Data Management Workgroup

BURT.WALSH@AST.MYFLORIDA.COM

APRIL 6, 2018

Overview

Progress on Enterprise Data Inventory (meetings, contacts)

Azure Data Catalog (First step agency sharing)

Work group efforts, agency efforts, tool user groups

Leverage vendors where appropriate

Data Quality Management

Next topics

Data Quality Management

BURT.WALSH@AST.MYFLORIDA.COM

Data Quality Management

The planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meets the needs of data consumers.

(Mosley, M., & Brackett, M. (2015). *The DAMA guide to the data management body of knowledge (DAMA-DMBOK guide), second edition*. Bradley Beach, N.J.: Technics Publications.)

Data Quality

Data has quality to an end consumer if it is accurate, useable, timely, consistent; in short, if it is fit for purpose.

Two other factors affecting data quality are **believability** and **interpretability**. [1]

- Believability reflects how much the data is trusted by users
- interpretability reflects how easy the data is understood

(Many companies have challenges with data quality: https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards -- 97% threshold)

[1] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

DATA LAKE

- Structured Data
- Unstructured Data
- OLAP

Extract
Transform
Load
(ETL)

Create
Read
Update

N-Tier App

N-Tier App

BI / Reporting App

Create
Read
Update

Is what I'm looking at based on data that is accurate, timely, consistent, usable, and "fit for purpose?"

End Consumer

Data Quality

Data quality standards need to be measurable.

One way to do this is to leverage data dimensions and thresholds – examples include:

- Completeness
- Uniqueness (non-duplicative)
- Timeliness (reality based)
- Validity (syntax value range)
- Accuracy (precision, format)
- Consistency (same representation/same value)

Data quality rules are a function of data dimensions, business rules/organizational needs, and impacts of non-compliance.

Context of evaluation of the data is the need of the business.

Metrics can be and should be evaluated over time (trends).

(Primary Dimensions for Data Quality Assessment with examples and descriptive format: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf)

(CDC dimensions with descriptive format: https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf)

Business Assets Driven

Business Rules (functional high-level system requirements)

Data Flow diagrams

Interviews with business (subject matter) experts

UML (activity, sequence diagrams)

Database constraints

Code/Comments/Unit Test Cases/Test Case Documents

Enterprise architecture documents

Business Glossary (Data Stewards)

More than a data dictionary

Goal to promote common understanding of core business concepts

- Term, definition
- Business unit that has ownership
- Data Steward contact information
- Business function association/context
- Issues with term/conflicting definitions
- Algorithms supporting definition
- Lineage

(Mosley, M., & Brackett, M. (2015). *The DAMA guide to the data management body of knowledge (DAMA-DMBOK guide), second edition*. Bradley Beach, N.J.: Technics Publications.)

Data Steward Quality Activities

Parsing and standardization-- breakup up fields, standardizing data against tables such as International Organization for Standardization

Cleansing– cleaning data to meet domain restrictions, business rules, constraints

Matching – record linkage/entity resolution

Profiling – data statistics (metadata) outlier analysis

Monitoring – conformance to governance rules

Enrichment – adding/associating new attributes (ex. GIS!) to data

(Reference: https://www.gartner.com/it-glossary/data-quality-tools/)

Simple Data Quality Example

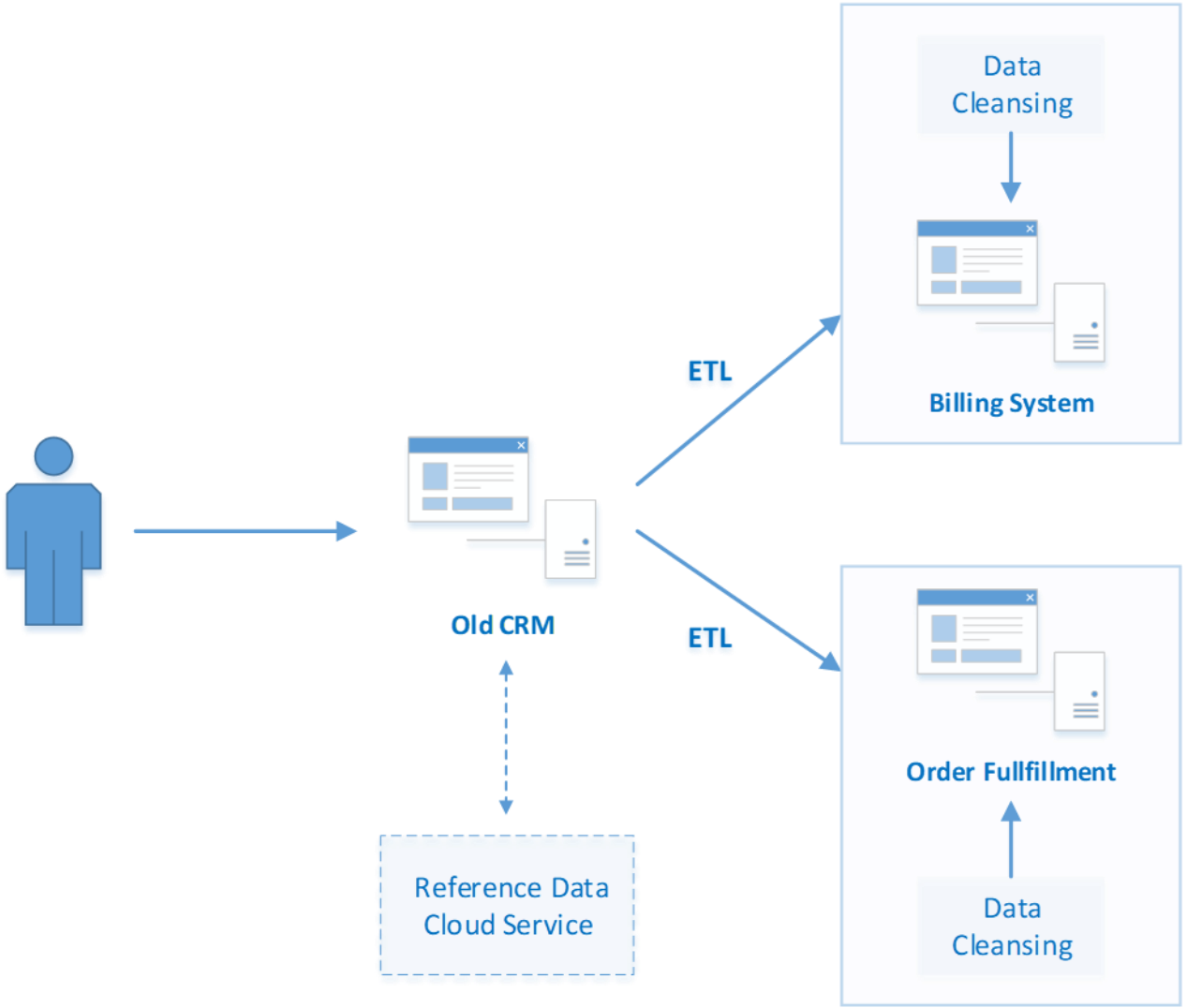
Data challenge: Address standardization and zip code validation

Business rules

- Zip codes need to match the address with which they are associated
- Zip codes need to be five valid digits
- Zip codes need to be a distinct field

Risks: Incorrect data bills go to wrong address, packages go to incorrect address

Schedule for improvement to get to 100% correctness


Data
Cleansing
Billing System
ETL
Old CRM
ETL
Order Fullfillment
Data
Cleansing
Reference Data
Cloud Service

Tools Used

The AST does not advocate specific tools and vendors!

