Data Wrangling using Power BI Desktop

Phil Robinson
Email: sqldbdev@gmail.com
Blog: sqldbdev.com
LinkedIn: linkedin.com/in/sqldbdev

Data Wrangling using Power BI Desktop

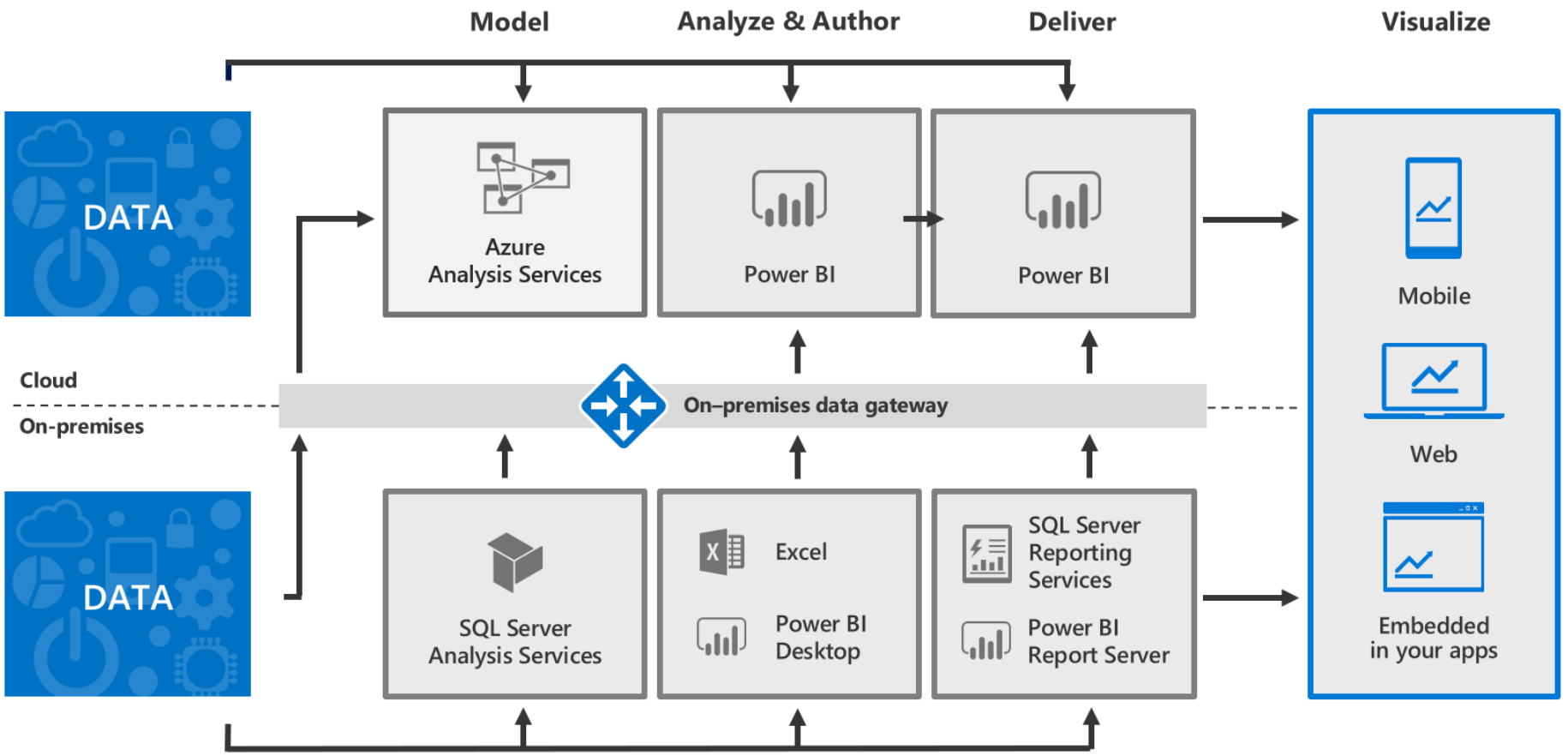
- Independent MS Consultant since 1997
- ASP/ASP.NET
- Current focus is Business Intelligence development using SQL Server tools.
- PASS Regional Mentor – Southwest Region
- User Group Leader – SD SQL BI Group
- Co-founder - SQL Saturday – San Diego

