

WSCAD 2023

Porto Alegre/RS

**XXIV Simpósio em
Sistemas
Computacionais de
Alto Desempenho**





Processamento e Análise de *Big Data* e Aplicação de Algoritmos de *Machine Learning* através da utilização da Plataforma *HPCC Systems*

Minicurso: Objetivo

Processamento e Análise de *Big Data* e Aplicação de Algoritmos de *Machine Learning* através da utilização da Plataforma *HPCC Systems*

Objetivo: Ao longo do minicurso, os participantes terão a oportunidade de conhecer os conceitos essenciais de processamento e análise de volumes massivos de dados (*Big Data*) e o processo de desenvolvimento de um serviço de consulta fazendo uso da plataforma *open-source* composta por um *Cluster* Computacional de Alto Desempenho (*HPCC Systems*) e, também, a utilização de algoritmos de Aprendizado de Máquina, bem como terão a possibilidade de aplicar os conhecimentos adquiridos em um ambiente de treinamento disponibilizado em sala de aula.

Informações Técnicas: Curso de nível básico. O curso requer apenas um computador com acesso à Internet e uma conta no [Github](#).

Minicurso: Autores

Autores



- **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
12 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
04 anos como engenheiro de software e mentor em projetos acadêmicos
06 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;
 - Desenvolvimento de um serviço de consulta;
 - Utilização de algoritmos de Aprendizado de Máquina.
- **Próximos passos**
 - Cursos *online*;
 - Projetos de pesquisa;
 - Oportunidades profissionais.
- **Considerações Finais**

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;
 - Desenvolvimento de um serviço de consulta;
 - Utilização de algoritmos de Aprendizado de Máquina.
- **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 de um grupo contendo um portfólio de marcas, abrangendo vários setores que fornecem aos clientes tecnologias inovadoras, análises baseadas em informações e ferramentas de decisão e serviços de dados.

LexisNexis Risk Solutions: Corporação



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

RELX Group



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



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;
- Desenvolvimento de um serviço de consulta;
- Utilização de algoritmos de Aprendizado de Máquina.

➤ **Próximos passos**

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

➤ **Considerações Finais**

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)

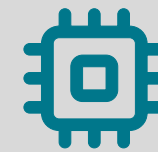
2012 - 16



Melhorias contínuas com **FOCO NA QUALIDADE**

Suporte e treinamento aprimorado

2017-Presente



Aprimoramentos de arquitetura (*Cloud*)
Desenvolvimentos em *Machine Learning*

HPCC Systems: Ativos e Clientes

| 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 |

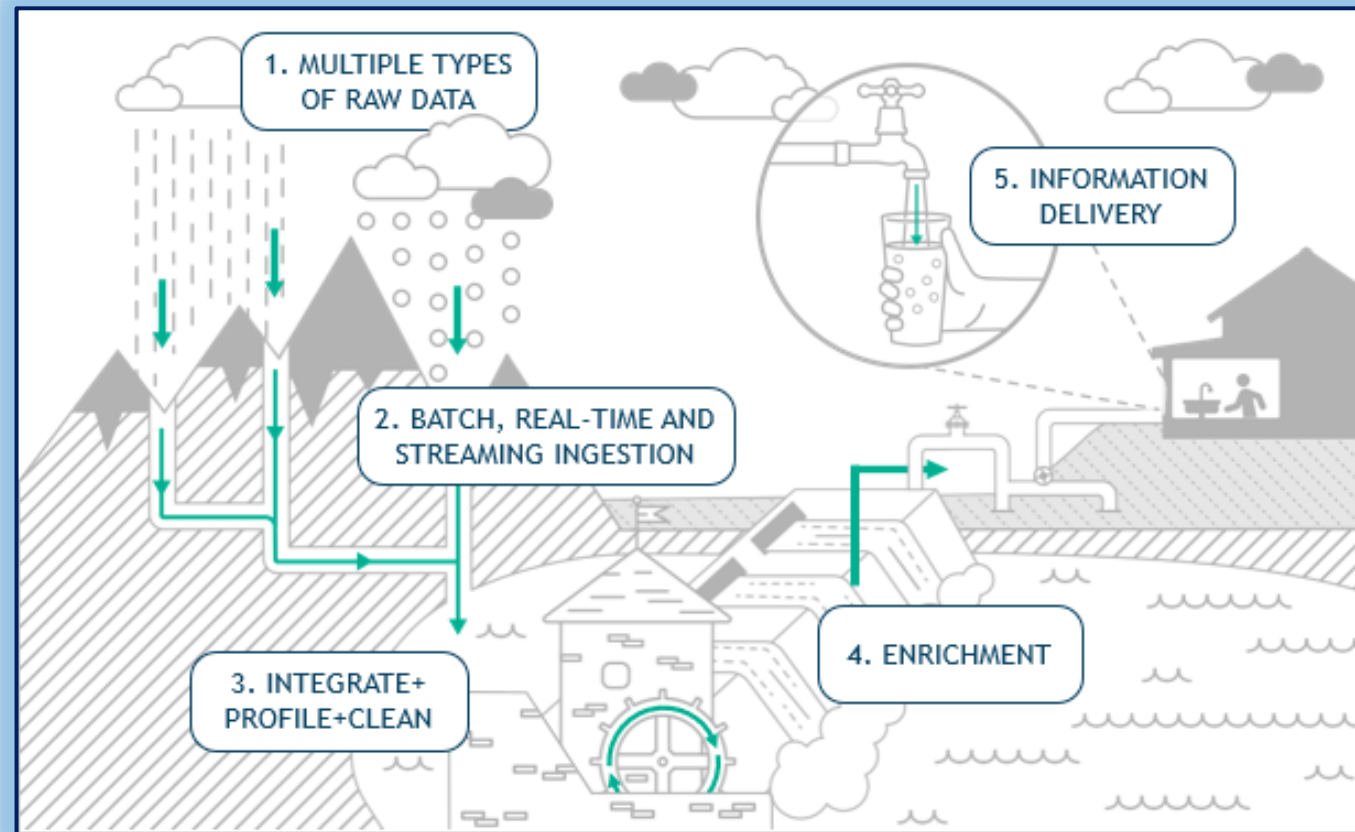
- 12 petabytes de dados públicos e privados
- 270 milhões de transações por hora
- Clientes em mais de **100** países
- **76%** de todas as empresas Fortune 500
- **7** dos 10 maiores bancos do mundo
- **100%** dos 50 maiores bancos americanos
- **95 das 100** maiores seguradoras
- **Mais de 7.500** órgãos governamentais: locais, estaduais e federais

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;
 - Desenvolvimento de um serviço de consulta;
 - Utilização de algoritmos de Aprendizado de Máquina.
- **Próximos passos**
 - Cursos *online*;
 - Projetos de pesquisa;
 - Oportunidades profissionais.
- **Considerações Finais**

HPCC Systems: Visão Geral

- Apresentação de conceitos



<https://youtu.be/FDuCuDRy1wU>

HPCC Systems

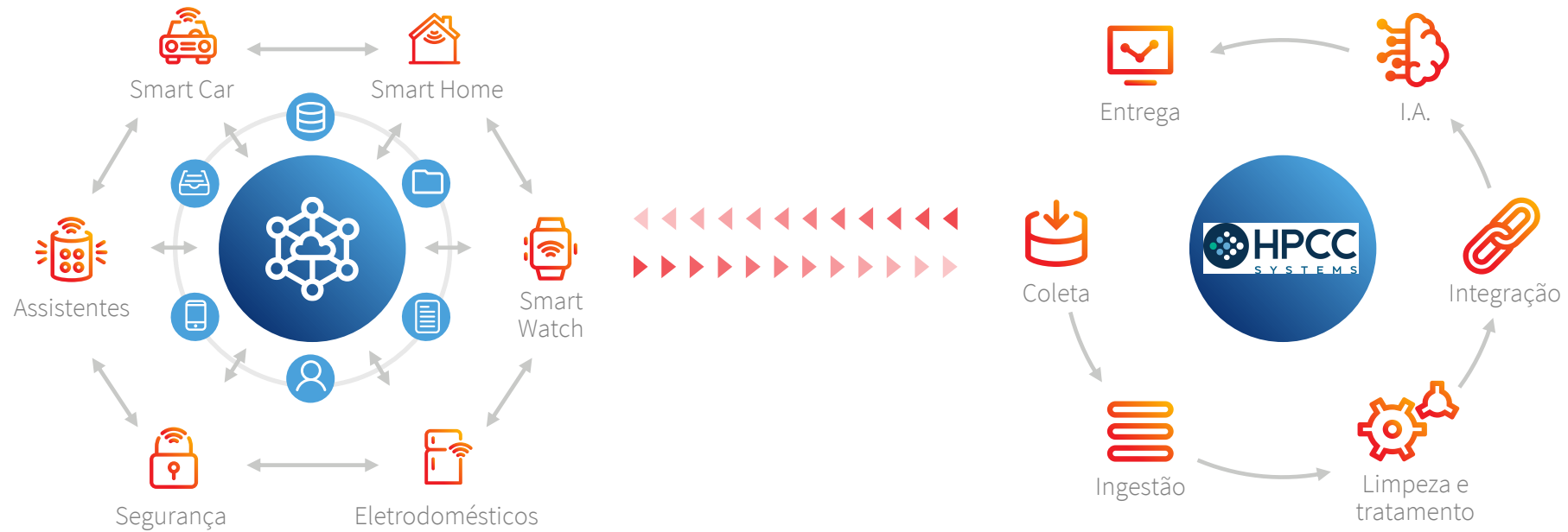
Processamento e Análise de *Big Data*

Trade Off do Big Data

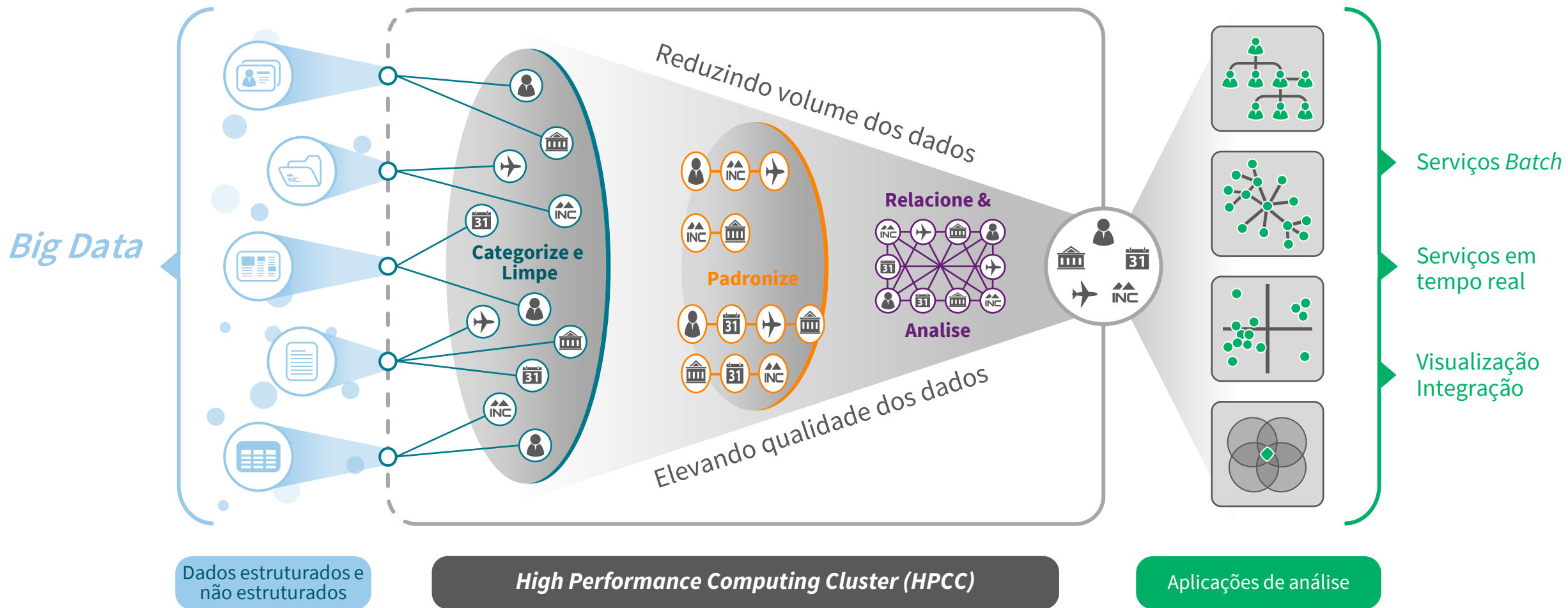
- Problema **N²**
- Quantidade de dados X Recursos computacionais
- Como processar bilhões de registros em segundos?
- Como analisar múltiplas fontes de dados e transformá-las em informação e conhecimento?

Saiba mais em: <https://www.internetlivestats.com/>

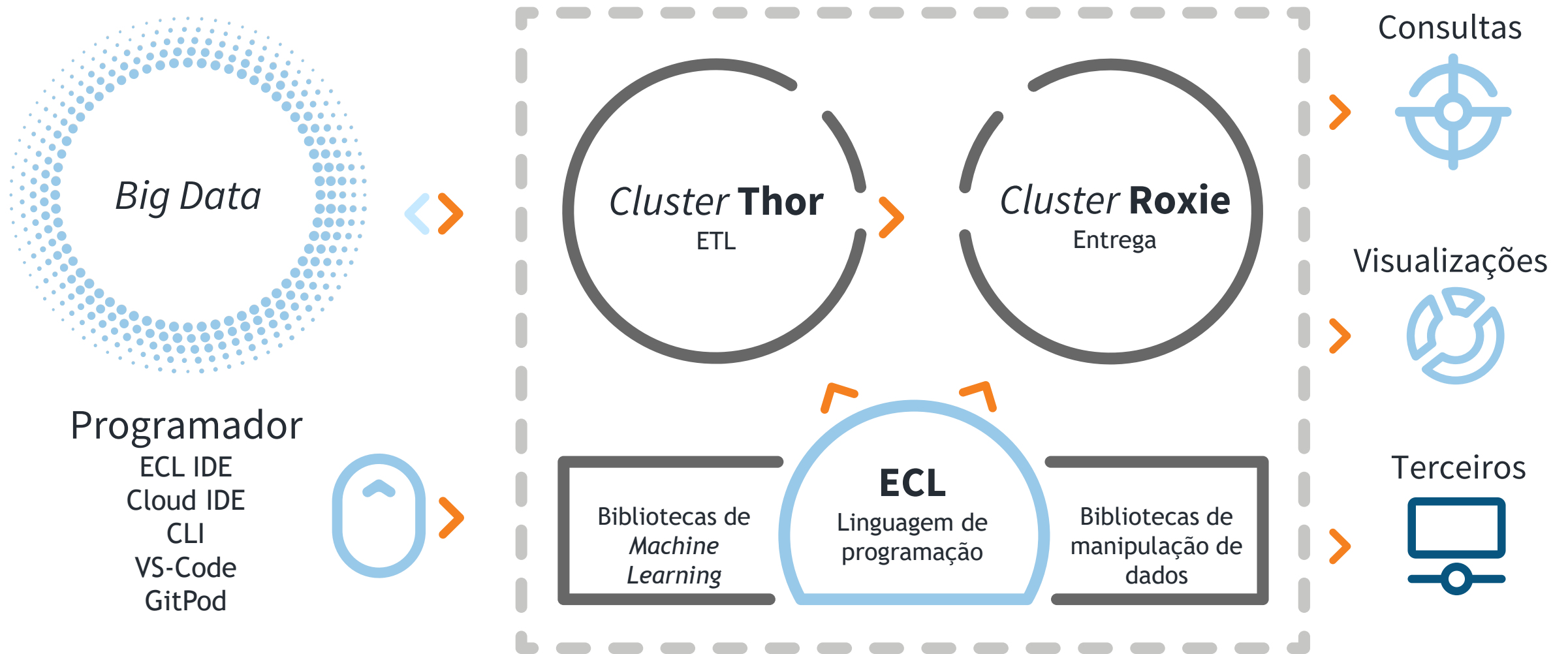
Como trabalha a plataforma *HPCC Systems*?



