Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas

Aula 3

LexisNexis® RISK SOLUTIONS

RELX Group

PRO

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

USP UNIVERSIDADE DE SÃO PAULO

Agenda da aula 3

- ✓Revisão
  - ✓Aula 2
  - ✓Resolução Chicago Crime
- ✓Extração de dados
  - ✓Pefilamento de dados
  - ✓Filtragem de dados/agregações
- ✓Exercicio prático e Desafio

Desafio: Chicago Crimes

Desafio:

- Definir uma estrutura de dados para o dataset de crimes da polícia de Chicago (data.cityofchicago.org)

##	id	case_number	date	block	iucr	primary_type	description
1	11718445	JC301146	06/10/2019 11:55:00 PM	022XX S SAWYER AVE	0312	ROBBERY	ARMED:KNIFE/CUTTING I
2	11718423	JC301185	06/10/2019 11:55:00 PM	003XX N PINE AVE	0890	THEFT	FROM BUILDING
3	11718364	JC301127	06/10/2019 11:55:00 PM	033XX S MICHIGAN AVE	2093	NARCOTICS	FOUND SUSPECT NARCOTI
4	11718476	JC301140	06/10/2019 11:50:00 PM	057XX S ABERDEEN ST	0497	BATTERY	AGGRAVATED DOMESTIC B
5	11718619	JC301294	06/10/2019 11:47:00 PM	050XX W DIVISION ST	031A	ROBBERY	ARMED: HANDGUN
6	11718392	JC301160	06/10/2019 11:45:00 PM	003XX E 118TH ST	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
7	11718384	JC301137	06/10/2019 11:40:00 PM	080XX S LANGLEY AVE	0486	BATTERY	DOMESTIC BATTERY SIMP
8	11718398	JC301118	06/10/2019 11:31:00 PM	047XX N KEYSTONE AVE	051A	ASSAULT	AGGRAVATED: HANDGUN
9	11718368	JC301109	06/10/2019 11:31:00 PM	096XX S MERRION AVE	1310	CRIMINAL DAMAGE	TO PROPERTY
10	11718393	JC301135	06/10/2019 11:24:00 PM	105XX S SANGAMON ST	0497	BATTERY	AGGRAVATED DOMESTIC B
11	11718444	JC301119	06/10/2019 11:21:00 PM	013XX N PAULINA ST	0860	THEFT	RETAIL THEFT
12	11718411	JC301116	06/10/2019 11:15:00 PM	065XX N FAIRFIELD AVE	0486	BATTERY	DOMESTIC BATTERY SIMP
13	11718351	JC301114	06/10/2019 11:15:00 PM	057XX N WINTHROP AVE	0820	THEFT	$500 AND UNDER
14	11722011	JC301099	06/10/2019 11:15:00 PM	015XX N WASHTENAW AVE	051A	ASSAULT	AGGRAVATED: HANDGUN
15	11718459	JC301152	06/10/2019 11:15:00 PM	054XX W WALTON ST	3710	INTERFERENCE WITH PU...	RESIST/OBSTRUCT/DISAR

Solução proposta:

```
EXPORT File_crime_raw := MODULE
  EXPORT Layout:= RECORD
    STRING ID;
    STRING Case_Number;
    STRING Date;
    STRING Block;
    STRING IUCR;
    STRING Primary_Type;
    STRING Description;
    STRING Location_Description;
    STRING Arrest;
    STRING Domestic;
    STRING Beat;
    STRING District;
    STRING Ward;
    STRING Community_Area;
    STRING FBI_Code;
    STRING X_Coordinate;
    STRING Y_Coordinate;
    STRING Year;
    STRING Updated_On;
    STRING Latitude;
    STRING Longitude;
    STRING Location;
  END;
  EXPORT File:=DATASET('~chicago::hmw::crimes_2001_to_present',Layout,CSV(heading(1)));
END;
```

Solução proposta (cont.):

