



Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas

Aula 4

Agenda da aula 4

- ✓ Resolução Chicago Crime
- ✓ Transformação de dados
 - ✓ Identificadores de registros
 - ✓ Padronização de registros
- ✓ Exercício prático e Desafio

Desafio: Chicago Crimes

Solução proposta:

```
EXPORT File_crime_optimized := MODULE
  EXPORT Layout:=      RECORD
    UNSIGNED ID;
    STRING Case_Number;
    STRING Date;
    STRING Block;
    STRING IUCR;
    STRING Primary_Type;
    STRING Description;
    STRING Location_Description;
    BOOLEAN Arrest;
    BOOLEAN Domestic;
    UNSIGNED2 Beat;
    UNSIGNED2 District;
    UNSIGNED2 Ward;
    UNSIGNED2 Community_Area;
    STRING FBI_Code;
    UNSIGNED6 X_Coordinate;
    UNSIGNED6 Y_Coordinate;
    UNSIGNED2 Year;
    STRING Updated_On;
    DECIMAL11_9 Latitude;
    DECIMAL11_9 Longitude;
    STRING Location;
  END;
  EXPORT File:=DATASET('~chicago::hmw::crimes_2001_to_present',Layout,CSV(heading(1)));
END;
```

Transformação de dados

124.38 MB

id	firstname	lastname	middlename	namesuffix	filedate	bureaucode	marital	gender	dependentcount	birthdate	streetaddress	city	state	zipcode
91082180...	Cherianne	Khatchatourian	N		19990922	24		M	0		69 BOULDER RIDGE RD # 25A	HAWKINS	WI	54530
16505326...	Muyesser	Raplee	X		20001111	353		F	0		55 SWAMP RD	DISTRICT HEIGHT	MD	20747
24548180...	Roselin	Viceconte			19990325	344		F	0	19800113	107 HILL TER	ENTERPRISE	OR	97828
15880908...	Inda	Provines			20000909	13		U	0		290 W MOUNT PLEASANT AVE	LAVACA	AR	72941
65127056...	Inderdeep	Laurence	D		20001228	344		M	0		44 PROSPECT PL	GREENSBORO	FL	32330
91939895...	Chrystine	Mangiapane			19990827	315		F	0	19780306	1806 1ST AVE APT 8F	ARVADA	CO	80007
12286552...	Adelene	Stock	R		20000827	252		M	0		1117 FARM RD	DOVER	DE	19901
11459575...	Mendy	Rufenblanchette			20000903	24		M	0		3 W 83RD ST APT 4C	WILLIAMSTON	SC	29697
80539064...	Lannie	Amerantes	I		20001219	313		U	0		200 W 20TH ST APT 909	CHARLESTON	WV	25312
48476875...	Tare	Gonyeau	T		19930807	48		F	0	19750801	6 CANDLE CT	EL PASO	TX	79924
16156125...	Finney	Aristilde	P		19900621	344		M	0	19560920	222 1ST AVE APT 2B	MACON	GA	31220
13804468...	Oreoluwa	Marthaler			19931006	358		F	0	19731201	176 CLAREMONT GDNS	AUBURN	ME	04210
11995825...	Surge	Abbottkrepp	D		20000308	13		F	0		22 LE PARC CT	TWINSBURG	OH	44087
15714117...	Dave	Mcjury			20001129	238		U	0		510 COOPER RD # 1	TACOMA	WA	98402