- Talend
- Straight SQL
- SAP Microservices
- Free Python tools with support libraries

Desktop Tool

One possible tool for Data Quality checks:

Talend Open Studio
https://www.talend.com/products/talend-open-studio/data-quality-open-studio/

The free version does not have remedy piece for data errors, nor does it have a scheduler as the paid version does.

Talend Open Studio for Data Quality (6.5.1.20180116_1512)

Learn Ask Upgrade!

DQ Repository

- Data Profiling
  - Analyses (1)
    - Look at addresses 0.1
  - Libraries
  - Metadata
    - Db connections
      - BurtTest 0.1
        - employees
          - Tables (1)
            - employees
              - Columns (8)
                - address(VARCHAR)
                - birth_date(DATE)
                - emp_no(INT)
                - first_name(VARCHAR)
                - gender(ENUM)
                - hire_date(DATE)
                - last_name(VARCHAR)
                - zipcode(VARCHAR)
          - Views
      - FileDelimited connections
  - Recycle Bin

Detail View

General

No detail available

BurtTest 0.1

Connection Settings

Connection Metadata

Set the properties of connnection.

Name: BurtTest

Purpose:

Description:

Author: talend@talend.com

Status:

Connection Information

The information of connection.

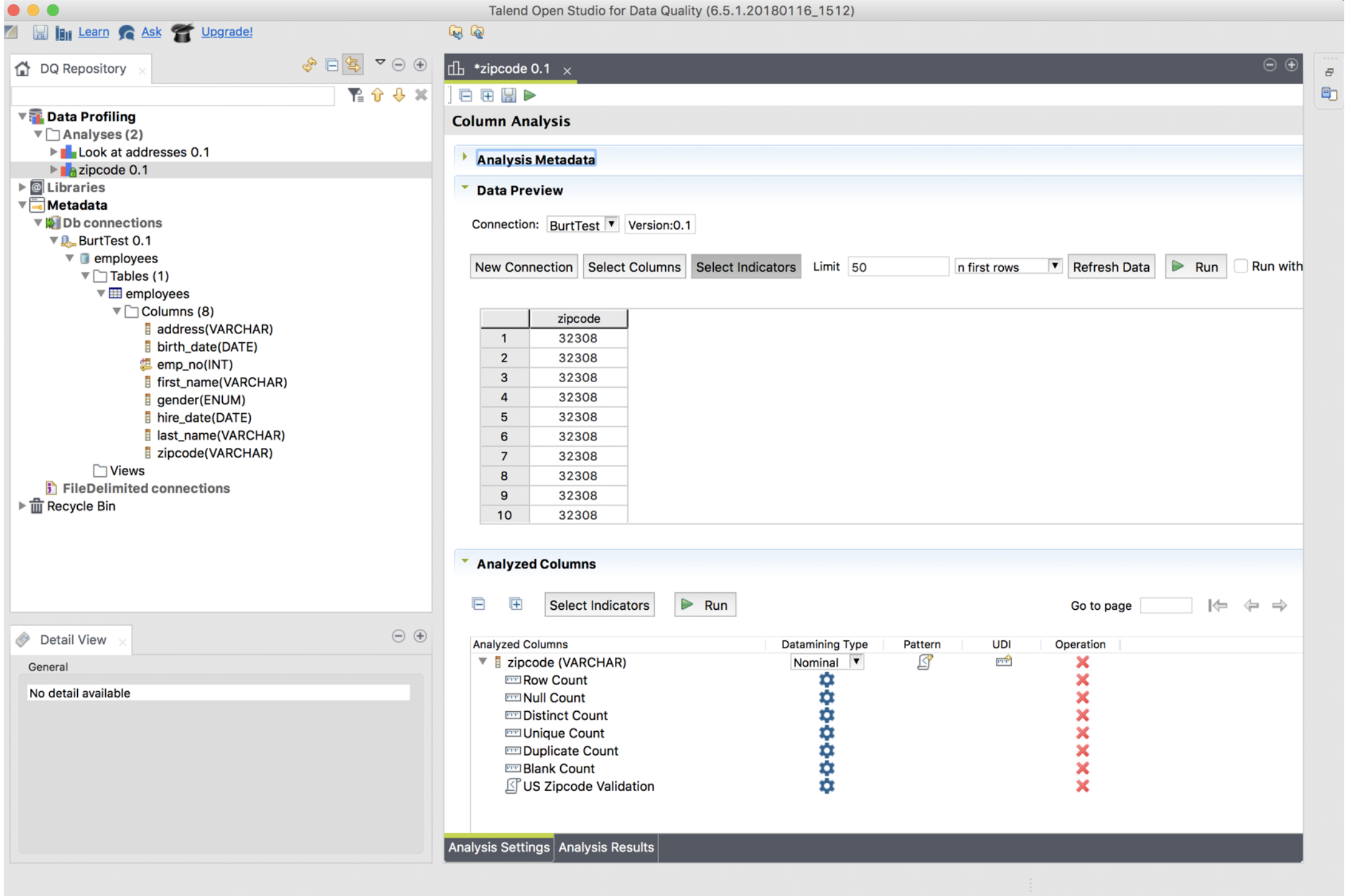
Login: root

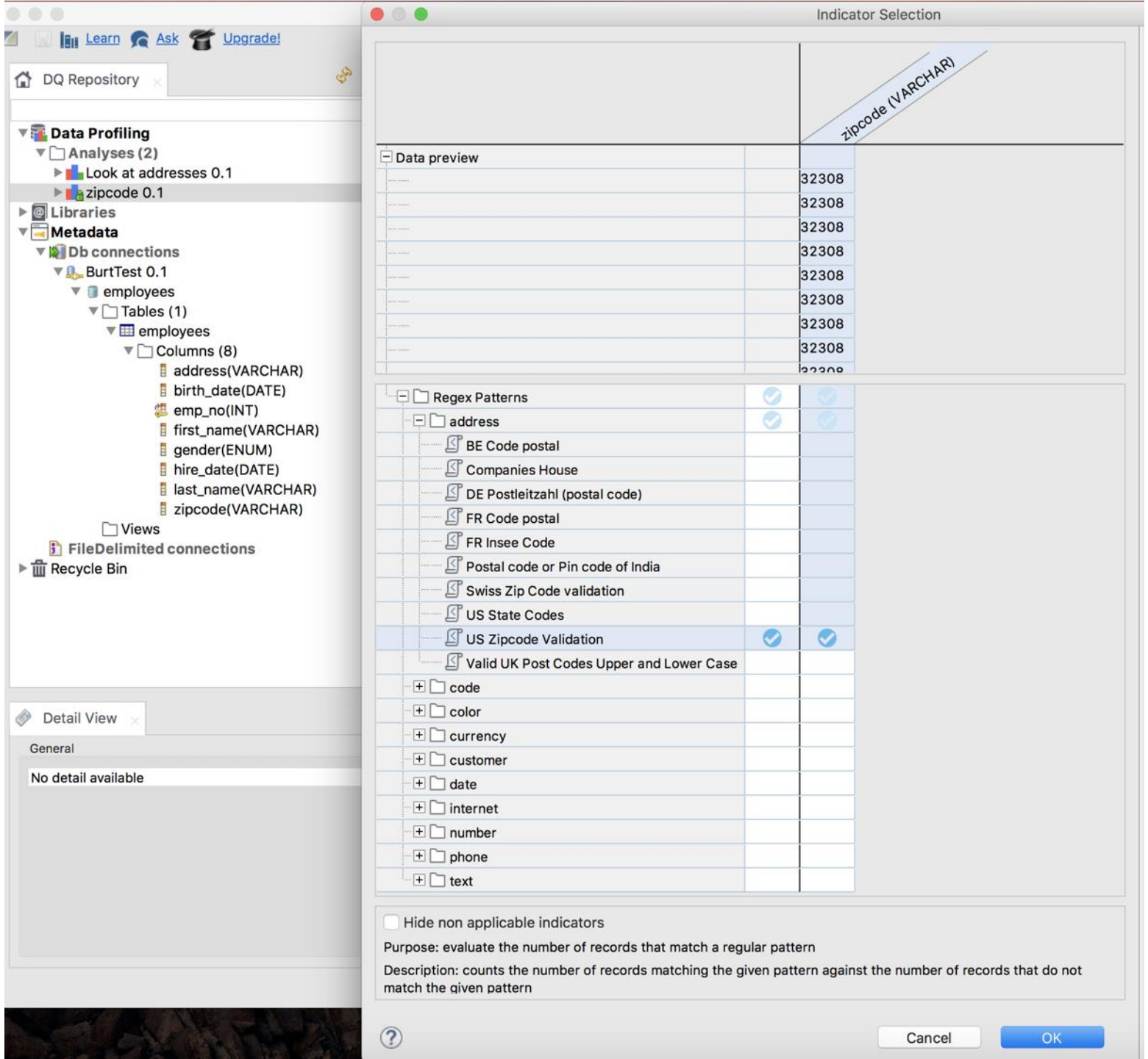
Password: •••••••

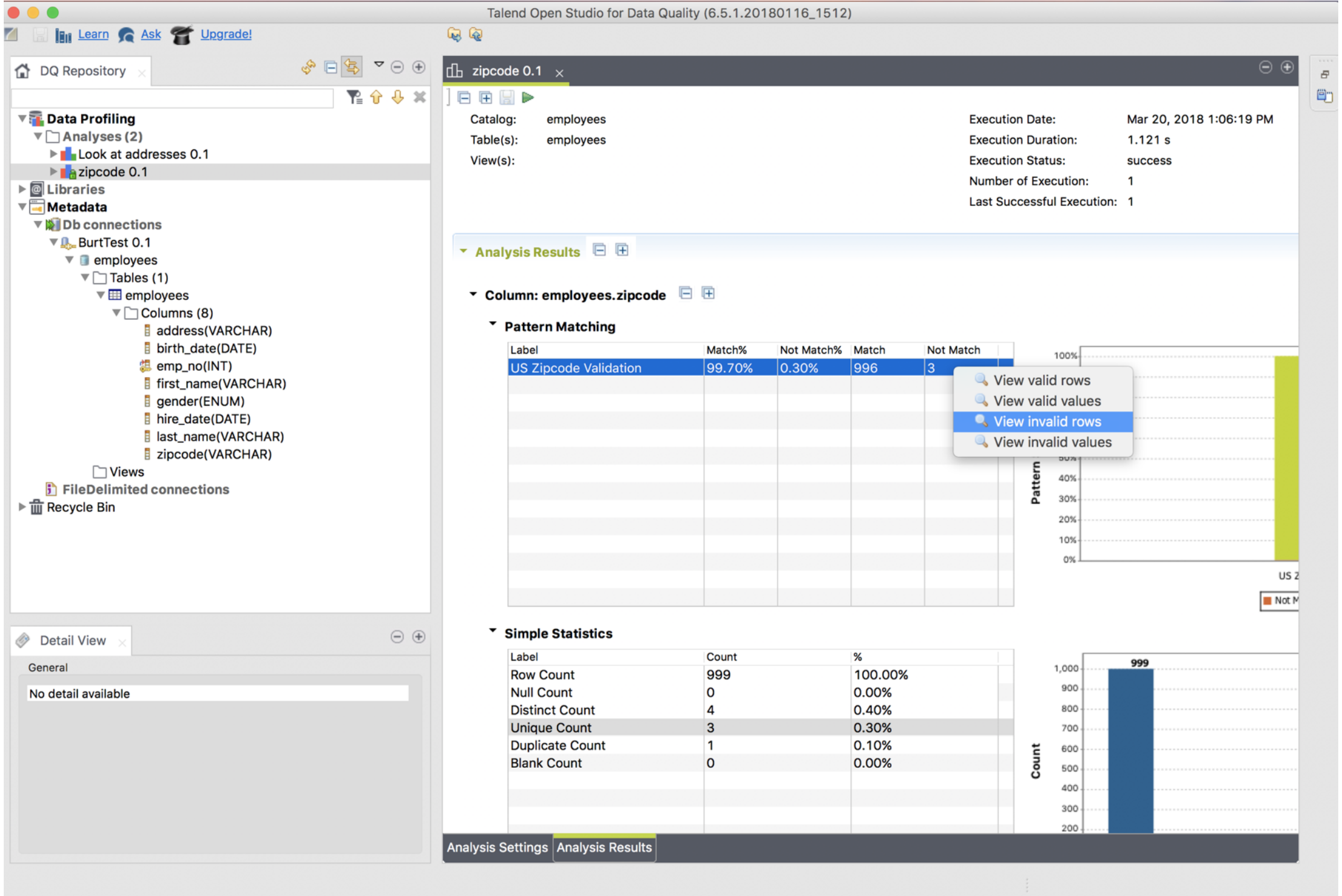
Url: jdbc:mysql://localhost:3306/employees?noDatetimeStringSync=true Edit...