##	id	case_number	date	block	iucr	primary_type	description
1	11718445	JC301146	06/10/2019 11:55:00 PM	022XX S SAWYER AVE	0312	ROBBERY	ARMED:KNIFE/CUTTING INSTRUMENT
2	11718423	JC301185	06/10/2019 11:55:00 PM	003XX N PINE AVE	0890	THEFT	FROM BUILDING
3	11718364	JC301127	06/10/2019 11:55:00 PM	033XX S MICHIGAN AVE	2093	NARCOTICS	FOUND SUSPECT NARCOTICS
4	11718476	JC301140	06/10/2019 11:50:00 PM	057XX S ABERDEEN ST	0497	BATTERY	AGGRAVATED DOMESTIC BATTERY: OTHE
5	11718619	JC301294	06/10/2019 11:47:00 PM	050XX W DIVISION ST	031A	ROBBERY	ARMED: HANDGUN
6	11718392	JC301160	06/10/2019 11:45:00 PM	003XX E 118TH ST	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
7	11718384	JC301137	06/10/2019 11:40:00 PM	080XX S LANGLEY AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
8	11718398	JC301118	06/10/2019 11:31:00 PM	047XX N KEYSTONE AVE	051A	ASSAULT	AGGRAVATED: HANDGUN
9	11718368	JC301109	06/10/2019 11:31:00 PM	096XX S MERRION AVE	1310	CRIMINAL DAMAGE	TO PROPERTY
10	11718393	JC301135	06/10/2019 11:24:00 PM	105XX S SANGAMON ST	0497	BATTERY	AGGRAVATED DOMESTIC BATTERY: OTHE
11	11718444	JC301119	06/10/2019 11:21:00 PM	013XX N PAULINA ST	0860	THEFT	RETAIL THEFT
12	11718411	JC301116	06/10/2019 11:15:00 PM	065XX N FAIRFIELD AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
13	11718351	JC301114	06/10/2019 11:15:00 PM	057XX N WINTHROP AVE	0820	THEFT	$500 AND UNDER
14	11722011	JC301099	06/10/2019 11:15:00 PM	015XX N WASHTENAW AVE	051A	ASSAULT	AGGRAVATED: HANDGUN
15	11718459	JC301152	06/10/2019 11:15:00 PM	054XX W WALTON ST	3710	INTERFERENCE WIT...	RESIST/OBSTRUCT/DISARM OFFICER
16	11718395	JC301123	06/10/2019 11:14:00 PM	029XX N CENTRAL AVE	0610	BURGLARY	FORCIBLE ENTRY

Extração de dados

Definição de Extração:

Importação e limpeza de dados brutos provenientes de diferentes fontes

- Importação dos dados brutos
- Definição da estrutura de dados
- Análise do perfil dos dados

Análise do perfil dos dados

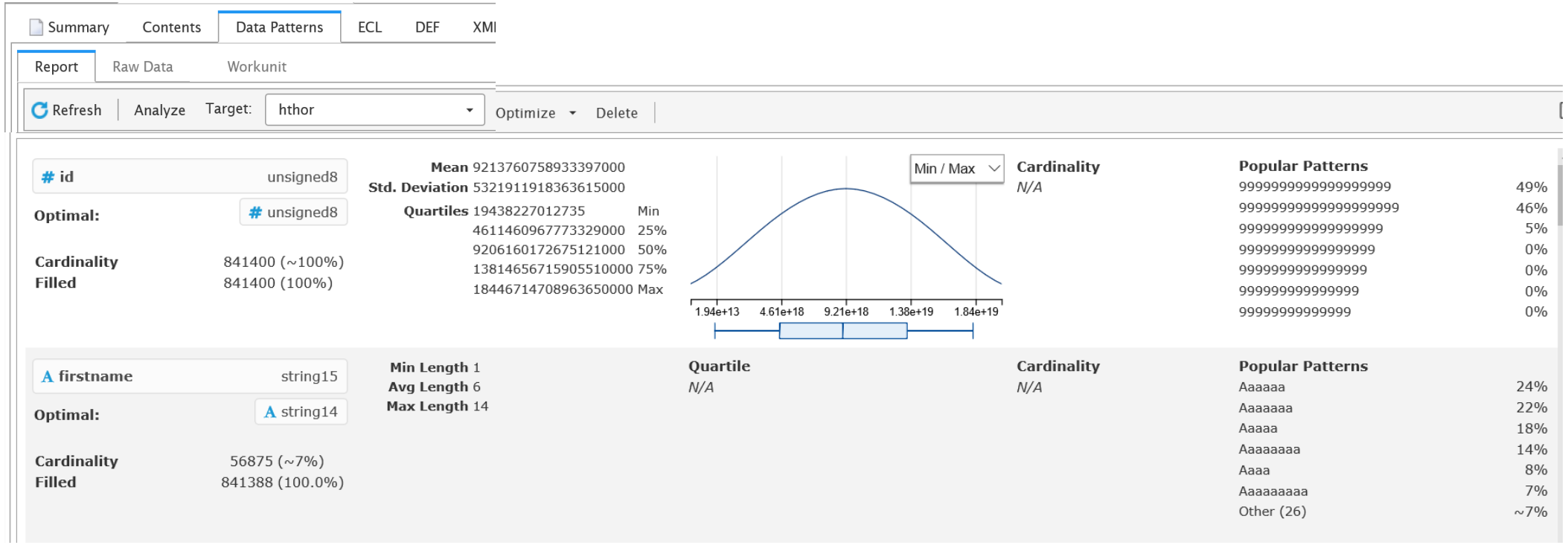
- **Perfilamento de dados (Data Patterns)**
- **Filtragem de dados / funções agregativas**
- **Tabulação cruzada (TABLE)**

Perfilamento de dados

Perfilamento de dados: Data Patterns

Um relatório detalhado dos dados é disponibilizado no ECL Watch para todos os arquivos lógicos.

