Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas

Aula 1

LexisNexis® RISK SOLUTIONS

RELX Group

PRO

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

USP UNIVERSIDADE DE SÃO PAULO

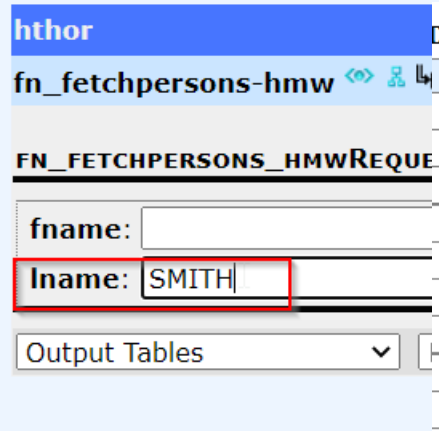
Agenda da aula 1

- ✓ Apresentação do curso
- ✓ Conceitos gerais de Big Data
- ✓ Iniciando a extração de dados

Apresentação do curso

Objetivo final do curso!

- Elaborar um serviço de consulta de dados pessoais:



Dataset: Result 1

	lastname	firstname	recid	middlename	namesuffix	filedate	bureaucode	gender	dependentcount	birthdate	streetaddress	city	state	zipcode
1	SMITH	DETELIN	596872	M		19861027	171	U	0	19650301	338 W 89TH ST APT 3R	ARLINGTON	MN	55307
2	SMITH	CELENIA	644126			19920319	199	F	0	19221201	2355 FOREST HILLS DR	TRANSFER	PA	16154
3	SMITH	TEH	623354	L	SR	19871001	238	M	0	19561201	18 CROSS RIDGE RD	EAST ORANGE	NJ	7017
4	SMITH	YAMROT	341777			20000810	168	F	0	19611214	280 W 25TH ST # C	FAYETTEVILLE	NC	28314
5	SMITH	GANIJA	44886	Z		19870909	13	F	0	19260301	2090 POTTS HILL RD	POMPTON PLAINS	NJ	7444
6	SMITH	CASIANO	643584	N		19871210	78	F	0	19620219	1754 PALISADE AVE	PERU	IL	61354
7	SMITH	RUEI	727071	S		19950921	252	M	0	19751201	43 HILLANDO DR	GLOUCESTER POIN	VA	23062
8	SMITH	ANNONAN	533969			19850821	376	F	0	19530108	134 E 17TH ST APT 63	JAMESON	MO	64647
9	SMITH	NAMIT	347861			19860401	24	M	0	19640928	221 CLERMONT AVE # 1	GLEN ELLYN	IL	60138
10	SMITH	MONTAKARN	277642	Q		19870309	13	M	0	19640929	45 MALLARD RD	DUNCANSVILLE	PA	16635
11	SMITH	VALDINA	609590	T		19940913	352	F	0	19730712	122 N BROAD ST	PITTSFIELD	NH	3263

➢**Persons (~841 k)**

- Desafio de dados: data.cityofchicago.org/

CHICAGO DATA PORTAL

Fundação Vanzolini

Treinamento em ECL/HPCC: learn.lexisnexis.com/hpcc

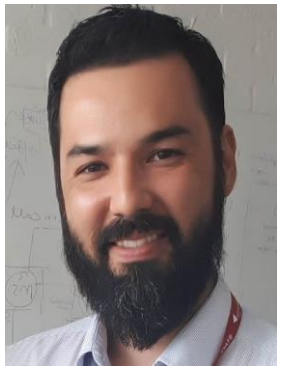
- Introdução ao ECL (parte 1)
  - Conceitos e consultas
- Introdução ao ECL (parte 2)
  - ETL com ECL
- ECL Avançado (parte 1)
  - Dados relacionais
- ECL Avançado (parte 2)
  - Superarquivos, XML/JSON e PLN
- ECL Aplicado
  - Geração e automação de código ECL
- ROXIE ECL (parte 1)
  - Índices e consultas
- ROXIE ECL (parte 2)
  - Otimização de consultas
- Machine Learning com HPCC Systems
  - Tutoriais para uso de plugins
- Administração de Sistemas
  - Conceitos e operação básica
- HPCC para gestores
  - Visão geral e aplicações da plataforma

Recursos e operação

- Aulas: das 19:00hs às 22:00hs
- Dias:
  - 22, 24 e 26 de Agosto (semana 1)
  - 29 e 31 de Agosto e 02 de Setembro (semana 2)
- Computador pessoal (ECL IDE v8.2.18/VSCode/GitHub)
- Cluster: http://trainingcluster.us-hpccsystems-dev.azure.lnrsg.io:8010/
- Slides de aula e livro de exercícios
- Moodle: https://ead.vanzolini.org.br/course/view.php?id=881
- 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



