Relatorio de criação do ETL

Daniel Nonato Fernandes PUC Minas

Conteúdo

I	Introdução	1							
II	Definição do Problema	1							
Ш	Proposta de soluçãoIII-APrimeira soluçãoIII-BSegunda solução	1 1 1							
IV	Criteiro de aceitação	1							
V	Metodo de pesquisa								
VI	Analise e Interpretação								
VII VIII	ETL VII-A Carga Stage VII-B Carga das Dimensões VII-C Carga da Fato Lista de Dimensões VIII-A DIM STATUS	1 1 2 2 2							
	VIII-B DIM EMPLOYER VIII-C DIM SOCIAL VIII-D DIM JOB VIII-E DIM FTP VIII-F DIM PREVAILING VIII-G DIM CITY VIII-H DIM STATE VIII-I DIM STATE FULL VIII-J DIM YEAR	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2							
IX	Fato	3							
X	Automatização do processo de ETL	3							
ΧI	GIT	3							
	Lista de Figuras								
1 2 3 4 5	Dados brutos . Projeto ETL . STAGE . Extração para Dimensão . Extração para Fato .								

LISTA DE TABELAS

Relatorio de criação do ETL

Resumo—O projeto busca facilitar a retirada de vistos e criar padrões de tomada de decisão para aprovação de vistos para os EUA. A busca de dados massivos com o resultado de solicitações de visto do ano de 2015 até o ano de 2019, para ser criado um modelo preditivo para a retirada de vistos, tanto para o solicitante tanto para o solicitado.

I. Introdução

A proposta pretende facilitar a tomada de decisão para aprovação e solicitação de vistos para os Estados Unidos da America. A base de dados criada para o mesmo deverá ser insumo de informação para Machine Learning, sendo assim facilitar ou até mesmo pre aprovar solicitações de vistos para os Estados Unidos da America

II. DEFINIÇÃO DO PROBLEMA

O problema apresentado é a falta pré analise do perfil de um cidadão para ter um visto para os Estados Unidos da America.