Fluxo de dados no *HPCC Systems*: “Funil” de dados

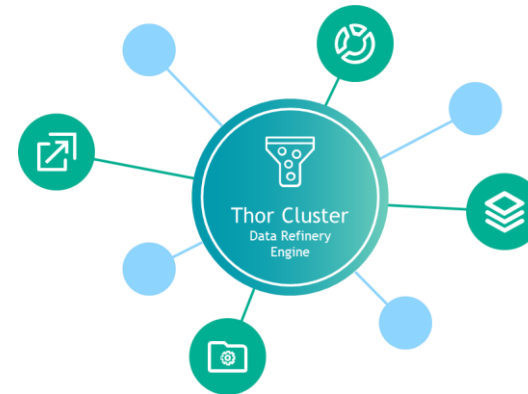


Arquitetura da plataforma *HPCC Systems*



Thor e Roxie: Objetivos específicos

Thor:
“Identificar e catalogar todos os seres vivos nos oceanos”



ROXIE:
“Disponibilizar todas as informações sobre uma espécie”





Enterprise Control Language (ECL)

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

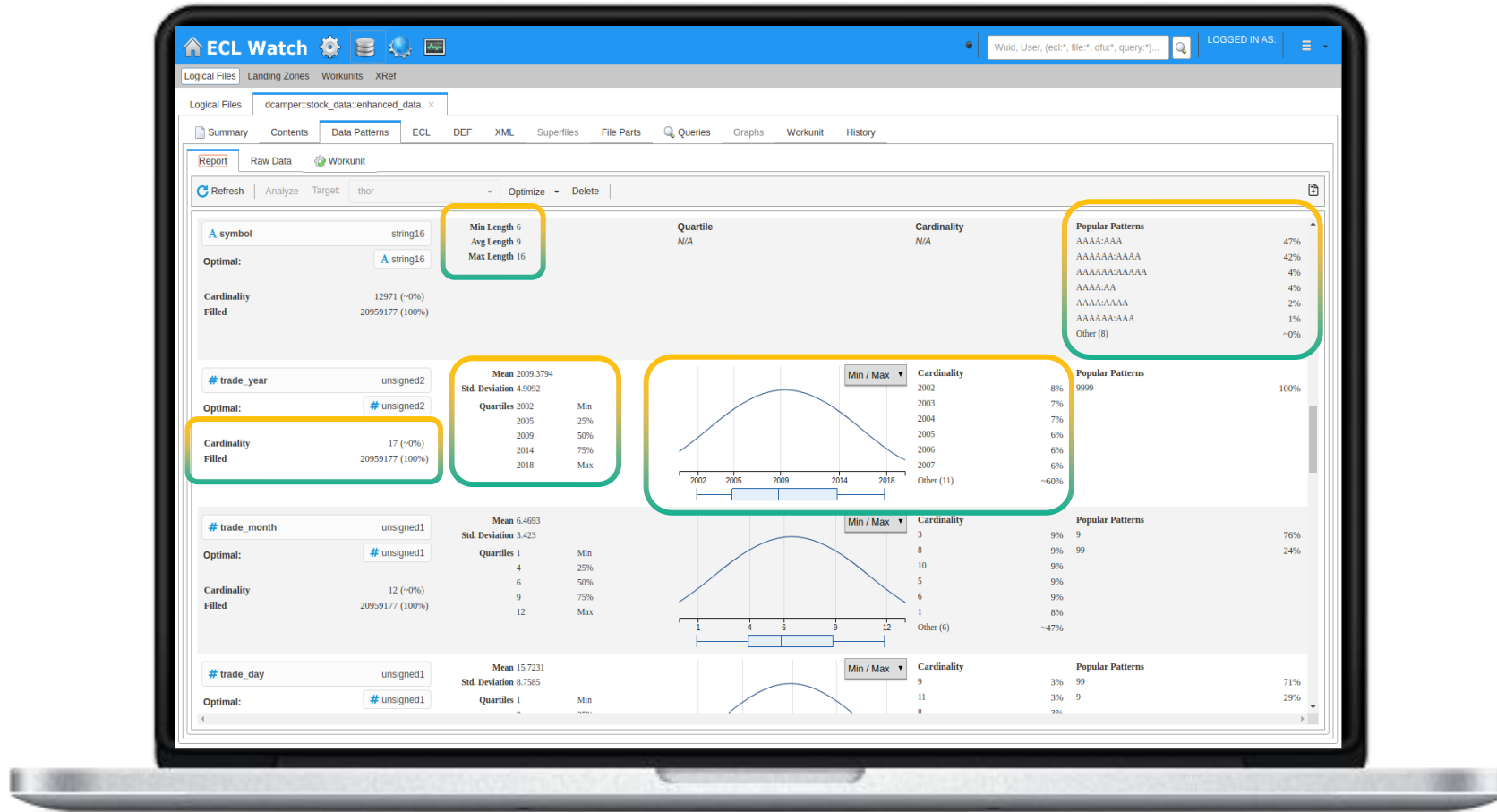
- Declarativa e não-procedural
- Códigos menores e reutilizáveis
- Biblioteca para manipulação de dados

Compilador

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

Como fazer  vs.  **O que fazer**

Bibliotecas de "Profiling" de datos



Bibliotecas de “Machine Learning”



Não Supervisionado

Clusterização

DBSCAN
K-Means

PLN

Text Vectors



Supervisionado

Classificação

SVM
Árvores de decisão
Regression logística
Classification Forest

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

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;

▪ Desenvolvimento de um serviço de consulta;

▪ Utilização de algoritmos de Aprendizado de Máquina.

➤ **Próximos passos**

▪ Cursos *online*;

▪ Projetos de pesquisa;

▪ Oportunidades profissionais.

➤ **Considerações Finais**

HPCC Systems: Visão Geral

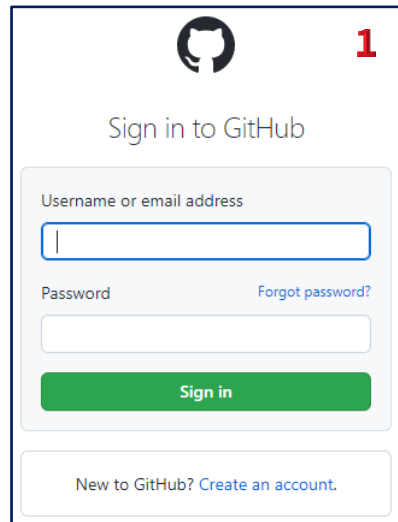
■ Aplicação de conhecimentos

The screenshot displays the HPCC Systems ECL Watch interface. On the left, the 'Tarefas' (Tasks) panel shows a list of tasks with WUIDs. The main area is split into a file explorer on the left and a code editor on the right. The code editor shows ECL code for a record structure. Below the code editor, a data table is displayed with 8 rows of results.

| firstname | lastname | gender | age | nationality | occupation |
|-----------|----------|--------|-----|-------------|------------|
| Isaac | Newton | M | 84 | english | scientist |
| Albert | Einstein | M | 76 | german | scientist |
| Marie | Curie | F | 66 | polish | scientist |
| Victor | Hugo | M | 83 | french | writer |
| Jane | Austen | F | 41 | english | writer |
| Emily | Bronte | F | 30 | english | writer |
| Jane | Doe | | 25 | brazilian | unemployed |
| John | Doe | U | 65 | american | retired |

A) Configuração do ambiente: *Github* / *Gitpod* sem “Fork”

1



Sign in to GitHub

Username or email address

Password [Forgot password?](#)

Sign in

New to GitHub? [Create an account.](#)

5

Popular repositories

WSCAD_2022 Public


ECL 1

6

During the workshop GitPod will be used as main environment:


1. By using your GitHub credentials, just click on the following link for instantiate a environment via GitPod:
<https://gitpod.io/#https://github.com/mauromarx/WSCAD>

7

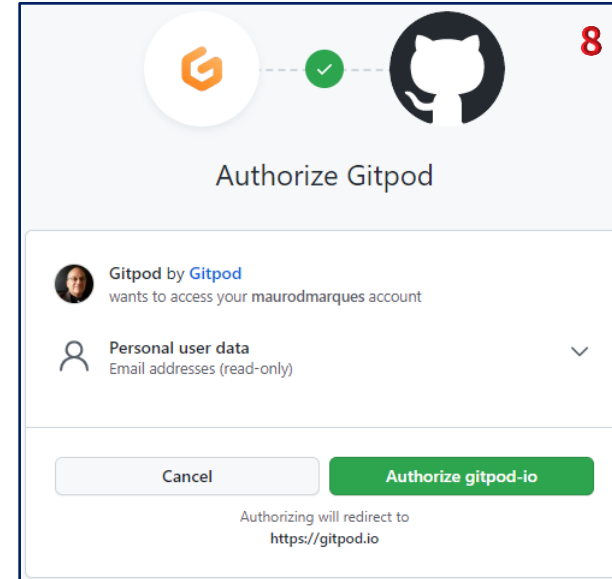


Gitpod

Open a cloud-based developer environment for the repository mauromarx/WSCAD

 Continue with GitHub

8



Authorize Gitpod

Gitpod by Gitpod wants to access your mauromarques account

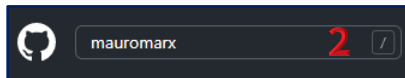
Personal user data
Email addresses (read-only)

Authorize gitpod-io

Cancel

Authorizing will redirect to <https://gitpod.io>

2



mauromarx

3


Wikis 0

Users 1

[Advanced search](#) [Cheat sheet](#)

4

1 user

 Mauro D. Marques mauromarx
Software Engineer
São Paulo - Brazil

9

Select Editor

Choose the editor for opening workspaces. You can always change later the editor in [user preferences](#).

VS Code BROWSER

VS Code

IntelliJ IDEA


GoLand

10

User Validation Required

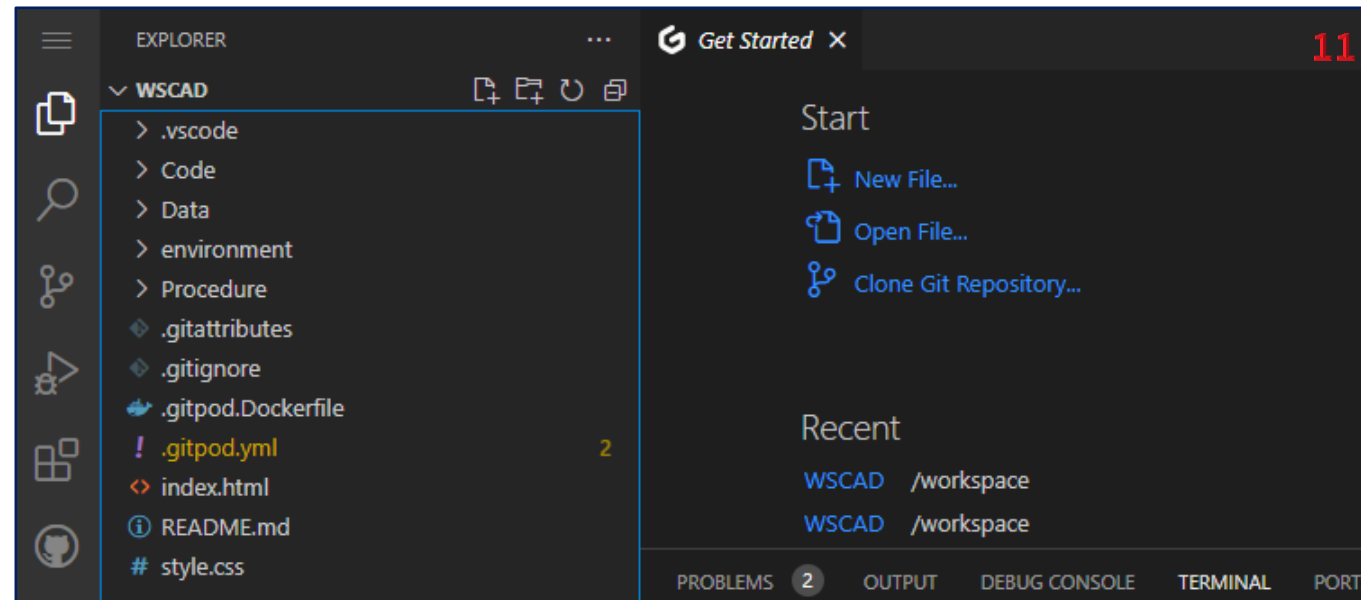
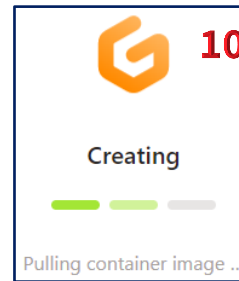
Enter a mobile phone number you would like to use to verify your account. Having trouble? [Contact support](#)