Data Wrangling using Power BI Desktop

Agenda

- Microsoft Business Intelligence Roadmap
- What is Data Wrangling ?
- Working with Data
- Data Governance
- Tools
- Resources

Data Wrangling using Power BI Desktop



Data Wrangling using Power BI Desktop

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 using Power BI Desktop

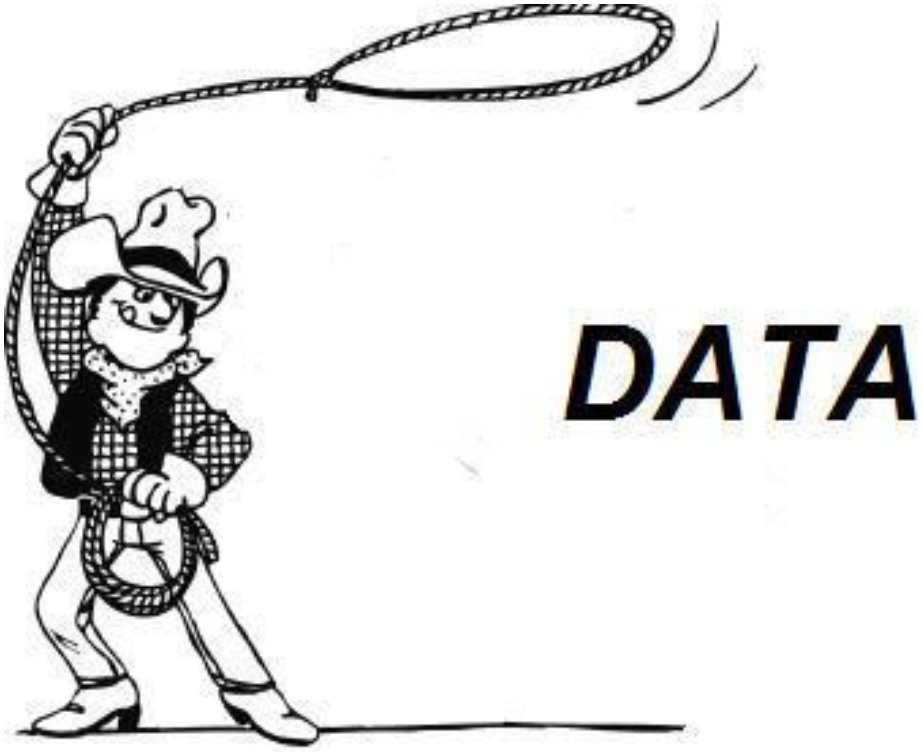
What is Data Wrangling ?

- Gartner

  Data preparation is an iterative-agile process for exploring, combining, cleaning and transforming raw data into curated datasets for self-service data integration, data science, data discovery, and BI/analytics.

"Market Guide for Data Preparation" Gartner, December 2017

Data Wrangling using Power BI Desktop



Data Wrangling using Power BI Desktop

Data preparation — the most time-consuming task in analytics and BI — is evolving from a self-service activity to an enterprise imperative.

By 2019, data and analytics organizations that provide agile, curated internal and external datasets for a range of content authors will realize twice the business benefits as those that do not.

“Market Guide for Data Preparation” Gartner, December 2017

Data Wrangling using Power BI Desktop

Percentage of time spent on data preparation by data scientists?

“Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says”
Gil Press , Forbes Contributor
Based on a CrowdFlower survey of 179 data scientists conducted in February and March of 2017.

