DATA X: FUZZY JOINS

Project 38, Group 2

MEET THE TEAM

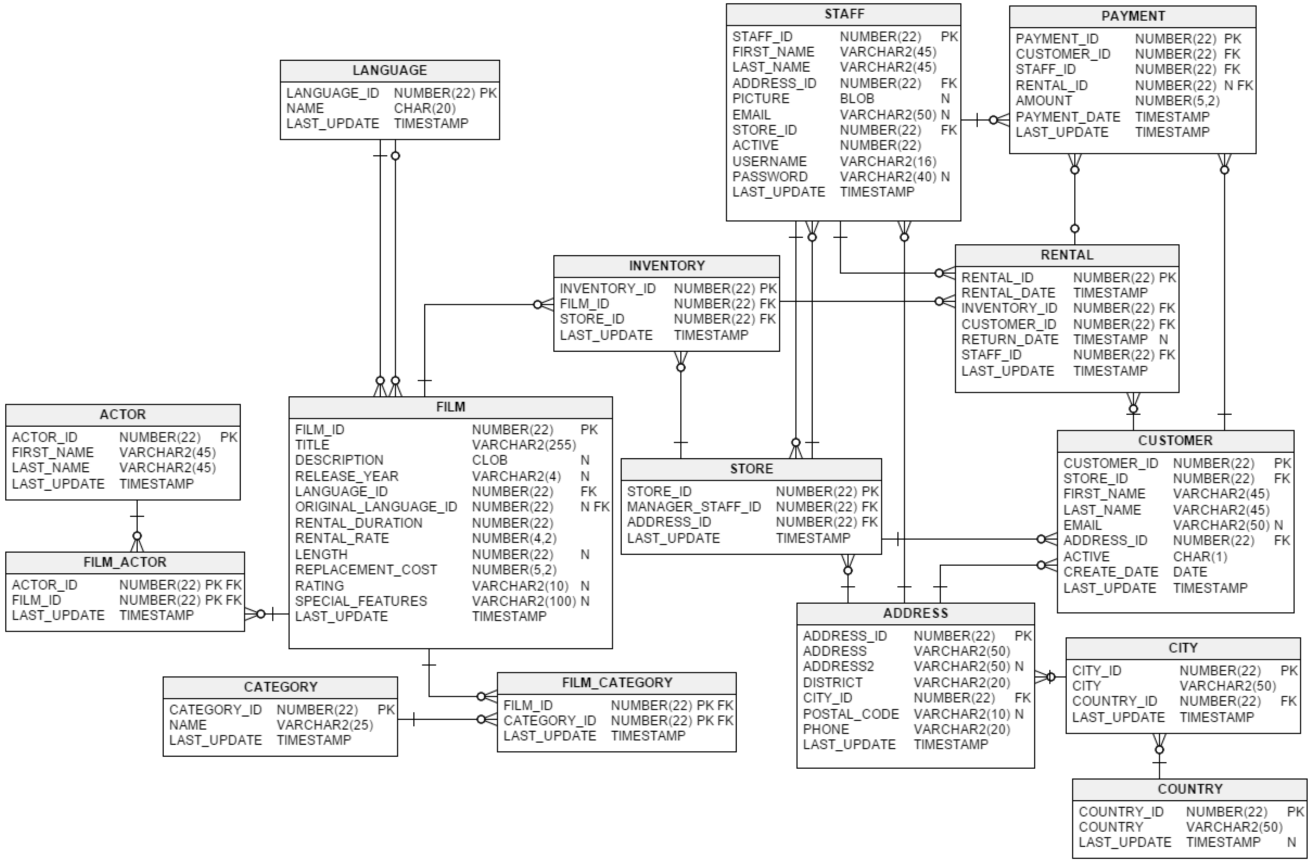
Jake Mainwaring

Chase Smith

Cyril Tamraz

Kochise Bennett

HIGH LEVEL DESCRIPTION



STAFF

STAFF_ID	NUMBER(22)	PK
FIRST_NAME	VARCHAR2(45)	
LAST_NAME	VARCHAR2(45)	
ADDRESS_ID	NUMBER(22)	FK
PICTURE	BLOB	N
EMAIL	VARCHAR2(50)	N
STORE_ID	NUMBER(22)	FK
ACTIVE	NUMBER(22)	
USERNAME	VARCHAR2(16)	
PASSWORD	VARCHAR2(40)	N
LAST_UPDATE	TIMESTAMP	