Mobile Phone Number

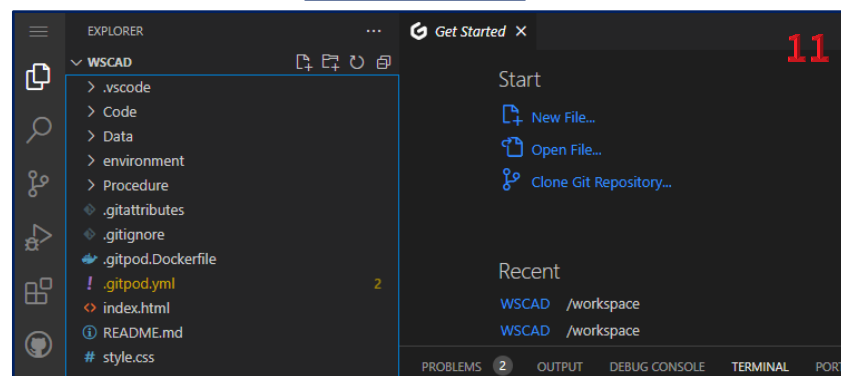
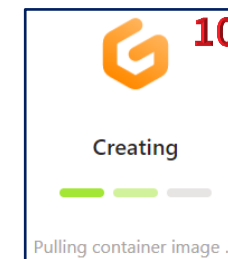
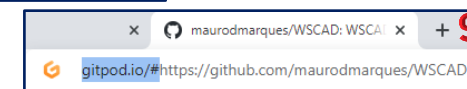
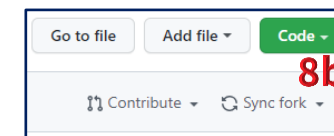
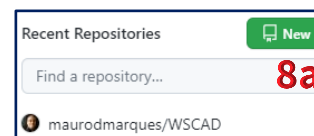
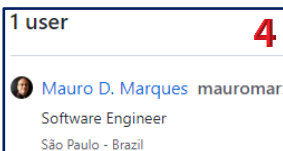
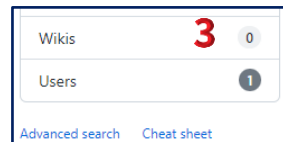
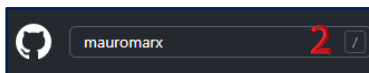
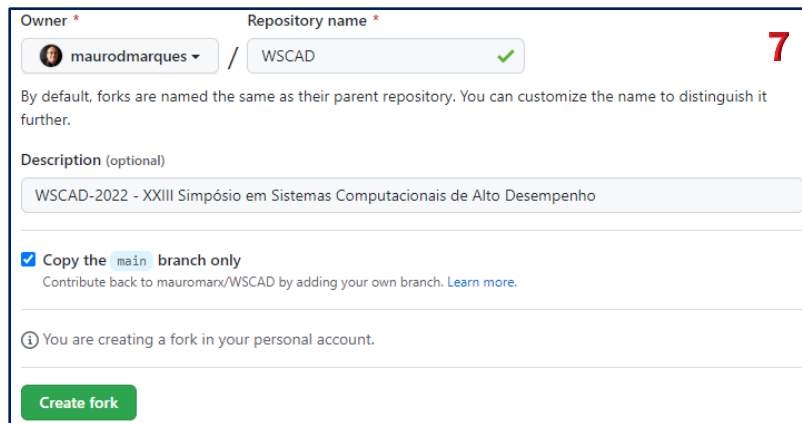
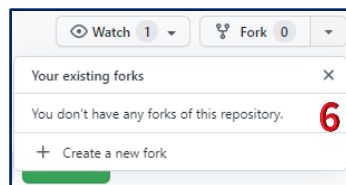
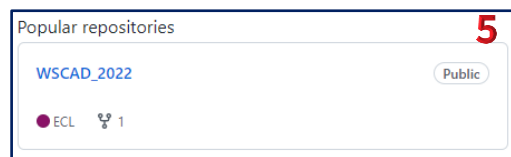
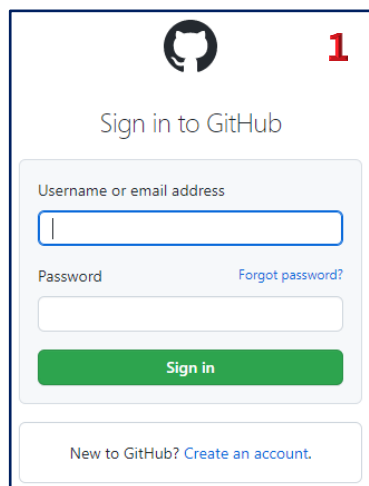
 (11) 9

Send Code via SMS

A) Configuração do ambiente: *Github* / *Gitpod* sem “Fork”



A) Configuração do Ambiente: *Github* / *Gitpod* com “Fork”



B) 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

// referenciando um campo em um *dataset*

NomedoDiretorio.Definicao

// referenciando uma definição em outro módulo

B) Enterprise Control Language (ECL)

✓ O código ECL é constituído de Definições e Ações

✓ Definições estabelecem *o que* as coisas são, não *como* fazê-las

(arquivos de definição ECL)

```
MyDef := 'Hello World';           // não inicia uma WU
```

✓ Ações em ECL resultam em compilação e execução

(arquivos BWR)

```
OUTPUT(MyDef);                   // inicia uma WU  
OUTPUT('Hello World, again...'); // inicia uma WU
```

B) Enterprise Control Language (ECL)

- “*Inline*” datasets utilizados durante a primeira parte da demonstração:

| 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 |

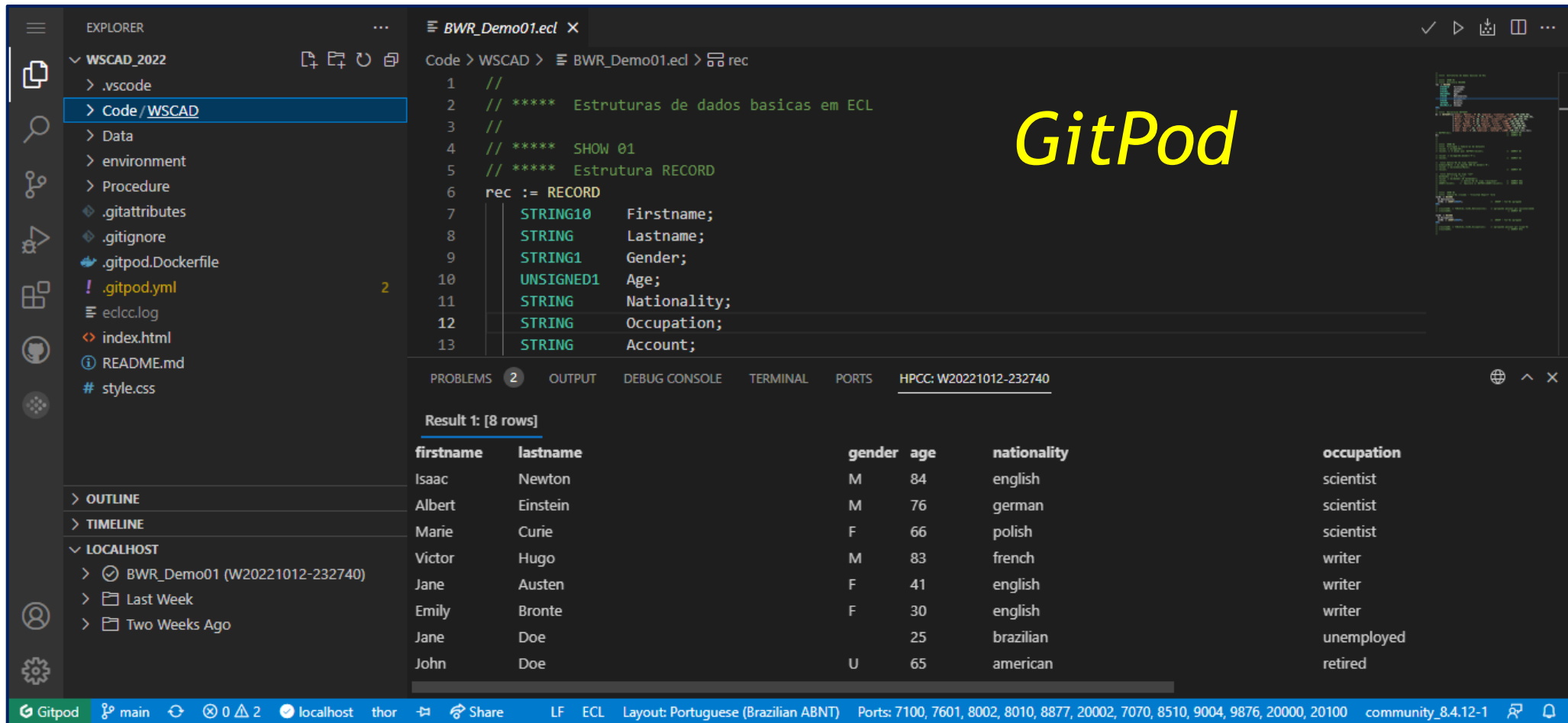
| Firstname | Lastname | Email | Phone |
|-----------|----------|-------------------------------|----------|
| ISAAC | NEWTON | isaac.newton@cam.ac.uk | 16431727 |
| ALBERT | EINSTEIN | albert.einstein@princeton.edu | 18791955 |
| MARIE | CURIE | marie.curie@sorbonne.fr | 18671934 |
| VICTOR | HUGO | 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 |

RoadMap 01 - Sequência dos códigos ECL:

❑ Demonstração - ECL codes :

1. BWR_Tests.ecl (submit)
2. BWR_Demo00.ecl (submit)
3. BWR_Demo01.ecl (submit)
4. BWR_Demo02.ecl (submit)
5. BWR_Demo03.ecl (submit)
6. BWR_Demo04.ecl (submit)
7. BWR_Demo05.ecl (submit)
8. modInline01.ecl
9. modInline02.ecl
10. BWR_BrowseData01.ecl (submit)

B) Enterprise Control Language (ECL)



The screenshot displays the GitPod IDE interface. The Explorer on the left shows a project named 'WSCAD_2022' with a file 'BWR_Demo01.ecl' selected. The main editor shows the following ECL code:

```
Code > WSCAD > BWR_Demo01.ecl > rec
1 //
2 // ***** Estruturas de dados basicas em ECL
3 //
4 // ***** SHOW 01
5 // ***** Estrutura RECORD
6 rec := RECORD
7     STRING10  Firstname;
8     STRING    Lastname;
9     STRING1   Gender;
10    UNSIGNED1 Age;
11    STRING    Nationality;
12    STRING    Occupation;
13    STRING    Account;
```

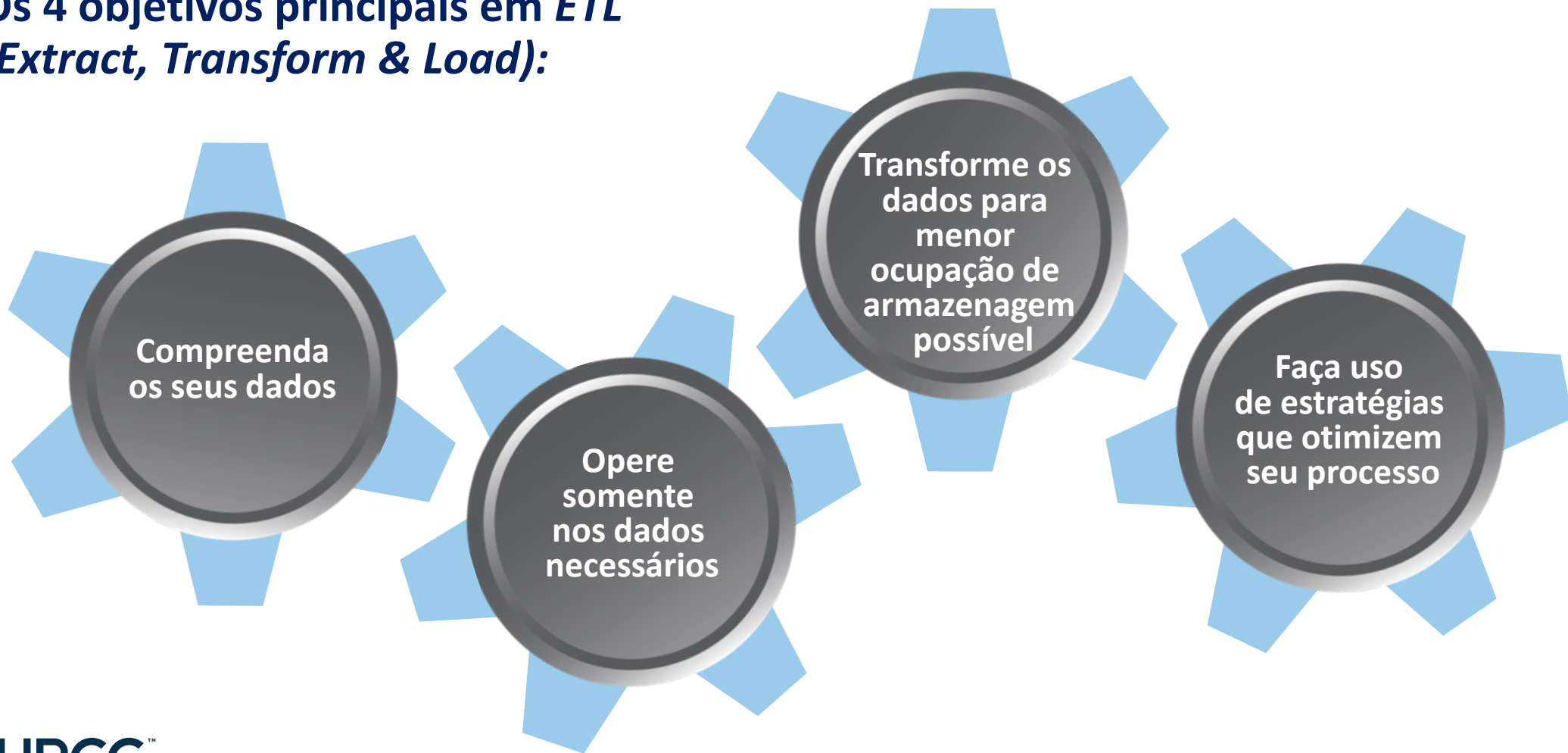
The output window shows the result of the execution, displaying a table with 8 rows:

| firstname | lastname | gender | age | nationality | occupation |
|-----------|----------|--------|-----|-------------|------------|
| Isaac | Newton | M | 84 | english | scientist |
| Albert | Einstein | M | 76 | german | scientist |
| Marie | Curie | F | 66 | polish | scientist |
| Victor | Hugo | M | 83 | french | writer |
| Jane | Austen | F | 41 | english | writer |
| Emily | Bronte | F | 30 | english | writer |
| Jane | Doe | | 25 | brazilian | unemployed |
| John | Doe | U | 65 | american | retired |

The status bar at the bottom indicates the current environment is 'main' on 'localhost', with ports 7100, 7601, 8002, 8010, 8877, 20002, 7070, 8510, 9004, 9876, 20000, and 20100 open. The layout is set to Portuguese (Brazilian ABNT).

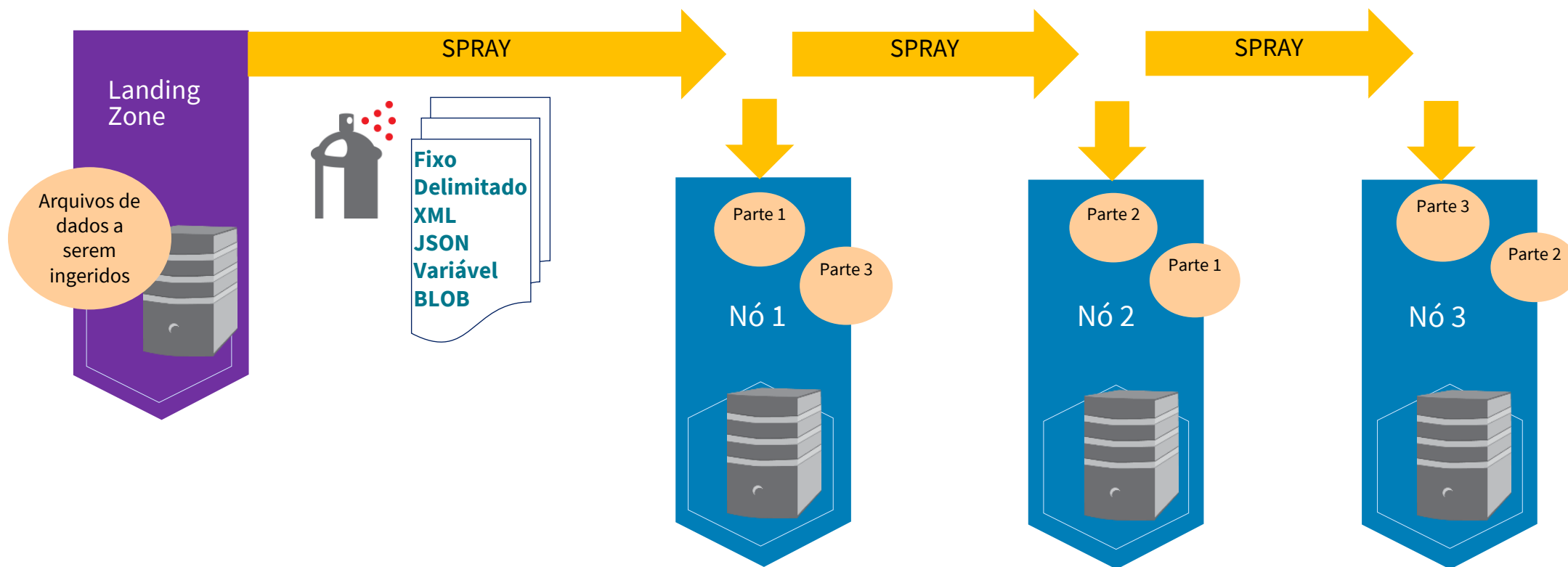
C) ETL

- Os 4 objetivos principais em *ETL* (*Extract, Transform & Load*):