Data Wrangling using Power BI Desktop

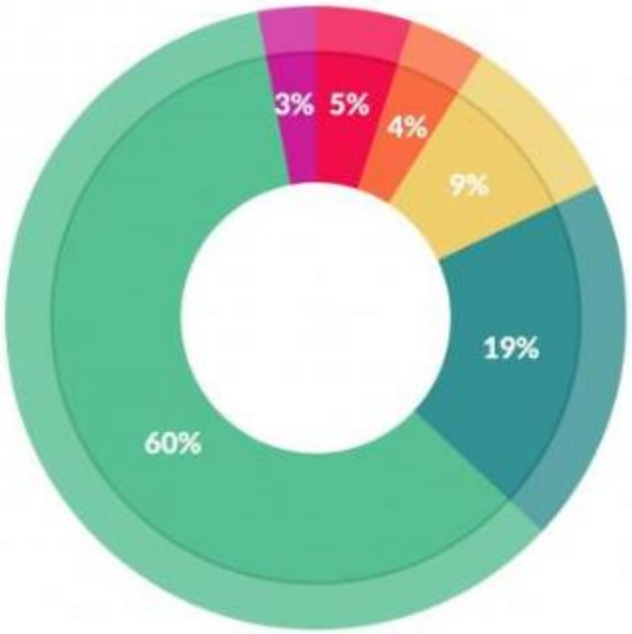
Percentage of time spent on data preparation by data scientists?

80%

“Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says”
Gil Press , Forbes Contributor
Based on a CrowdFlower survey of 179 data scientists conducted in February and March of 2017.

Data Wrangling using Power BI Desktop

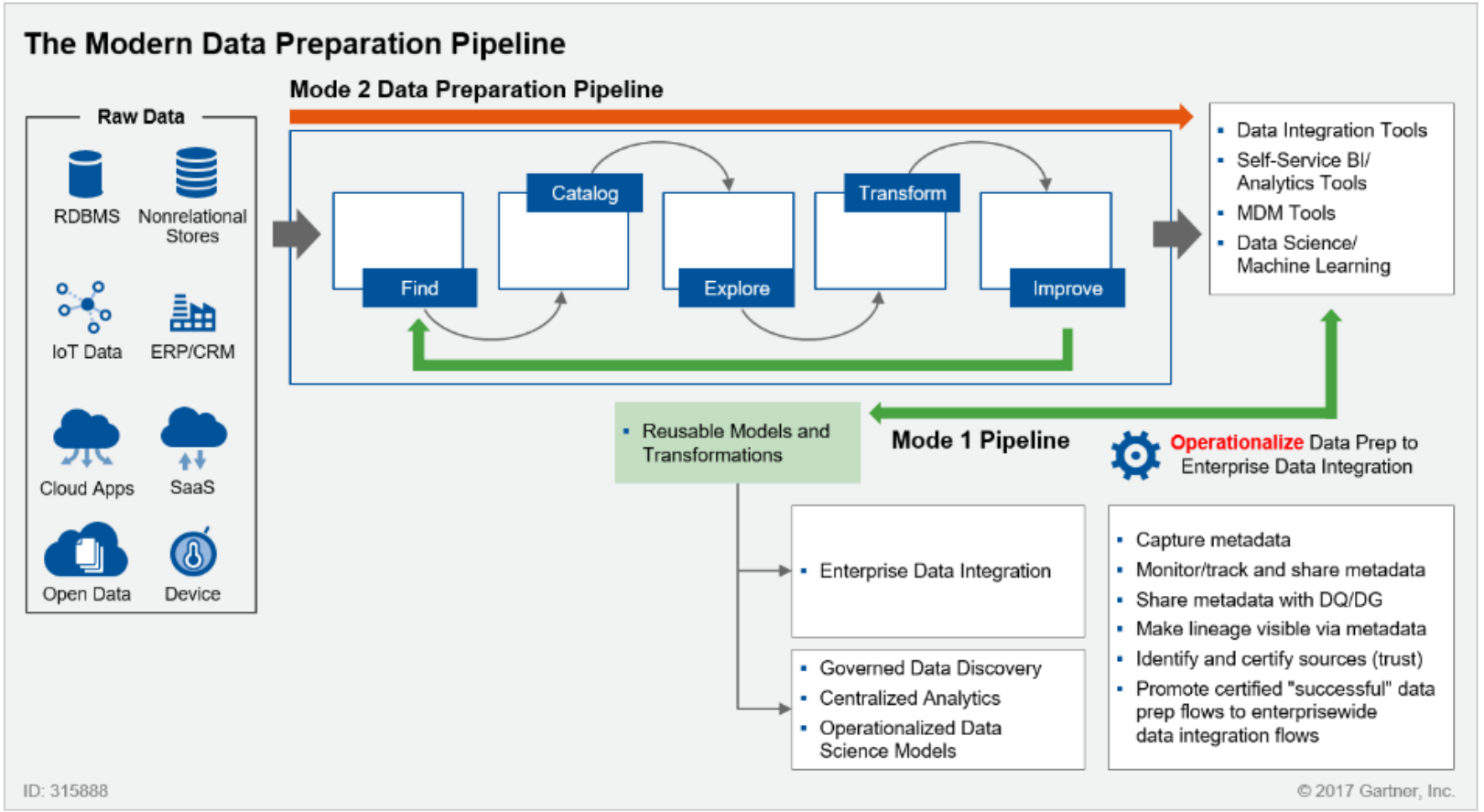
Data preparation *accounts for about 80% of the work of data scientists*