Apresentação dos participantes

- ✓Nome
- ✓Área de atuação
- ✓Experiência/Interesse com big data

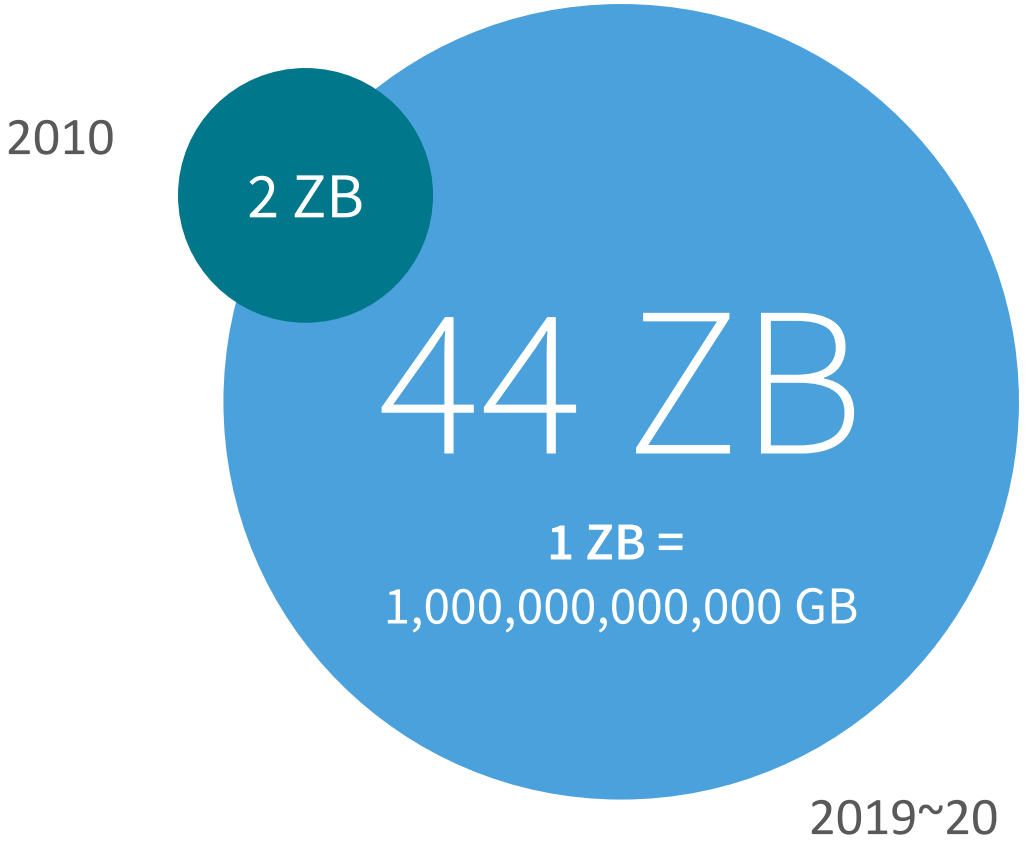
O que é Big Data?

Conceito de Big Data

- Cinco V´s?
  - Volume
  - Variedade
  - Velocidade
  - Veracidade
  - Valor

Fundação Vanzolini

Quão grande é "*Big*"?



https://www.internetlivestats.com/

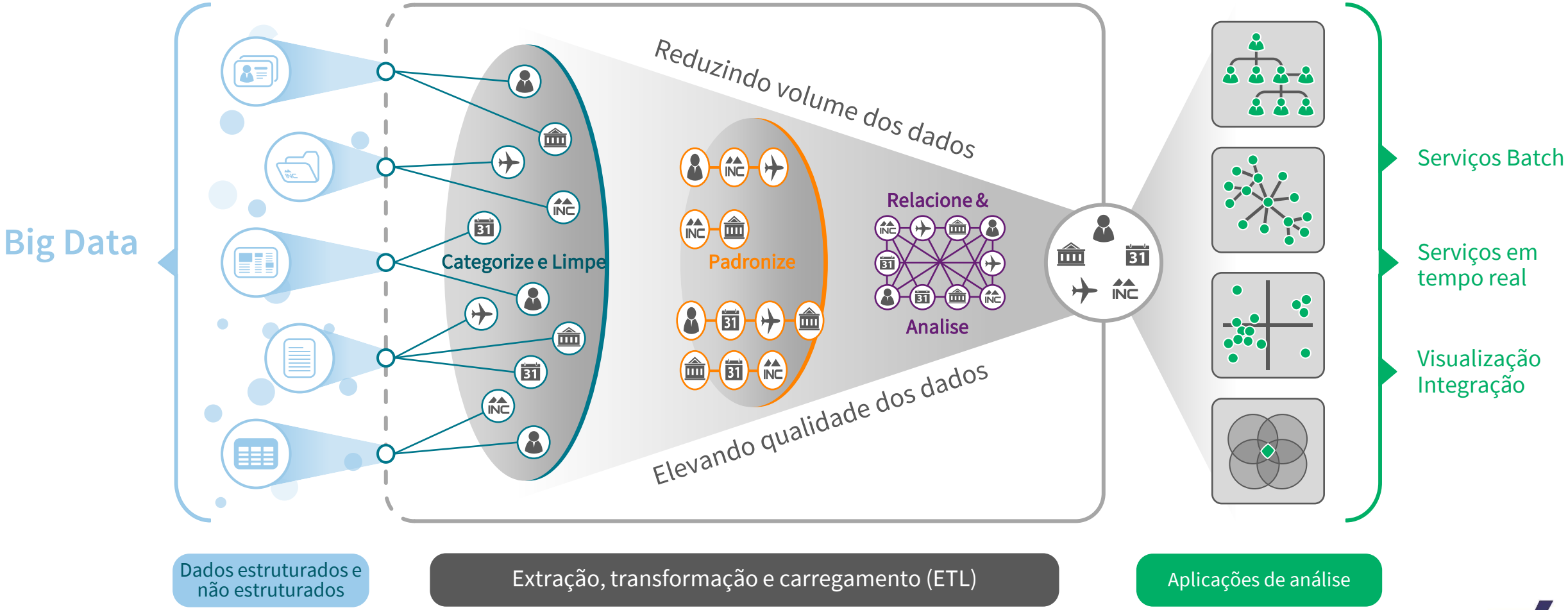
Fundação Vanzolini

Trade off do *Big Data*

- Problema N^2
- Quantidade de dados X Recursos computacionais
- Como processar bilhões de registros em segundos?
- Como analisar dados de múltiplas fontes e transformá-los em informação e conhecimento?