C) Extração dos dados

• Operação *spray*



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

C) Extração dos dados

▪ Escopo e nomes de arquivos lógicos

Nomes de arquivos sempre começam com um escopo (estrutura de diretórios) e terminando 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

C) 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

C) 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

DatasetInline01

Opções

Formato: ASCII ▾

Maximo tamanho do registro: 8192

Seperadores: \\

Omitir Separador:

Escapar:

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

Aspas: "

Sobrescrever: Replicar:

Sem Separador: Incomum:

Compactar: Falha em caso de arquivo sem fonte:

Estrutura de registro disponível: Terminador em Aspas:

Expira em (dias): Replicação atrasada:

Spray (Distribuir aos Nós)

C) Extração dos dados

- **“Raw” dataset utilizado durante a segunda parte da demonstração:**
 - Importação dos Dados Brutos (Upload = *Landing Zone*)

Persons

Fixo: 151

963.512
registros

Zona de Entrada de Arquivos

Largura: 151 EBCDIC:

| | | | | | | | | |
|----|------|----------------|-----------------|---|------------|---|---------------------------------|-----------------|
| 1 | 0000 | AB**Cherianne | Khatchatourian | N | 19990922** | M | 69 BOULDER RIDGE RD # 25A | HAWKINS |
| 2 | 0097 | BB**Muyesser | Raplee | X | 20001111a | F | 55 SWAMP RD | DISTRICT HEIGHT |
| 3 | 012e | CB**Roselin | Viceconte | | 19990325X | F | 19800113107 HILL TER | ENTERPRISE |
| 4 | 01c5 | DB**Inda | Provines | | 20000909** | U | 290 W MOUNT PLEASANT AVE | LAVACA |
| 5 | 025c | EB**Inderdeep | Laurence | D | 20001228X | M | 44 PROSPECT PL | GREENSBORO |
| 6 | 02f3 | FB**Chrystine | Mangiapane | | 19990827; | F | 197803061806 1ST AVE APT 8F | ARVADA |
| 7 | 038a | GB**Adelene | Stock | R | 20000827ü | M | 1117 FARM RD | DOVER |
| 8 | 0421 | HB**Mendy | Rufenblanchette | | 20000903** | M | 3 W 83RD ST APT 4C | WILLIAMSTON |
| 9 | 04b8 | IB**Lannie | Amerantes | I | 200012199 | U | 200 W 20TH ST APT 909 | CHARLESTON |
| 10 | 054f | JB**Tare | Gonyeau | T | 199308070 | F | 197508016 CANDLE CT | EL PASO |
| 11 | 05e6 | KB**Finney | Aristilde | P | 19900621X | M | 19560920222 1ST AVE APT 2B | MACON |
| 12 | 067d | LB**Oreoluwa | Marthaler | | 19931006f | F | 19731201176 CLAREMONT GDNS | AUBURN |
| 13 | 0714 | MB**Surge | Abbottkrepp | D | 20000308** | F | 22 LE PARC CT | TWINSBURG |
| 14 | 07ab | NB**Dave | Mcjury | | 20001129i | U | 510 COOPER RD # 1 | TACOMA |
| 15 | 0842 | OB**Ramsay | Ping | | 20001129i | M | 404 AVENUE L | MESQUITE |
| 16 | 08d9 | PB**Lacides | Wisniveskydr | Q | 20001227** | M | 23 JEFFERSON LN | ALTON BAY |
| 17 | 0970 | QB**Hazele | Scoggins | H | 20001102** | M | 43 RENAISSANCE DR | COTATI |
| 18 | 0a07 | RB**Laini | Mandrake | B | 20001205** | M | 5 SHEVCHENKO AVE | CROSS RIVER |
| 19 | 0a9e | SB**Setthaphon | Zeuli | N | 20000922ö | M | 196212011714 E WILLOW GROVE AVE | FORT WORTH |
| 20 | 0b35 | TB**Tomas | Drabick | U | 19830116** | F | 19550901155 E 34TH ST # 20 | LEES SUMMIT |
| 21 | 0bcc | UB**Ginabelle | Munkel | Q | 20000914** | M | 833 SUMMER ST STE 3B | NEW ALEXANDRIA |
| 22 | 0c63 | VB**Ornah | Aschermann | | 200012079 | M | 36 UNION ST | CHARTLEY |
| 23 | 0cfa | WB**Moisey | Shupp | R | 200012299 | M | 12 STEEPLECHASE LN | ATTALLA |
| 24 | 0d91 | XB**Jynevelyne | Hirschkind | | 19870618** | F | 19630324404 SUMMIT AVE | PEPPERELL |
| 25 | 0e28 | YB**Norli | Pisciotta | V | 19870320~ | F | 1933120123 SEVINOR RD # 0 | FORT GEORGE G M |
| 26 | 0ebf | ZB**Toai | Ibric | | 20000706** | M | 17 TOLKIEN PSGE | MADISONVILLE |
| 27 | 0f56 | [B**Murshell | Bykov | | 20001209a | F | 2142 MOUNTAIN VIEW AVE | PALM COAST |

C) 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: dfuserver_queue ▾

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

| Nome do Destino | Tamanho do Registro |
|-----------------|---------------------|
| persons | 151 |

Opções

Sobrescrever:

Sem Separador:

Compactar:

Expira em (dias):

Replicar:

Incomum:

Falha em caso de arquivo sem fonte:

Replicação atrasada:

Spray (Distribuir aos Nós)

C) Extração dos dados

▪ Definição da Estrutura de Dados & Análise do Perfil (*Profiling*) dos dados

```
EXPORT modPersons := MODULE
  EXPORT Layout := RECORD
    INTEGER4 RecID;
    STRING15 FirstName;
    STRING25 LastName;
    STRING15 MiddleName;
    STRING2 NameSuffix;
    STRING8 FileDate;
    UNSIGNED2 BureauCode;
    STRING1 MaritalStatus;
    STRING1 Gender;
    UNSIGNED1 DependentCount;
    STRING8 BirthDate;
    STRING42 StreetAddress;
    STRING20 City;
    STRING2 State;
    STRING5 ZipCode;
  END;
  EXPORT File := DATASET('~CLASS::MDM::DEMO::persons', Layout, THOR);
END;
```

| ## | recid | firstname | lastname | middlename | namesuffix | filedate | bureaucode | maritalstatus | gender | dependentcount | birthdate | streetaddress |
|----|---------|-----------|-----------------|------------|------------|----------|------------|---------------|--------|----------------|-----------|------------------------|
| 1 | 1000001 | Cherianne | Khatchatourian | N | | 19990922 | 24 | | M | 4 | | 69 BOULDER RIDGE RD # |
| 2 | 1000002 | Muyesser | Raplee | X | | 20001111 | 353 | | F | 0 | | 55 SWAMP RD |
| 3 | 1000003 | Roselin | Viceconte | | | 19990325 | 344 | | F | 4 | 19800113 | 107 HILL TER |
| 4 | 1000004 | Inda | Provines | | | 20000909 | 13 | | U | 3 | | 290 W MOUNT PLEASANT P |
| 5 | 1000005 | Inderdeep | Laurence | D | | 20001228 | 344 | | M | 3 | | 44 PROSPECT PL |
| 6 | 1000006 | Chrystine | Mangiapane | | | 19990827 | 315 | | F | 0 | 19780306 | 1806 1ST AVE APT 8F |
| 7 | 1000007 | Adelene | Stock | R | | 20000827 | 252 | | M | 0 | | 1117 FARM RD |
| 8 | 1000008 | Mendy | Rufenblanchette | | | 20000903 | 24 | | M | 1 | | 3 W 83RD ST APT 4C |
| 9 | 1000009 | Lannie | Amerantes | I | | 20001219 | 313 | | U | 0 | | 200 W 20TH ST APT 909 |
| 10 | 1000010 | Tare | Gonyeau | T | | 19930807 | 48 | | F | 0 | 19750801 | 6 CANDLE CT |
| 11 | 1000011 | Finney | Aristilde | P | | 19900621 | 344 | | M | 2 | 19560920 | 222 1ST AVE APT 2B |
| 12 | 1000012 | Oreoluwa | Marthaler | | | 19931006 | 358 | | F | 5 | 19731201 | 176 CLAREMONT GDNS |
| 13 | 1000013 | Surge | Abbottkrepp | D | | 20000308 | 13 | | F | 4 | | 22 LE PARC CT |
| 14 | 1000014 | Dave | Mcjury | | | 20001129 | 238 | | U | 1 | | 510 COOPER RD # 1 |

C) Transformação e carregamento dos dados

- Limpeza, padronização e consolidação de registros

| ## | recid | id | firstname | lastname | middlename | namesuffix | filedate | bureaucode | gender | birthdate | streetaddress | city | state | zipcode |
|----|-------|---------|-----------|-----------------|------------|------------|----------|------------|--------|-----------|---------------------------|-----------------|-------|---------|
| 1 | 1 | 1000001 | CHERIANNE | KHATCHATOURIAN | N | | 19990922 | 24 | M | 0 | 69 BOULDER RIDGE RD # 25A | HAWKINS | WI | 54530 |
| 2 | 2 | 1000002 | MUYESSER | RAPLEE | X | | 20001111 | 353 | F | 0 | 55 SWAMP RD | DISTRICT HEIGHT | MD | 20747 |
| 3 | 3 | 1000003 | ROSELIN | VICECONTE | | | 19990325 | 344 | F | 19800113 | 107 HILL TER | ENTERPRISE | OR | 97828 |
| 4 | 4 | 1000004 | INDA | PROVINES | | | 20000909 | 13 | U | 0 | 290 W MOUNT PLEASANT AVE | LAVACA | AR | 72941 |
| 5 | 5 | 1000005 | INDERDEEP | LAURENCE | D | | 20001228 | 344 | M | 0 | 44 PROSPECT PL | GREENSBORO | FL | 32330 |
| 6 | 6 | 1000006 | CHRISTINE | MANGIAPANE | | | 19990827 | 315 | F | 19780306 | 1806 1ST AVE APT 8F | ARVADA | CO | 80007 |
| 7 | 7 | 1000007 | ADELENE | STOCK | R | | 20000827 | 252 | M | 0 | 1117 FARM RD | DOVER | DE | 19901 |
| 8 | 8 | 1000008 | MENDY | RUFENBLANCHETTE | | | 20000903 | 24 | M | 0 | 3 W 83RD ST APT 4C | WILLIAMSTON | SC | 29697 |
| 9 | 9 | 1000009 | LANNIE | AMERANTES | I | | 20001219 | 313 | U | 0 | 200 W 20TH ST APT 909 | CHARLESTON | WV | 25312 |
| 10 | 10 | 1000010 | TARE | GONYEAU | T | | 19930807 | 48 | F | 19750801 | 6 CANDLE CT | EL PASO | TX | 79924 |
| 11 | 11 | 1000011 | FINNEY | ARISTILDE | P | | 19900621 | 344 | M | 19560920 | 222 1ST AVE APT 2B | MACON | GA | 31220 |
| 12 | 12 | 1000012 | OREOLUWA | MARTHALER | | | 19931006 | 358 | F | 19731201 | 176 CLAREMONT GDNS | AUBURN | ME | 4210 |
| 13 | 13 | 1000013 | SURGE | ABBOTTKREPP | D | | 20000308 | 13 | F | 0 | 22 LE PARC CT | TWINSBURG | OH | 44087 |

RoadMap 02 - Sequência dos códigos ECL:

❑ ETL - ECL codes:

1. modPersons.ecl
2. BWR_Persons_CSV (submit = opcional para CSV testes)
3. modPersons_CSV.ecl (opcional para CSV testes)
4. BWR_BrowseData02.ecl (submit)
5. BWR_DP_Persons.ecl (submit = opcional para CSV testes)
6. BWR_Demo06.ecl (submit)
7. BWR_Demo07.ecl (submit)
8. UID_Persons.ecl
9. STD_Persons.ecl
10. BWR_BrowseData03.ecl (submit)

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;

▪ Desenvolvimento de um serviço de consulta;

▪ Utilização de algoritmos de Aprendizado de Máquina.

➤ **Próximos passos**

▪ Cursos *online*;

▪ Projetos de pesquisa;

▪ Oportunidades profissionais.

➤ **Considerações Finais**

HPCC Systems: Visão Geral

- Desenvolvimento de um serviço de consulta: Dados Pessoais

The screenshot displays a web interface for a service named 'hthor'. The main heading is 'hthor'. Below it, the service name 'fetch_persons_mdm' is shown with several icons (code, data, refresh, save) and a dropdown menu set to 'Dynamic Form'. A checkbox labeled 'FETCH_PERSONS_MDMREQUEST' is checked. The form contains three input fields: 'firstname_value', 'lastname_value', and 'state_value'. At the bottom, there are four buttons: 'Output Tables' (with a dropdown arrow), 'FORM POST' (with a dropdown arrow), 'Submit', and 'Clear All'.

RoadMap 03 - Sequência dos códigos ECL:

Roxie - ECL codes:

1. IDX_Persons.ecl
2. BWR_BuildIndexes.ecl (submit)
3. Fetch_Persons.ecl (compilar em Roxie)
4. BWR_TestQueries.ecl (submit)

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;
 - ✓ Desenvolvimento de um serviço de consulta;
 - Utilização de algoritmos de Aprendizado de Máquina.
- **Próximos passos**
 - Cursos *online*;
 - Projetos de pesquisa;
 - Oportunidades profissionais.
- **Considerações Finais**



Machine Learning

HPCC Systems: Visão Geral

- Utilização de algoritmos de *ML* – Serviço de consulta: Preço de Imóveis

The screenshot displays a web interface for a dynamic form titled "roxie". The form is identified as "fn_getprice_roxiequery_web.1" and is set to "Dynamic Form". The form name "FN_GETPRICE_ROXIEQUERY_WEB_1REQUEST" is checked. The form contains several input fields with the following values:

| | |
|---------------|---------|
| assess_val: | 1188720 |
| bedrooms: | 3 |
| full_baths: | 2 |
| half_baths: | 1 |
| land_sq_ft: | 14774 |
| living_sq_ft: | 1437 |
| year_acq: | 2011 |
| year_built: | 1968 |
| zip: | 95451 |

At the bottom of the form, there are checkboxes for "Capture Log Info." and "No Timeout", a "Trace Level:" input field, and a "FORM POST" dropdown menu. The "Call Query" dropdown is set to "Output Tables". There are "Submit" and "Clear All" buttons at the bottom right.

