



Fundação Vanzolini

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

Aula 1

Agenda da aula 1

- ✓ Apresentação do curso
- ✓ Revisão de conceitos: HPCC Systems e ECL
- ✓ Configuração do ambiente

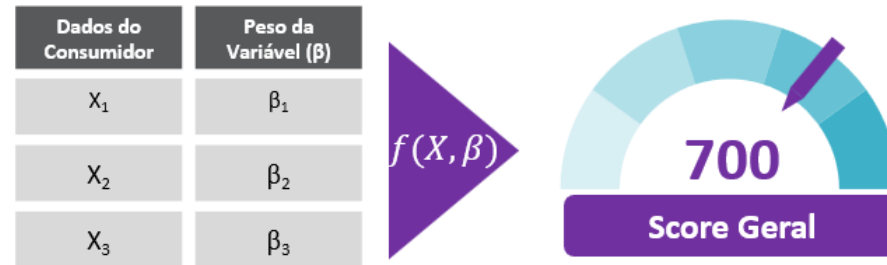
Apresentação do curso

Treinamento em ECL/HPCC: learn.lexisnexus.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
 - Tutoriais para uso de plugins
- Administração de Sistemas
 - Conceitos e operação básica
- HPCC para gestores
 - Visão geral e aplicações da plataforma

Objetivo final do curso!

- Identificar um modelo de predição baseado em atributos de operações financeiras que permitam aferir o risco de um pedido de crédito



- Dados (~3Mi):

- <https://www.kaggle.com/ethon0426/lending-club-20072020q1>



	field1	id	loan amnt	funded amnt	funded amnt inv	term	int rate	installment	grade	sub grade	emp title	emp length	home ownership	annual inc	verification status	issue d	loan status	pymnt plan	url	purpose	title	zip code	addr state
1	0	1077501	5000	5000	4975	36 months	10.65%	162.87	B	B2		10+ years	RENT	24000	Verified	Dec-2011	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?loan_id=1077501	credit_card	Computer	860xx	AZ
2	1	1077430	2500	2500	2500	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	RENT	30000	Source Verified	Dec-2011	Charged Off	n	https://lendingclub.com/browse/loanDetail.action?loan_id=1077430	car	bike	309xx	GA
3	2	1077175	2400	2400	2400	36 months	15.96%	84.33	C	C5		10+ years	RENT	12252	Not Verified	Dec-2011	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?loan_id=1077175	small_business	real estate business	606xx	IL
4	3	1076863	10000	10000	10000	36 months	13.49%	339.31	C	C1	AIR RESOURCES BOARD	10+ years	RENT	49200	Source Verified	Dec-2011	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?loan_id=1076863	other	personel	917xx	CA
5	4	1075358	3000	3000	3000	60 months	12.69%	67.79	B	B5	University Medical Group	1 year	RENT	80000	Source Verified	Dec-2011	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?loan_id=1075358	other	Personal	972xx	OR
6	5	1075269	5000	5000	5000	36 months	7.90%	156.46	A	A4	Veolia Transportaton	3 years	RENT	36000	Source Verified	Dec-2011	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?loan_id=1075269	wedding	My wedding loan I promise to pay back	852xx	AZ

Recursos e operação

- Aulas: das 19:00hs às 22:00hs
- Dias:
 - 5, 6 e 8 de Dezembro (semana 1)
 - 12, 13 e 15 de Dezembro (semana 2)
- Computador pessoal (ECL IDE v8.2.18/VSCode/GitHub)
- Cluster: <http://52.8.201.222:8010/>
- Moodle: <https://ead.vanzolini.org.br/course/view.php?id=949>
- Repositório: <https://github.com/HWatanuki/TrainingDominandoBigDataInter>
- Certificado USP e badges HPCC Systems

Coordenação

- Prof. Hugo Watanuki (hwatanuki@usp.br)
 - Doutor em engenharia de produção POLI-USP
 - Engenheiro de software na LexisNexis Risk Solutions
- Prof. Renato Moraes (remo@usp.br)
 - Doutor em administração pela FEA-USP
 - Professor do depto. de engenharia de produção POLI-USP