Processamento end to end de Big Data

“Funil” de dados

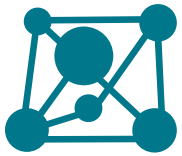


Fundação Vanzolini

“*Stack*” tecnológico

Ferramentas de consulta e visualização

Entrega online de consultas em *Big Data*



Bibliotecas de *Machine Learning*

Supervisionado, não-supervisionado, aprendizagem profunda

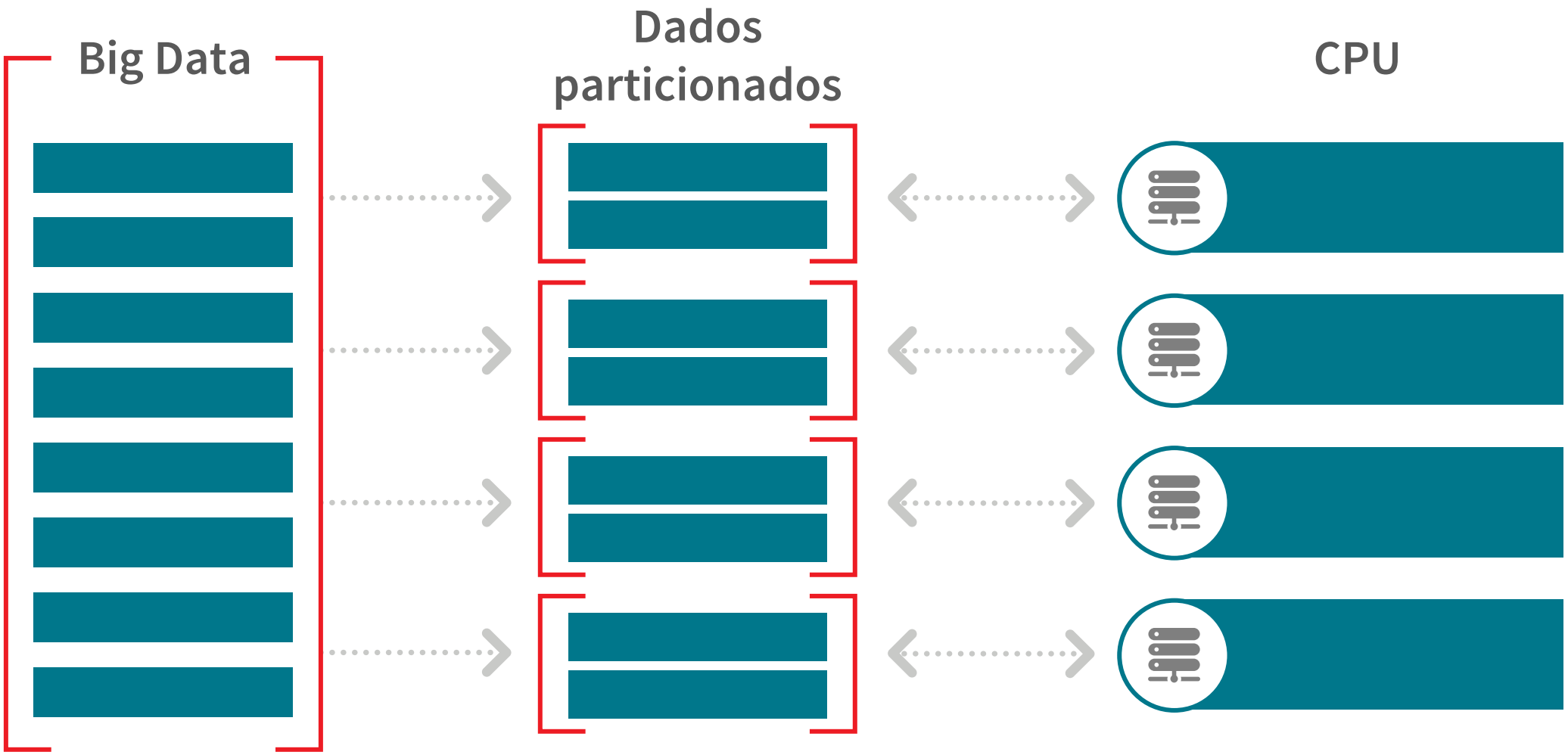
Ferramentas para manipulação de dados

Perfilamento, limpeza, consolidação de dados

ETL

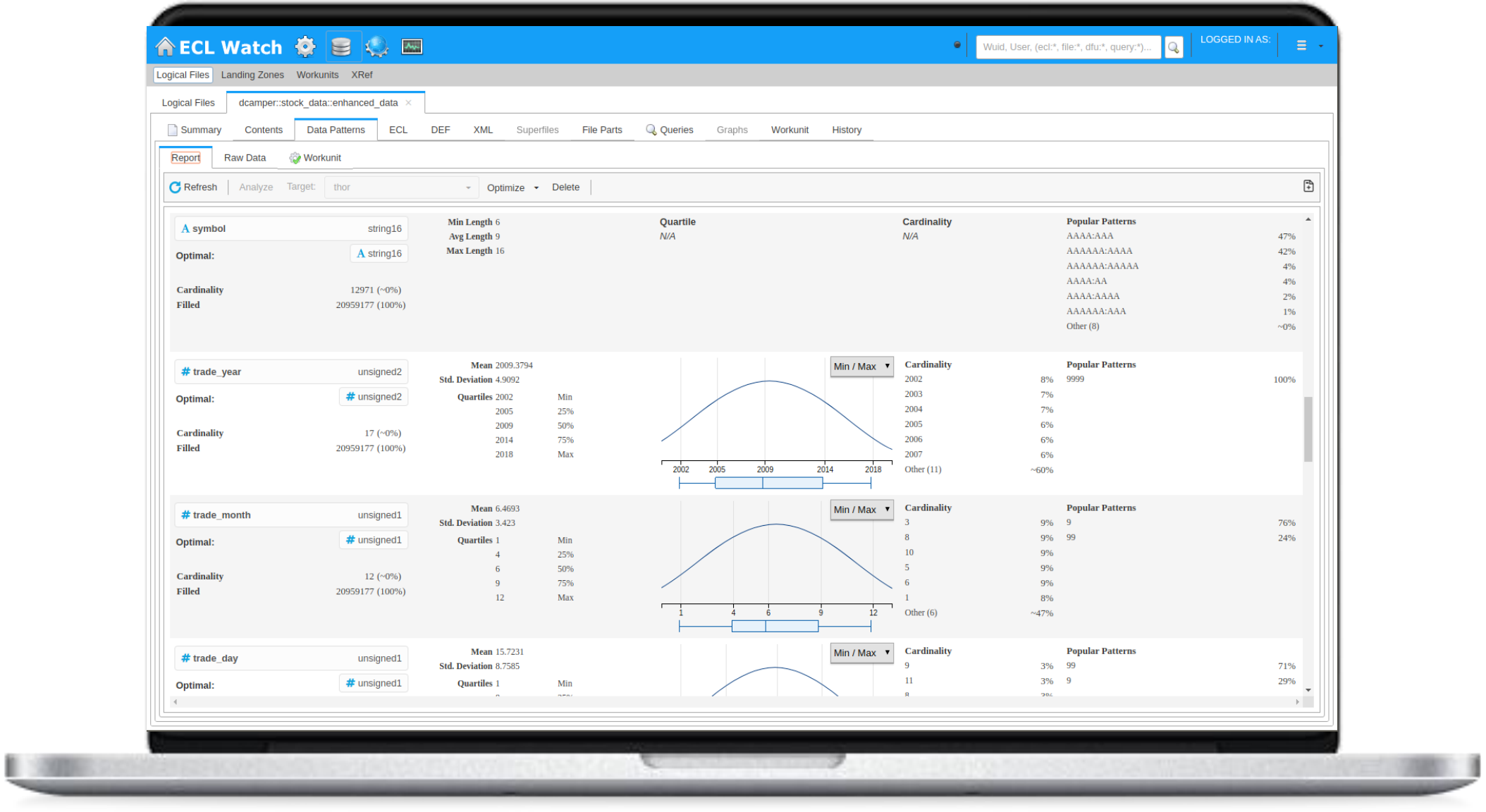
Extração, transformação e carregamento de dados