89.87 MB

recid	id	firstname	lastname	middlename	namesuffi	filedate	bureaucode	gender	birthdate	streetaddress	csz_id
1	91082180...	CHERIANNE	KHATCHATOURIAN	N		19990922	24	M	0	69 BOULDER RIDGE RD # 25A	1
2	16505326...	MUYESSER	RAPLEE	X		20001111	353	F	0	55 SWAMP RD	2
3	24548180...	ROSELIN	VICECONTE			19990325	344	F	19800113	107 HILL TER	3
4	15880908...	INDA	PROVINES			20000909	13	U	0	290 W MOUNT PLEASANT AVE	4
5	65127056...	INDERDEEP	LAURENCE	D		20001228	344	M	0	44 PROSPECT PL	5
6	91939895...	CHRYSTINE	MANGIAPANE			19990827	315	F	19780306	1806 1ST AVE APT 8F	6
7	12286552...	ADELENE	STOCK	R		20000827	252	M	0	1117 FARM RD	7
8	11459575...	MENDY	RUFENBLANCHETTE			20000903	24	M	0	3 W 83RD ST APT 4C	8
9	80539064...	LANNIE	AMERANTES	I		20001219	313	U	0	200 W 20TH ST APT 909	9
10	48476875...	TARE	GONYEAU	T		19930807	48	F	19750801	6 CANDLE CT	10
11	16156125...	FINNEY	ARISTILDE	P		19900621	344	M	19560920	222 1ST AVE APT 2B	11
12	13804468...	OREOLUNA	MARTHALER			19931006	358	F	19731201	176 CLAREMONT GDNS	12
13	11995825...	SURGE	ABBOTTKREPP	D		20000308	13	F	0	22 LE PARC CT	13
14	15714117...	DAVE	MCJURY			20001129	238	U	0	510 COOPER RD # 1	14

586.32 KB

csz_id	city	state	zipcode
1	HAWKINS	WI	54530
2	DISTRICT HEIGHT	MD	20747
3	ENTERPRISE	OR	97828
4	LAVACA	AR	72941
5	GREENSBORO	FL	32330
6	ARVADA	CO	80007
7	DOVER	DE	19901
8	WILLIAMSTON	SC	29697
9	CHARLESTON	WV	25312
10	EL PASO	TX	79924
11	MACON	GA	31220
12	AUBURN	ME	4210
13	TWINSBURG	OH	44087
14	TACOMA	WA	98402

Designação de identificadores (recid's)

Definição de Transformação:

Mapeamento e conversão de dados para layouts de registros padronizados

- Designação de identificadores (recid's)
- Padronização de campos
- Ordenação (e remoção de duplicidade) de registros
- Normalização/desnormalização

Opere somente nos dados necessários...

Função TABLE

A função **TABLE** function é similar ao OUTPUT, mas disponibiliza os registros somente em memória.

TABLE(*recordset*, *formato* [,*expressão*])

- *recordset* – Conjunto de registros a processar.
- *formato* – Estrutura RECORD de saída.
- *expressão* – Diretriz de agrupamento para tabulação. Múltiplas expressões separadas por vírgula criam uma única cláusula lógica.

Exemplo de TABLE (vertical slice):

```
Layout_Name_State := RECORD  
    Persons.LastName;  
    Persons.FirstName;  
    Persons.State;  
END;
```

v#	lastname	firstname	state
1	Khatchatourian	Cherianne	WI
2	Raplee	Muyesser	MD
3	Viceconte	Roselin	OR
4	Provines	Inda	AR
5	Laurence	Inderdeep	FL
6	Mangiapane	Chrystine	CO
7	Stock	Adelene	DE
8	Rufenblanchette	Mendy	SC
9	Amerantes	Lannie	WV
10	Gonyeau	Tare	TX

```
Per_Name_State := TABLE(Persons, Layout_Name_State);
```

Tabulação cruzada

A palavra reservada **GROUP** substitui o parâmetro *recordset* em qualquer função de agregação utilizada na estrutura RECORD de uma definição de TABLE que contenha uma *expressão* de agrupamento.

```
R := RECORD
```

```
  Persons.State;
```

```
  Persons.Gender;
```

```
  cnt := COUNT(GROUP);
```

```
END;
```

```
CTOut := TABLE(Persons, R, State, Gender);
```

##	state	gender	cnt
1	AA	M	139
2	AA	N	12
3	AE	F	1503
4	AK	F	1563
5	AK	M	1540
6	AL	M	7910
7	AL	N	407
8	AP	M	814
9	AP	N	49
10	AR	N	252

Exemplo de CROSSTAB:

BWR_Training_Examples.Crosstab_Example

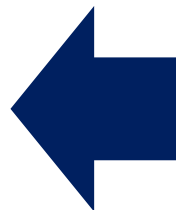
Designação de identificadores (recid's)