PAYMENT

PAYMENT_ID	NUMBER(22)	PK
CUSTOMER_ID	NUMBER(22)	FK
STAFF_ID	NUMBER(22)	FK
RENTAL_ID	NUMBER(22)	N FK
AMOUNT	NUMBER(5,2)	
PAYMENT_DATE	TIMESTAMP	
LAST_UPDATE	TIMESTAMP	

LANGUAGE

LANGUAGE_ID	NUMBER(22)	PK
NAME	CHAR(20)	
LAST_UPDATE	TIMESTAMP	

INVENTORY

INVENTORY_ID	NUMBER(22)	PK
FILM_ID	NUMBER(22)	FK
STORE_ID	NUMBER(22)	FK
LAST_UPDATE	TIMESTAMP	

RENTAL

RENTAL_ID	NUMBER(22)	PK
RENTAL_DATE	TIMESTAMP	
INVENTORY_ID	NUMBER(22)	FK
CUSTOMER_ID	NUMBER(22)	FK
RETURN_DATE	TIMESTAMP	N
STAFF_ID	NUMBER(22)	FK
LAST_UPDATE	TIMESTAMP	

ACTOR

ACTOR_ID	NUMBER(22)	PK
FIRST_NAME	VARCHAR2(45)	
LAST_NAME	VARCHAR2(45)	
LAST_UPDATE	TIMESTAMP	

FILM

FILM_ID	NUMBER(22)	PK
TITLE	VARCHAR2(255)	
DESCRIPTION	CLOB	N
RELEASE_YEAR	VARCHAR2(4)	N
LANGUAGE_ID	NUMBER(22)	FK
ORIGINAL_LANGUAGE_ID	NUMBER(22)	N FK
RENTAL_DURATION	NUMBER(22)	
RENTAL_RATE	NUMBER(4,2)	
LENGTH	NUMBER(22)	N
REPLACEMENT_COST	NUMBER(5,2)	
RATING	VARCHAR2(10)	N
SPECIAL_FEATURES	VARCHAR2(100)	N
LAST_UPDATE	TIMESTAMP	

STORE

STORE_ID	NUMBER(22)	PK
MANAGER_STAFF_ID	NUMBER(22)	FK
ADDRESS_ID	NUMBER(22)	FK
LAST_UPDATE	TIMESTAMP	

CUSTOMER

CUSTOMER_ID	NUMBER(22)	PK
STORE_ID	NUMBER(22)	FK
FIRST_NAME	VARCHAR2(45)	
LAST_NAME	VARCHAR2(45)	
EMAIL	VARCHAR2(50)	N
ADDRESS_ID	NUMBER(22)	FK
ACTIVE	CHAR(1)	
CREATE_DATE	DATE	
LAST_UPDATE	TIMESTAMP	

FILM_ACTOR

ACTOR_ID	NUMBER(22)	PK FK
FILM_ID	NUMBER(22)	PK FK
LAST_UPDATE	TIMESTAMP	

ADDRESS

ADDRESS_ID	NUMBER(22)	PK
ADDRESS	VARCHAR2(50)	
ADDRESS2	VARCHAR2(50)	N
DISTRICT	VARCHAR2(20)	
CITY_ID	NUMBER(22)	FK
POSTAL_CODE	VARCHAR2(10)	N
PHONE	VARCHAR2(20)	
LAST_UPDATE	TIMESTAMP	

