Data Wrangling for Reporting & Analytics

Phil Robinson
sqldbdev@gmail.com

Data Wrangling for Reporting & Analytics

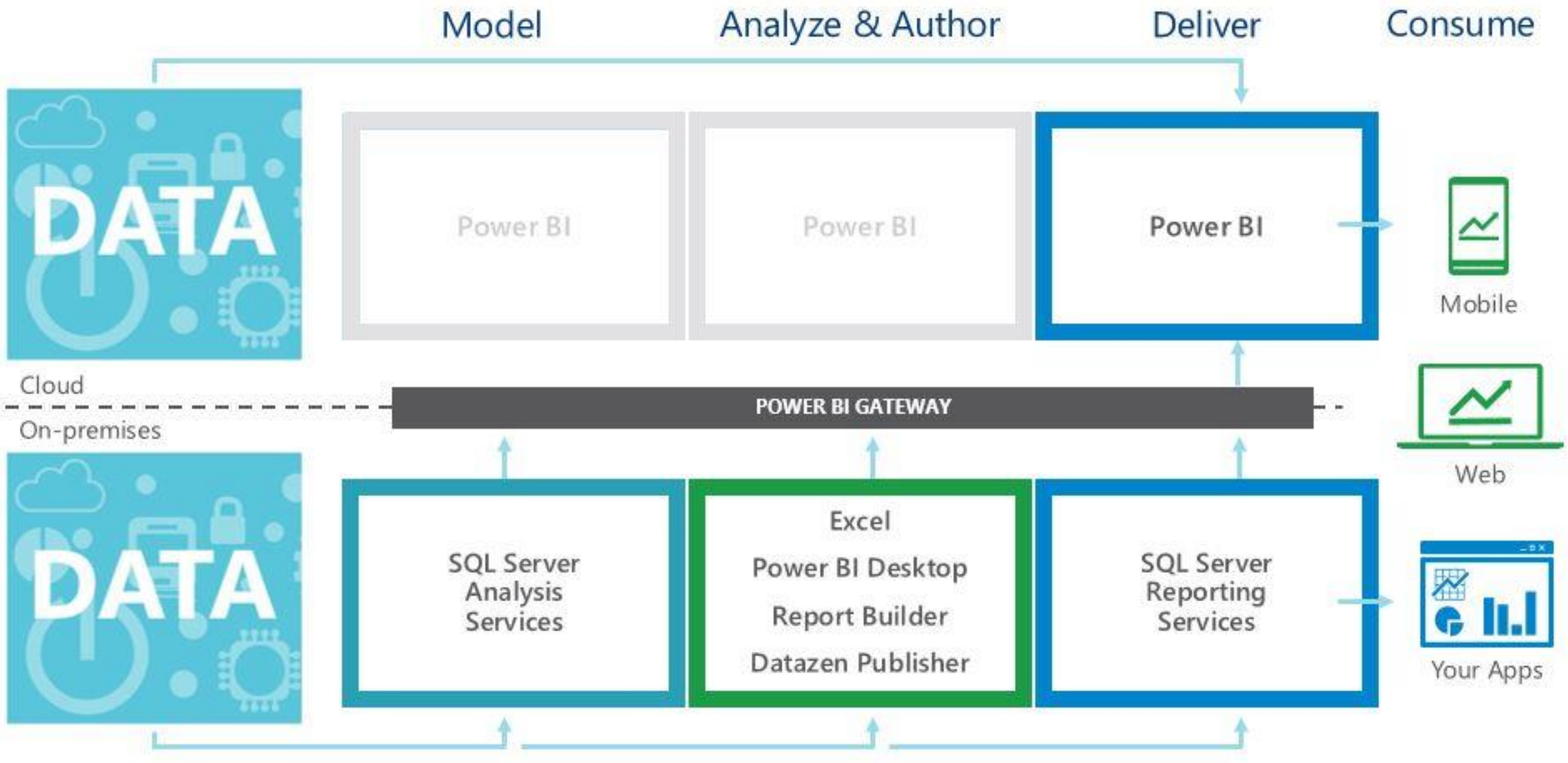
- Independent MS Consultant since 1997
- ASP/ASP.NET
- Current focus is Business Intelligence development using SQL Server tools.
- PASS Regional Mentor – Southwest Region
- Chapter Leader – SD SQL BI Group
- Co-founder - SQL Saturday – San Diego

Data Wrangling for Reporting & Analytics

Agenda

- Microsoft Roadmap 2016
- What is Data Wrangling ?
- Working with Data
- Tools
- Resources