Identificadores de registros

- Estrutura TRANSFORM
- Função PROJECT
- Serviço PERSIST

Como eu transformo estruturas de dados?

```
New_Layout := RECORD  
    UNSIGNED8 recid;  
    UNSIGNED8 ID;  
    STRING15   FirstName;  
    STRING25   LastName;  
    STRING15   MiddleName;  
END;
```



```
Layout := RECORD  
    UNSIGNED8 ID;  
    STRING15   FirstName;  
    STRING25   LastName;  
    STRING15   MiddleName;  
END;
```

```
New_Layout AddRecID(Layout Le, UNSIGNED cnt) := TRANSFORM  
    SELF. recid := cnt;  
    SELF := Le;  
END;
```

```
YP_Slim := PROJECT(Persons, AddRecID(LEFT,COUNTER)): PERSIST('~CLASS::HMW::NewPersons');
```


Estrutura TRANSFORM

```
formatosaida nome( parâmetros ) := TRANSFORM  
    SELF.campo := transformação;  
END;
```

- *formatosaida* – Nome da estrutura RECORD especificando o formato de saída.
- *nome* – Nome da estrutura TRANSFORM.
- *parâmetros* – Tipos de valores e rótulos dos parâmetros a serem passados para a transformação.
- **SELF** – Referência à estrutura de saída.
- *campo* – Nome do campo na estrutura de saída.
- *transformação* – Expressão que especifica como produzir o valor a ser designado no campo de saída.