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Data Wrangling using Power BI Desktop



"Market Guide for Data Preparation" Gartner, December 2017

Data Wrangling using Power BI Desktop

Data Wrangling Steps

- Find

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

Data Wrangling using Power BI Desktop

Data Source Challenges

- Diverse/Disparate
- Source reliability
- Update Frequency

Data Wrangling using Power BI Desktop

Data Wrangling Steps

- Catalog
  - Identify the data set source
  - Determine trustworthiness of source
  - Capture metadata

Data Wrangling using Power BI Desktop

Data Governance

- Data Source Trust Documentation
- Data Set Metadata Descriptions
- Transformation Process Details
  - Naming Conventions
    - ColumnName_Function_Description/Details

Data Wrangling using Power BI Desktop

CSVEasy

Catalog Data

Data Wrangling using Power BI Desktop

Data Wrangling Steps

- Explore
  - Check for data quality and consistency
  - Remove or repair data that might distort the analysis or report

Data Wrangling using Power BI Desktop

Dirty Data Challenges

- Corrupt Fields or Records
- Missing or Incorrect Values
- FUBAR Formats
- High Cardinality Columns

Data Wrangling using Power BI Desktop

Fixing Dirty Data

- Corrupt Fields or Records
  - Reload the data from the original source
  - Delete corrupt records
  - Treat corrupt columns as incorrect values

Data Wrangling using Power BI Desktop

Fixing Dirty Data

- Missing or incorrect values
  - Predict the missing value
  - Leave the record as is
  - Delete the record
  - Replace with the mean or median value
  - Create a new category
  - Introduce a new variable/column

Data Wrangling using Power BI Desktop

Fixing Dirty Data

- Missing or incorrect values
  - Nulls, Spaces, Special characters
  - Inconsistent format or value

Data Wrangling using Power BI Desktop

Fixing Dirty Data

- Missing or incorrect values
  - Gender
    - Invalid characters
    - Blank or U
      - Suffix – Sr, Jr, II, III
      - Prefix – Mr, Mrs, Miss, Ms

Data Wrangling using Power BI Desktop

Fixing Dirty Data

- Missing or incorrect values
  - 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 using Power BI Desktop

Fixing Dirty Data

- FUBAR
  - Multiple value columns
    - “123 Any Street, San Diego CA 91901”
    - “537 Some Ave, Apt 301, San Diego CA 91901”
    - “537 Some Ave Apt 222, San Diego CA 91901”