O relatório pode ser acessado na aba Data Patterns:



Fundação Vanzolini

Geração do relatório: Data Patterns

O relatório pode ser gerado de três maneiras:

- ECL Watch (na aba Data Patterns)
- Biblioteca padrão (STD.DataPatterns)
- Bundle (https://github.com/hpcc-systems/DataPatterns.git)

Via código ECL:

```
IMPORT STD;

filePath := '~aaa::bmf::reptest::accounts';
ds := DATASET(filePath, RECORDOF(filePath), csv);
profileResults := STD.DataPatterns.Profile(ds);
OUTPUT(profileResults, ALL, NAMED('profileResults'));
```

Via ECL Watch:

1. Arquivo lógico deve possuir uma estrutura RECORD
2. Evitar o uso da interface web para arquivos muito grandes.

Filtragem de dados / Funções agregativas

Recapitulando: tipos básicos de definição

Booleana (*boolean*)

```
IsSeniorCitizen := People.Age >= 65;
```

Valor único (*value*)

```
ValueTrue := 1;
```

Conjunto de valores (*set*)

```
SetTrueFalseValues := [0, 1];
```

Conjunto de registros (*recordset*)

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

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

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

Operadores aritméticos:

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

Nota: Qualquer divisão por (0) resulta em zero (0).

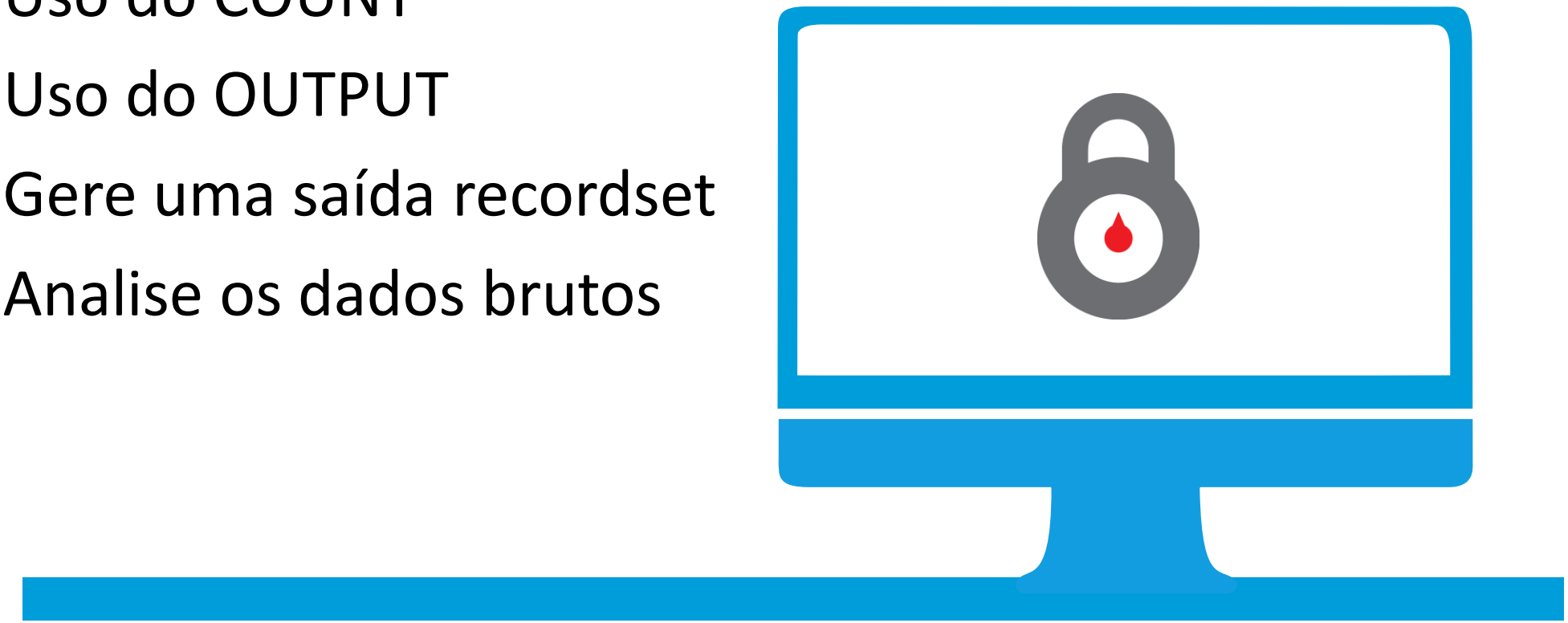
Esse comportamento pode ser alterado especificando-se