CITY

CITY_ID	NUMBER(22)	PK
CITY	VARCHAR2(50)	
COUNTRY_ID	NUMBER(22)	FK
LAST_UPDATE	TIMESTAMP	

CATEGORY

CATEGORY_ID	NUMBER(22)	PK
NAME	VARCHAR2(25)	
LAST_UPDATE	TIMESTAMP	

FILM_CATEGORY

FILM_ID	NUMBER(22)	PK FK
CATEGORY_ID	NUMBER(22)	PK FK
LAST_UPDATE	TIMESTAMP	

COUNTRY

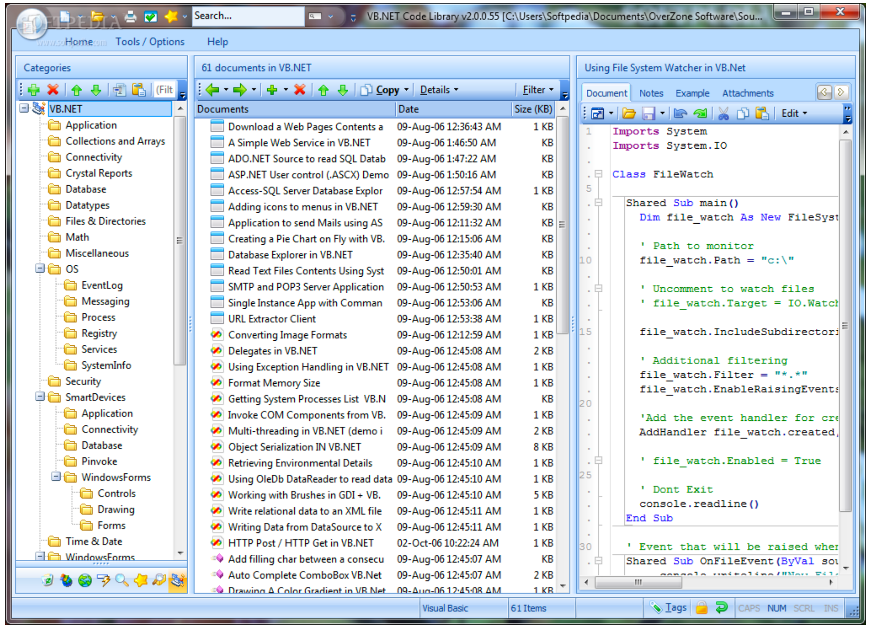
COUNTRY_ID	NUMBER(22)	PK
COUNTRY	VARCHAR2(50)	
LAST_UPDATE	TIMESTAMP	N

Current problem – companies rely on a large number of databases to store information, but the majority of time is spent cleaning and aggregating rather than conducting analysis

Our solution – create a tool that helps users standardize and aggregate data based on fuzzy comparisons. This could include minor differences in column name (e.g. “User_ID” vs. “UserID”) specific values (e.g. “123 MLK Jr. Way” vs. “123 Martin Luther King Jr. Way”), or other comparisons

USER PERSPECTIVE

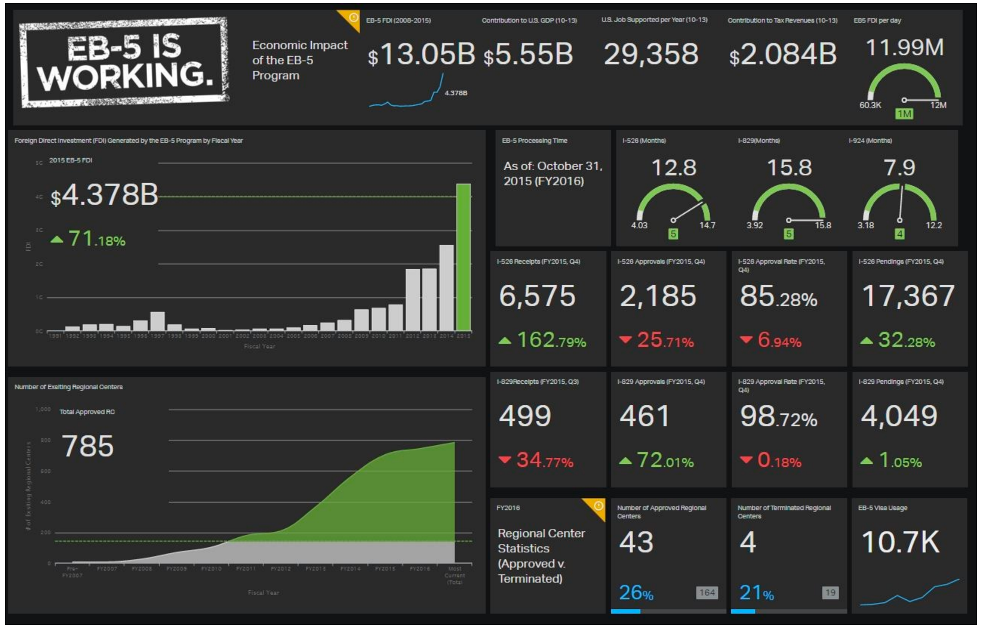
Option 1 (library):