Data Wrangling using Power BI Desktop

Fixing Dirty Data

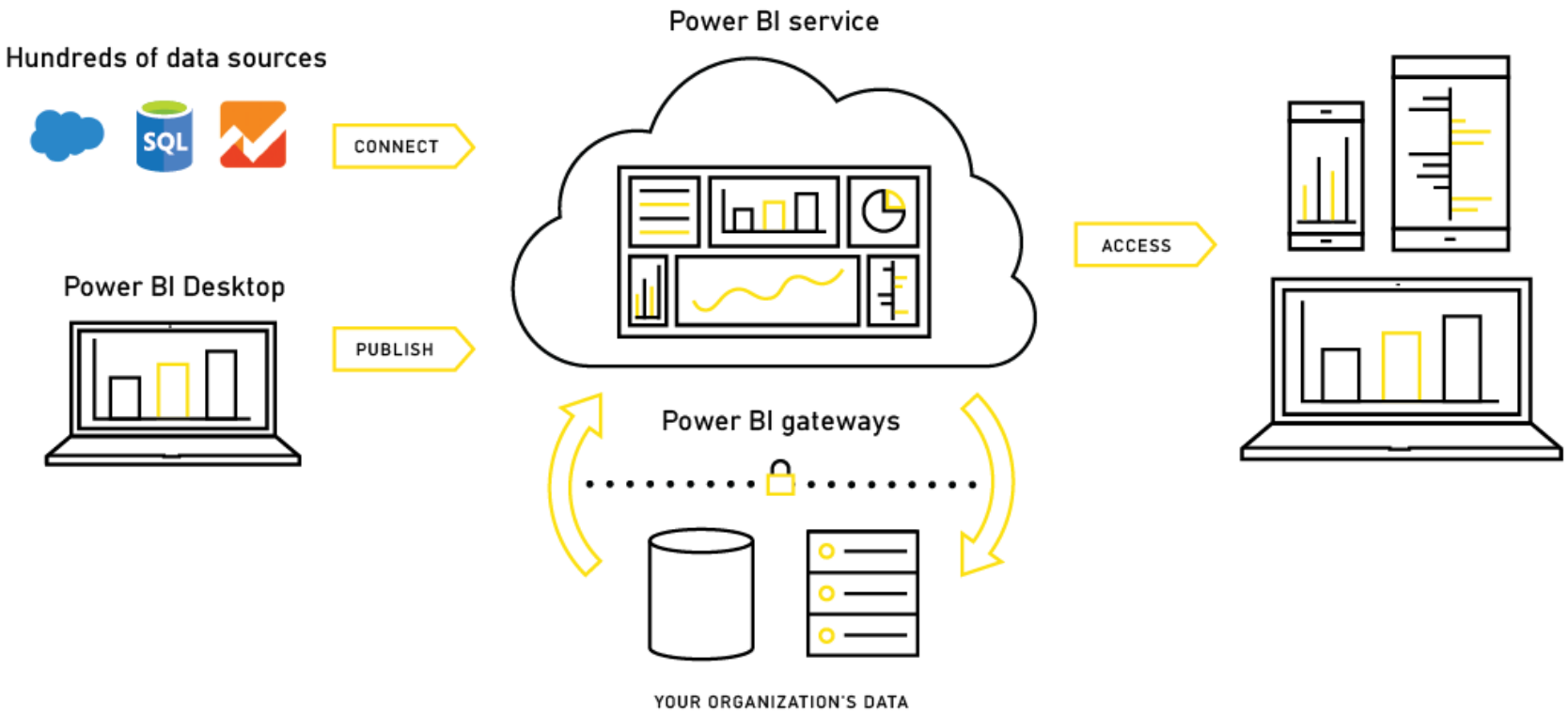
- High cardinality columns
  - Address columns
  - Date columns