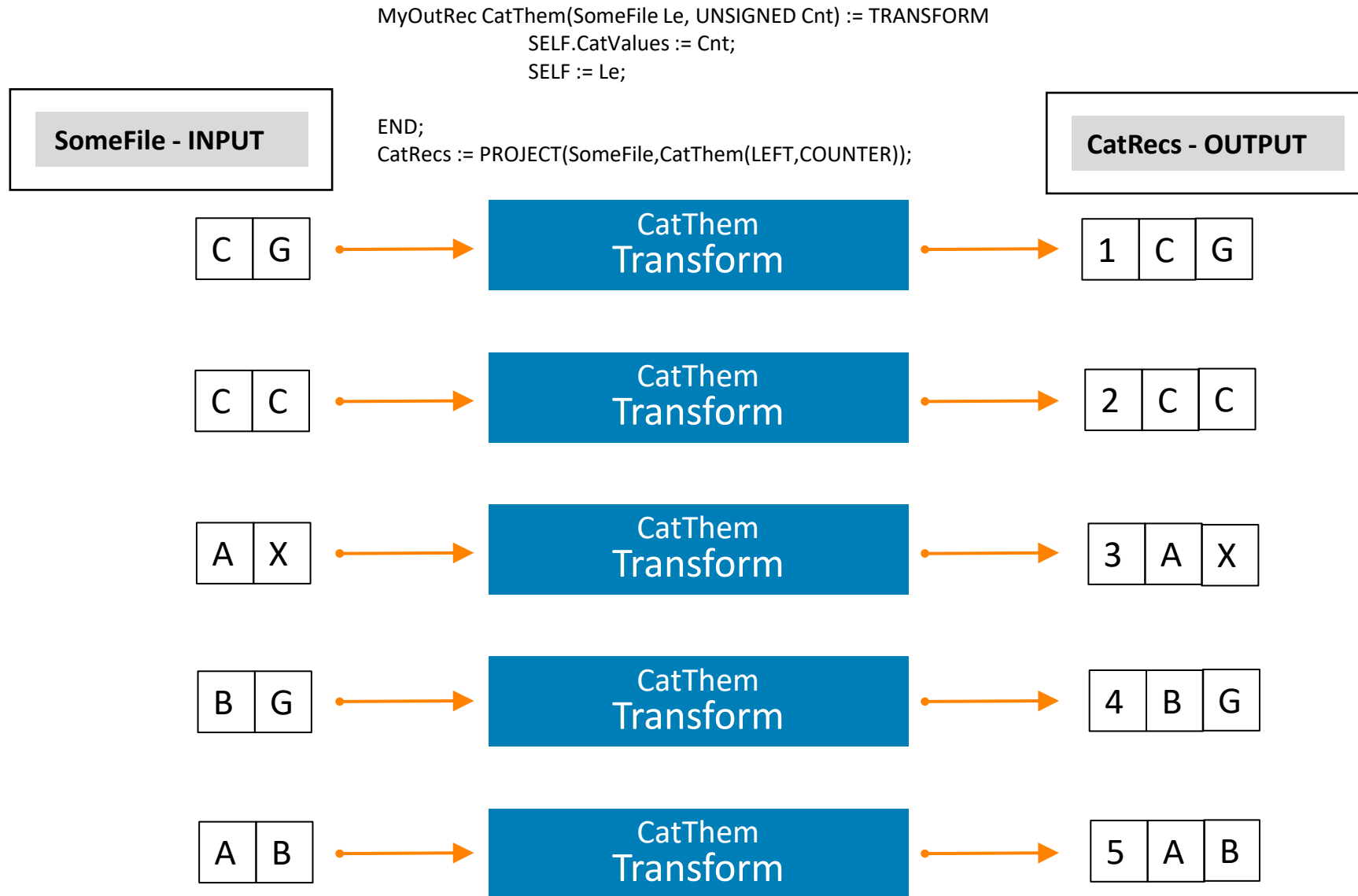
Função PROJECT

A função **PROJECT** processa todos os registros do *recordset* chamando a estrutura *transform* para sua execução em cada registro.

PROJECT(*recordset*, *transform*)

- *recordset* – Conjunto de registros a serem processados.
- *transform* – Nome da estrutura TRANSFORM a ser chamada para cada registro do *recordset*.

Diagrama funcional de um PROJECT



Serviço :PERSIST

O serviço **PERSIST** armazena o resultado de uma *expressão* em um arquivo lógico.

nome := *exp* : **PERSIST**(*arquivo* [,**EXPIRE**(*dias*)]);

- *nome* – Nome da definição.
- *exp* – Expressão da definição.
- *arquivo* – Nome do arquivo lógico onde o resultado do PERSIST será armazenado.
- EXPIRE – Indica um valor customizado para remoção do arquivo.
- Dias – Número de dias para remoção automática do arquivo.

Padronização de registros

Padronização de registros

- TABLE ou PROJECT
- Bibliotecas-padrão e CASTING
- Função SIZEOF

Elementos para padronização de registros:

- **Utilize projeções verticais para trabalhar somente com os campos necessários**
- **Padronize nomes, endereços, datas, horários, etc.**
 - `A := STD.Str.ToUpperCase('abcde');` //A contains 'ABCDE'
 - `D2 := STD.Date.FromStringToDate('4/29/1974', '%m/%d/%Y');` //D2 contains 19740429
- **STRING: selecione o menor tamanho apropriado à faixa de valores de trabalho**
- **Números: UNSIGNED é a melhor escolha, utilize INTEGER somente se valores negativos forem necessários**

Conversão de tipos

Conversão Explícita –

O novo tipo é especificado entre parênteses na expressão que precede o elemento a ser convertido.

```
STRING10 Value      := '34658';  
Cnt                  := (INTEGER4) Value;  
StrCnt               := (STRING5) Cnt;
```


Função SIZEOF

A função **SIZEOF** retorna o número total de bytes necessários para armazenamento da estrutura ou campo de dados.

SIZEOF(*dado*)

- *dado* – Nome de um dataset, estrutura RECORD, nome de campo totalmente qualificado ou expressão de valor constante STRING.

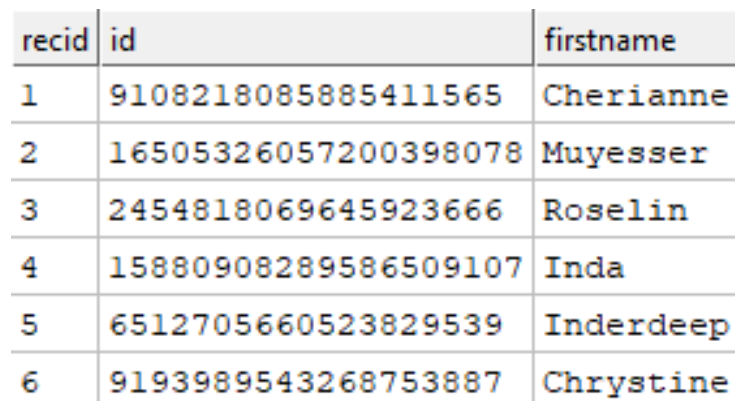
```
MyRec := RECORD
  INTEGER5      F2;
END;
MyData := DATASET([1],MyRec);
```