Documents	Date	Size (KB)
Download a Web Pages Contents a	09-Aug-06 12:36:43 AM	1 KB
A Simple Web Service in VB.NET	09-Aug-06 1:46:50 AM	KB
ADO.NET Source to read SQL Datab	09-Aug-06 1:47:22 AM	KB
ASP.NET User control (.ASCX) Demo	09-Aug-06 1:50:16 AM	KB
Access-SQL Server Database Explor	09-Aug-06 12:57:54 AM	1 KB
Adding icons to menus in VB.NET	09-Aug-06 12:59:30 AM	KB
Application to send Mails using AS	09-Aug-06 12:11:32 AM	KB
Creating a Pie Chart on Fly with VB.	09-Aug-06 12:15:06 AM	KB
Database Explorer in VB.NET	09-Aug-06 12:35:40 AM	KB
Read Text Files Contents Using Syst	09-Aug-06 12:50:01 AM	KB
SMTP and POP3 Server Application	09-Aug-06 12:50:53 AM	1 KB
Single Instance App with Comman	09-Aug-06 12:53:06 AM	KB
URL Extractor Client	09-Aug-06 12:53:38 AM	1 KB
Converting Image Formats	09-Aug-06 12:12:59 AM	1 KB
Delegates in VB.NET	09-Aug-06 12:45:08 AM	2 KB
Using Exception Handling in VB.NET	09-Aug-06 12:45:08 AM	1 KB
Format Memory Size	09-Aug-06 12:45:08 AM	1 KB
Getting System Processes List VB.N	09-Aug-06 12:45:08 AM	KB
Invoke COM Components from VB.	09-Aug-06 12:45:09 AM	1 KB
Multi-threading in VB.NET (demo i	09-Aug-06 12:45:09 AM	2 KB
Object Serialization IN VB.NET	09-Aug-06 12:45:09 AM	8 KB
Retrieving Environmental Details	09-Aug-06 12:45:10 AM	1 KB
Using OleDb DataReader to read data	09-Aug-06 12:45:10 AM	1 KB
Working with Brushes in GDI + VB.	09-Aug-06 12:45:10 AM	5 KB
Write relational data to an XML file	09-Aug-06 12:45:11 AM	1 KB
Writing Data from DataSource to X	09-Aug-06 12:45:11 AM	1 KB
HTTP Post / HTTP Get in VB.NET	02-Oct-06 10:22:24 AM	1 KB
Add filling char between a consecu	09-Aug-06 12:45:07 AM	KB
Auto Complete ComboBox VB.Net	09-Aug-06 12:45:07 AM	2 KB
Drawing A Color Gradient in VB.Net	09-Aug-06 12:45:08 AM	1 KB

```
Imports System
Imports System.IO

Class FileWatch

  Shared Sub main()
    Dim file_watch As New FileSyst

    ' Path to monitor
    file_watch.Path = "c:\"

    ' Uncomment to watch files
    ' file_watch.Target = IO.Watch

    file_watch.IncludeSubdirector:

    ' Additional filtering
    file_watch.Filter = "*.*"
    file_watch.EnableRaisingEvents

    'Add the event handler for cre
    AddHandler file_watch.created,

    ' file_watch.Enabled = True

    ' Dont Exit
    console.readline()
  End Sub

  ' Event that will be raised wher
  Shared Sub OnFileEvent(ByVal sou
```