Data Wrangling using Power BI Desktop

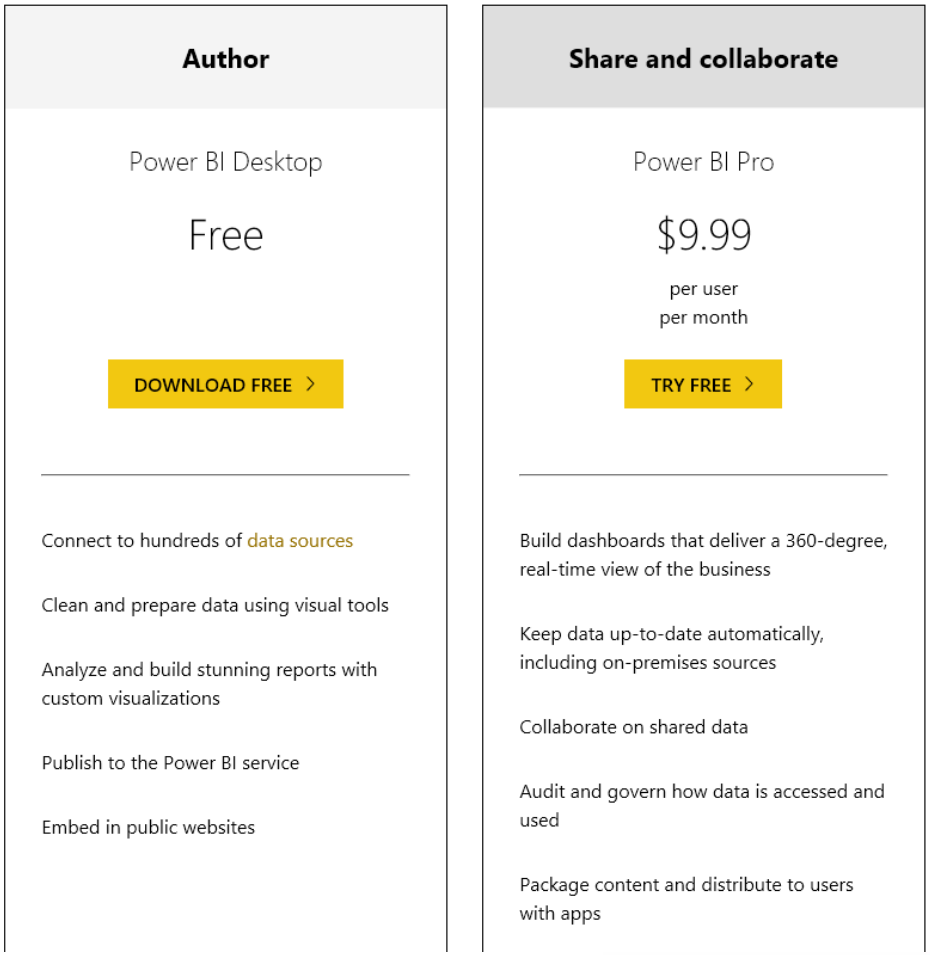
CSVEasy/MultiEdit
Data Exploration

Data Wrangling using Power BI Desktop

Data Wrangling using Power BI Desktop



Data Wrangling using Power BI Desktop



Data Wrangling using Power BI Desktop

Power BI
Data Exploration

Data Wrangling using Power BI Desktop

Data Wrangling Steps

- Transform
  - Fix or remove dirty data
  - Remove, split or combine columns
  - Join or pivot datasets

Data Wrangling using Power BI Desktop

Power BI
Data Transformation

Data Wrangling using Power BI Desktop

Data Wrangling Steps

- Improve
  - Identify and add other data which might be useful in this analysis
  - Can additional data be derived from the existing data

Data Wrangling using Power BI Desktop

Power Query
The M Language

Data Wrangling using Power BI Desktop

Tools

- CSVEasy
  - http://csveasy.com
- Multi-Edit Lite 2008
  - http://multieditsoftware.com/product/multi-edit-lite-2008

Data Wrangling using Power BI Desktop

Tools

- Power BI Desktop
  - https://powerbi.microsoft.com/en-us/desktop
- Power BI in the Cloud
  - https://powerbi.microsoft.com/en-us/pricing

Data Wrangling using Power BI Desktop

Resources

- Power Query (M) Formula Reference
  - https://msdn.microsoft.com/en-us/library/mt211003.aspx
- Chris Webb's BI Blog
  - https://blog.crossjoin.co.uk
- Melissa Coats - SQL Chick Blog
  - https://www.sqlchick.com/entries/2011/5/7/documenting-a-reporting-project.html

Data Wrangling using Power BI Desktop

Resources

- SQL Server 2017 – Developer Edition
  - https://www.microsoft.com/en-us/sql-server/sql-server-downloads

Data Wrangling using Power BI Desktop

?