CASE_STATUS	EMPLOYER NAME	SOC_NAME	500,0006		FULL TIME POSITION	PREVAILING_WASE	WORKSITE_CITY
CERTIFIED	HAGIC CIPCLE FILHS INTERNATIONAL LLC	FLM AND VIDGO EDITORS	27-4002	VIDEO EDITOR	Y	29332	ALRESSN.
WITHORAWN	SAPHRE SOLUTIONS INC	NETWORK AND COMPUTER SYSTEMS ADMINISTRATORS	15-1142	SYSTEMS ADMINISTRATOR	Y	60630	SETHPAGE
CERTIFIED	LIMMS LLC	FINANCIAL ANALYSTS	13-2051	FININCIAL ANKLIST	N	49100.0	CARBONDALE
CENTIFIED	INSOVATIVE FASHION/SPOUP	FASHION DESIGNERS	27/3022	PRODUCT DEVELOPMENT SPECIALIST	N	41849.5	NEW YORK
CERTIFIED	AMC INC	COMPUTER SYSTEMS ANALYSTS	15-1121	PROGRAMMER ANALYST	Y	42431	BUFFALO
CERTIFIED-WITHORANIN	INTONE NETWORKS INC.	DATABASE ADMINISTRATORS	15-1141	TERROATA DATABASE ADMINISTRATOR	N	43171	WARRENDALE
CERTIFIED	AMCINC	COMPUTER SYSTEMS ANALYSTS	15/1121	PROGRAMMER ANALYST	Y	47401	BUFFALO
CERTIFIED WITHORANIN	WEST FOURTH & PERRY LTD.	INTERIOR DESIGNERS	27/1025	INTERIOR DESIGNER	N	42764.0	NEW YORK
CERTIFIED WITHORAWN	CT CORPORATION SYSTEM	ACCOUNTANTS AND AUDITORS	13-2011	BUSINESS PROCESS ANKLYST II	Y	74110	NEW YORK
WITHORNWIN	CHICAGO BRIDGE & IRON CONPANY (DELAWARE)	ELECTRICAL ENGINEERS	17/2071.00	SENOR ELECTRICAL ENSINEER I	Y	77738	PLANFIELD
WITHOMAN	NCO HOLDINGS LLC	OPERATIONS RESEARCH ANALYSTS	15-2000	FINANCIAL ENGINEER - MODELING	Y	55002	CHICAGO
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPUTER PROGRAMMERS	15-1131	DEVELOPER	Y	65740	SCHENGCTADY
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPLITER PROGRAMMERS	15/1120	DEVELOPER	Y	66602	COLLEGEWILLE
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPLITER SYSTEMS ANALYSTS	15-1121	FINANCE & CONTROLLING SYSTEMS ANALYST	Y	66579	WOLFFE
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPLITER PROGRAMMERS	15-1131	DEVELOPER	Y	61509	WESTERVILLE
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	WEB DEVELOPERS	15 1134	DEVELOPER USER INTERPACE	Y	54330	RICHERN
CENTRED	TATA CONSULTANCY SERVICES LIMITED	WEB DEVELOPERS	15-1134	DEVELOPER USER INTERPACE	Y	54330	RECEPT
CERTIFIED	HOSLADREY LLP	ACCOUNTANTS AND AUDITORS	13-2011	ASSURANCE HANAGER	Y	74464	CHICAGO
CERTIFIED-WITHORANIN	PROGRESSIVE STEP CORP.	PHYSICAL THERAPISTS	29-1122	PHYSICAL THERAPIST	Y	64147.2	WEST READING
CERTIFIED	SAPHRE SOLUTIONS INC	NETWORK AND COMPUTER SYSTEMS ADMINISTRATORS	15-1142	SYSTEMS ADMINISTRATOR	Y	69632	SETHPAGE
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPUTER SYSTEMS ANALYSTS	15-1121	ANALYST	Y	61339	WARSKW
CERTIFIED WITHDRAWN	JPMORSAN CHASE & CO.	FININCIAL AMILITATS	13-2051.00	ANALYST	Y	61942	NEW YORK
CENTRED	TATA CONSULTANCY SERVICES LIMITED	COMPUTER PROGRAMMERS	15-1131	DEVELOPER	Y	52754	INFERVILLE
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	WEB DEVELOPERS	151134	DEVELOPER USER INTERFACE	Y	66862	DENSALEM
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPLITER PROGRAMMERS	15/1120	DEVELOPER	Y	66602	PHILADELPHIA
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPLITER PROGRAMMERS	15-1121	DEVELOPER	Y	66602	MACHEN
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPLITER PROGRAMMERS	15-1131	DEVELOPER	Y	54764	COLUMBUS
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPUTER PROSPAMMERS	15/1130	COMPUTER PROGRAMMER	Y	56347	8L00MMS70%
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPUTER PROGRAMMENS	15-1131	DEVELOPER	Y	61505	COLUMBUS
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPUTER PROGRAMMERS	15-1131	DEVELOPER	Y	62754	CHICAGO
CERTIFIED	BEHAVIOR ANALYST	REHABILITATION COUNSELORS	21-1015	BEHAVOR ANKLYST	Y	27010	HURPHYSBORO
DENED	DELL MARKETING LP	COMPLITER SUPPORT SPECIFILISTS	15-1041.00	NETWORK SUPPORT ANALYST	Y	47070	NEW YORK
CERTIFIED	TATA CONSULTANCY SERVICES LIMITED	COMPLITER PROGRAMMERS	15-1131	DEVELOPER	Y	65416	SPOOKLYN PA.
CERTIFIED WITHORAWIN	ALTO SHAMING.	MECHANICAL ENGINEERS	17/2141	PROJECT ENGINEER	Y	74214	MENOMONEE F
CENTRED	TATA CONSULTANCY SERVICES LIMITED	WEB DEVELOPERS	15-1134	DEVELOPER USER INTERPACE	Y	68112	UNCOUNSHIRE

Figura 1. Dados brutos

III. PROPOSTA DE SOLUÇÃO

A proposta de criar um meio buscar solicitações anteriores de visto, e verificar um padrão de acordo com o perfil de cada solicitação. Essa base com os padrões é para servir de auxilio na tomada de decisão das proximas solicitações de visto para os EUA.

A. Primeira solução

Criar uma base de dados com todas as solicitações de vistos anteriores.

B. Segunda solução

Criar um sistema de apoio a decisão de acordo com o solicitante do visto..