SIZEOF(MyRec);	//resultado é 5
SIZEOF(MyData);	//resultado é 5
SIZEOF(MyData.F2);	//resultado é 5
SIZEOF('abc' + '123');	//resultado é 6

Exercício prático

Exercício 4a – Adição de identificadores (Persons)

- Uso do TRANSFORM
- Uso do PROJECT
- Uso do PERSIST

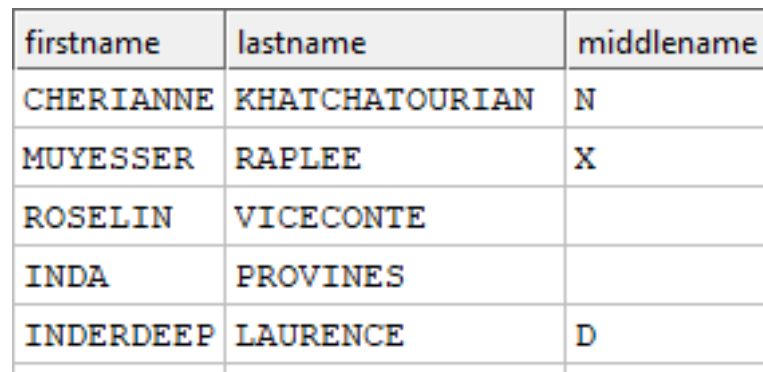


recid	id	firstname
1	9108218085885411565	Cherianne
2	16505326057200398078	Muyesser
3	2454818069645923666	Roselin
4	15880908289586509107	Inda
5	6512705660523829539	Inderdeep
6	9193989543268753887	Chrystine

Exercício prático

Exercício 5a – Padronização e otimização (Persons)

- Estrutura MODULE
- Novo RECORD
- TABLE/PERSIST



firstname	lastname	middlename
CHERIANNE	KHATCHATOURIAN	N
MUYESSER	RAPLEE	X
ROSELIN	VICECONTE	
INDA	PROVINES	
INDERDEEP	LAURENCE	D

Desafio: Chicago Crimes

Desafio Chicago Crimes:

- Gere um dataset padronizado
 - Adicione um campo de identificador de registro
 - Padronize os campos de hora e data

##	row_id	day	time	id	case_number	block	iucr	primary_type	description
1	1	20190610	235500	11718445	JC301146	022XX S SAWYER AVE	0312	ROBBERY	ARMED:KNIFE/CUTTING INSTRUMENT
2	2	20190610	235500	11718423	JC301185	003XX N PINE AVE	0890	THEFT	FROM BUILDING
3	3	20190610	235500	11718364	JC301127	033XX S MICHIGAN AVE	2093	NARCOTICS	FOUND SUSPECT NARCOTICS
4	4	20190610	235000	11718476	JC301140	057XX S ABERDEEN ST	0497	BATTERY	AGGRAVATED DOMESTIC BATTERY: OTHER DANG WEAPC
5	5	20190610	234700	11718619	JC301294	050XX W DIVISION ST	031A	ROBBERY	ARMED: HANDGUN
6	6	20190610	234500	11718392	JC301160	003XX E 118TH ST	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
7	7	20190610	234000	11718384	JC301137	080XX S LANGLEY AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
8	8	20190610	233100	11718398	JC301118	047XX N KEYSTONE AVE	051A	ASSAULT	AGGRAVATED: HANDGUN
9	9	20190610	233100	11718368	JC301109	096XX S MERRION AVE	1310	CRIMINAL DAMAGE	TO PROPERTY
10	10	20190610	232400	11718393	JC301135	105XX S SANGAMON ST	0497	BATTERY	AGGRAVATED DOMESTIC BATTERY: OTHER DANG WEAPC

Até a próxima aula!!!