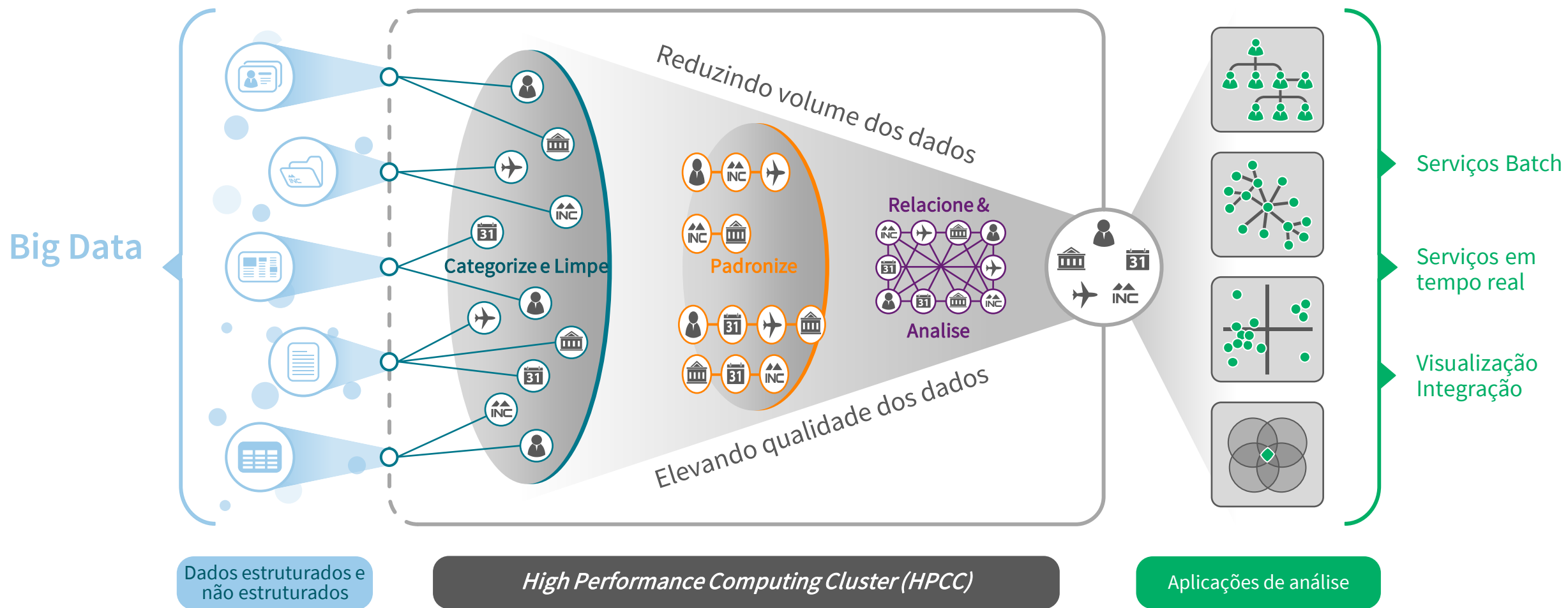


Introdução

- ✓ Nome
- ✓ Área de atuação
- ✓ Experiência e interesse em Big Data

Revisão de conceitos: HPCC Systems e ECL

Extract, Transform, Load

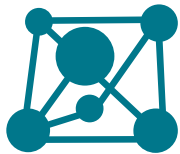


“Stack” tecnológico



Ferramentas de consulta e visualização

Entrega online de consultas em *Big Data*



Bibliotecas de *Machine Learning*

Supervisionado, não-supervisionado, aprendizagem profunda



Ferramentas para manipulação de dados

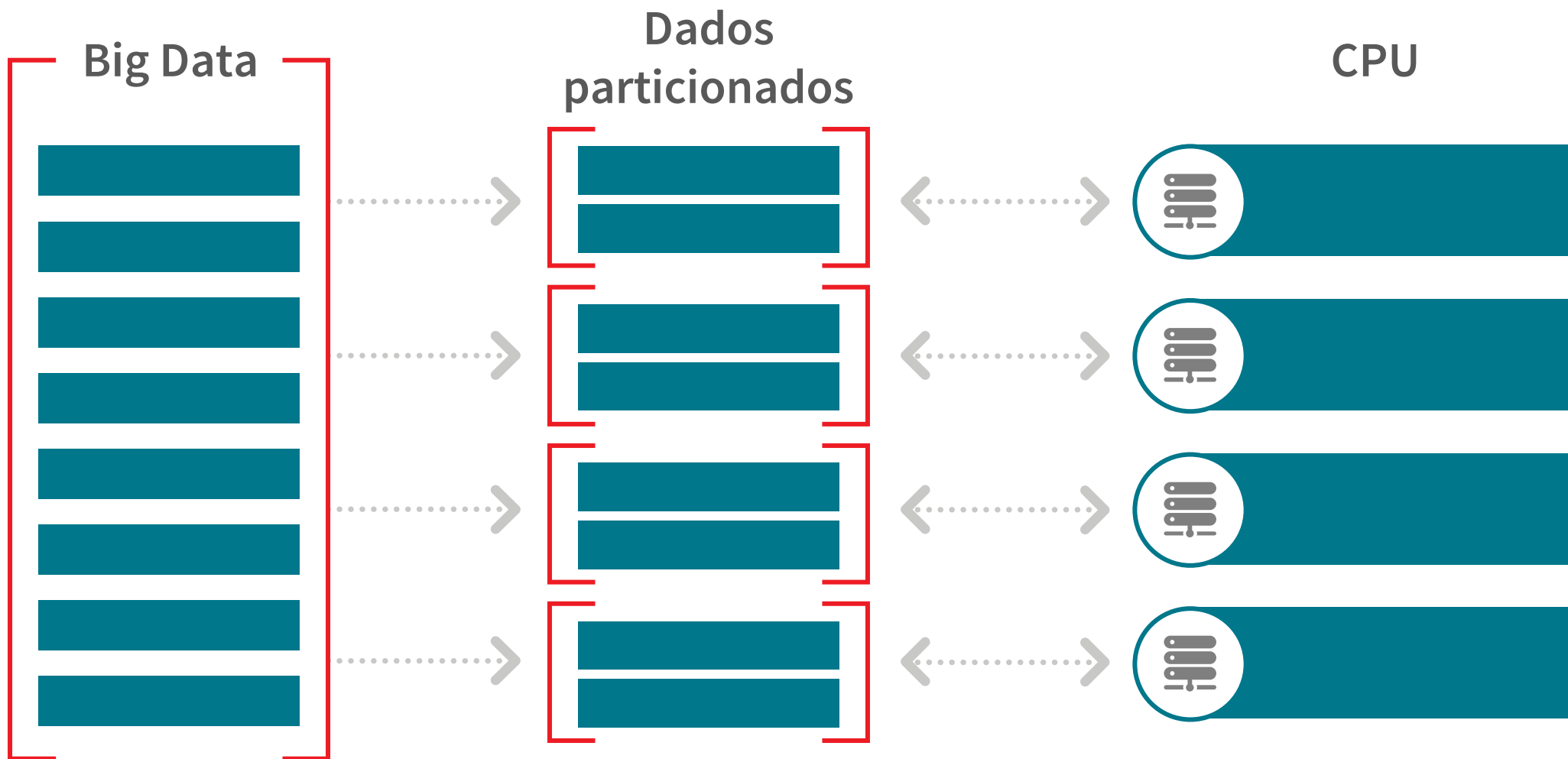
Perfilamento, limpeza, consolidação de dados