Aprendizado de Máquina

- **“Raw” dataset utilizado durante a terceira parte da demonstração:**
 - Importação dos Dados Brutos (Upload = *Landing Zone*)

Property.csv

Delimitado

1.662.959
registros

| ## | propertyid | house_number | house_number_suffix | predir | street | streettype | postdir | apt | city | state | zip | total_value | assessed_value |
|----|------------|--------------|---------------------|--------|-------------|------------|---------|--------|-------------------|-------|-------|-------------|----------------|
| 1 | 828195 | 144 | | | MCKIERNAN | DR | | | WALNUT CREEK | CA | 94597 | 62614 | 62614 |
| 2 | 1144455 | 281 | | | CENTER | ST | | | BALTIMORE | MD | 21136 | 105500 | 10550 |
| 3 | 1494347 | 483 | | | NEWTON | RD | | | FLAGSTAFF | AZ | 86011 | 2220 | 2220 |
| 4 | 1910847 | 802 | | | HATCHERY | CT | | | WOODLAND | WA | 98674 | 356000 | 356000 |
| 5 | 4267562 | 5007 | | E | ROY ROGERS | RD | | | TROY | MI | 48085 | 327253 | 327253 |
| 6 | 4888602 | 7607 | | | PEBBLESTONE | DR | | 000009 | KERNVILLE | CA | 93238 | 732179 | 732179 |
| 7 | 54135 | 4 | | | WAINWRIGHT | DR | | | NORTH FORT MYERS | FL | 33917 | 159724 | 87848 |
| 8 | 762012 | 125 | | | SHIPYARD | DR | | 000150 | MELBOURNE VILLAGE | FL | 32904 | 96300 | 96300 |
| 9 | 2331721 | 1190 | | | LITTLEOAK | DR | | | HOUSTON | TX | 77011 | 238854 | 217810 |
| 10 | 3276109 | 2506 | | | MEADOW | DR | | | LA QUINTA | CA | 92253 | 30977 | 30660 |

Aprendizado de Máquina

■ Visão geral – Conceituação

A característica principal do *Machine Learning* é a capacidade de inferir sobre relacionamentos e, visa prever uma resposta razoável quando apresentado com dados nunca antes vistos.

O "aprendizado" do *Machine Learning (ML)* tem várias categorias:

- **Supervisionado** - O tipo mais comum de *ML*. Esse método envolve o treinamento do sistema em que os *recordset*, juntamente com o padrão de saída de destino, são fornecidos ao sistema para executar uma tarefa.
- **Não Supervisionado** - Este método não envolve a saída de destino, o que significa que nenhum treinamento é fornecido ao sistema. O sistema precisa aprender por meio da determinação e adaptação de acordo com as características estruturais nos padrões de entrada.
- **Deep Learning** - Move-se para a área dos métodos *ML* de Redes Neurais (*Neural Networks - NNs*). O *Deep Learning* implica várias camadas maiores que '2' e, também, implica em técnicas utilizadas com dados complexos, como análise de vídeo ou áudio.

Aprendizado de Máquina

▪ **Aprendizado Supervisionado**

A premissa básica do *ML* Supervisionado é que, dado um conjunto de “amostras de dados” (registros) e um conjunto de “valores-alvo” em campo e formato de registro, que ele aprenda como prever “valores-alvo para novas amostras”.

Amostras de dados: conhecidas como variáveis “Independentes”, porque são as informações fornecidas e não dependem de nenhum outro dado. Variáveis Independentes também são conhecidas como ‘Características’ (*Features*) dos dados.

Valores-alvo: conhecidos como variáveis “Dependentes”, porque eles são de alguma forma dependentes das amostras de dados.

As variáveis ‘Independentes’ e ‘Dependentes’ juntas são conhecidas como “conjunto de treinamento”.

Dois tipos básicos de Modelos de Análise em Aprendizado Supervisionado:

Quantitativo - conhecido como “Regressão” - Implica um valor numérico;

Qualitativo - conhecido como “Classificação” - Implica uma categoria ou, às vezes, um resultado binário.

Aprendizado de Máquina

▪ **Machine Learning Bundles da plataforma HPCC Systems**

Os *bundles* de produção (excluindo os *bundles* de suporte **ML_Core** e **PBblas**) fornecem uma interface principal muito semelhante para o *Machine Learning*. No entanto, cada algoritmo tem suas próprias peculiaridades (suposições e restrições) que devem ser levadas em consideração.

É importante ler a documentação que acompanha cada bundle para usá-lo efetivamente.

➤ Link definitivo para todos os *bundles* de produção, sua documentação e tutoriais, quando aplicável:

[\[https://hpccsystems.com/download/free-modules/machine-learning-library\]](https://hpccsystems.com/download/free-modules/machine-learning-library).

Aprendizado de Máquina

▪ **Machine Learning Bundles**

a. *Bundles* Principais:

- ***ML_Core - Machine Learning Core***

Fornecer as principais definições de dados para *ML*. É um pré-requisito para todos os outros pacotes configuráveis de produção.

➤ Mais informações: [https://github.com/hpcc-systems/ML_Core].

- ***PBblas - Parallel Block Basic Linear Algebra Subsystem***

Fornecer operações de matriz escalonáveis e distribuídas usadas por vários dos outros pacotes configuráveis. Também pode ser usado diretamente sempre que as operações da matriz estiverem em ordem. Essa é uma dependência para vários dos outros pacotes configuráveis.

➤ Mais informações: [<https://github.com/hpcc-systems/PBblas>].

Aprendizado de Máquina

▪ **Machine Learning Bundles**

b. *Bundles de Aprendizado Supervisionado:*

- **LearningTrees** (baseado no algoritmo clássico “*ML RandomForest*”)

Classificação e Regressão baseadas em Árvores de Decisão. Um dos melhores métodos de *ML* "prontos para uso", pois faz poucas suposições sobre a natureza dos dados e é resistente ao “*overfitting*” (*). Capaz de lidar com um grande número de variáveis independentes. Cria uma “floresta” de diversas Árvores de Decisão e calcula a média das respostas das diferentes Árvores.

➤ Mais informações: [<https://github.com/hpcc-systems/LearningTrees>].

(*) **Superajuste** (*overfitting*): Muitos algoritmos tendem a superestimar o conjunto de treinamento. Isso significa que ele está reduzindo o erro de previsão ajustando-se ao ruído que ocorre nesse conjunto de dados. Um modelo de excesso de ajuste tratará esse ruído como sinal e usará o que aprendeu para prever os próximos dados apresentados. Infelizmente, esse próximo conjunto de dados estará sujeito a um conjunto de ruído completamente diferente, o que não fornecerá bons resultados.

Aprendizado de Máquina

▪ **ML Supervisionado: Árvores de Decisão (*Learning Trees - Random Forests*)**

As Árvores de Decisão têm sido utilizadas, pelo menos, desde a década de 1930 como uma forma de estruturar o conhecimento usando um conjunto de regras em cascata. Elas são conceitualmente simples e razoavelmente fáceis de entender e interpretar.

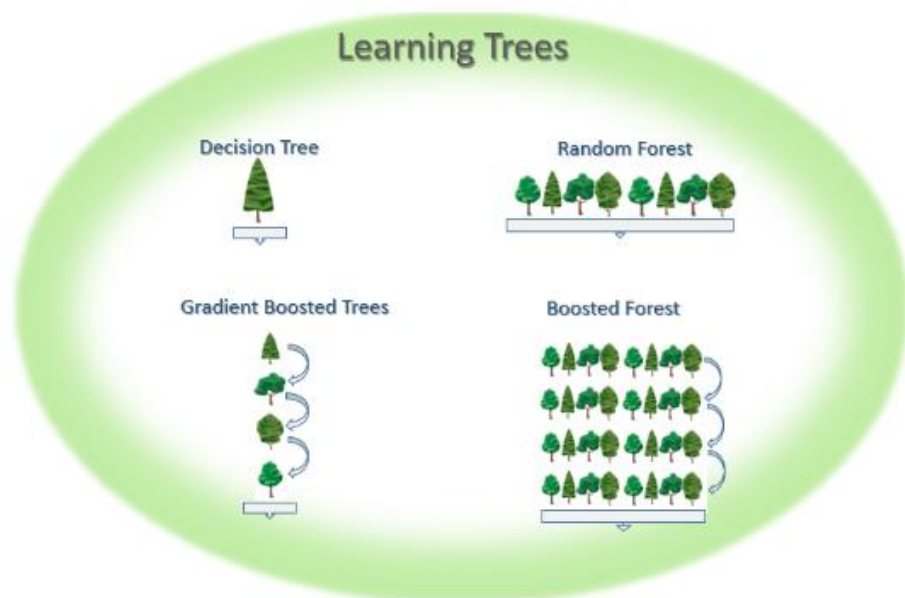
Random Forest são facilmente “paralelizáveis”, sendo, portanto, um algoritmo ideal para uso em *clusters* do **HPCC Systems**.

O *LearningTrees bundle* fornece uma implementação eficiente e escalável dos métodos *Learning Trees*. Atualmente, fornece algoritmos de "Decision Trees", "Random Forest", "Gradient Boosted Trees" e "Boosted Forest”.

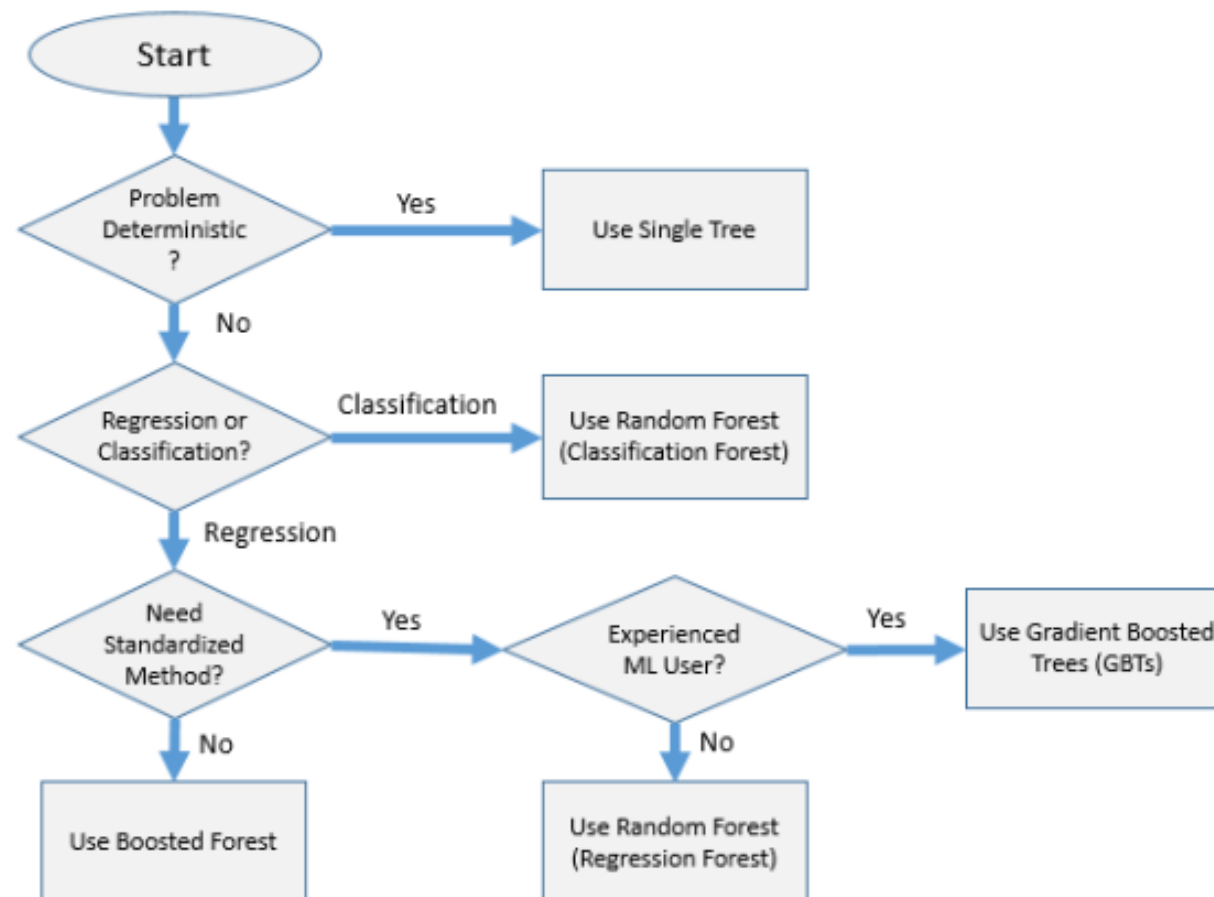
Diante dos diversos algoritmos disponíveis, qual algoritmo se deve escolher?

Aprendizado de Máquina

ML Supervisionado: Árvores de Decisão (*Learning Trees - Random Forests*)

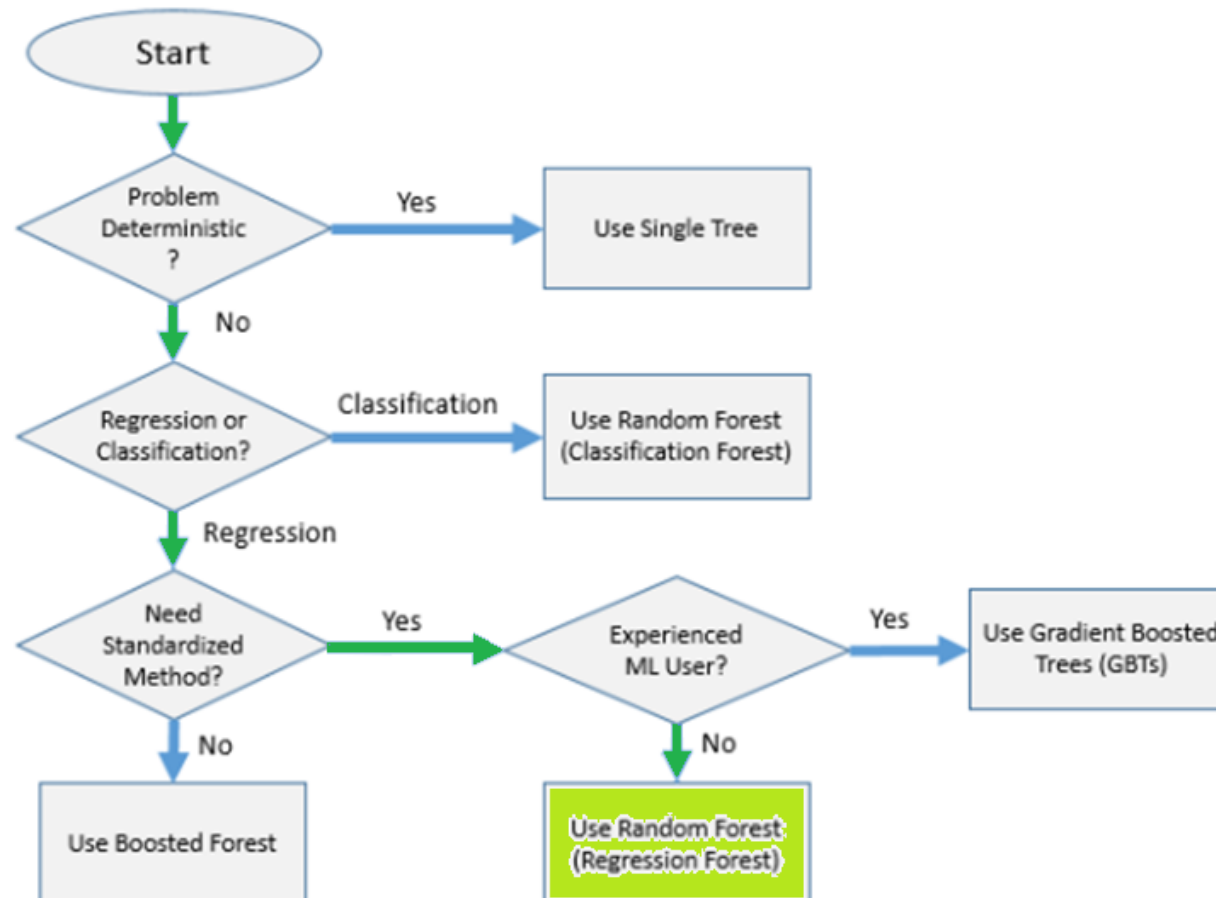


Esse fluxograma pode ser usado para auxiliar nessa escolha



Aprendizado de Máquina

■ ML Supervisionado: Árvores de Decisão (*Learning Trees - Random Forests*)



