Profiling Your Data

Angela Henry

Angela Henry

Angela is a BI Developer, living in High Point, NC

Solution Architect at BlueGranite

Data Platform MVP

Local group leader for PASS User Group in Greensboro, NC

Tweets at @SQLSwimmer
Blogs at SQLSwimmer.com

Who Are You?

DBA

Developer

Business Analyst

Other

Overview

What

Why

How

What does it mean?

What - Minimum Needs

Distinct/Percentage Count

Percentage Zero, Blank & NULL

Min, Max, AVG string lengths

Numerical Ranges

Date Ranges

What – Great to Have

Key Integrity

Cardinality

Patterns

Frequencies

Why

Makes Your Life Easier!

Analyze Source Data More Effectively

Understand Source Data Better

Prevent Data Quality Problems Before They Are Introduced

Kimball Four Step Method

Step 1: **Profile** at **project start** to determine project viability

Step 2: **Identify and correct** data quality issues in source data

Step 3: Identify data quality issues that can be **corrected by ETL**

Step 4: Identify unanticipated business rules, hierarchical structures and foreign key / private key relationships

How

Non-SQL Server Based

R

Azure Data Catalog

Informatica

Talend Open Studio for Data Quality

SAS

Collibra

Power BI

SQL Server Based

Data Quality Services (DQS)

SSIS – Data Profiling Task

Demo

Power BI

Power BI Limitations

No String Length Statistics

Default First 1000 Rows

Data Quality Services

SSIS Data Profiling Task

Requirements

Set Up

Analyze

Requirements

SQL Server is only source

Tempdb Permissions

- Read
- Write
- Create Table

Setup

Which profiles to compute

Where to output the profiles

Demo

SSIS Data Profiling Task

SSIS Data Profiling Task Review

SQL Server Only Data Source

View Results Within Visual Studio

Use Stand Alone Viewer to View Results

Script Task With XPath Query to Make Decisions

Dynamic XML Example

```
"<?xml version=\"1.0\" encoding=\"utf-16\"?>

<DataProfile xmlns:xsi=\http://www.w3.org/2001/XMLSchema-instance\
   xmlns:xsd=\http://www.w3.org/2001/XMLSchema\
   xmlns=\"http://schemas.microsoft.com/sqlserver/2008/DataDebugger/\">

  <DataSources />

  <DataProfileInput>

    <ProfileMode>Exact</ProfileMode>

    <Timeout>0</Timeout>

    <Requests>

      <ColumnNullRatioProfileRequest ID=\"NullRatioReq\">

        <DataSourceID>DatabaseConn</DataSourceID>

        <Table Schema=\"" + @[User::SchemaName] + "\" Table=\"" + @[User::TableName] + "\" />

        <Column IsWildCard=\"true\" />

      </ColumnNullRatioProfileRequest>

      <ColumnStatisticsProfileRequest ID=\"StatisticsReq\">

        <DataSourceID>DatabaseConn</DataSourceID>

        <Table Schema=\"" + @[User::SchemaName] + "\" Table=\"" + @[User::TableName] + "\" />

        <Column IsWildCard=\"true\" />

      </ColumnStatisticsProfileRequest>

      <ColumnLengthDistributionProfileRequest ID=\"LengthDistReq\">
```

SQL Swimmer

Summary

Just Do It!

References

Profile Task & Viewer Documentation

Azure Data Catalog Documentation

Talend Information

Power BI Data Profiling

Collibra

References

Data Quality Services Profiling Information

Data Quality Services Installation & Setup

Automate Multiple Dynamic Tables Blog Post

Data Profiling in Informatica

SAS Dataflux

Data Profile Viewer

Stand alone profile viewer can be found at

C:\Program Files (x86)\Microsoft SQL Server\140\DTS\Binn

Note* The 140 may be different depending on the version of SQL Sever you are using.

Thank You

Twitter: @SQLSwimmer

Blog: sqlswimmer.com

Linked In: \in\angelahenrydba

Email: angela@sqlswimmer.com