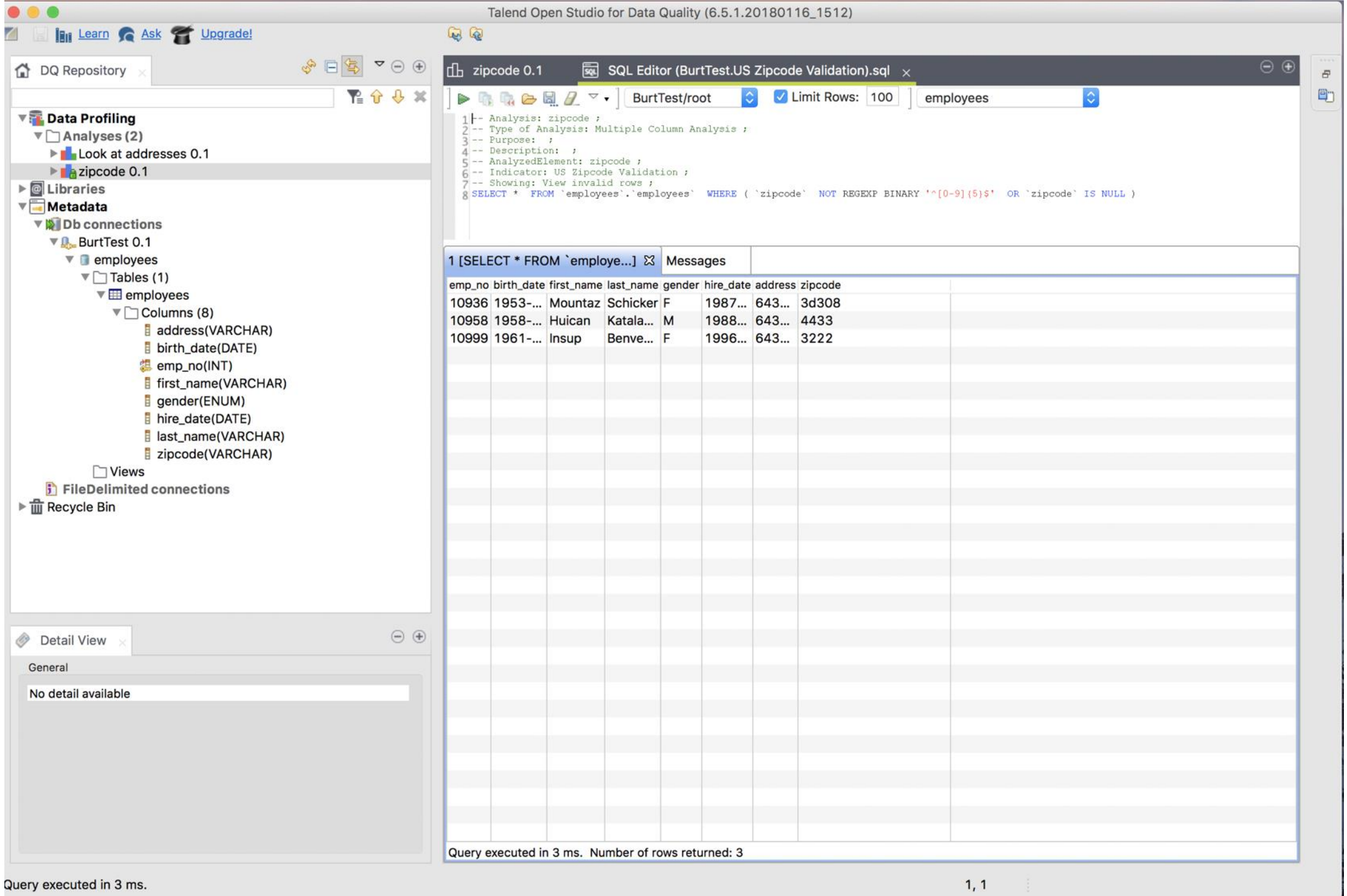
Check

Connection Settings


Talend Open Studio for Data Quality (6.5.1.20180116_1512)
Learn
Ask
Upgrade!
DQ Repository
Data Profiling
Analyses (2)
Look at addresses 0.1
zipcode 0.1
Libraries
Metadata
Db connections
BurtTest 0.1
employees
Tables (1)
employees
Columns (8)
address(VARCHAR)
birth_date(DATE)
emp_no(INT)
first_name(VARCHAR)
gender(ENUM)
hire_date(DATE)
last_name(VARCHAR)
zipcode(VARCHAR)
Views
FileDelimited connections
Recycle Bin
Detail View
General
No detail available
*zipcode 0.1
Column Analysis
Analysis Metadata
Data Preview
Connection: BurtTest
Version:0.1
New Connection
Select Columns
Select Indicators
Limit 50
n first rows
Refresh Data
Run
Run with
zipcode
1 32308
2 32308
3 32308
4 32308
5 32308
6 32308
7 32308
8 32308
9 32308
10 32308
Analyzed Columns
Select Indicators
Run
Go to page
Analyzed Columns
Datamining Type
Pattern
UDI
Operation
zipcode (VARCHAR)
Nominal
Row Count
Null Count
Distinct Count
Unique Count
Duplicate Count
Blank Count
US Zipcode Validation
Analysis Settings
Analysis Results


Learn
Ask
Upgrade!
DQ Repository
Data Profiling
Analyses (2)
Look at addresses 0.1
zipcode 0.1
Libraries
Metadata
Db connections
BurtTest 0.1
employees
Tables (1)
employees
Columns (8)
address(VARCHAR)
birth_date(DATE)
emp_no(INT)
first_name(VARCHAR)
gender(ENUM)
hire_date(DATE)
last_name(VARCHAR)
zipcode(VARCHAR)
Views
FileDelimited connections
Recycle Bin
Detail View
General
No detail available
Indicator Selection
zipcode (VARCHAR)
Data preview
32308
32308
32308
32308
32308
32308
32308
32308
Regex Patterns
address
BE Code postal
Companies House
DE Postleitzahl (postal code)
FR Code postal
FR Insee Code
Postal code or Pin code of India
Swiss Zip Code validation
US State Codes
US Zipcode Validation
Valid UK Post Codes Upper and Lower Case
code
color
currency
customer
date
internet
number
phone
text
Hide non applicable indicators
Purpose: evaluate the number of records that match a regular pattern
Description: counts the number of records matching the given pattern against the number of records that do not match the given pattern
Cancel
OK


Talend Open Studio for Data Quality (6.5.1.20180116_1512)
Learn
Ask
Upgrade!
DQ Repository
Data Profiling
Analyses (2)
Look at addresses 0.1
zipcode 0.1
Libraries
Metadata
Db connections
BurtTest 0.1
employees
Tables (1)
employees
Columns (8)
address(VARCHAR)
birth_date(DATE)
emp_no(INT)
first_name(VARCHAR)
gender(ENUM)
hire_date(DATE)
last_name(VARCHAR)
zipcode(VARCHAR)
Views
FileDelimited connections
Recycle Bin
Detail View
General
No detail available
zipcode 0.1
Catalog: employees
Table(s): employees
View(s):
Execution Date: Mar 20, 2018 1:06:19 PM
Execution Duration: 1.121 s
Execution Status: success
Number of Execution: 1
Last Successful Execution: 1
Analysis Results
Column: employees.zipcode
Pattern Matching
Label Match% Not Match% Match Not Match
US Zipcode Validation 99.70% 0.30% 996 3
View valid rows
View valid values
View invalid rows
View invalid values
Pattern
100%
40%
30%
20%
10%
0%
Simple Statistics
Label Count %
Row Count 999 100.00%
Null Count 0 0.00%
Distinct Count 4 0.40%
Unique Count 3 0.30%
Duplicate Count 1 0.10%
Blank Count 0 0.00%
Count
1,000
900
800
700
600
500
400
300
200
999
Analysis Settings
Analysis Results