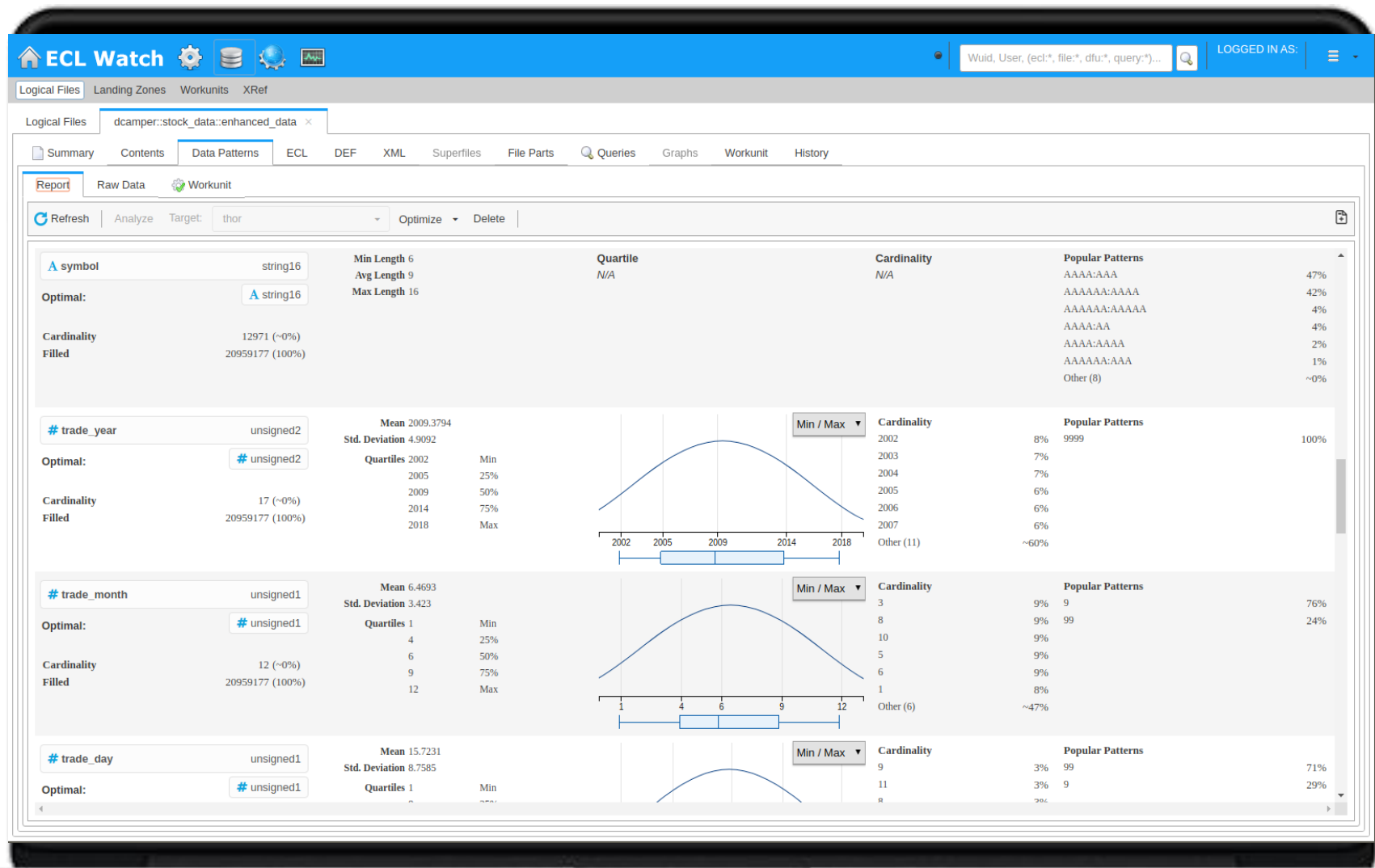
ETL

Extração, transformação e carregamento de dados

ETL: Supercomputação



Ferramentas de *Profiling*



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



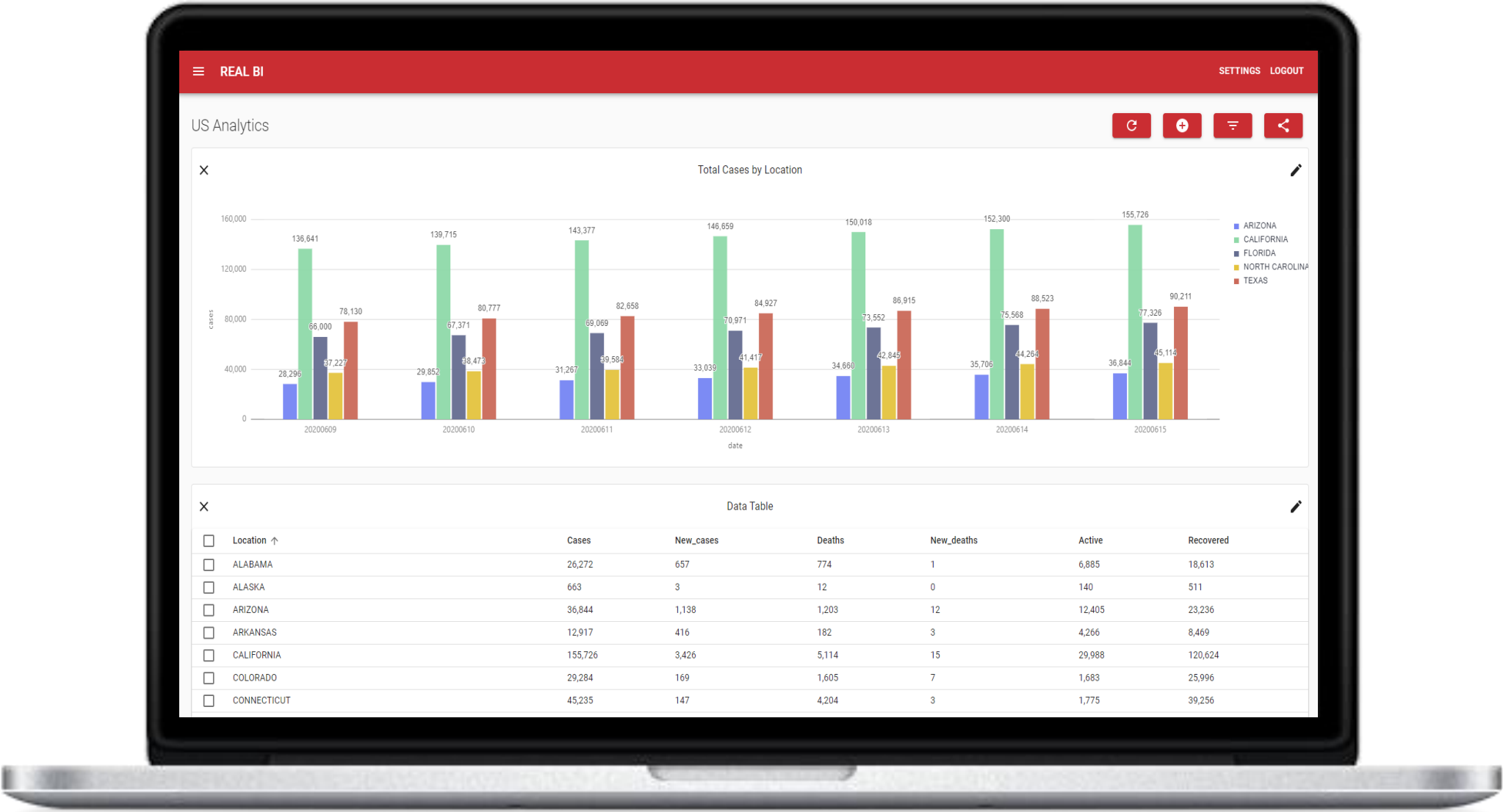
Métodos ensemble

Random Forest

Gradient Boosted
Forest

Gradient Boosted
Trees

Ferramentas de Consulta e Visualização



Conceitos básicos de ECL

- Paradigma declarativo (não-procedural)
- Estrutura básica: **Nome := Expressão ;**
- ECL não é sensível a caixa alta/baixa
- Espaço em branco é ignorado
- 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 diretório