Aprendizado de Máquina

Fluxo de Aprendizado de Máquina Supervisionado

1. Definição do problema



2. Extração dos dados



3. Preparação dos dados



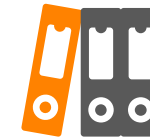
4. Segregação dos dados



5. Treinamento e avaliação do modelo



6. Implantação do modelo



1. Definição do problema

- “Dado um conjunto de atributos de uma propriedade (localização, m², ano de construção/aquisição, n^o cômodos, etc), como predizer o seu valor real de venda?”

| propertyid | house_number | house_number | predir | street | street | postdir | apt | city | state | zip | total_value | assessed_value | year_acquired | land_square_foot | living_square_feet | bedrooms | full_baths |
|------------|--------------|--------------|--------|-----------------|--------|---------|--------|-----------------|-------|-------|-------------|----------------|---------------|------------------|--------------------|----------|------------|
| 828195 | 144 | | | MCKIERNAN | DR | | | WALNUT CREEK | CA | 94597 | 62614 | 62614 | 2006 | 20418 | 2485 | 3 | 2 |
| 1144455 | 281 | | | CENTER | ST | | | BALTIMORE | MD | 21136 | 105500 | 10550 | 2007 | 4807 | 1368 | 0 | 0 |
| 1494347 | 483 | | | NEWTON | RD | | | FLAGSTAFF | AZ | 86011 | 2220 | 2220 | 0 | 5654 | 1011 | 3 | 1 |
| 1910847 | 802 | | | HATCHERY | CT | | | WOODLAND | WA | 98674 | 356000 | 356000 | 0 | 6094 | 0 | 2 | 1 |
| 4267562 | 5007 | | E | ROY ROGERS | RD | | | TROY | MI | 48085 | 327253 | 327253 | 2007 | 3484 | 0 | 3 | 0 |
| 4888602 | 7607 | | | PEBBLESTONE | DR | | 000009 | KERNVILLE | CA | 93238 | 732179 | 732179 | 2010 | 19597 | 6132 | 6 | 6 |
| 48725 | 4 | | | LONG | AVE | | | SUNRISE | FL | 33323 | 271000 | 271000 | 2008 | 6880 | 2392 | 4 | 2 |
| 83528 | 6 | | | TRILLUM | LN | | | WAYLAND | MA | 02193 | 79889 | 79889 | 2007 | 7657 | 1657 | 4 | 1 |
| 94604 | 7 | | | PARMENTER | AVE | | | PLYMOUTH | MN | 55441 | 23800 | 23800 | 2005 | 19994 | 1754 | 3 | 2 |
| 220326 | 17 | | | TIMBER | RD | | | LOS ANGELES | CA | 90063 | 89000 | 89000 | 2008 | 7840 | 954 | 3 | 1 |
| 994609 | 212 | | | FREYER | DR | NE | | PHILOMONT | VA | 20131 | 59800 | 59800 | 2009 | 11199 | 1241 | 3 | 0 |
| 1836173 | 724 | | | EASTER | ST | | | ALLEN TOWN | PA | 18102 | 191600 | 191600 | 0 | 9100 | 2534 | 4 | 2 |
| 2910797 | 1903 | | | SADDLE BROOK | DR | | | CLIO | CA | 96106 | 61610 | 61610 | 2007 | 0 | 0 | 0 | 0 |
| 3083959 | 2158 | | | RIVERSIDE | DR | | | UPPER MORELA... | PA | 19006 | 90300 | 0 | 0 | 0 | 1235 | 3 | 2 |
| 3952189 | 4040 | | | GRAND VIEW | BLVD | | 000054 | RIO LINDA | CA | 95673 | 0 | 0 | 0 | 2700720 | 0 | 0 | 0 |
| 4186238 | 4726 | | | LAS PALMAS | CT | | | WAE LDER | TX | 78959 | 18816 | 18816 | 2009 | 2159 | 1320 | 0 | 0 |
| 4597143 | 6213 | | | WILSON | RD | | | ZOLFO SPRINGS | FL | 33890 | 72600 | 0 | 0 | 8496 | 0 | 3 | 1 |
| 4624905 | 6321 | | | STONEWALL | LN | | | PATERSON | NJ | 07514 | 139880 | 139880 | 2008 | 10454 | 1391 | 4 | 2 |
| 92326 | 7 | | | KNOLL CREST | DR | | | NARANJA | FL | 33032 | 76214 | 76214 | 2008 | 4800 | 930 | 2 | 0 |
| 1792852 | 704 | | | ERIN | DR | | | TRABUCO | CA | 92678 | 28010 | 28010 | 2007 | 5200 | 0 | 3 | 1 |
| 1843977 | 728 | | S | ARLINGTON HE... | RD | | | BLOOMING GRO... | TX | 76626 | 130400 | 130400 | 2007 | 36154 | 1629 | 3 | 1 |
| 4214872 | 4821 | | | MYRTLE OAK | DR | | 000025 | SAN BERNARDT | CA | 92376 | 22250 | 0 | 2007 | 93654 | 0 | 0 | 0 |

2. Extração dos dados

▪ Importação e análise dos dados

| ## | personid | propertyid | house_number | house_number_suffix | predir | street | streettype | postdir | apt | city | state | zip | total_value |
|----|------------------|------------|--------------|---------------------|--------|--------------|------------|---------|--------|----------------|-------|-------|-------------|
| 1 | 187522928604396 | 828195 | 144 | | | MCKIERNAN | DR | | | WALNUT CREEK | CA | 94597 | 62614 |
| 2 | 187522928604396 | 1144455 | 281 | | | CENTER | ST | | | BALTIMORE | MD | 21136 | 105500 |
| 3 | 187522928604396 | 1494347 | 483 | | | NEWTON | RD | | | FLAGSTAFF | AZ | 86011 | 2220 |
| 4 | 187522928604396 | 1910847 | 802 | | | HATCHERY | CT | | | WOODLAND | WA | 98674 | 356000 |
| 5 | 187522928604396 | 4267562 | 5007 | | E | ROY ROGERS | RD | | | TROY | MI | 48085 | 327253 |
| 6 | 187522928604396 | 4888602 | 7607 | | | PEBBLESTONE | DR | | 000009 | KERNVILLE | CA | 93238 | 732179 |
| 7 | 1258313199446079 | 48725 | 4 | | | LONG | AVE | | | SUNRISE | FL | 33323 | 271000 |
| 8 | 1258313199446079 | 83528 | 6 | | | TRILLUM | LN | | | WAYLAND | MA | 02193 | 79889 |
| 9 | 1258313199446079 | 94604 | 7 | | | PARMENTER | AVE | | | PLYMOUTH | MN | 55441 | 23800 |
| 10 | 1258313199446079 | 220326 | 17 | | | TIMBER | RD | | | LOS ANGELES | CA | 90063 | 89000 |
| 11 | 1258313199446079 | 994609 | 212 | | | FREYER | DR | NE | | PHILOMONT | VA | 20131 | 59800 |
| 12 | 1258313199446079 | 1836173 | 724 | | | EASTER | ST | | | ALLENTOWN | PA | 18102 | 191600 |
| 13 | 1258313199446079 | 2910797 | 1903 | | | SADDLE BROOK | DR | | | CLIO | CA | 96106 | 61610 |
| 14 | 1258313199446079 | 3083959 | 2158 | | | RIVERSIDE | DR | | | UPPER MORELAND | PA | 19006 | 90300 |
| 15 | 1258313199446079 | 3952189 | 4040 | | | GRAND VIEW | BLVD | | 000054 | RIO LINDA | CA | 95673 | 0 |

3. Preparação dos dados

- **Limpeza, padronização e consolidação de registros**

| ## | propertyid | zip | assessed_value | year_acquired | land_square_footage | living_square_feet | bedrooms | full_baths | half_baths | year_built | total_value |
|----|------------|-------|----------------|---------------|---------------------|--------------------|----------|------------|------------|------------|-------------|
| 1 | 79784 | 33424 | 76440 | 2015 | 4299 | 1255 | 3 | 2 | 0 | 2010 | 76440 |
| 2 | 3924129 | 20601 | 95900 | 2013 | 11224 | 1468 | 3 | 2 | 1 | 2007 | 95900 |
| 3 | 413843 | 8803 | 76000 | 2015 | 57000 | 1858 | 3 | 2 | 0 | 1970 | 76000 |
| 4 | 608224 | 98370 | 39340 | 2012 | 7405 | 1066 | 3 | 1 | 1 | 1967 | 39340 |
| 5 | 942963 | 72032 | 278400 | 2008 | 9600 | 2459 | 3 | 2 | 0 | 1963 | 278400 |
| 6 | 2237271 | 79935 | 143600 | 2011 | 8430 | 1008 | 2 | 1 | 1 | 1961 | 143600 |
| 7 | 4443742 | 84065 | 166934 | 2013 | 9317 | 1700 | 4 | 2 | 0 | 1991 | 166934 |
| 8 | 3834707 | 66227 | 348350 | 2012 | 15300 | 2663 | 4 | 2 | 1 | 2002 | 348350 |
| 9 | 3592739 | 19606 | 54000 | 2015 | 15060 | 2292 | 4 | 2 | 1 | 1980 | 90000 |
| 10 | 2916349 | 34639 | 119050 | 2015 | 6947 | 1709 | 3 | 2 | 0 | 2009 | 140950 |

4. Segregação dos dados

- Seleção aleatória de amostras de treinamento e validação com a distinção de variáveis independentes e dependentes

| ## | wi | id | number | value |
|----|----|---------|--------|---------|
| 1 | 1 | 79784 | 1 | 33424.0 |
| 2 | 1 | 79784 | 2 | 76440.0 |
| 3 | 1 | 79784 | 3 | 2015.0 |
| 4 | 1 | 79784 | 4 | 4299.0 |
| 5 | 1 | 79784 | 5 | 1255.0 |
| 6 | 1 | 79784 | 6 | 3.0 |
| 7 | 1 | 79784 | 7 | 2.0 |
| 8 | 1 | 79784 | 8 | 0.0 |
| 9 | 1 | 79784 | 9 | 2010.0 |
| 10 | 1 | 3924129 | 1 | 20601.0 |

| ## | wi | id | number | value |
|----|----|---------|--------|----------|
| 1 | 1 | 79784 | 1 | 76440.0 |
| 2 | 1 | 3924129 | 1 | 95900.0 |
| 3 | 1 | 413843 | 1 | 76000.0 |
| 4 | 1 | 608224 | 1 | 39340.0 |
| 5 | 1 | 942963 | 1 | 278400.0 |
| 6 | 1 | 2237271 | 1 | 143600.0 |
| 7 | 1 | 4443742 | 1 | 166934.0 |
| 8 | 1 | 3834707 | 1 | 348350.0 |
| 9 | 1 | 3592739 | 1 | 90000.0 |
| 10 | 1 | 2916349 | 1 | 140950.0 |

5. Treinamento e avaliação do modelo

- **Obtenção de modelo a partir da amostra de treinamento e validação na amostra de teste**

| ## | wi | id | number | value |
|----|----|--------|--------|-------------------|
| 1 | 1 | 3634 | 1 | 59055.31318837311 |
| 2 | 1 | 5840 | 1 | 126151.3283316611 |
| 3 | 1 | 12721 | 1 | 150876.4676173128 |
| 4 | 1 | 47045 | 1 | 233897.4086392291 |
| 5 | 1 | 91757 | 1 | 111950.2604939628 |
| 6 | 1 | 117238 | 1 | 81157.13156934927 |
| 7 | 1 | 149746 | 1 | 75868.58107175257 |
| 8 | 1 | 239046 | 1 | 39961.17077444747 |
| 9 | 1 | 246517 | 1 | 128203.9088547347 |
| 10 | 1 | 252615 | 1 | 69009.47259550788 |

| ## | wi | regressor | r2 | mse | rmse |
|----|----|-----------|--------------------|-------------------|------------------|
| 1 | 1 | 1 | 0.7304899830671003 | 7982069594.129144 | 89342.4288573416 |

6. Implantação do modelo

- Carregamento de dados e disponibilização de serviço de consulta

roxie

fn_getprice_roxiequery_web.1 Dynamic Form ▾

FN_GETPRICE_ROXIEQUERY_WEB_1REQUEST

| | |
|---------------|--------------------------------------|
| assess_val: | <input type="text" value="1188720"/> |
| bedrooms: | <input type="text" value="3"/> |
| full_baths: | <input type="text" value="2"/> |
| half_baths: | <input type="text" value="1"/> |
| land_sq_ft: | <input type="text" value="14774"/> |
| living_sq_ft: | <input type="text" value="1437"/> |
| year_acq: | <input type="text" value="2011"/> |
| year_built: | <input type="text" value="1968"/> |
| zip: | <input type="text" value="95451"/> |

Capture Log Info. Trace Level: No Timeout

Call Query ▾ Output Tables ▾ FORM POST ▾ Submit Clear All

RoadMap 04 - Sequência dos códigos ECL:

☐ ML - ECL codes:

1. modProperty.ecl
2. BWR_BrowseData.ecl (submit)
3. isCleanFilter.ecl
4. CleanProperty.ecl
5. modPrepData.ecl
6. BWR_ViewData1.ecl (submit)
7. modSegConvData.ecl
8. BWR_ViewData2.ecl (submit)
9. BWR_TrainReg.ecl (submit)
10. fn_GetPriceReg.ecl (compilar em Roxie)
11. BWR_TestQueriesReg.ecl (submit)

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;
 - ✓ Desenvolvimento de um serviço de consulta;
 - ✓ Utilização de algoritmos de Aprendizado de Máquina.
- **Próximos passos**
 - Cursos *online*;
 - Projetos de pesquisa;
 - Oportunidades profissionais.
- **Considerações Finais**

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

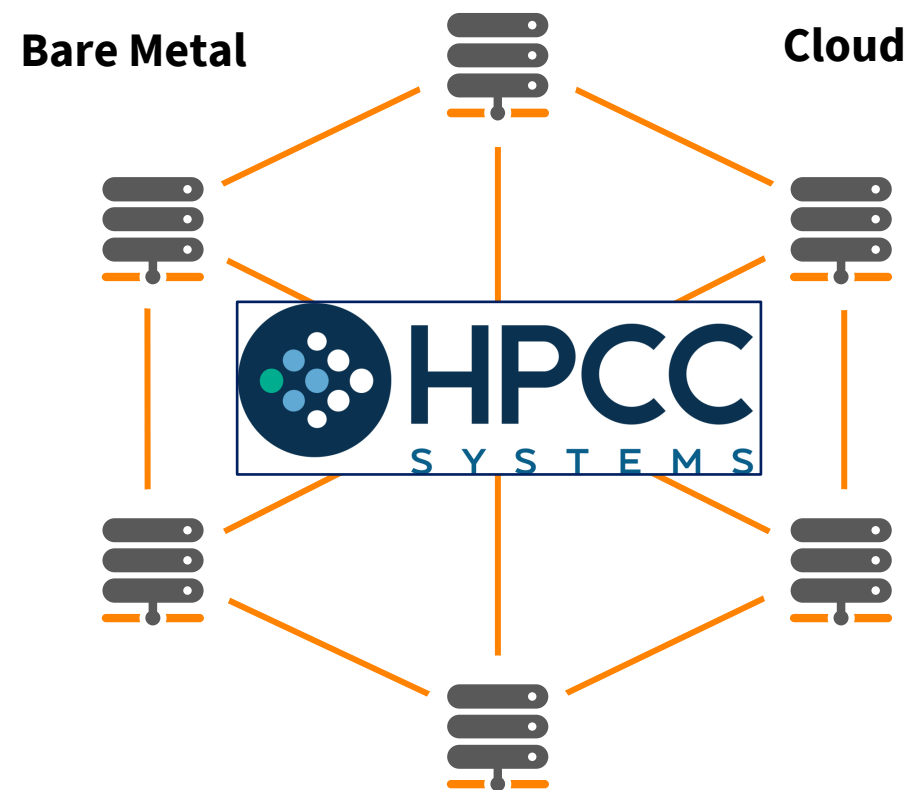
- Opções de utilização: *Playground* (<http://play.hpccsystems.com:18010/>)

... um único desktop ou laptop.