Talend Open Studio for Data Quality (6.5.1.20180116_1512)
Learn
Ask
Upgrade!
DQ Repository
Data Profiling
Analyses (2)
Look at addresses 0.1
zipcode 0.1
Libraries
Metadata
Db connections
BurtTest 0.1
employees
Tables (1)
employees
Columns (8)
address(VARCHAR)
birth_date(DATE)
emp_no(INT)
first_name(VARCHAR)
gender(ENUM)
hire_date(DATE)
last_name(VARCHAR)
zipcode(VARCHAR)
Views
FileDelimited connections
Recycle Bin
Detail View
General
No detail available
zipcode 0.1
SQL Editor (BurtTest.US Zipcode Validation).sql
BurtTest/root
Limit Rows: 100
employees
-- Analysis: zipcode ;
-- Type of Analysis: Multiple Column Analysis ;
-- Purpose: ;
-- Description: ;
-- AnalyzedElement: zipcode ;
-- Indicator: US Zipcode Validation ;
-- Showing: View invalid rows ;
SELECT * FROM `employees`.`employees` WHERE (`zipcode` NOT REGEXP BINARY '^[0-9]{5}$' OR `zipcode` IS NULL)
1 [SELECT * FROM `employe...]
Messages
emp_no birth_date first_name last_name gender hire_date address zipcode
10936 1953-... Mountaz Schicker F 1987... 643... 3d308
10958 1958-... Huican Katala... M 1988... 643... 4433
10999 1961-... Insup Benve... F 1996... 643... 3222
Query executed in 3 ms. Number of rows returned: 3
Query executed in 3 ms.
1, 1

Straight SQL

- Ideal is a business tool not a technical tool, but…
- Select (in stored procedure) can be run as jobs that generate reports on a schedule
- Fixes can be done also by leveraging same SQL

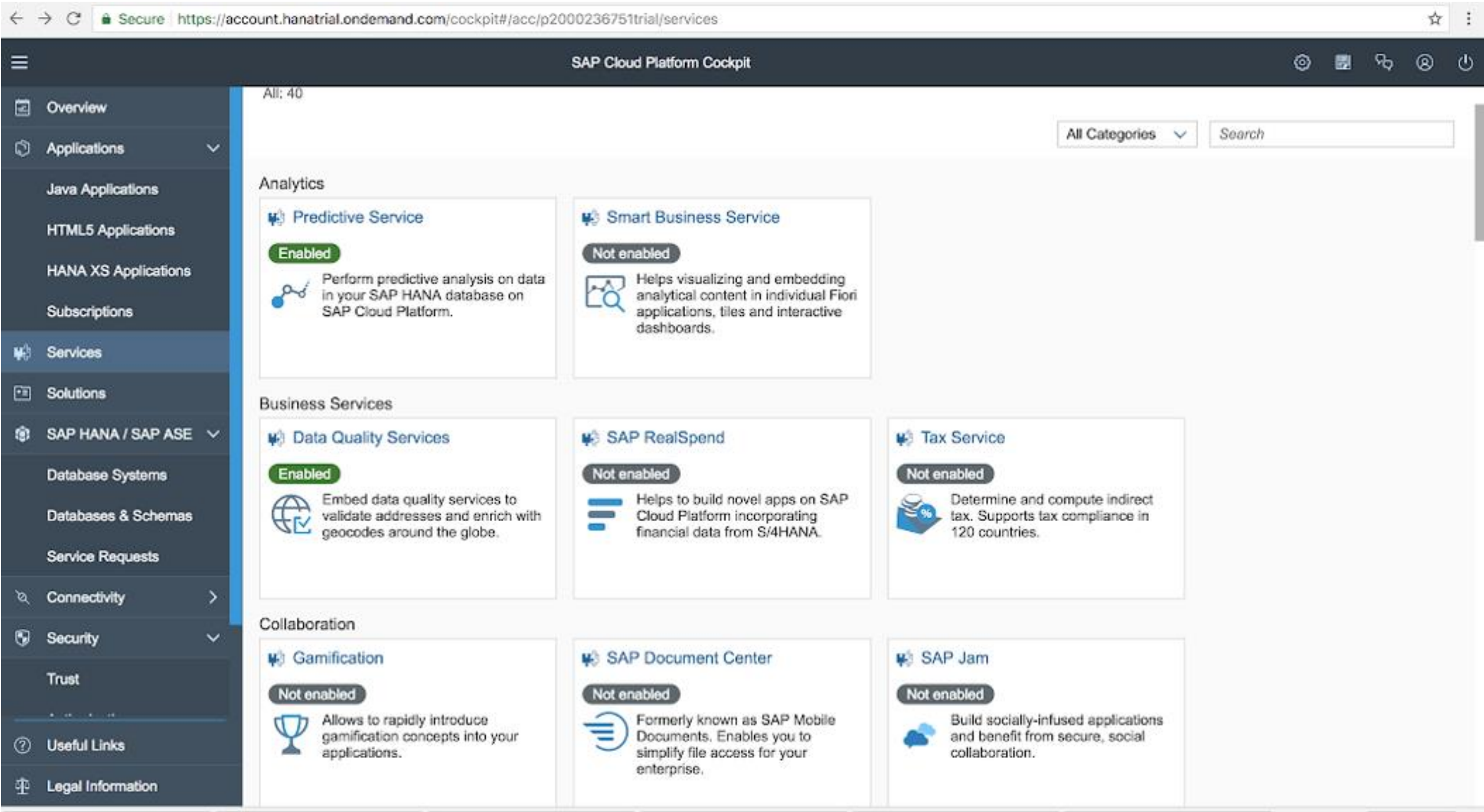
burtwalsh — docker exec -it b00908242b7f bash — 167×45

```
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql>
mysql> use employees;
Database changed
mysql> select * from employees where zipcode NOT REGEXP '[[:digit:]]{5}';
+--------+------------+------------+----------------+--------+------------+---------------------------------------+---------+
| emp_no | birth_date | first_name | last_name      | gender | hire_date  | address                               | zipcode |
+--------+------------+------------+----------------+--------+------------+---------------------------------------+---------+
|  10936 | 1953-04-08 | Mountaz    | Schicker       | F      | 1987-10-27 | 643 Chancey Lane, Tallahassee Florida | 3d308   |
|  10958 | 1958-11-07 | Huican     | Katalagarianos | M      | 1988-11-27 | 643 Chancey Lane, Tallahassee Florida | 4433    |
|  10999 | 1961-12-04 | Insup      | Benveniste     | F      | 1996-03-05 | 643 Chancey Lane, Tallahassee Florida | 3222    |
+--------+------------+------------+----------------+--------+------------+---------------------------------------+---------+
3 rows in set (0.00 sec)

mysql>
mysql>
mysql> update employees set zipcode = '32301' where zipcode = '3d308';
Query OK, 1 row affected (0.01 sec)
Rows matched: 1  Changed: 1  Warnings: 0

mysql> commit;
Query OK, 0 rows affected (0.01 sec)

mysql> select * from employees where zipcode NOT REGEXP '[[:digit:]]{5}';
+--------+------------+------------+----------------+--------+------------+---------------------------------------+---------+
| emp_no | birth_date | first_name | last_name      | gender | hire_date  | address                               | zipcode |
+--------+------------+------------+----------------+--------+------------+---------------------------------------+---------+
|  10958 | 1958-11-07 | Huican     | Katalagarianos | M      | 1988-11-27 | 643 Chancey Lane, Tallahassee Florida | 4433    |
|  10999 | 1961-12-04 | Insup      | Benveniste     | F      | 1996-03-05 | 643 Chancey Lane, Tallahassee Florida | 3222    |
+--------+------------+------------+----------------+--------+------------+---------------------------------------+---------+
2 rows in set (0.00 sec)

mysql>
```