Ações vs. Definições

✓ O código ECL é constituído de:

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

A := 'People' ; // não inicia uma WU

✓ Ações resultam em compilação e execução (arquivos BWR)

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

Tipos de dados primitivos

BOOLEAN

```
BOOLEAN IsFloridian := TRUE;
```

STRING[n]

```
STRING1 Gender := 'M';
```

INTEGER[n], UNSIGNED[n],

```
INTEGER1 ictr := -100;          // -128 to 127
```

```
UNSIGNED1 ctr := 0;            // 0 - 255
```

REAL[n], DECIMALn[_y]

```
REAL4 PI := 3.14159;
```

```
DECIMAL7_2 Salary := 75000.00;
```

Tipos básicos de definição ECL

Booleana (*boolean*)

```
IsSeniorCitizen := People.birthdate>19600101;
```

Valor único (*value*)

```
MaleValue := 'M';
```

Conjunto de valores (*set*)

```
GenderValues := ['M', 'F'];
```

Conjunto de registros (*recordset*)

```
SeniorPeople := People(IsSeniorCitizen);
```

```
MalePeople := People(Gender = MaleValue);
```

```
FemaleMalePeople := People(Gender IN GenderValues);
```

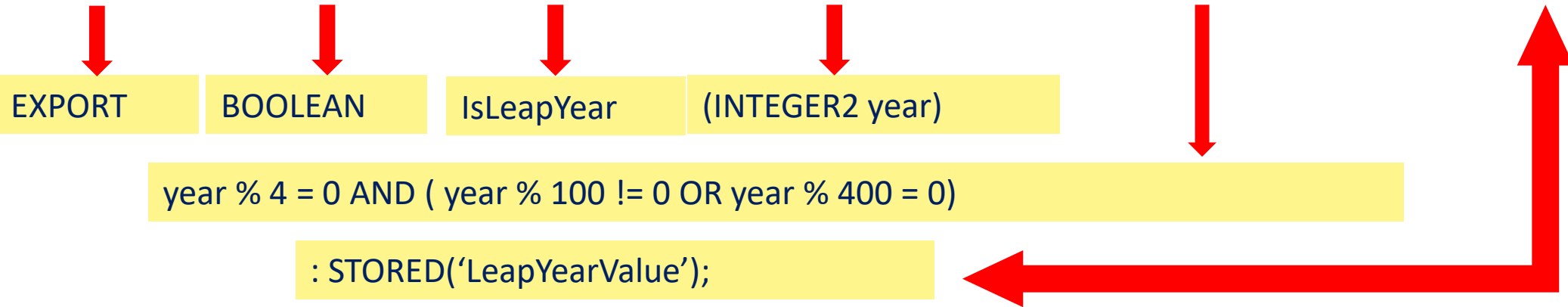
People

##	firstname	lastname	middlename	namesuffix	filedate	bureaucode	maritalstatus	gender	dependentcount	birthdate	streetaddress
1	Cherianne	Khatchatourian	N		19990922	24		M	0		69 BOULDER RIDGE RD # 25
2	Muyesser	Raplee	X		20001111	353		F	0		55 SWAMP RD
3	Roselin	Viceconte			19990325	344		F	0	19800113	107 HILL TER
4	Inda	Provines			20000909	13		U	0		290 W MOUNT PLEASANT AVE
5	Inderdeep	Laurence	D		20001228	344		M	0		44 PROSPECT PL
6	Chrystine	Mangiapane			19990827	315		F	0	19780306	1806 1ST AVE APT 8F
7	Adelene	Stock	R		20000827	252		M	0		1117 FARM RD
8	Mendy	Rufenblanchette			20000903	24		M	0		3 W 83RD ST APT 4C
9	Lannie	Amerantes	I		20001219	313		U	0		200 W 20TH ST APT 909
10	Tare	Gonyeau	T		19930807	48		F	0	19750801	6 CANDLE CT

Sintaxe Completa de uma Definição ECL

Nome := Expressão ;

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



Escopo da definição (Visibilidade)

Global –

A palavra-chave **EXPORT** torna a definição disponível “globalmente” no repositório

EXPORT PeopleCount := COUNT(People);

Módulo –

A palavra-chave **SHARED** torna a definição disponível somente no modulo/diretório que a contém

SHARED StateCount := 50;

Local –

A **ausência dessas palavras-chave** torna a definição disponível somente no arquivo que a contém e até a próxima definição ECL que contenha EXPORT ou SHARED

Num5 := 5;

EXPORT NumTotal := Num5 + 10 + StateCount;

Escopo da definição (Visibilidade)

IMPORT listadiretorios

- listadiretorios – Uma lista de diretórios separados por vírgula.