Máquina Virtual

... um cluster com nós trabalhando em conjunto como uma única entidade.



Próximos passos

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



Universidade de São Paulo
Brasil



UNIVERSIDADE FEDERAL
DE SANTA CATARINA



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

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;
 - ✓ Desenvolvimento de um serviço de consulta;
 - ✓ Utilização de algoritmos de Aprendizado de Máquina.
- **Próximos passos**
 - ✓ Cursos *online*;
 - Projetos de pesquisa;
 - Oportunidades profissionais.
- **Considerações Finais**

Próximos passos

- **Projetos de pesquisa**



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

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;
 - ✓ Desenvolvimento de um serviço de consulta;
 - ✓ Utilização de algoritmos de Aprendizado de Máquina.
- **Próximos passos**
 - ✓ Cursos *online*;
 - ✓ Projetos de pesquisa;
 - Oportunidades profissionais.
- **Considerações Finais**

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
📞 +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>

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;
 - ✓ Desenvolvimento de um serviço de consulta;
 - ✓ Utilização de algoritmos de Aprendizado de Máquina.
- ✓ **Próximos passos**
 - ✓ Cursos *online*;
 - ✓ Projetos de pesquisa;
 - ✓ Oportunidades profissionais.
- **Considerações Finais**

CONSIDERAÇÕES FINAIS



PERGUNTAS & RESPOSTAS

Material complementar

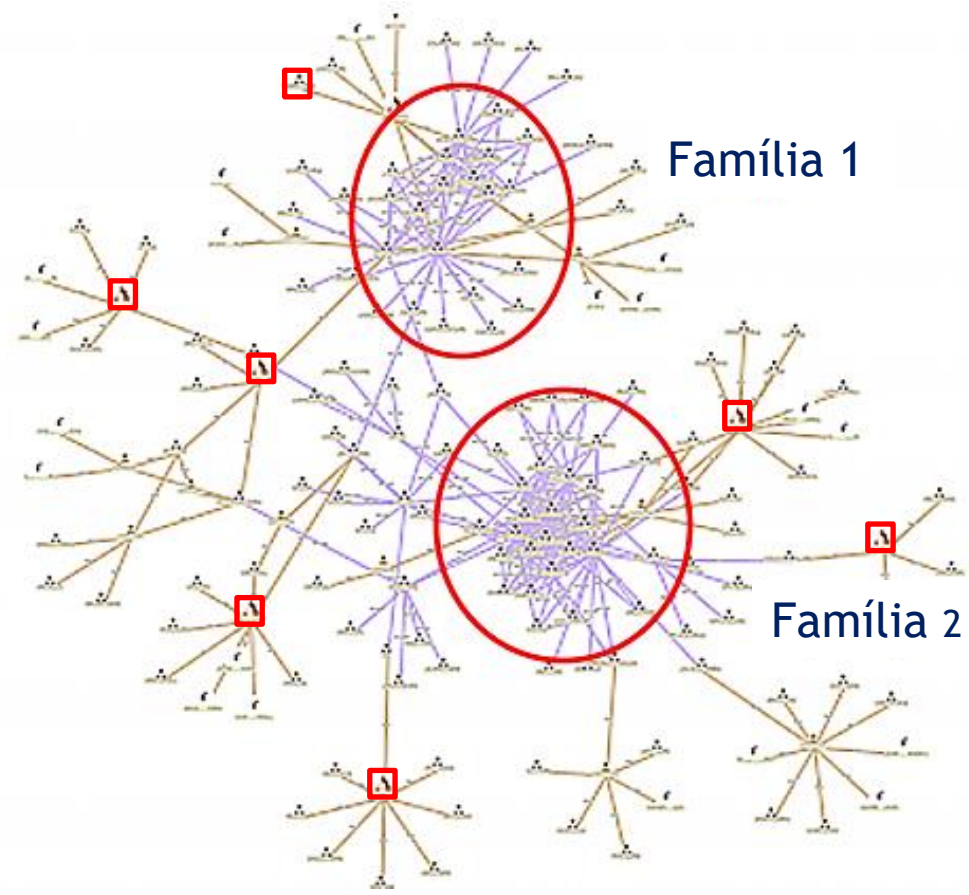
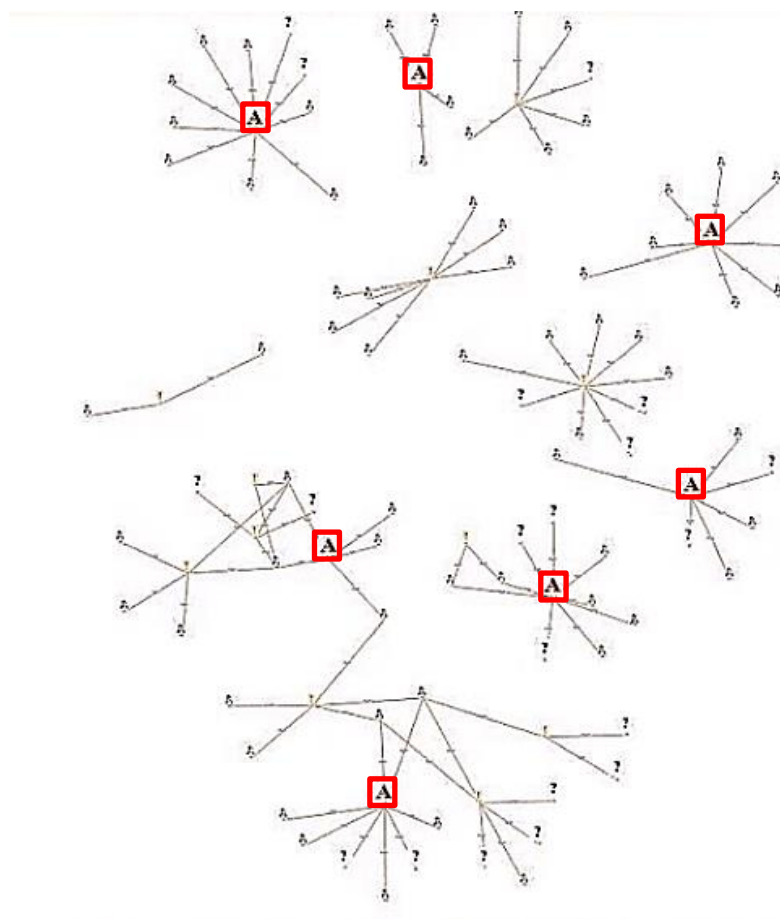
Apresentação de conceitos

BIG DATA

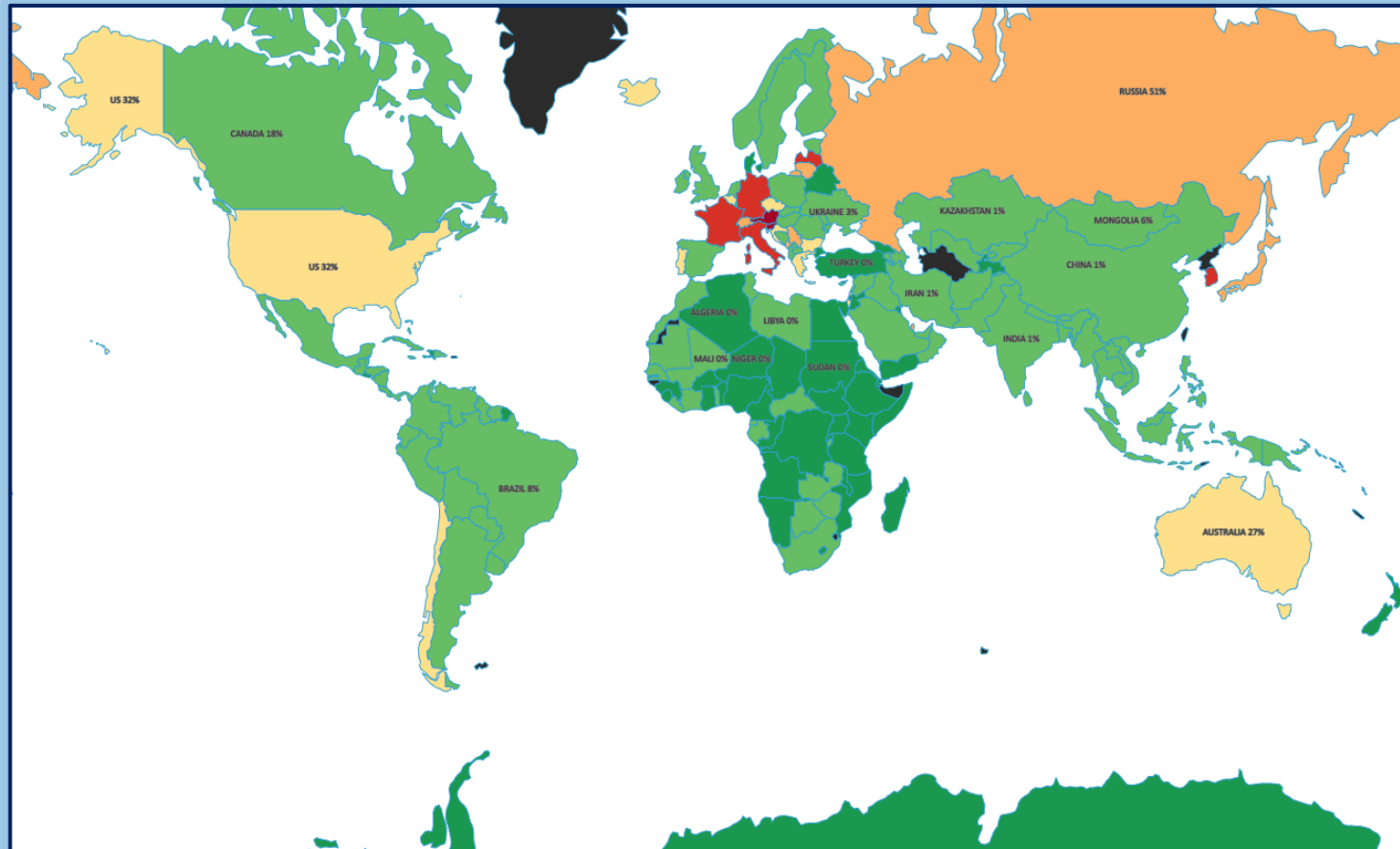
- Os cinco V's:
 - Volume
 - Variedade
 - Velocidade
 - Veracidade
 - Valor



Exemplo de uso da plataforma: “Detecção de Fraude”

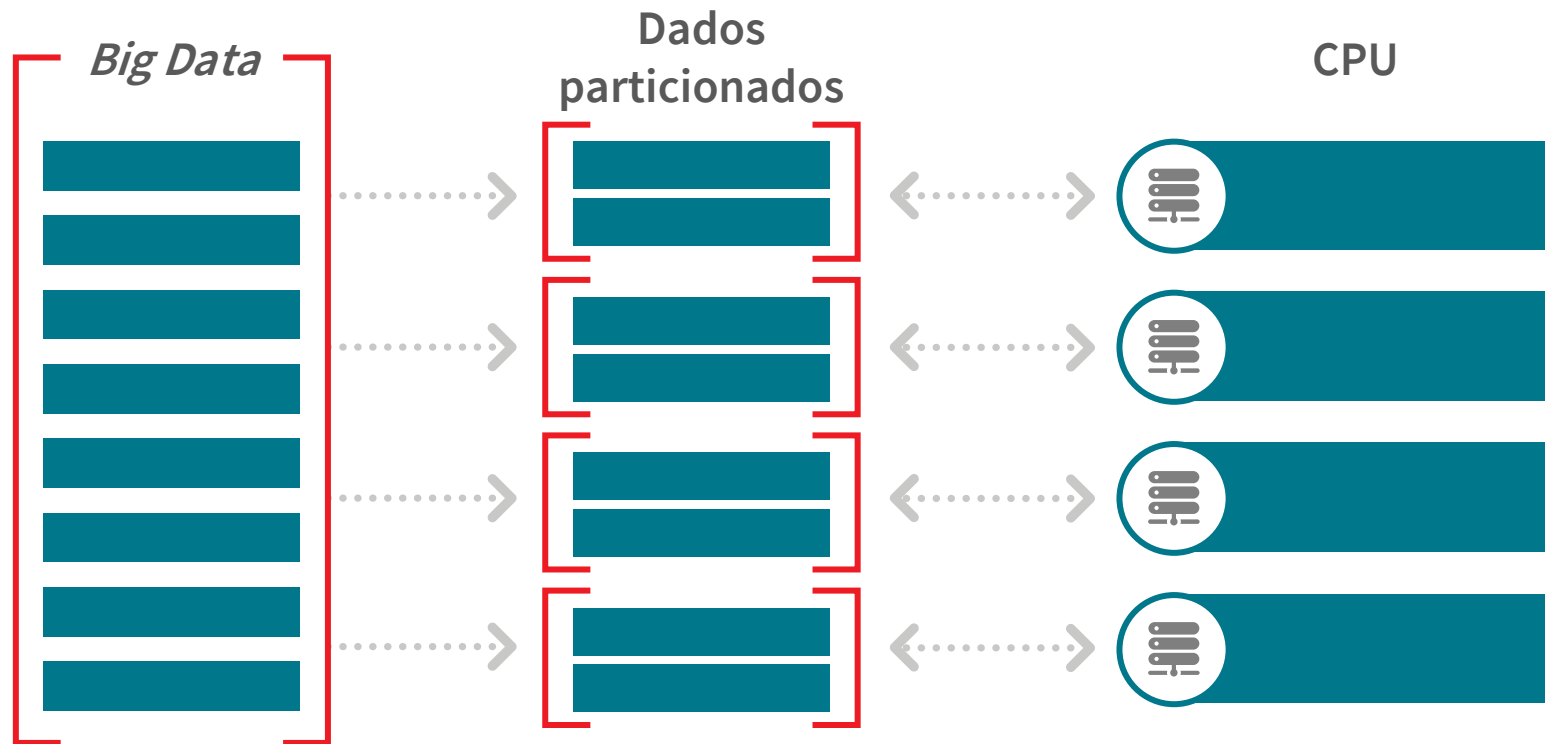


Exemplo de uso da plataforma: Uma situação mais atual...