#OPTION ('divideByZero', ‘fail'); //Aborta e reporta erro

Exercício prático

Exercício 6 – Consultas básicas (somente Persons)

- Uso do IMPORT
- Uso do COUNT
- Uso do OUTPUT
- Gere uma saída recordset
- Analise os dados brutos



Exercício prático

Exercício 7a – Filtros

- Filtros básicos
- Mais usos do COUNT
- Uso de operadores lógicos

Exercício prático

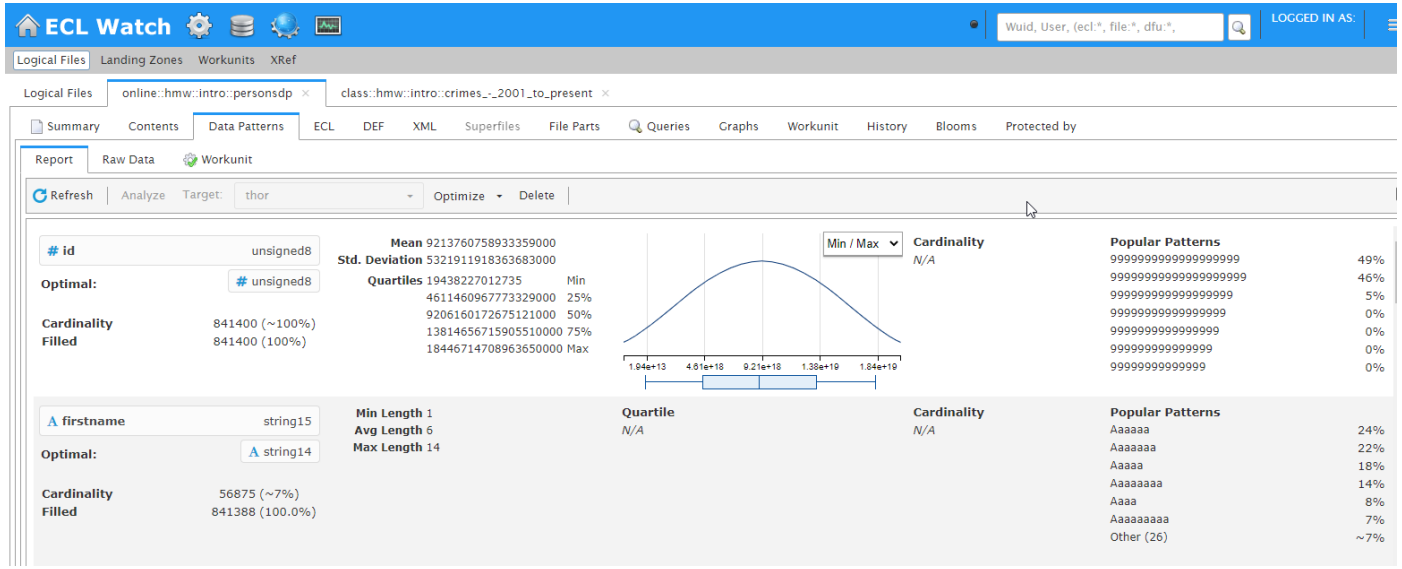
Exercícios 8a e 10a – Combinação de filtros

- Definições booleanas simples
- Combinação de booleanas
- Aplicação de filtro

Desafio: Chicago Crimes

Desafio Chicago Crimes:

- Analise o perfil dos dados brutos e procure por problemas de qualidade dos dados:
  - Campos vazios, preenchimento incosistente, tipos de campos não otimizados, etc
- Gere uma nova estrutura RECORD com tipos de campos otimizados e a utilize para definir o DATASET



Fundação Vanzolini

Resumo da 1º semana:

- ✓HPCC Systems é uma plataforma para solucionar desafios de *Big Data*
  - ✓ECL IDE: manipulação de dados e criação de consultas
  - ✓ECL Watch: gestão de arquivos e workunits

- ✓ECL é uma *linguagem* para consulta e ETL
  - ✓Definições estabelecem o que as coisas são
  - ✓Ações resultam em compilação e execução

- ✓Extração de dados:
  - ✓Importação via spray/ECL Watch
  - ✓Definição da estrutura de dados: RECORD, DATASET, MODULE
  - ✓Análise de perfil dos dados: Output, Data Patterns e filtragem/agregações

Links úteis

- Site principal: **hpccsystems.com**
- Primeiros passos: **hpccsystems.com/Why-HPCC-Systems**
- Treinamento Online: **learn.lexisnexis.com/hpcc**
- Download: **hpccsystems.com/download**
- Fórum da Comunidade: **hpccsystems.com/forums**

Faça parte da Comunidade

Registre-se em hpccsystems.com

Até a próxima aula!!!