A palavra-chave **IMPORT** define uma lista de diretórios cujos arquivos de definições exportados tornam-se disponíveis para uso no código.

```
IMPORT Companies;           // Definições Exportadas em Companies estão disponíveis  
FloridaCompanies := Companies.File_Company(state='FL');
```

```
IMPORT $;                   // Definições Exportadas no Módulo atual estão disponíveis  
FloridaCompanies := $.File_Company(state='FL');
```

Estrutura **MODULE**

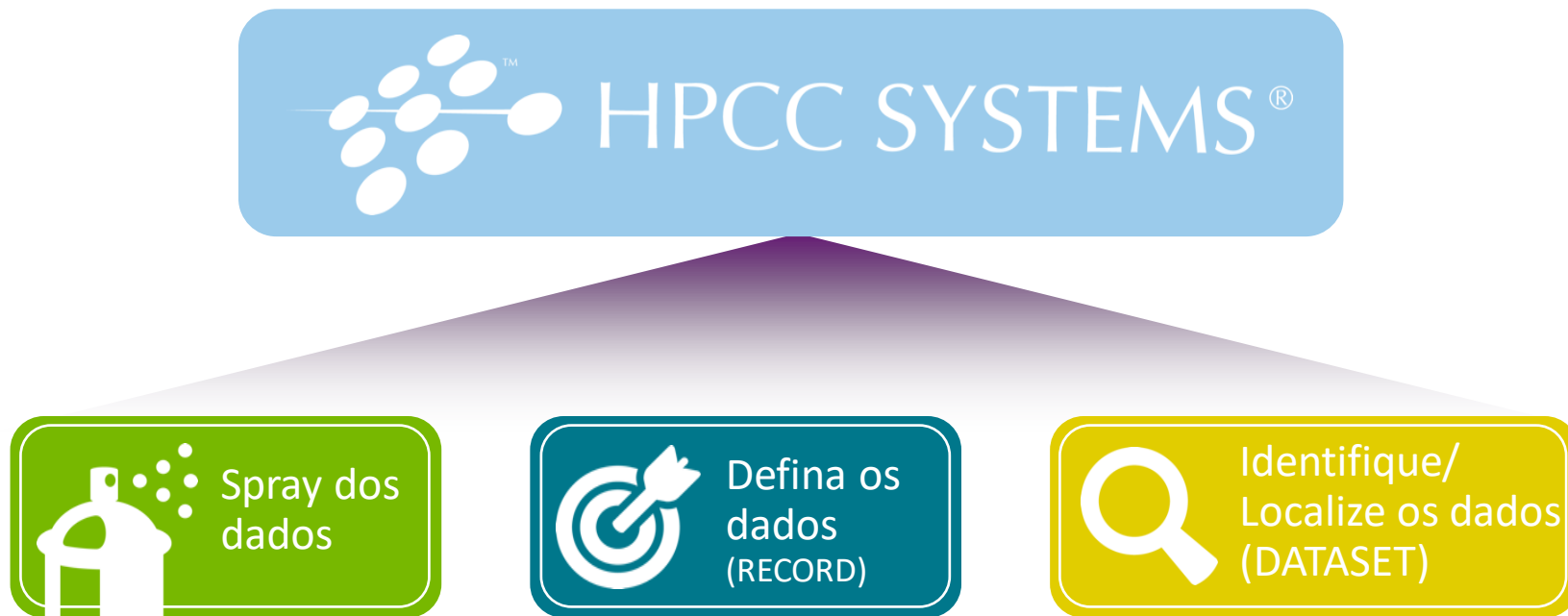
A estrutura **MODULE** permite agrupar e fornecer parâmetros para um conjunto de definições ECL relacionadas.

```
nome [ ( parametros ) ] := MODULE  
    definições;  
END;
```

- *Nome* – O nome da definição ECL do modulo.
- *parametros* – Os parâmetros disponíveis para todas as *definições*.
- *definições* – As definições ECL que compõem o módulo.

Começando a trabalhar com dados

Antes de começar a trabalhar com qualquer arquivo de dados na plataforma HPCC Systems, três passos devem ser executados:



Exemplo de estrutura de dados

```
EXPORT File_Persons := MODULE
```

```
  EXPORT Layout := RECORD
```

```
    UNSIGNED8 ID;
```

```
    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::hmw::Intro::Persons', Layout, FLAT);
```

```
END;
```