ETL: Supercomputação



Fundação Vanzolini

Ferramentas de *Profiling*



Fundação Vanzolini

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

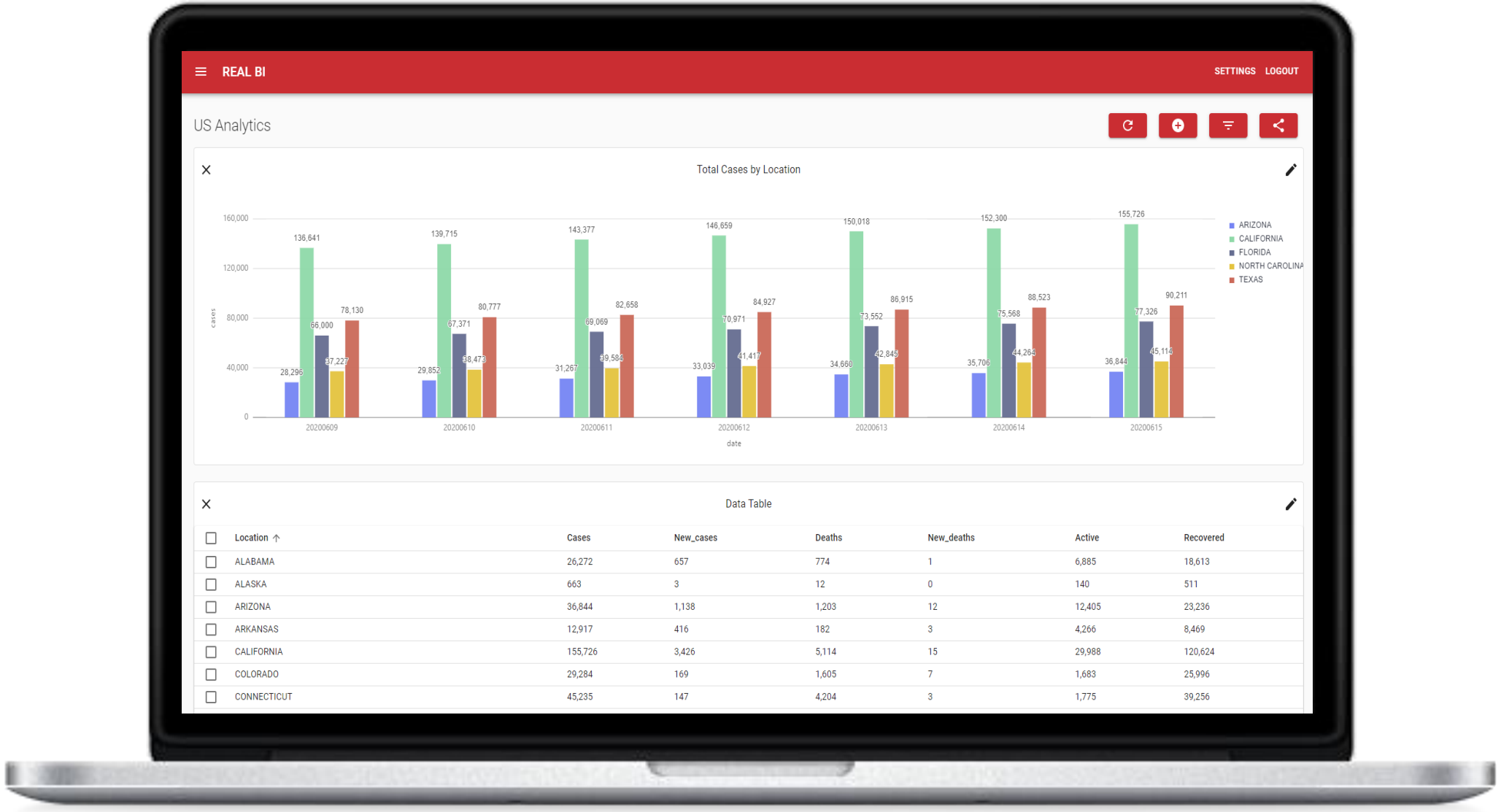
Métodos ensemble

Random Forest

Gradient Boosted Forest

Gradient Boosted Trees

Ferramentas de Consulta e Visualização

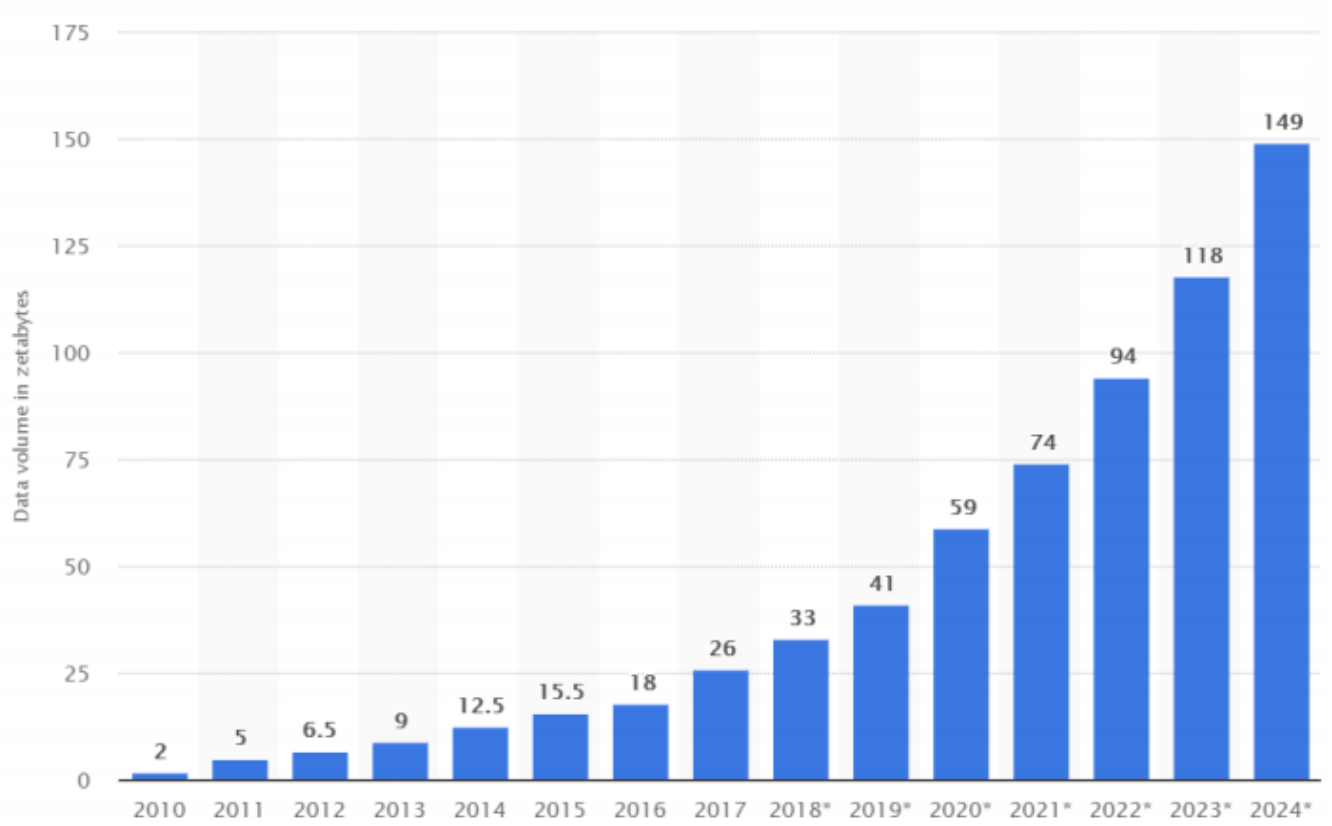


Location	Cases	New_cases	Deaths	New_deaths	Active	Recovered
ALABAMA	26,272	657	774	1	6,885	18,613
ALASKA	663	3	12	0	140	511
ARIZONA	36,844	1,138	1,203	12	12,405	23,236
ARKANSAS	12,917	416	182	3	4,266	8,469
CALIFORNIA	155,726	3,426	5,114	15	29,988	120,624
COLORADO	29,284	169	1,605	7	1,683	25,996
CONNECTICUT	45,235	147	4,204	3	1,775	39,256

Fundação Vanzolini

E o futuro?

Para o alto e avante!