Data Wrangling for Reporting & Analytics



Data Wrangling for Reporting & Analytics

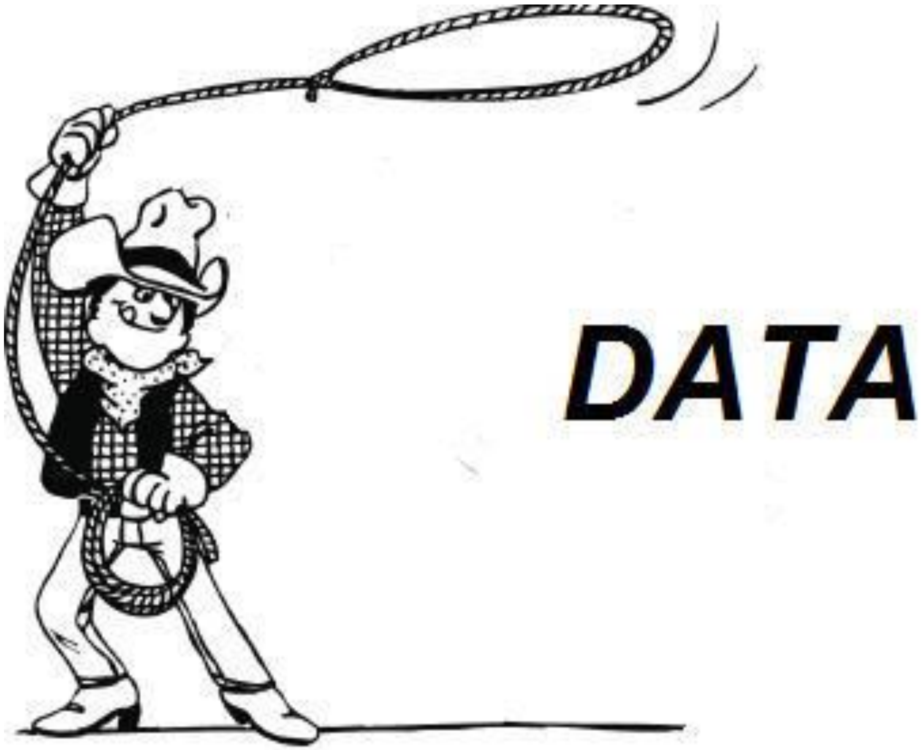
What is Data Wrangling ?

- Wikipedia

  Data wrangling is the process of taking data in its native format & making it usable for analysis.

  A data wrangler is the person performing the wrangling. In the scientific research context, the term often refers to a person responsible for gathering and organizing disparate data sets collected by many different investigators, often as part of a field campaign.

Data Wrangling for Reporting & Analytics



Data Wrangling for Reporting & Analytics

Data Wrangling Steps

- Discovering
- Structuring
- Cleaning and Validation
- Enrichment
- Deployment

Data Wrangling for Reporting & Analytics

Data Wrangling Steps

- Discovering

  Finding data that may be useful in answering the business question associated with the project.

Data Wrangling for Reporting & Analytics

Data Wrangling Steps

- Structuring

  Join or pivot datasets. Remove, split or combine columns

Data Wrangling for Reporting & Analytics

Data Wrangling Steps

- Cleaning and Validation

  Remove or repair data that might distort the analysis or report

  Check for data quality and consistency

Data Wrangling for Reporting & Analytics

Data Wrangling Steps

- Enrichment

  Identify and add other data which might be useful in this analysis

  Can additional data be derived from the existing data

Data Wrangling for Reporting & Analytics

Data Wrangling Steps

- Deployment

  What is the useful life span of the data

  When and how is the data updated

  Who needs access to the data

Data Wrangling for Reporting & Analytics

Data Source Challenges

- Diverse/Disparate
- Source reliability
- Update Frequency

Data Wrangling for Reporting & Analytics

Data Set Challenges

- Schema consistency
- Text formatting, abbreviations and case
- Missing or incorrect values
- Date formats and cardinality
- Multiple value columns
- Duplicate records

Data Wrangling for Reporting & Analytics

Data Set Challenges

- Schema consistency
  - – Missing field separators
  - – Lack of field qualifiers
  - – Missing or non-Windows line breaks
  - – Too many columns

Data Wrangling for Reporting & Analytics

Data Set Challenges

- Text formatting, abbreviations and case
  - "123-45-6789" or "123456780" or "123 45 6789“
  - "IBM" or "I.B.M." or "Int. Bus. Machines"
  - “VISTA” or “Vista” or “vista”

Data Wrangling for Reporting & Analytics

Data Set Challenges

- Missing or incorrect values
  - Nulls
  - Spaces
  - Special characters
  - Outliers

Data Wrangling for Reporting & Analytics

Data Set Challenges

- Date formats and cardinality

Data Wrangling for Reporting & Analytics

Data Set Challenges

- Multiple value columns
  - – “123 Any Street, San Diego CA 91901”

Data Wrangling for Reporting & Analytics

Data Set Challenges

- Duplicate records
  - – Incomplete records
  - – No unique identifiers

Data Wrangling for Reporting & Analytics

Demos

Data Wrangling for Reporting & Analytics

Tools

- CSVEasy
  - http://csveasy.com
- Multi-Edit Lite 2008
  - http://multieditsoftware.com/product/multi-edit-lite-2008
- Talend Data Preparation
  - https://www.talend.com/products/data-preparation#free-desktop
- Trifacta Wrangler
  - https://www.trifacta.com/products/wrangler

Data Wrangling for Reporting & Analytics

Resources

- Power BI Desktop
  - https://powerbi.microsoft.com/en-us/desktop
- Power BI in the Cloud
  - https://powerbi.microsoft.com/en-us/pricing
- Power Query (M) Formula Reference
  - https://msdn.microsoft.com/en-us/library/mt211003.aspx

Data Wrangling for Reporting & Analytics

Resources

- SQL Server 2016 – Developer Edition
  - https://www.microsoft.com/en-us/sql-server/sql-server-downloads

Data Wrangling for Reporting & Analytics

?