ +55.11.97075.5659
ana.vieira@lexisnexisrisk.com

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
- Poster Competition: [Link](#)



Faça parte da Comunidade

Registre-se em:
<https://hpccsystems.com/pt-br>