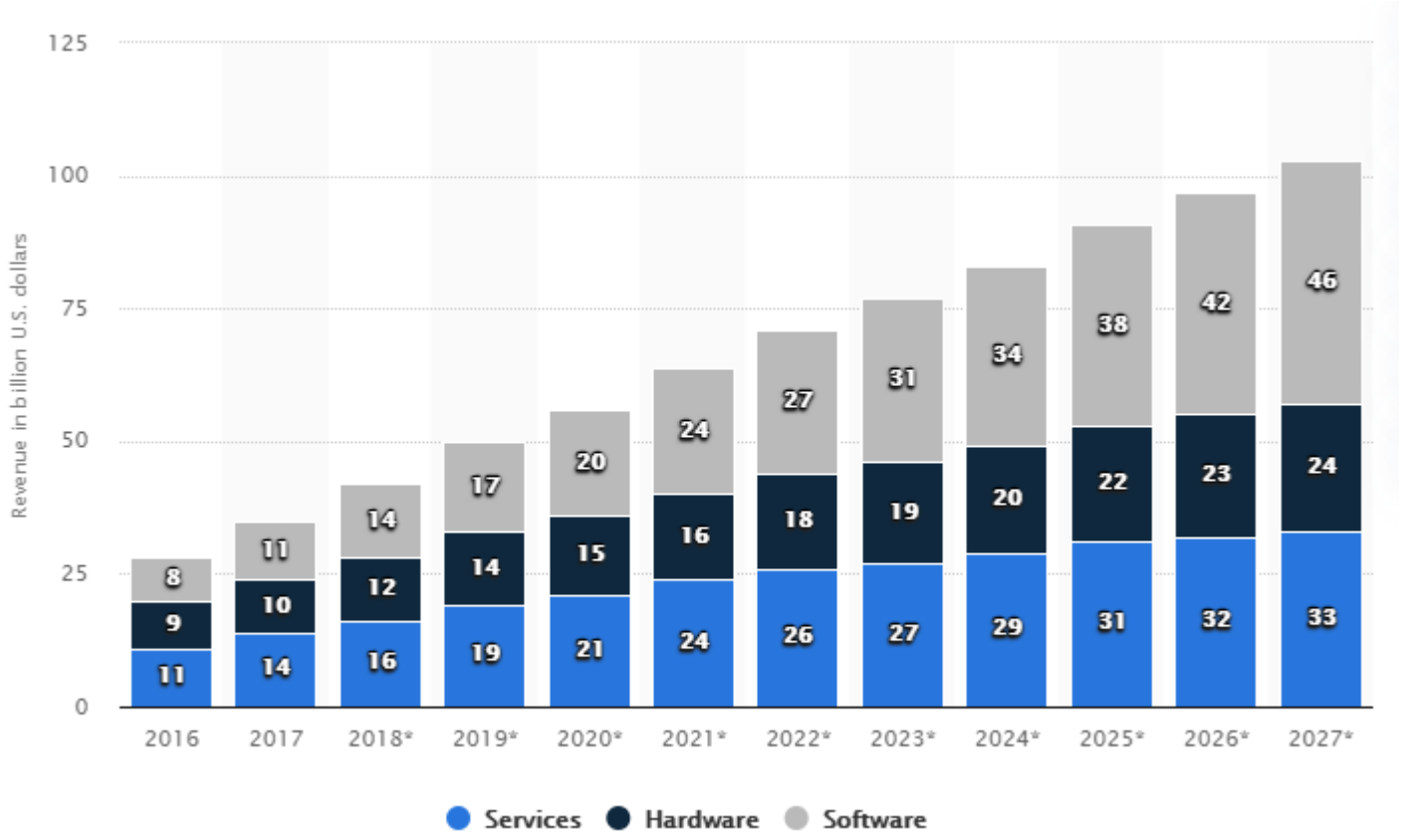


https://www.statista.com/statistics/871513/worldwide-data-created/

Fundação Vanzolini

Recursos humanos

- Demanda profissional: engenharia de dados, ciência de dados, regulamentação...
- Multidisciplinar
- Migração / mudança de carreira