##	id	firstname	lastname	middl...	n...	filedate	bureaucode	marit...	gender	dep...	birthdate	streetaddress	city	state	zipcode
1	9108218085885411565	Cherianne	Khatchatourian	N		19990922	24		M	0		69 BOULDER RIDGE RD # 25A	HAWKINS	WI	54530
2	16505326057200398078	Muyesser	Raplee	X		20001111	353		F	0		55 SWAMP RD	DISTRICT HEIGHT	MD	20747
3	2454818069645923666	Roselin	Viceconte			19990325	344		F	0	19800113	107 HILL TER	ENTERPRISE	OR	97828
4	15880908289586509107	Inda	Provines			20000909	13		U	0		290 W MOUNT PLEASANT AVE	LAVACA	AR	72941
5	6512705660523829539	Inderdeep	Laurence	D		20001228	344		M	0		44 PROSPECT PL	GREENSBORO	FL	32330
6	9193989543268753887	Chrystine	Mangiapane			19990827	315		F	0	19780306	1806 1ST AVE APT 8F	ARVADA	CO	80007
7	12286552293562700162	Adelene	Stock	R		20000827	252		M	0		1117 FARM RD	DOVER	DE	19901
8	11459575736386985069	Mendy	Rufenblanchette			20000903	24		M	0		3 W 83RD ST APT 4C	WILLIAMSTON	SC	29697
9	8053906447536575038	Lannie	Amerantes	I		20001219	313		U	0		200 W 20TH ST APT 909	CHARLESTON	WV	25312
10	484768759680234166	Tare	Gonyeau	T		19930807	48		F	0	19750801	6 CANDLE CT	EL PASO	TX	79924
11	16156125023194932930	Finney	Aristilde	P		19900621	344		M	0	19560920	222 1ST AVE APT 2B	MACON	GA	31220
12	13804468446718957143	Oreoluwa	Marthaler			19931006	358		F	0	19731201	176 CLAREMONT GDNS	AUBURN	ME	04210
13	11995825474648190448	Surge	Abbottkrepp	D		20000308	13		F	0		22 LE PARC CT	TWINSBURG	OH	44087
14	15714117310244664573	Dave	Mcjury			20001129	238		U	0		510 COOPER RD # 1	TACOMA	WA	98402
15	12587451362606486546	Ramsay	Ping			20001129	238		M	0		404 AVENUE L	MESQUITE	NV	89024

Estrutura RECORD

Uma estrutura **RECORD** define o layout de campos do DATASET.

```
Nome := RECORD  
    campos;  
END;
```

- *Nome* – O nome da estrutura RECORD.
- *campos* – O tipo e o nome de cada campo.

Nota: As palavras-chave RECORD e END podem ser substituídas com chaves ({}) e os delimitadores de campos (;) podem ser substituídos por vírgulas (,).

Declaração DATASET

DATASET introduz um novo arquivo de dados no sistema com o layout *record* especificado.

```
nome := DATASET( arquivo, record, FLAT[THOR] [opções] );  
nome := DATASET(arquivo, record, CSV [ ( opções ) ] );  
nome := DATASET(arquivo, record, XML( caminho, [opções] ) );  
nome := DATASET(arquivo, record, JSON( caminho, [opções] ) );
```

- ✓ *nome* – O nome da definição pelo qual o arquivo passará a ser referenciado.
- ✓ *arquivo* – Uma constante string contendo o nome do arquivo lógico.
- ✓ *record* – A estrutura RECORD do dataset.

Nota: Um conjunto de registros pode ser definido inline entre colchetes (indicando uma definição set). Dentro dos colchetes, cada registro é delimitado por chaves ({}) e separado por vírgulas. Os campos dentro de cada registro são delimitados por vírgula.

```
Names := DATASET([{'John','Jones'}, {'Jane','Smith'}], {STRING first_name, STRING last_name});
```

Atenção! 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

Já posso ver os meus dados?

A ação **OUTPUT** grava o *recordset* em um arquivo e formatos especificados.

OUTPUT(*recordset* [,*formato*] [,*arquivo* [,OVERWRITE]])

- *recordset* – O conjunto de registros a processar.
- *formato* – O formato de saída dos registros: uma estrutura RECORD previamente definida, ou um layout de registros "on-the-fly" entre chaves ({ }).
- *arquivo* – Nome opcional do arquivo onde os registros serão gravados. Caso seja omitido, os dados formatados são mostrados na linha de comando ou no ECL IDE.
- OVERWRITE – Permite sobreescrever o arquivo, caso ele já exista.

Exemplos de OUTPUT:

```
OUTPUT(File_Accounts.File);
```

```
OUTPUT(Persons,{FirstName, LastName}, NAMED('Names_Only'));
```

```
OUTPUT(MyRecordset,, '~CLASS::BMF::NewData', OVERWRITE);
```

//THOR é o formato padrão, mas também é possível gerar saída como:

```
OUTPUT(MyRecordset,, '~CLASS::BMF::NewData', CSV);
```

```
OUTPUT(MyRecordset,, '~CLASS::BMF::NewData', XML);
```

```
OUTPUT(MyRecordset,, '~CLASS::BMF::NewData', JSON);
```

Filtragem simples de dados

- Uma expressão booleana entre parênteses após um Dataset/Recordset é um **filtro**
- Múltiplos filtros podem ser especificados usando uma vírgula (,) ou usando “AND”

```
ValidNames := People(Lastname >= 'T', Lastname < 'U');
```

```
ValidTrades := Trades(Rate >= 7);
```

```
ValidPeople := People(NOT IsSeniorCitizen AND Lastname < 'U');
```

```
ValidPeople2 := People(state IN ['FL','NY']);
```