Cloud Based Solutions

- Can be used in an automated manner
- Reference type data approach
- Service based REST call
- JSON format based (structured open data format)
- Can buy the service in the cloud (AWS)

(Example https://github.com/SAP/cloud-dqm-sample-payloads)


Secure https://account.hanatrial.ondemand.com/cockpit#/acc/p2000236751trial/services
SAP Cloud Platform Cockpit
Overview
Applications
Java Applications
HTML5 Applications
HANA XS Applications
Subscriptions
Services
Solutions
SAP HANA / SAP ASE
Database Systems
Databases & Schemas
Service Requests
Connectivity
Security
Trust
Useful Links
Legal Information
All: 40
All Categories
Search
Analytics
Predictive Service
Enabled
Perform predictive analysis on data in your SAP HANA database on SAP Cloud Platform.
Smart Business Service
Not enabled
Helps visualizing and embedding analytical content in individual Fiori applications, tiles and interactive dashboards.
Business Services
Data Quality Services
Enabled
Embed data quality services to validate addresses and enrich with geocodes around the globe.
SAP RealSpend
Not enabled
Helps to build novel apps on SAP Cloud Platform incorporating financial data from S/4HANA.
Tax Service
Not enabled
Determine and compute indirect tax. Supports tax compliance in 120 countries.
Collaboration
Gamification
Not enabled
Allows to rapidly introduce gamification concepts into your applications.
SAP Document Center
Not enabled
Formerly known as SAP Mobile Documents. Enables you to simplify file access for your enterprise.
SAP Jam
Not enabled
Build socially-infused applications and benefit from secure, social collaboration.

Postman

New | Import | Runner | My Workspace | SYNC OFF | Sign In

Filter

History | Collections

samplesAddressCleanse
336 requests

Basic

- POST United States Fielded
- POST United States Free-form
- POST Argentina Fielded
- POST Argentina Free-form
- POST Australia Fielded
- POST Australia Free-form
- POST Austria Fielded
- POST Austria Free-form
- POST Belgium Fielded
- POST Belgium Free-form
- POST Brazil Fielded
- POST Brazil Free-form
- POST Bulgaria Fielded
- POST Bulgaria Free-form
- POST Canada Fielded
- POST Canada Free-form
- POST China Fielded
- POST China Free-form
- POST Colombia Fielded

United States | Request oAuth token | https://dqmmicrou23 | United States Fielded | samplesEnvironment

POST https://dqmmicrou23133c2f-p2000236751trial.hanatrial.ondemand.com/dq/addressCleanse Params Send Save

Authorization | Headers (3) | Body | Pre-request Script | Tests | Cookies | Code

form-data | x-www-form-urlencoded | raw | binary | JSON (application/json)

```
{
  "addressInput": {
    "country": "US",
    "mixed": "875 NORTH MICHIGAN AVENUE, SUITE 104",
    "locality": "CHICAGO",
    "region": "IL",
    "postcode": ""
  },
  "outputFields": [
    "std_addr_address_delivery",
    "std_addr_locality_full",
    "std_addr_region_full",
    "std_addr_postcode_full",
    "std_addr_country_2char",
    "addr_asmt_info",
    "addr_asmt_level",
    "addr_info_code",
    "addr_info_code_msg",
    "addr_change_sig"
```

Body | Cookies (3) | Headers (5) | Test Results | Status: 200 OK | Time: 1307 ms | Size: 548 B