https://www.statista.com/statistics/301566/big-data-factory-revenue-by-type/

Fundação Vanzolini

Iniciando a extração de dados

Dataset

##	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

➢ **Persons (~841 k)**

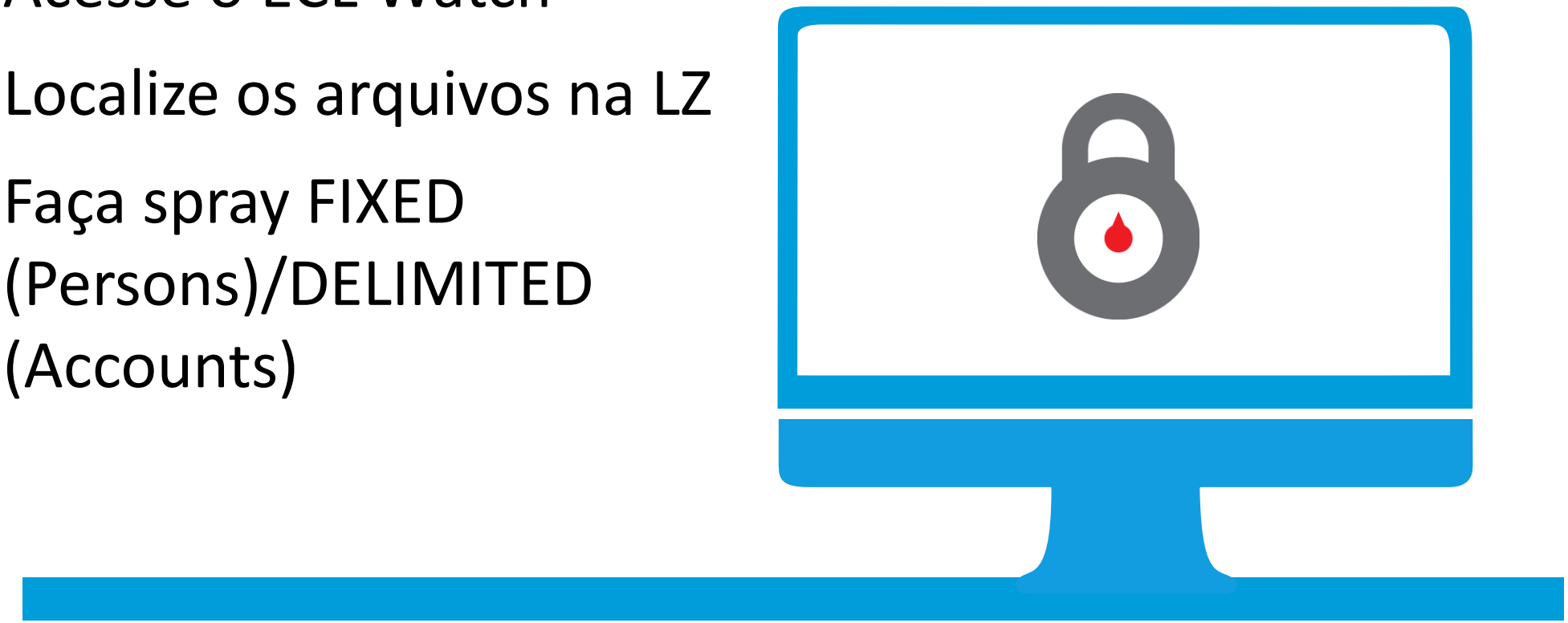
##	personid	reportdate	industrycode	member	opendate	tradetype	traderate	narr1	narr2	highcredit	balance	terms	termtyper	accountnumber	lastactivitydate	late30day
1	9108218085885411565	20000101	ZZ	2121728	19990901	O	9	101	0	9999999	438	0	0	SEARS 9999 -99999999	19990901	0
2	2454818069645923666	20001201	FZ	28868655	20000801	I	0	88	0	1313	1329	0	0	27489999999999999	20001101	0
3	2454818069645923666	20001201	FZ	28868655	19990801	I	0	88	0	533	1145	0	0	27489999999999999	20001101	0
4	2454818069645923666	20001201	FZ	28868655	19990801	I	0	88	0	780	1560	0	0	27489999999999999	20001101	0
5	2454818069645923666	20001201	FZ	28868655	19990201	I	0	88	0	331	371	0	0	27489999999999999	20001101	0
6	2454818069645923666	20001201	FZ	28868655	19990201	I	0	88	0	982	982	0	0	27489999999999999	20001101	0