Operadores de comparação

Equivalência	=
Diferente de	<>
Diferente de	!=
Menor que	<
Maior que	>
Menor ou igual que	<=
Maior ou igual que	>=
Comparação de equivalência	<=> retorna -1, 0, or 1

Operadores aritméticos:

Divisão	/
Divisão Inteira	DIV
Divisão Módulo	%
Multiplicação	*
Adição	+
Subtração	-

**Nota: Qualquer divisão por (0) resulta em zero (0).
Esse comportamento pode ser alterado especificando-se
`#OPTION ('divideByZero', 'fail');` //Aborta e reporta erro**

Funções de agregação

COUNT(*recordset*)

COUNT(*listavalores*)

MAX(*recordset* , *campo*)

MAX(*listavalores*)

MIN(*recordset* , *campo*)

MIN(*listavalores*)

SUM(*recordset* , *campo*)

SUM(*listavalores*)

AVE(*recordset* , *campo*)

AVE(*listavalores*)

- *recordset* – O set ou conjunto de registros a serem processados.
- *campo* – O campo ou expressão a partir dos quais o valor deve ser calculado.
- *listavalores* – Uma lista de expressões separadas por vírgula a partir dos quais o valor deve ser calculado. Também pode ser um SET de valores.

```
OldCount:=COUNT(People(IsSeniorCitizen));
```

```
MaxVal := MAX(People, People.age);
```

```
MinVal1 := MIN(People, People.age);
```

Tarefa: Lending Club

Lending Club

- Foi a maior plataforma de empréstimos peer-to-peer do mercado, com sede em São Francisco, Califórnia.
- US\$ 15,98 bilhões em empréstimos e 3,8 milhões de usuários.
- Fonte: <https://www.kaggle.com/ethon0426/lending-club-20072020q1>
- ~3 milhões de registros (pedidos de empréstimo) e 141 atributos

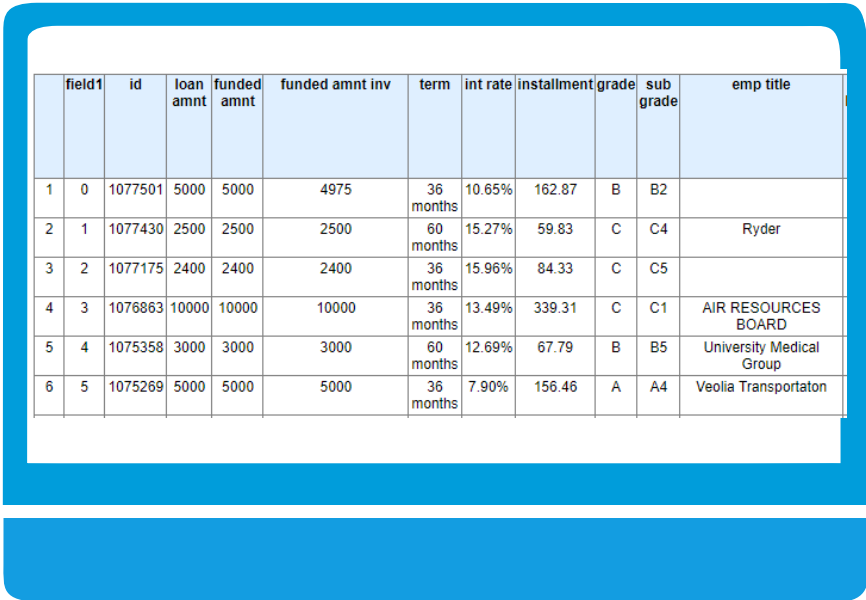
O que esperamos descobrir

- Problema: pedidos de empréstimo podem ser honrados (adimplentes) ou não (inadimplentes)
- Quais **atributos** podem ser utilizados para estabelecer o risco de inadimplência de alguém que contratou o empréstimo?
- Quais as características do **modelo** capaz de prever o risco de inadimplência do empréstimo?
- É possível disponibilizar o modelo via uma interface para **consultas online**?

Exercício prático:

Faça a extração do dataset do Lending Club

- Spray
- Estrutura RECORD
- Declaração DATASET



	field1	id	loan amnt	funded amnt	funded amnt inv	term	int rate	installment	grade	sub grade	emp title
1	0	1077501	5000	5000	4975	36 months	10.65%	162.87	B	B2	
2	1	1077430	2500	2500	2500	60 months	15.27%	59.83	C	C4	Ryder
3	2	1077175	2400	2400	2400	36 months	15.96%	84.33	C	C5	
4	3	1076863	10000	10000	10000	36 months	13.49%	339.31	C	C1	AIR RESOURCES BOARD
5	4	1075358	3000	3000	3000	60 months	12.69%	67.79	B	B5	University Medical Group
6	5	1075269	5000	5000	5000	36 months	7.90%	156.46	A	A4	Veolia Transportaton

Até a próxima aula!!!