Pretty | Raw | Preview | JSON | Save Response

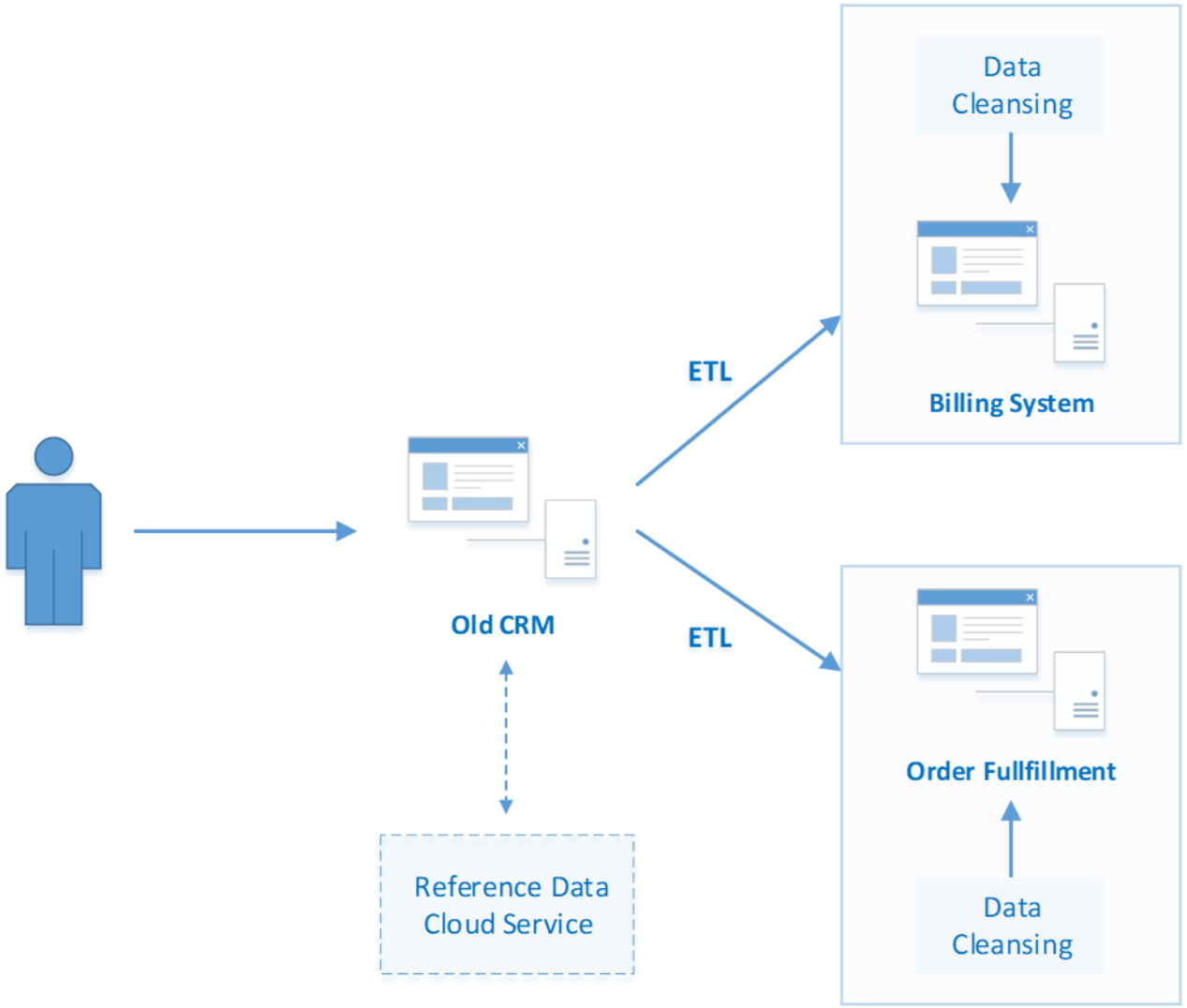
```
{
    "std_addr_locality_full": "Chicago",
    "addr_info_code_msg": "",
    "addr_info_code": "",
    "addr_asmt_info": "C",
    "std_addr_region_full": "IL",
    "addr_change_sig": "H",
    "std_addr_address_delivery": "875 N Michigan Ave Ste 104",
    "std_addr_postcode_full": "60611-1882",
    "std_addr_country_2char": "US",
    "addr_asmt_level": "S"
}
```

Programmatic Solutions

- Python (language)
- Pandas/NumPy/MXNet/Gluon (libraries)
- Highly leveraged in industry
- Powerful but requires programming

```
      name  age gender  ytr
0     Burt  481      M   30
1    Kathy   48      F   30
2  Lincoln   10      m   50
3     Liam    9      M   50
4   Hubert   72      M    0
5  Barbara   71      F    0
6   Brenda   44      f   10


      name  age gender  ytr
0     Burt   51      M   30
1    Kathy   48      F   30
2  Lincoln   10      M   50
3     Liam    9      M   50
4   Hubert   72      M    0
5  Barbara   71      F    0
6   Brenda   44      F   10
```

```
import pandas as pd
import numpy as np

#read the file again into a data frame (we can use ODBC with Pandas)
k = pd.read_csv('ok.csv', names=["name","age","gender","ytr"])
print(k)

print('')
print('')

#make sure the age is a numeric
k['age']=k['age'].apply(pd.to_numeric)

#make sure the gender is UPPER CASE this is IMPORTANT for RECORD LINKAGE
k['gender']=k['gender'].apply(lambda p : p.upper())

#if the age is greater than 120 make the age 21 plus the number of years till retirement (ytr)
k['age']=k.apply(lambda row: (21 + row['ytr']) if row['age'] > 120 else row['age'], axis=1)
print(k)
```


Data
Cleansing
Billing System
ETL
Old CRM
ETL
Order Fullfillment
Reference Data
Cloud Service
Data
Cleansing

Requirements/Remedies/Governance

- Requirements for Address
- Importance of data to the business
- Risk of address issue and “good enough” error tolerance

- Guidance to data entry staff/training/monitoring
- Data input into new secondary system with ingestion into old system; could be first step to modernization
- Fix data in ETL

- Long term fix, value to the business, cost of work arounds

Tools

- What tools do you have?
  - Can you share your experiences to help others? AST can present if needed
  - Enterprise licensing agreements/shared model
  - Are you fully using your tool(s)? Can we help?
  - What are your needs and new tools and approaches

Data Lineage

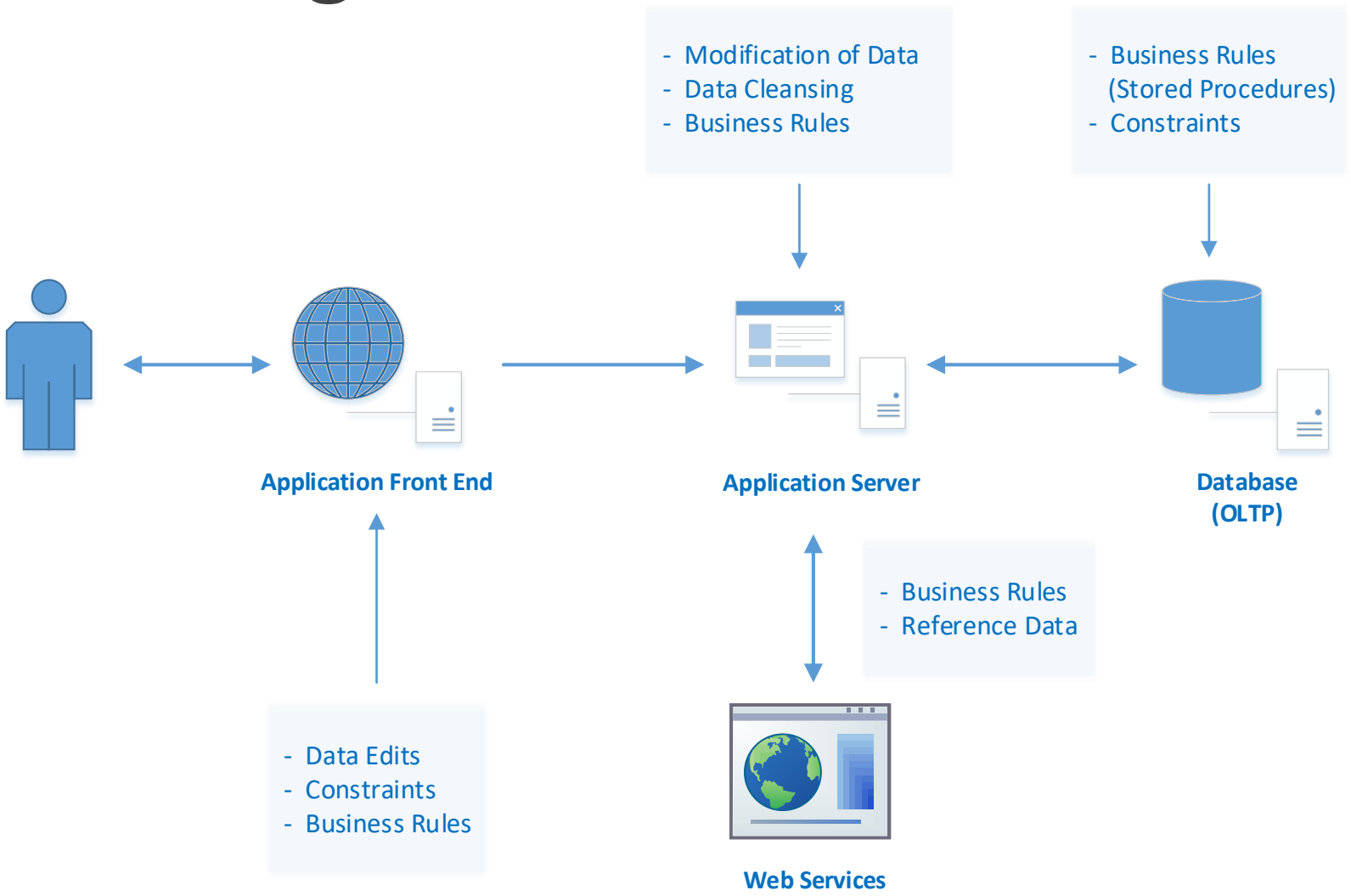
Data lineage is concerned with the flow of data through systems to end users