7	2454818069645923666	20001101	FC	157905	19981201	I	1	214	0	2800	0	165	2	7217999999	19990901	0
8	2454818069645923666	20001001	BB	594148	19990901	I	1	214	0	23089	19477	458	2	20019999999	20001001	0
9	2454818069645923666	20000701	FZ	11564684	19990801	I	0	123	248	1560	0	120	1	27489999999	20000601	0
10	2454818069645923666	20000701	FZ	11564684	19990201	I	*	123	248	1964	0	120	1	27489999999		0
11	2454818069645923666	20000601	FZ	11564684	19990801	I	0	88	0	1065	9999999	120	1	27489999999	20000601	0
12	2454818069645923666	20000501	FZ	11564684	19990201	I	0	88	0	0	9999999	120	1	274899999B	20000501	0
13	2454818069645923666	20000501	FZ	11564684	19990201	I	0	88	0	0	9999999	120	1	274899999A	20000501	0
14	6512705660523829539	20001101	ON	8302668	20000701	R	1	2	233	5000	2080	38	2	5491999999999999	20001101	0
15	9193989543268753887	20001201	FS	363906	19980301	I	1	127	150	70859	66424	638	2	1195999	20001201	0

➢ **Accounts (~8.4 M)**

Exercício prático:

Exercícios 1 e 2 – Spray de dados

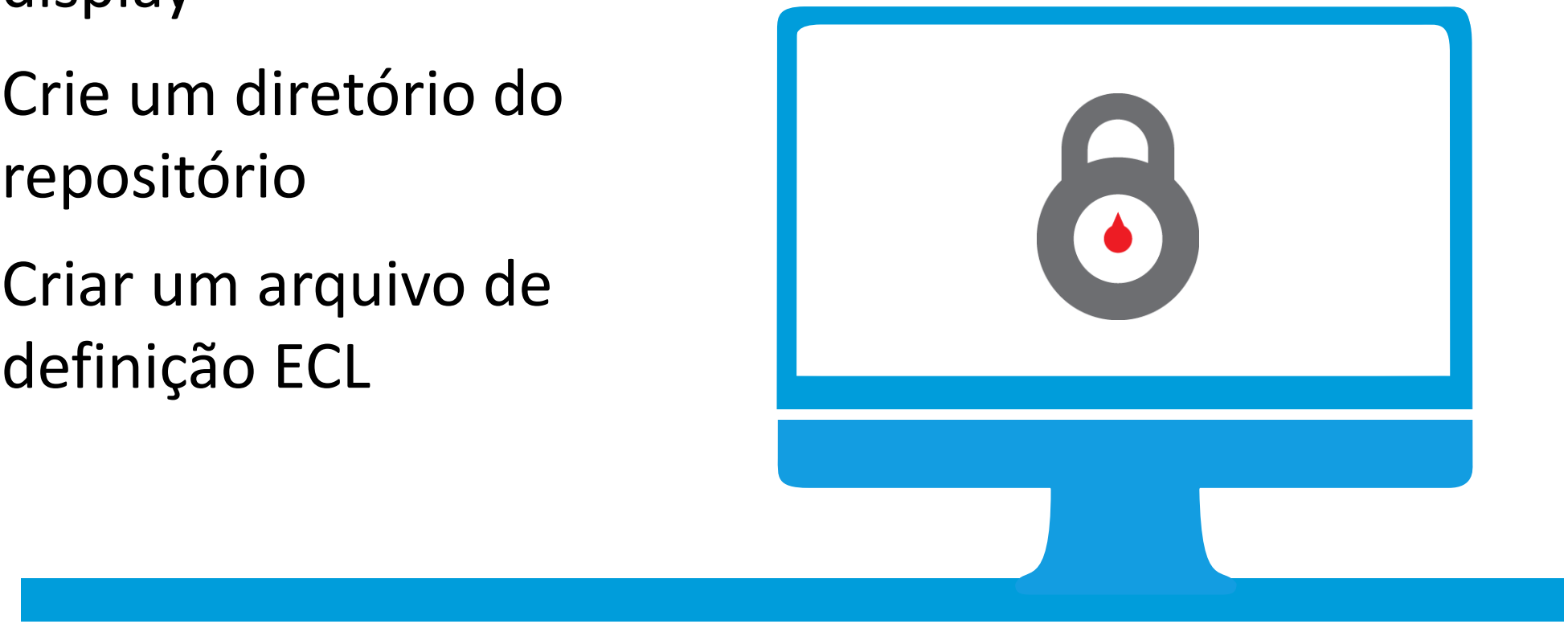
- Acesse o ECL Watch
- Localize os arquivos na LZ
- Faça spray FIXED (Persons)/DELIMITED (Accounts)