Option 2 (dashboard):



EB-5 IS WORKING.
Economic Impact of the EB-5 Program
EB-5 FDI (2008-2015)
$13.05B
4.378B
Contribution to U.S. GDP (10-13)
$5.55B
U.S. Job Supported per Year (10-13)
29,358
Contribution to Tax Revenues (10-13)
$2.084B
EB5 FDI per day
11.99M
60.3K
12M
1M
Foreign Direct Investment (FDI) Generated by the EB-5 Program by Fiscal Year
2015 EB-5 FDI
$4.378B
▲ 71.18%
FDI
5C
4C
3C
2C
1C
0C
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Fiscal Year
EB-5 Processing Time
As of: October 31, 2015 (FY2016)
I-526 (Months)
12.8
4.03
5
14.7
I-829(Months)
15.8
3.92
5
15.8
I-924 (Months)
7.9
3.18
4
12.2
I-526 Receipts (FY2015, Q4)
6,575
▲ 162.79%
I-526 Approvals (FY2015, Q4)
2,185
▼ 25.71%
I-526 Approval Rate (FY2015, Q4)
85.28%
▼ 6.94%
I-526 Pendings (FY2015, Q4)
17,367
▲ 32.28%
Number of Exsiting Regional Centers
Total Approved RC
785
# of Exsiting Regional Centers
1,000
800
600
400
200
0
Pre-FY2007
FY2007
FY2008
FY2009
FY2010
FY2011
FY2012
FY2013
FY2014
FY2015
FY2016
Most Current (Total)
Fiscal Year
I-829Receipts (FY2015, Q3)
499
▼ 34.77%
I-829 Approvals (FY2015, Q4)
461
▲ 72.01%
I-829 Approval Rate (FY2015, Q4)
98.72%
▼ 0.18%
I-829 Pendings (FY2015, Q4)
4,049
▲ 1.05%
FY2016
Regional Center Statistics (Approved v. Terminated)
Number of Approved Regional Centers
43
26%
164
Number of Terminated Regional Centers
4
21%
19
EB-5 Visa Usage
10.7K

TECHNICAL COMPONENTS

- Development of fuzzy join/inexact matching criteria:
  - Column name comparison
  - String values within columns
  - Dictionary based on standardized column names, where users can input “User-id”, “User-ID”, etc. and it maps to the standard name for that column
- Joining data tables on cleaned column names and values
- UI-database connector

SAMPLE ARCHITECTURE/DATA MODEL

Input

1. Collection of two or more data tables

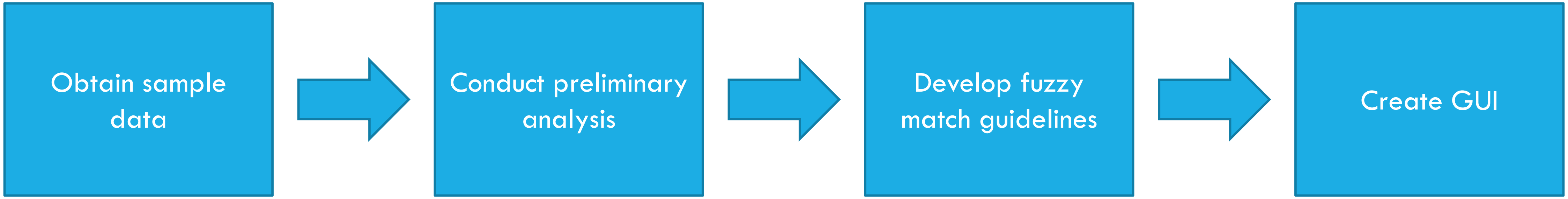
Process

1. Fuzzy match foreign keys with corresponding primary key column names
2. Fuzzy match column names between primary key tables with overlapping column names
3. Fuzzy match specific values based on similarity metrics

Output

1. Updated data tables with dictionary of column names

NEXT STEPS



Obtain sample data

Conduct preliminary analysis

Develop fuzzy match guidelines

Create GUI