- Where it is created and by whom
- How it is transformed (business processes, users, ETL)
- Data Flow Diagrams (annotations for entities and business rules)

Where it is used

- Root Cause Analysis
- Impact Analysis

Data Lineage



DATA LAKE

- Structured Data
- Unstructured Data
- OLAP

Extract
Transform
Load
(ETL)

Create
Read
Update

N-Tier App

N-Tier App

BI / Reporting App

Create
Read
Update

Is what I'm looking at based on data that is accurate, timely, consistent, usable, and "fit for purpose?"

End Consumer

Data Quality Principals/Goals

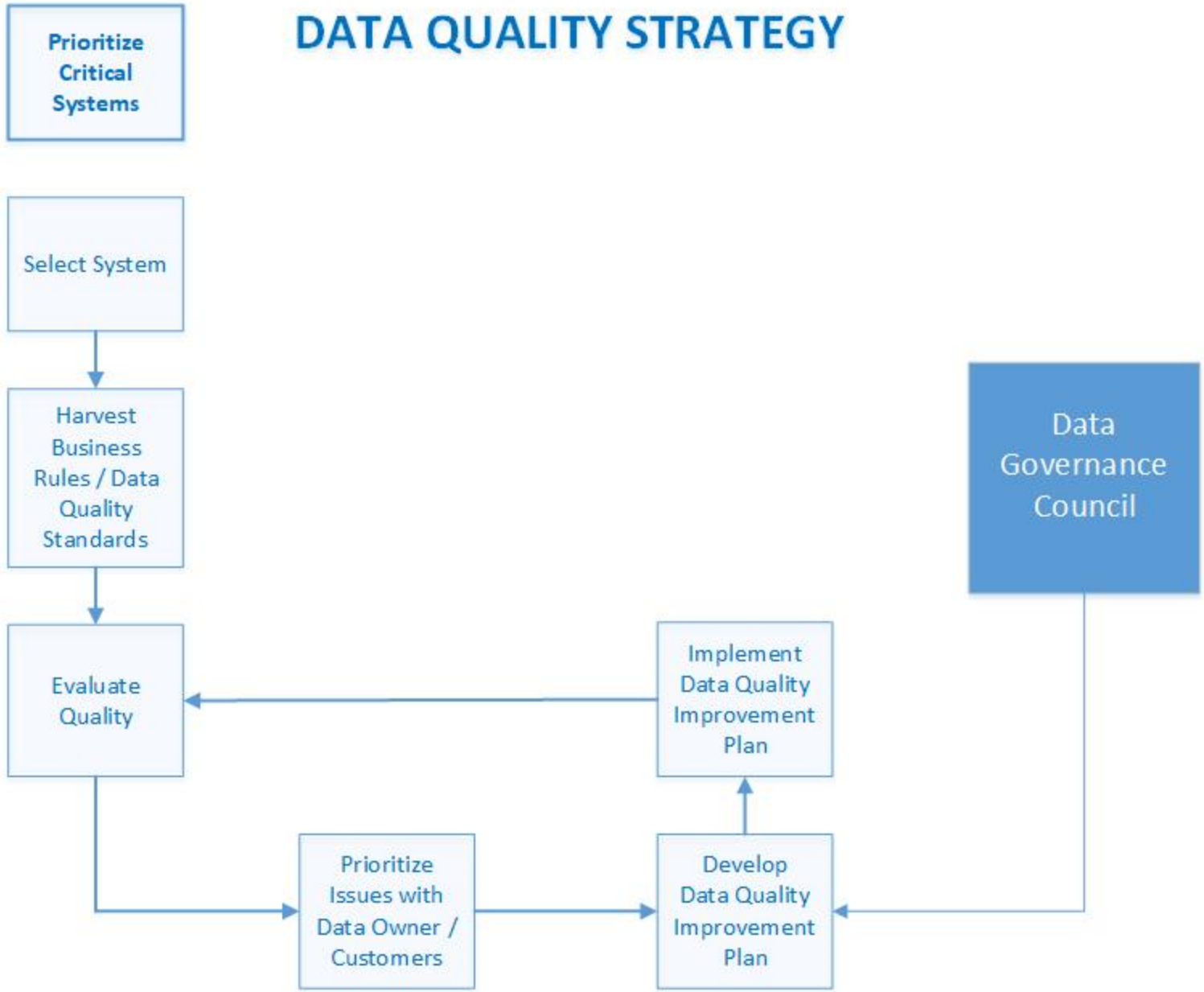
Focus on data which is most critical to business (KPI)

Quality efforts should focus on root causes of data issues (Data Lineage)

Data quality should be measurable and based upon business standards and rules

Data Quality should be enforced and improved throughout the data lifecycle (requirements gathering, system design, testing, integration, system improvement)

DATA QUALITY STRATEGY



Some Conclusions

- Data quality is critical to business success
- Data quality efforts should focus first on the most important data to the business in the form of profit and risk (non-compliance)
- Data Quality results and benefits should be measurable (ROI)

- Data moves through various systems/processes to consumers
- Data Quality is a full lifecycle and enterprise system concern
- Governance—roles, responsibilities, contacts, guidance and escalation processes

Possible Next Topic

- Record Linkage/Entity Resolution
  - https://en.wikipedia.org/wiki/Record_linkage
  - Data Quality is a key first step to Record Linkage (from previous example—address, gender)
- Some source examples
  - http://recordlinkage.readthedocs.io/en/latest/notebooks/link_two_dataframes.html
  - https://cran.r-project.org/web/packages/RecordLinkage/index.html
- Data sharing (MOU/DUA, closer look at FERPA, HIPPA, CJIS)
- Data Governance/rule/process

South Carolina