Exercício prático:

Exercício 3 – Primeiros passos

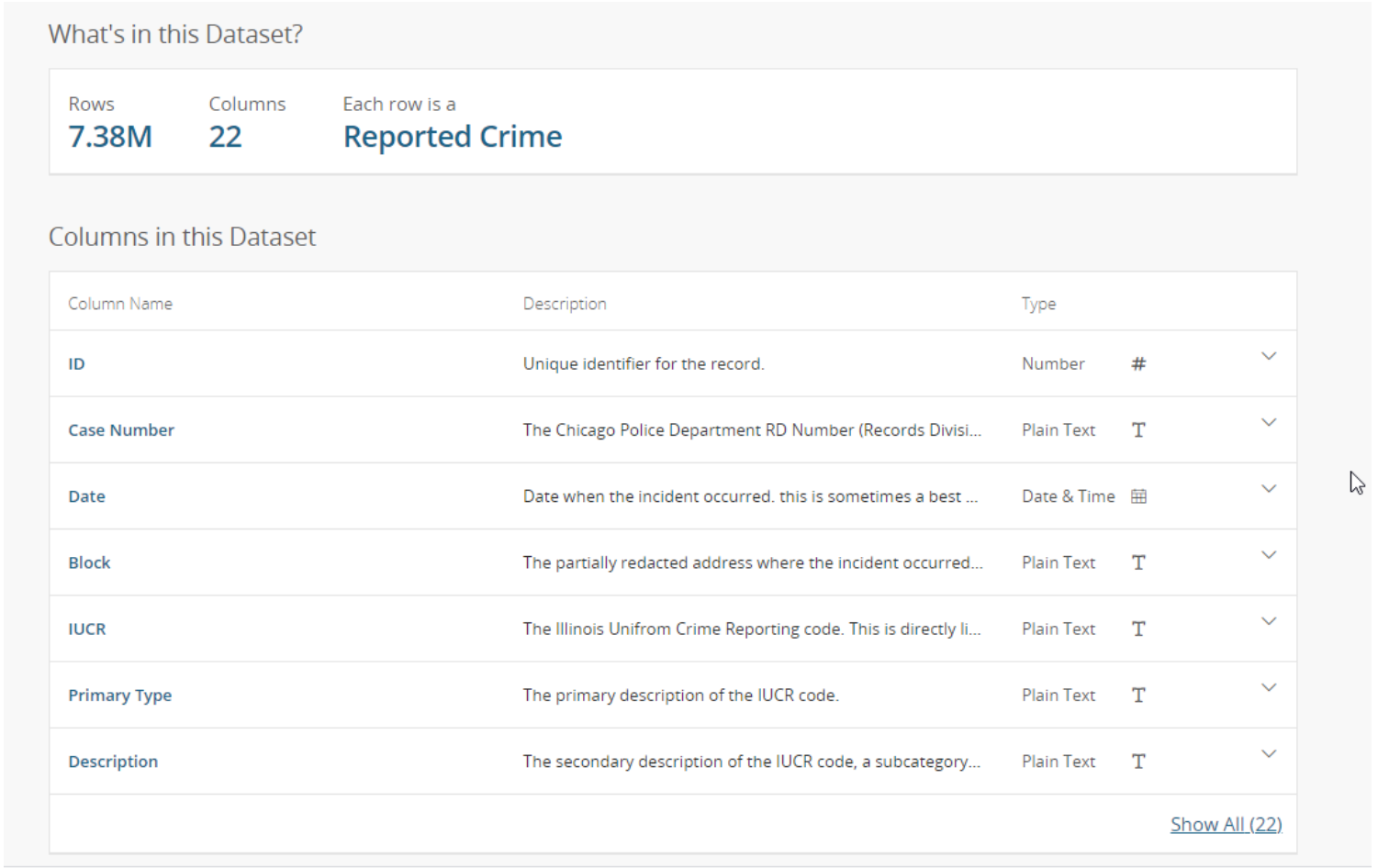
- Acessando – Opções do display
- Crie um diretório do repositório
- Criar um arquivo de definição ECL



Desafio: Chicago Crimes

Desafio:

- Fazer spray do dataset de crimes da polícia de Chicago (data.cityofchicago.org)



What's in this Dataset?

Rows	Columns	Each row is a
7.38M	22	Reported Crime

Columns in this Dataset

Column Name	Description	Type
ID	Unique identifier for the record.	Number #
Case Number	The Chicago Police Department RD Number (Records Divisi...	Plain Text T
Date	Date when the incident occurred. this is sometimes a best ...	Date & Time
Block	The partially redacted address where the incident occurred...	Plain Text T
IUCR	The Illinois Unifrom Crime Reporting code. This is directly li...	Plain Text T
Primary Type	The primary description of the IUCR code.	Plain Text T
Description	The secondary description of the IUCR code, a subcategory...	Plain Text T

Show All (22)

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!!!