Saiba mais em: <https://hpccsystems.com/resources/covid-19-is-there-a-better-way-to-tell-the-story/>

Dados Distribuídos e Processamento Paralelo



Arquitetura da plataforma *HPCC Systems*

Componentes da plataforma *HPCC Systems*:

- *Clusters*:

- Thor
- Roxie
- hThor (Agente ECL)

- *Middleware* – Servidores de sistema:

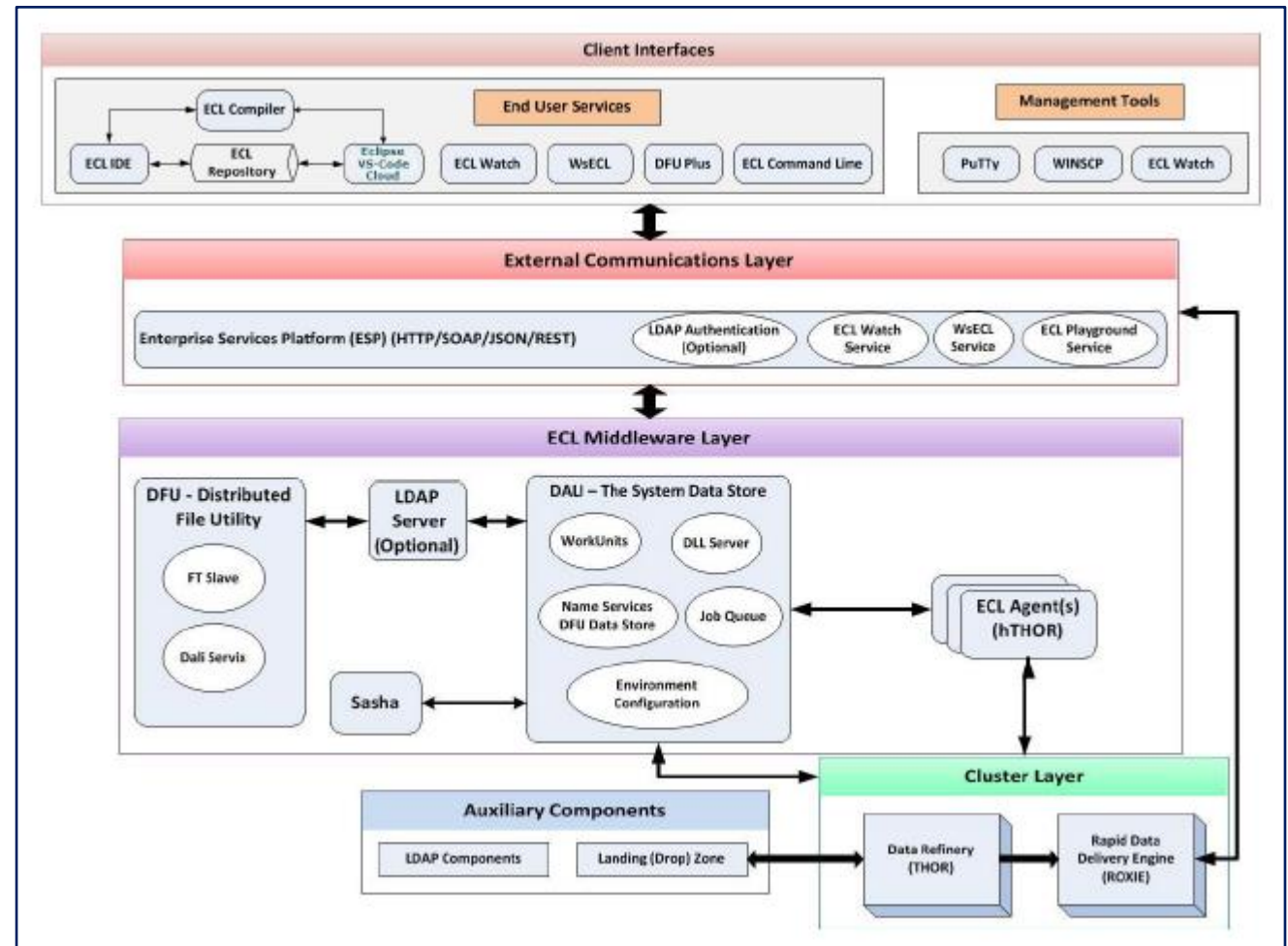
- DFU
- ESP
- ECLCC
- Dali
- Sasha

- Outros servidores:

- Agente ECL (hThor)
- LDAP

- Interfaces de usuário (*client*):

- ECL IDE
- ECL Watch
- CLI - Command line tools



“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*

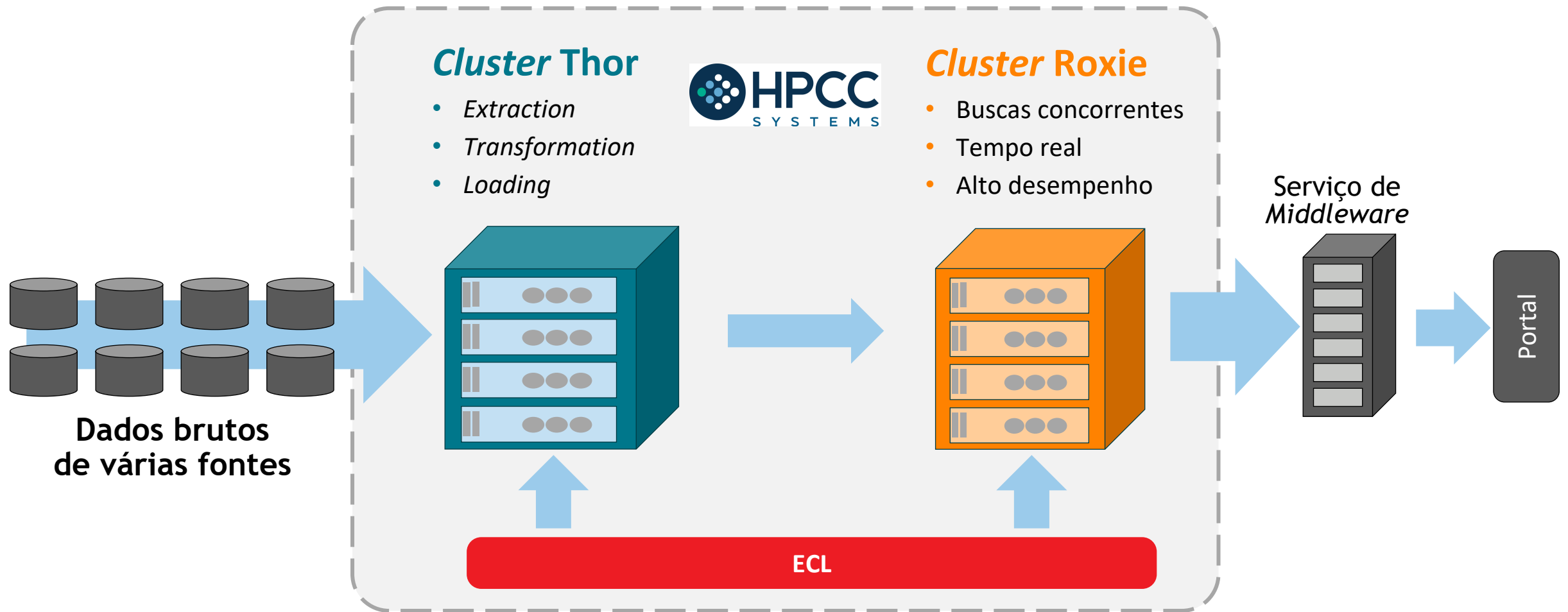
Supervisionado, não-supervisionado, aprendizagem profunda



Conectividade

Plugins de integração com outros sistemas

O Power Trio da plataforma: Thor, Roxie e ECL



B) Enterprise Control Language (ECL)

Sintaxe completa de uma Definição ECL

Nome := Expressão ;

[Escopo] [TipoValor] Nome [(parâmetros)] := Expressão [:ServiçoWorkflow] ;

EXPORT

BOOLEAN

IsLeapYear

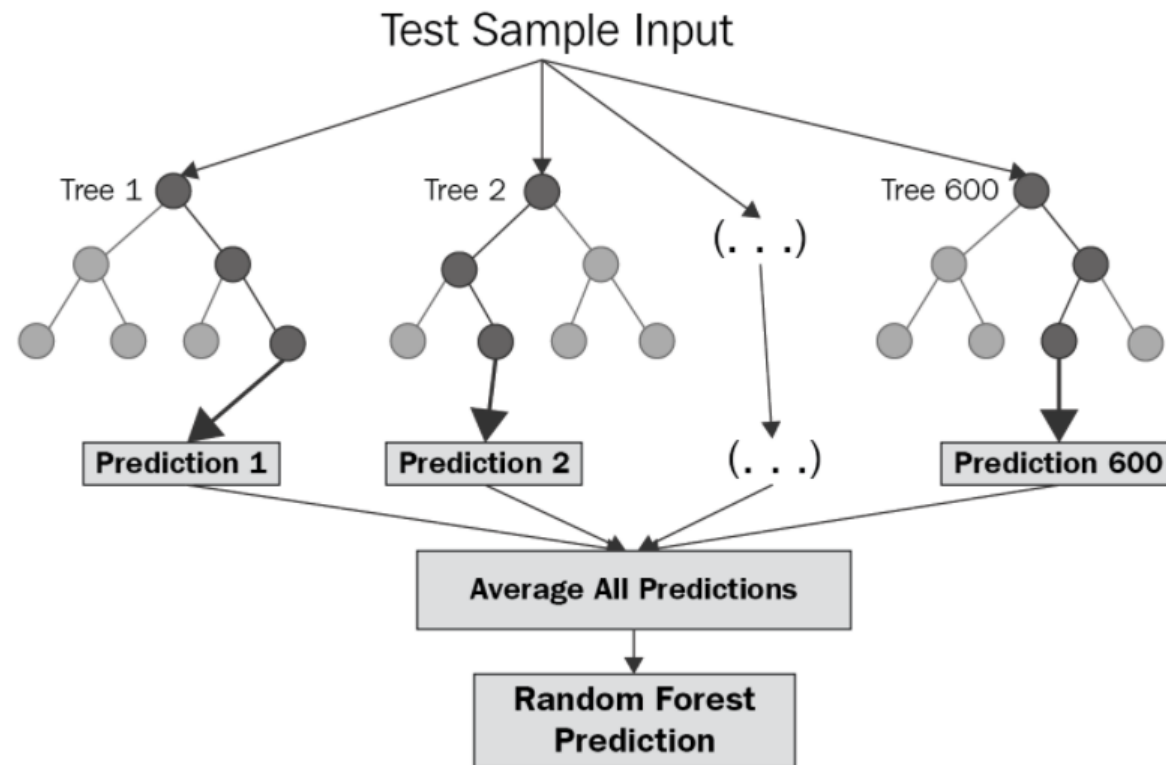
(INTEGER2 year)

year % 4 = 0 AND (year % 100 != 0 OR year % 400 = 0)

: STORED('LeapYearValue');

1. Definição do problema (cont.)

- **Árvores de Decisão – *Learning Trees***



Próximos passos

■ Treinamento e Suporte



Fácil de
Atualizar



Fácil de
Aprender

HPCC Systems: It's easier.



Fácil de
Programar



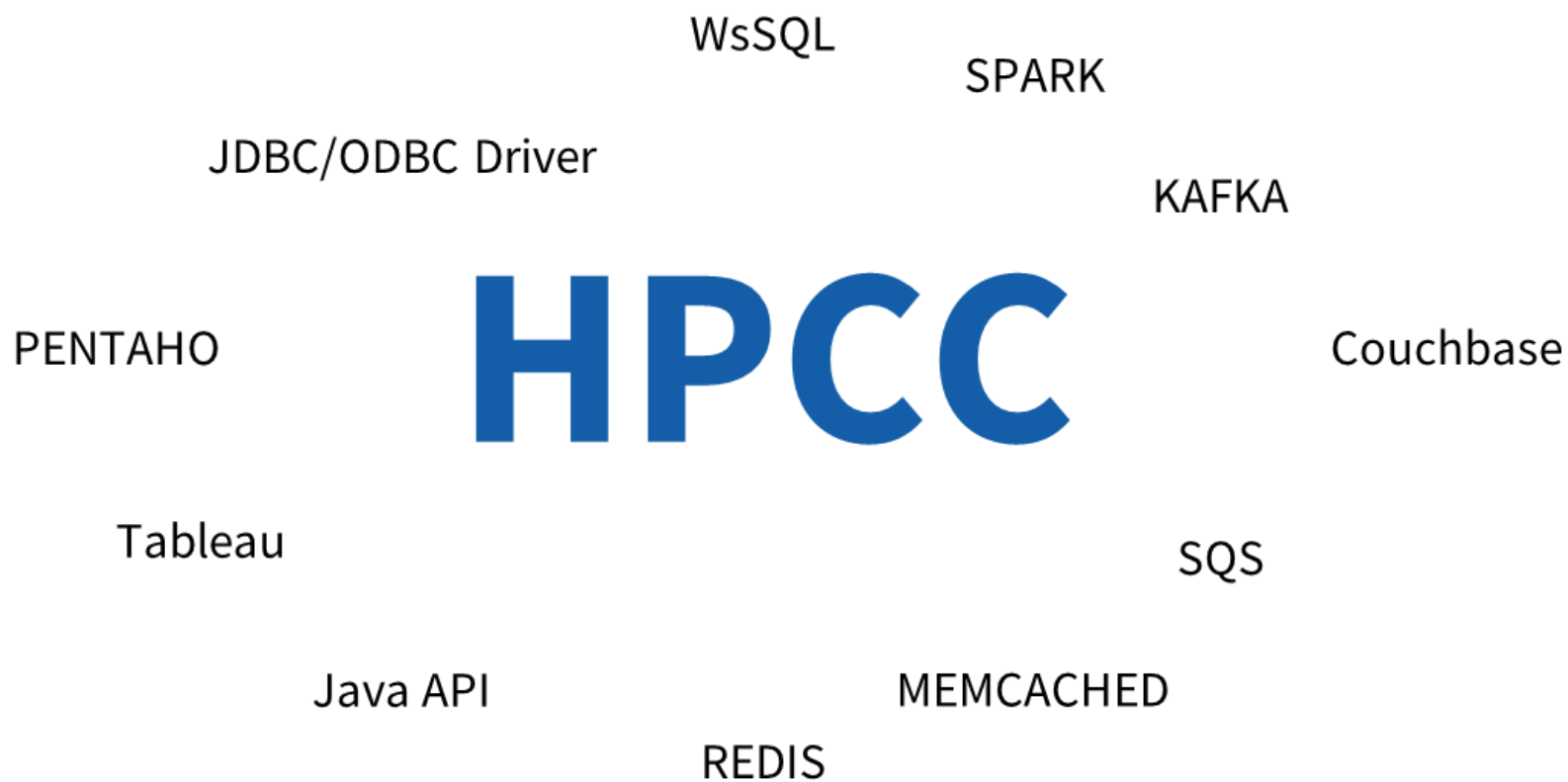
Fácil para
Integração de Dados



Fácil para
Gerenciamento de *Clusters*

Conectividade

- **Plugins**



Linguagens suportadas

■ Exemplos

- C++
- Java
- R
- CQL
- Python
- SQL

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 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 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);
```