IV. CRITEIRO DE ACEITAÇÃO

A solução possui um criterio de ser um apoiador de decisão, por esse motivo o mesmo é baseado em probabilidade.

V. METODO DE PESQUISA

1

Para buscar os dados para ser criado o projeto, foi feito uma busca por um data set com os dados das solicitações de vistos para os EUA, e criado uma base relacional para a mesma.

Foi criado um crowler para buscar os dados de solicitações de visto para os EUA. *Do not include your findings in this section.*

VI. ANALISE E INTERPRETAÇÃO

Os dados levantados indica que a solução é econômica, tecnicamente viavel e acessivel, por ser um sistema baseado em historico de solicitações e em banco de dados relacionais.

VII. ETL

Foi criado um ETL para buscar os dados da massa bruta de dados e colocados em uma tabela STAGE e apos isso padronizar em um banco relacional todos os dados.

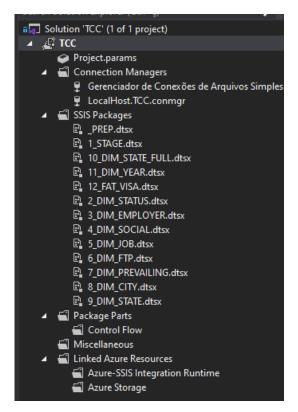


Figura 2. Projeto ETL

A. Carga Stage

Foi feita uma carga para uma tabela STAGE com os dados brutos.



Figura 3. STAGE

B. Carga das Dimensões

Foi feita uma extração da tabela STAGE para limpeza dos dados e gerar as dimensões.

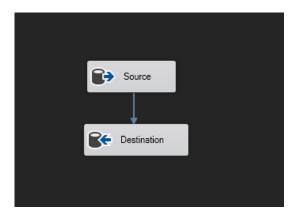


Figura 4. Extração para Dimensão

C. Carga da Fato

Foi feita uma extração da tabela STAGE para limpeza dos dados e gerar a tabela Fato.

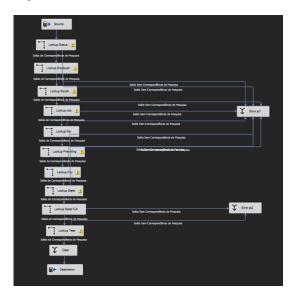


Figura 5. Extração para Fato

VIII. LISTA DE DIMENSÕES

As tabelas são compostas por:

- PK (Primary Key) = Chave primaria, o indentificador da tabela, está em auto incremente.
- DESC (Description) = Descrição do item da dimensão.
- START DATE = Data em que o dado foi inserido na tabela
- END DATE = Data em que o dado foi excluido logicamente da tabela

A. DIM STATUS

Tabela onde se encontra a dimensão de status da solicitação do visto.

B. DIM EMPLOYER

Tabela onde se encontra a dimensão do empregador do solicitante do visto.

C. DIM SOCIAL

Tabela onde se encontra a dimensão de profissão do solicitante do visto.

D. DIM JOB

Tabela onde se encontra a dimensão de cargo do solicitante do visto.

E. DIM FTP

Tabela onde se encontra a dimensão de tipo de carga horarario do solicitante do visto.

F. DIM PREVAILING

Tabela onde se encontra a dimensão de salario do solicitante do visto.

G. DIM CITY

Tabela onde se encontra a dimensão de cidade da solicitação do visto.

H. DIM STATE

Tabela onde se encontra a dimensão do estado da solicitação do visto.

I. DIM STATE FULL

Tabela onde se encontra a dimensão do estado completo da solicitação do visto.

J. DIM YEAR

Tabela onde se encontra a dimensão do ano da solicitação do visto.

IX. FATO

A tabela é compostas por:

- FK (Foreign Key) = Chave estrangeira, para todas as dimensões.
- CASE NUMBER = Numero de solicitação de visto.
- START DATE = Data em que o dado foi inserido na tabela
- END DATE = Data em que o dado foi excluido logicamente da tabela

X. AUTOMATIZAÇÃO DO PROCESSO DE ETL

É gerado um arquivo com o padrão ISPAC onde o mesmo é implantado no banco de dados, e criado uma job para execução do mesmo.

XI. GIT

https://github.com/eazydans/TCC.git